Ontology Based Personalized Modeling for Chronic Disease Risk Evaluation and Knowledge Discovery: An Integrated Approach

Anju Verma

A thesis submitted to Auckland University of Technology in

fulfillment of the requirements for degree of

Doctor of Philosophy (PhD)

2009

School of Computer & Information Sciences

Primary supervisor: Prof. Nikola Kasabov

Other supervisors: Dr. Qun Song, Prof. Elaine Rush, Dr. Neil Domigan

Table of Contents

List of Figures

List of Tables

dedicated to my hubby……

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in acknowledgements) nor material which to a substantial extent has been accepted for the award of any other degree or diploma of a university or other institution of higher learning.

Anju Verma

Acknowledgements

This thesis arose in part out of research that has been done over the last few years. Over this time, I have worked with a great number of people whose contributions to the research and the development of the thesis deserve special mention. It is my pleasure to convey my gratitude to them all in my acknowledgments.

In the first place I would like to acknowledge my gratitude to Prof. Nikola Kasabov for his supervision, guidance, and advice from the very early stage of this research and for providing me with extraordinary experiences throughout the research. Without his unwavering faith and advice, and stimulating suggestions and discussions, the completion of my study would not have been possible. His experience as a true scientist made him as an oasis of ideas and passion for research, which has inspired and enriched my growth as a student and a researcher. I am indebted to him more than he knows.

I would like to express my sincere gratitude to Brent Ogilvie, who showed faith in me and helped me all the way through my studies. Words fail to express my appreciation for him, as he has always given me moral support and professional guidance with his words of encouragement.

I would like to thank Prof. Elaine Rush for her motivation, invaluable comments, generosity of time and willingness to help. I would also like to thank my other supervisor Dr. Qun Song, who was always there to provide his valuable suggestions. My special thanks to Dr. Neil Domigan, for his constant encouragement to write thesis. I would like to thank Prof. Ajit Narayanan for the help and support. I would also like to thank Russel Pears for his help and assistance during my research.

Armaan, my hubby, the guiding star of my life, deserves all the credit for me finishing my studies as he made room in our lives for me to study, caught the worst of my down at times when this project felt insurmountable, and had the faith that I would get there. Thank you from the bottom of my heart for all the support and sacrifices.

My darling dolls, Monalisa and Alisha deserve a special mention here as they never demanded special time or care though both were going through transition phases of life. Monalisa had a sister and also started primary school during my studies and Alisha even though a small baby never bothered me during my studies and used to play on her own and spent most of her time at day care when she was supposed to be under my care. I appreciate Monalisa's understanding as well all the way during my studies. I would also like to thank 'Kindercare' especially Angela for looking after Alisha so well.

I have reached at this stage with the blessings of my parents. My parents and parents-in law have been an integral part of my research journey through their continuous support and encouragement. My father-in-law always encouraged and showed faith in me and always showered his most precious blessings on me. My acknowledgements would be incomplete if I do not mention my Mum, as she always wanted me to finish my research and during the course of my study she was diagnosed with a serious disease. As she did not want to disturb me during my research, she never informed me about her sickness. I wish everyone could be blessed with a great mum as mine. I thank my mum for her optimism, unfailing love and encouragement.

I am much indebted to Joyce D'Mello, who always encouraged me to finish my research and thesis in time. I would like to thank Peter Hwang for all his

support and encouragement especially with Matlab as well as for our useful discussions during the last phase. I would also like to thank Dr. Ilkka Havukkala for his advice and willingness to share his bright thoughts with me, which was very fruitful in shaping up my ideas and research. I would also like to thank other KEDRI members namely, Paulo, Vishal and Alex for their help.

I would also like to thank Cherry Gordon, who was my first contact person for administration work at AUT. I would also like to thank postgraduate office (Annette, Elena, Martin Wilson) and other AUT admin staff for their help during the period of research. I would also like to acknowledge the help and services provided by the library staff.

This work was funded by FidelityGenetic (an affiliation of Pacific Channel Limited) and the Foundation for Research, Science and Technology of New Zealand under grant number CSHA0401. I would also like to thank the Ministry of Health, New Zealand for providing me with NNS97 data. I would also like to thank Maurizio, PhD. Student from DIMET, University Mediterranea of Reggio Calabria, Italy and Transplant Regional Center of Stem Cells and Cellular Therapy, “A. Neri”, Reggio Calabria, Italy for sharing diabetes data.

Finally, I would like to thank everybody who played important role in the successful completion of this thesis, as well as express my apologies that I can not mention each one by name.

And to God, who makes all things possible.

Abstract

Populations are aging and the prevalence of chronic disease, persisting for many years, is increasing. The most common, non-communicable chronic diseases in developed countries are; cardiovascular disease (CVD), type 2 diabetes, obesity, arthritis and specific cancers. Chronic diseases such as cardiovascular disease, type 2 diabetes and obesity have high prevalence and develop over the course of life due to a number of interrelated factors including genetic predisposition, nutrition and lifestyle. With the development and completion of human genome sequencing, we are able to trace genes responsible for proteins and metabolites that are linked with these diseases.

A computerized model focused on organizing knowledge related to genes, nutrition and the three chronic diseases, namely, cardiovascular disease, type 2 diabetes and obesity has been developed for the Ontology-Based Personalized Risk Evaluation for Chronic Disease Project. This model is a Protégé-based ontological representation which has been developed for entering and linking concepts and data for these three chronic diseases. This model facilitates to identify interrelationships between concepts.

The ontological representation provides the framework into which information on individual patients, disease symptoms, gene maps, diet and life history can be input, and risks, profiles, and recommendations derived. Personal genome and health data could provide a guide for designing and building a medical health administration system for taking relevant annual medical tests, e.g. gene expression level changes for health surveillance.

One method, called transductive neuro-fuzzy inference system with weighted data normalization is used to evaluate personalized risk of chronic disease. This

personalized approach has been used for two different chronic diseases, predicting the risk of cardiovascular disease and predicting the risk of type 2 diabetes. For predicting the risk of cardiovascular disease, the National Nutrition Health Survey 97 data from New Zealand population has been used. This data contains clinical, anthropometric and nutritional variables. For predicting risk of type 2 diabetes, data from the Italian population with clinical and genetic variables has been used. It has been discovered that genes responsible for causing type 2 diabetes are different in male and female samples.

A framework to integrate the personalized model and the chronic disease ontology is also developed with the aim of providing support for further discovery through the integration of the ontological representation in order to build an expert system in genes of interest and relevant dietary components.

Chapter 1. Introduction

Chapter 1 gives the brief introduction about the motivation behind my PhD study. This chapter is divided into 5 sections. The first section of the chapter 1 gives the background of the study; second section describes the goals of my PhD study. This chapter also presents the organization of the thesis in third section of this chapter along with brief outline of each chapter. Fourth section lists the major achievements made through my PhD study and last section of chapter 1 gives the list of publications during my PhD study.

1.1 Background

A chronic disease is a serious illness that is caused by many factors such as genes, environment, life style (e.g. tobacco-use, lack of physical activity, and poor eating habits). Chronic diseases generally cannot be prevented by vaccines or cured by medication, nor do they simply disappear. But they may be cured by changes in lifestyle.

The prevalence of chronic diseases is increasing worldwide. The most common chronic diseases of the developed world are arthritis, cardiovascular disease such as heart attacks and stroke, cancers such as breast and colon cancer, type 2 diabetes, obesity, epilepsy, seizures and oral health problems. The chronic disease such as cardiovascular disease, type 2 diabetes and obesity are the most common chronic diseases and are caused by a number of common factors (Figure 1.1).

The main factors causing these three chronic diseases are a number of common genes and nutrition. Understanding of the human genome has shown that each individual is unique and has a different genotype (excluding identical

twins). So each individual reacts differently to environmental factors such as diet and nutrition. The study of how foods may interact with specific genes to increase the risk of common chronic diseases such as type 2 diabetes, obesity, heart disease, stroke and cancers is called nutrigenomics.

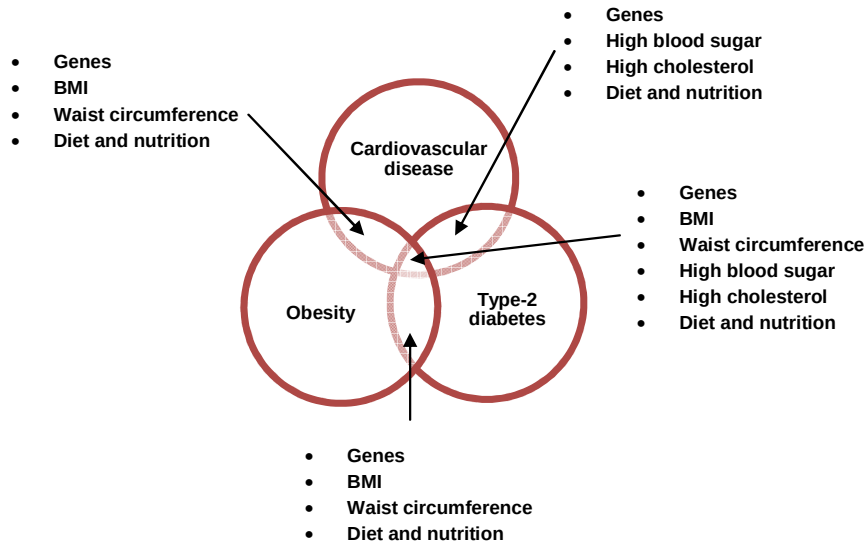


Figure 1.1. Venn diagram showing three chronic diseases with overlapping causes.

Nutrigenomics is a new and emerging field. Nutrigenomics has the potential to reduce the future risk of some human diseases by helping to specify nutritional guidelines based on the gene profile of the individual.

Nutrigenomics seeks to provide a molecular understanding of how common chemicals in the diet affect health by altering the expression of genes and the structure of an individual's genome. Nutrigenomics is the combination of genetics and nutrition science and involves the overlapping of health, genomics and diet (Figure 1.2).

The basic principles of nutrigenomics as explained by Jim Kaput (Kaput, 2004; Kaput and Rodriguez, 2004) are:

- Improper diets are risk factors for disease;
- Dietary chemicals alter gene expression and/or change genome structure;
- The influence of diet on health and disease susceptibility depends upon an individual's genetic makeup;
- Genes regulated by diet play a role in chronic diseases;
- "Intelligent nutrition"; that is, diets based upon genetics, nutritional requirements and health status may prevent and mitigate chronic diseases.

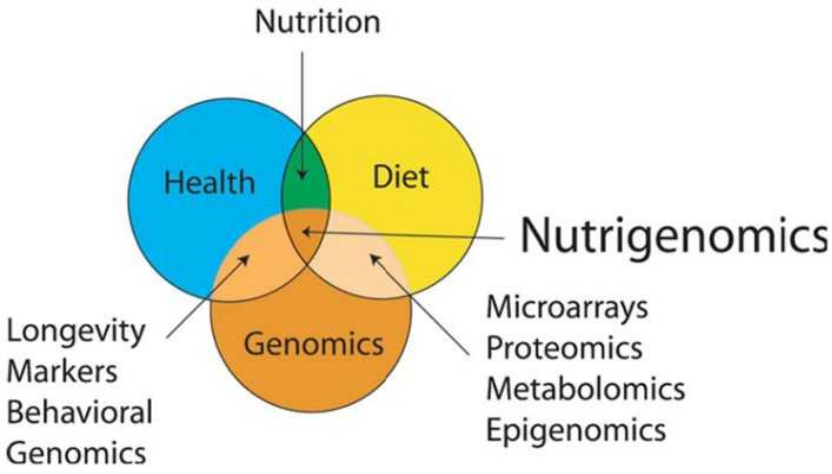


Figure 1.2. Venn diagram illustration of nutrigenomics as the intersection between health, diet, and genomics (Picture taken from Ruden et al, 2005).

Nutrigenomics aims to:

- establish scientifically-based “recommended dietary allowance” (RDA) (Stover, 2004).
- establish dietary recommendations that have a high predictive value with respect to disease prevention, minimize the risk of unintended consequences, and account for the modifying effects of human genetic variation.
- design effective dietary regimens for the management of complex chronic disease (Stover, 2004).
- generate recommendations regarding the risks and benefits of specific diets or dietary components to the individual. It has been also termed ‘personalized nutrition’ and ‘individualized nutrition’.
- be able to establish a person’s genetic profile in the hopes that it may help to determine his or her nutrient requirements as well as susceptibility to nutrition-related diseases.

The principles of nutrigenomics are used in this research to develop a risk evaluation system for chronic diseases by using nutritional and genetic variables along with clinical variables. There have been several models built to predict the risk of these diseases and to prevent the serious outcomes of the diseases such as the Anderson formula (Anderson et al, 1990). The methods developed to predict risk of cardiovascular disease and type 2 diabetes are detailed in Chapters 6 and 7 respectively.

The existing methods are global methods and only use clinical and general variables to predict the risk of cardiovascular disease and type 2 diabetes. Based on the principles of nutrigenomics, a personalized model to predict the

risk of these serious diseases using genetic and nutritional information is to be formulated.

1.2 Goals of the thesis

Chronic diseases, a major health problem throughout the world, are increasing and have very high prevalence. An effort has been made to create a method that can predict the risk of chronic diseases and inform lifestyle recommendations based on clinical, nutritional and genetic variables. The main aim is to discover new knowledge and build a personalized model using existing methods which can be used for disease prognosis and the improvement of human lifestyle and health. The project was to develop the following for both academic and overlapping commercial reasons:

General objectives:

- Design and build a knowledge acquisition tool to acquire and integrate knowledge from personal genome and metabolic information and also to integrate data from different domains of biology and medicine; including ontological information from Gene Ontology.
- To build a prediction model for risk evaluation of disease and associated disease prevention strategies of personal and commercial significance including dietary intake, and life style choices (such as nutritional advice).
- To build a framework for integration of ontology database and personalized model.

Specific objectives:

- Identify common causes and important genes for three chronic diseases; cardiovascular disease, type-2 diabetes and obesity;
- Build an ontology database for three chronic diseases such as CVD, type 2 diabetes and obesity in one domain in order to discover new knowledge from existing information;
- Identify relationships between different variables in National nutrition health survey data;
- Build a personalized risk prediction model for cardiovascular disease;
- Build a framework for integration of personalized risk prediction model for cardiovascular disease and chronic disease ontology;
- Identify the relationships between type 2 diabetes gene data and to find important genes responsible for causing type 2 diabetes in Italian type 2 diabetes data;
- Build a personalized risk prediction model for type 2 diabetes;
- Build a framework for integration of personalized risk prediction model for type 2 diabetes and chronic disease ontology.

The knowledge acquisition tool will also identify multi-factorial (from known single risk determinants) genetic and/or metabolic and lifestyle causes of disease risk prognosis that may be used in personal decision making for nutrition and lifestyle guidance. This tool will support further discovery by integration of the knowledge base and artificial intelligence methods to pinpoint further genes of interest and relevant diet components for advice on healthy lifestyle and disease-preventing nutrition and eating habits.

The project is sponsored through the Foundation for Research Science and Technology (TIF grant) through an affiliation of Pacific Channel Limited under grant number CSHA0401.

1.3 Organisation of the thesis

The present research in the form of a thesis has been structured in another seven chapters (Figure 1.3). Chapter 2 contains information about computational methods and systems available for risk evaluation. This chapter mainly contains information about reasoning methods, global, local and personalized risk evaluation methods such as Weighted-Weighted K Nearest Neighbour Algorithm for transductive Reasoning and personalized modeling (WWKNN) and the Transductive Neuro-Fuzzy Inference System with Weighted Data Normalization (TWNFI).

Chapter 3 defines “Ontology” and methods for building ontology. This chapter gives a brief description and comparative analysis of tools for developing ontology and also the existing methodologies for building ontology.

Chapter 4 explains structure and different domains of the chronic disease ontology. This chapter also presents the knowledge discoveries made through the chronic disease ontology.

Chapter 5 explains the integration framework for the ontology and the personalized risk evaluation system in general. This chapter presents the ways of knowledge discovery through the integration of the ontology and the personalized modeling framework with the help of three examples.

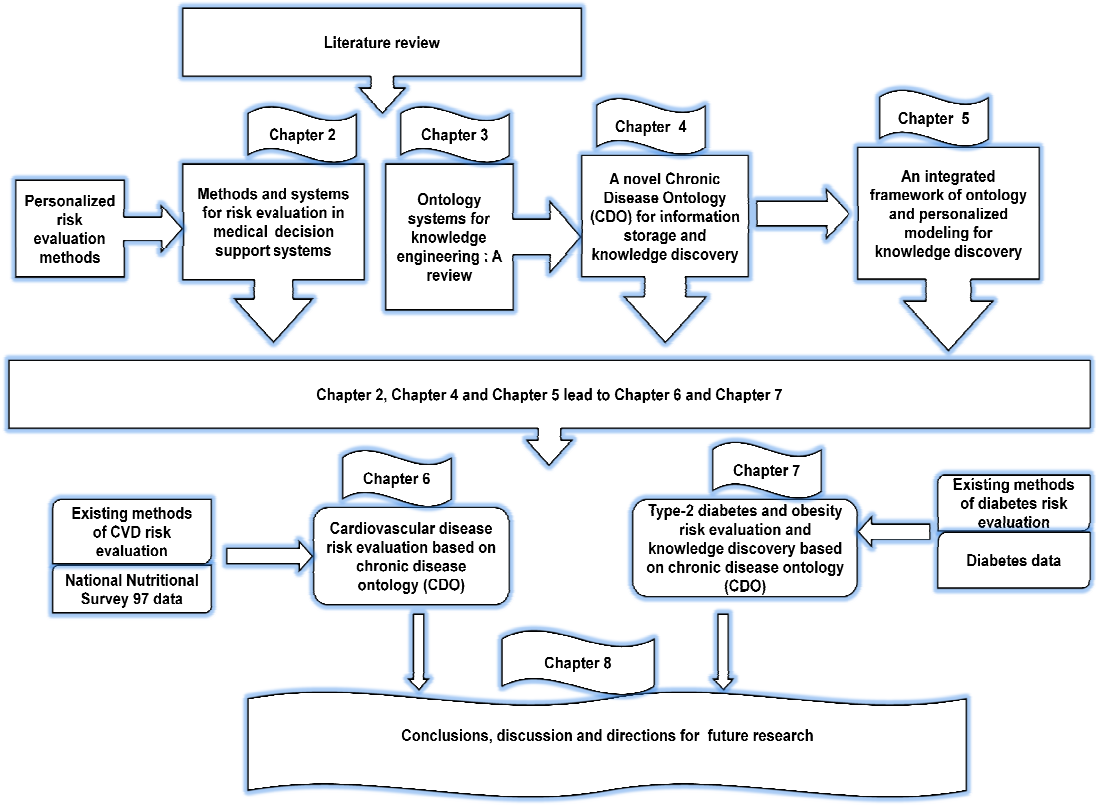


Figure 1.3. Structure and organization of the thesis.

Chapter 6 explains and defines cardiovascular disease and its prevalence in the world and in New Zealand. This chapter examines the literature available on existing methods for predicting risk of cardiovascular disease. A description of the data used for creating a personalized model for risk evaluation of cardiovascular disease is detailed in a subsection of Chapter 6. Different statistical methods used for data analysis are also mentioned in the following subsection of this chapter. The section 6.4 includes personalized risk prediction methods used to predict risk of cardiovascular disease such as WWKNN and TWNFI which have been used on data to discover new knowledge and build a personalized model. The last section of Chapter 6 includes the integration of the chronic disease ontology and the personalized risk evaluation method.

Chapter 7 presents the description and prevalence of type 2 diabetes and obesity in the world and in New Zealand. This chapter also explains the existing models for predicting risk of type 2 diabetes. The description and analysis of diabetes gene data is explained in the following section and this data has been used to create a personalized risk evaluation system. The last section of Chapter 7 explains the integration of the chronic disease ontology and the personalized risk evaluation model for type 2 diabetes.

Chapter 8 includes knowledge discovery through the current research, conclusions and a summary of current research. This chapter also gives directions for research in the near future, for the development and improvement of human health research.

1.4 Major contributions of the thesis

This thesis has made a major contribution to the field of bioinformatics as it resulted in knowledge discovery through the ontology as well as personalized modeling. The present research contributes to bioinformatics in the form of the following achievements and discoveries:

General achievements:

- A knowledge acquisition tool has been designed and developed to acquire and integrate knowledge from personal genome and metabolic information. This knowledge acquisition tool integrates data from different domains of biology and medicine; including genetic information from Gene Ontology.
- A personalized prediction model for risk evaluation of disease and associated disease prevention strategies of personal and commercial

significance has been developed in order to provide advice and recommendations on dietary intake, and life style choices (such as nutritional advice).

- A framework for integration of ontology database and personalized model has been designed.

Specific achievements:

- Chronic disease ontology: I have built the chronic disease ontology, a knowledge repository for the accumulation of clinical, genetic and nutritional information for three chronic diseases; cardiovascular disease, obesity and type 2 diabetes. There are about 71 genes in the ontology which are directly or indirectly related to these three diseases. The chronic disease ontology provides new knowledge of links between diseases and genes and can be used for integration with personalized modeling.
- National nutrition health survey data has been analyzed and further used for building personalized risk evaluation model for cardiovascular disease.
- Built a personalized risk evaluation system for cardiovascular disease: I have developed a personalized risk evaluation method for predicting risk of cardiovascular disease using clinical and nutritional variables with TWNFI (Personalized modeling). This method is unique as all the existing methods for determining risk of cardiovascular disease use clinical variables; none of the existing methods so far include nutritional variables. The personalized method (TWNFI) helps to retrieve more knowledge about the variables by generating sets of rules or profiles for

each sample based on nearest samples, hence resulting in better personalized disease risk evaluation.

- Analysis of type-2 diabetes gene data: Type-2 diabetes gene data has been analyzed and major genes causing type 2 diabetes have been identified.
- Personalized risk evaluation system for type 2 diabetes: I have also developed a personalized risk evaluation method for predicting risk of type 2 diabetes with TWNFI by using clinical and genetic variables. This method is unique as it uses clinical and genetic variables together, which have not been used so far to predict risk of type 2 diabetes. It has been discovered that genes responsible for causing type 2 diabetes are different in male and female populations.
- Integration framework for integration of the chronic disease ontology and the personalized risk evaluation system and knowledge discovery: This integration framework explains how new knowledge can be extracted from the evolving ontology and personalized model and can be used for better prediction. This framework has been explained with integration of two different personalized models (cardiovascular disease risk prediction and type 2 diabetes disease risk prediction) and chronic disease ontology.

1.5 Resulting publications

During the course of this study, the research work was presented at a few conferences in the form of poster and oral presentations. The list of publications is as follow:

Internationally refereed papers:

(1) Verma, A., Fiasche, M., Cuzzola, M., Iacopino, P., Morabito, F.C, Kasabov, N.(2009). Ontology Based Personalized Modeling for Type 2 Diabetes Risk Analysis: An Integrated Approach. International Conference on Neural Information Processing (ICONIP). 1-5 December, 2009. Paper accepted.

(2) Verma, A., Kasabov, N., Rush, E., Song, Q. (2009). Knowledge Discovery through Integration of Ontology and Personalized Modeling for Chronic Disease Risk Analysis. Paper accepted for Australian Journal of Intelligent Information Processing Systems.

(3) Verma, A., Kasabov, N., Rush, E., Song, Q. (2008) Ontology Based Personalized Modeling for Chronic Disease Risk Analysis: An Integrated Approach. International Conference on Neural Information Processing (ICONIP). 24 -28 November, 2008, Springer, LNCS No 5506/07, 2009.

(4) Kasabov, N., Song, Q., Benuskova, L., Gottgtroy, P., Jain, V., Verma, A., Havukkala, I., Rush, E., Pears, R., Tjahjana, A., Hu, R., MacDonell, S. (2008): Integrating Local and Personalized Modeling with Global Ontology Knowledge Bases for Biomedical and Bioinformatics Decision Support, chapter in: Smolin et al (eds) Computational Intelligence in Bioinformatics, Springer, 2008.

(5) Verma, A, Song, Q & Kasabov, N., (2006). Developing “Evolving Ontology” for Personalized Risk Evaluation for Type 2 diabetes Patients. At 6th International Conference on Hybrid Intelligence, Auckland, New Zealand.

Abstracts and posters presented: The current doctoral work has been presented at many conferences. The list of poster presentations and abstract publications is as follows:

(1) Verma, A., Song, Q. & Kasabov, N. (2008) Ontology based personalized modeling for chronic disease and risk evaluation in medical decision support system. At The 3rd Asia Pacific Nutrigenomics Conference 2008: Diet-Gene Interaction in Human Health and Disease, Melbourne, Australia, from May 6-9, 2008.

(2) Verma, A., Song, Q. & Kasabov, N. (2008) Ontology based personalized modeling for chronic disease risk evaluation in medical decision support systems. At New Zealand's Annual Biotechnology Conference, Sky City Convention Centre, Auckland, NZ. 31 March - 2 April 2008.

(3) Verma, A., Song, Q. & Kasabov, N. (2006) Developing "Evolving Ontology" for Personalized Risk Evaluation for Type 2 Diabetes Patients. At 6th International Conference on Hybrid Intelligent Systems (HIS'06) and 4th Conference on Neuro-Computing and Evolving Intelligence (NCEI'06) 13-15 December, 2006, Auckland, New Zealand.

(4) Verma, A., Song, Q. & Kasabov, N. (2006) Developing "Evolving Ontology" for Nutritional Advice for Type 2 diabetes Patients Poster at Queenstown Molecular Biology Meeting 29th August – 1 September 2006. Queenstown, New Zealand.

(5) Verma, A., Gottgtroy, P., Havukkala, I., and Kasabov, N. (2005) An Ontological Representation of Nutrigenomics Knowledge about Type 2 Diabetes. New Zealand's Annual Biotechnology Conference Sky City Convention Centre, Auckland, NZ. 27-28 February 2006.

(6) Verma, A., Gottgtroy, P., Havukkala, I., and Kasabov, N. (2005) An Ontological Representation of Nutrigenomics Knowledge about Type 2 diabetes. ILSI's First International conference on Nutrigenomics Opportunities in Asia 2005.

(7) Verma, A., Gottgtroy, P., Havukkala, I., and Kasabov, N. (2005) Understanding the Molecular Basis of Type 2 Diabetes by Means of Evolving Ontologies and Intelligent Modeling. The Queenstown Molecular Biology Meeting, 2005.

Chapter 2. Methods and Systems for Risk Evaluation in Medical Decision Support Systems

This chapter gives description of the background of reasoning methods, different methods of reasoning (inductive and transductive), the following sections explain modeling methods as global, local and personalized methods. The subsequent sections of the chapter explain different methods of personalized modeling using transductive reasoning, such as Weighted-weighted K nearest neighbour algorithm for transductive reasoning (WWKNN), Transductive neuro-fuzzy inference system with weighted data normalization (TWNFI).

Chronic diseases are widely spread all over the world. Chronic diseases such as cardiovascular disease, type 2 diabetes, arthritis, cancer and obesity are the most common chronic diseases throughout the world. There have been many efforts done to develop risk evaluation systems to predict the risk of occurrence of chronic diseases. The most traditional methods for predicting risk of cardiovascular disease is the Anderson formula (Anderson et al, 1990), QRISK2 (Hippisley-Cox et al, 2008) and also the New Zealand cardiovascular risk charts (Milne et al, 2003; Jackson, 2000), and the PREDICT-CVD-1 (Bannink et al, 2006) method, which is a web-based tool, have been widely used to predict risk of cardiovascular disease. The Framingham risk equation uses inductive reasoning (described in next section).

For predicting risk of type 2 diabetes, several methods have been developed such as the Global Diabetes Model (Brown et al, 2000(a), (b)), the Diabetes risk score (Lindstrom and Tuomilehto, 2003), the Archimedes model (Eddy and Schlessinger, 2003(a), (b)), the German risk score (Schulze et al, 2007)

and the Diabetes risk score in Oman (Al-Lawati and Tuomilehto, 2007). The global diabetes model uses the Monte Carlo method, the diabetes risk score (Lindstrom and Tuomilehto, 2003) uses the multivariate logistic regression method while the diabetes risk score in Oman (Al-Lawati and Tuomilehto, 2007) uses backward stepwise logistic regression and the German risk score (Schulze, 2007) uses Cox regression with forward selection. The details of the existing methods for cardiovascular and type 2 diabetes risk prediction are explained in Chapters 6 and 7.

With the completion of the human genome sequencing project, there is an abundance of genetic information which needs to be used for improving health and lifestyle. So, it becomes evident that data and knowledge need to be organized in a global knowledge repository and used in their complexity and richness for efficient profiling, prognosis, diagnosis and decision support for every individual who needs it. Nutrigenomics has made it possible for nutritional, genetic and computer science to explore the interactions and discover new relationships between nutrition and genes. This task requires both adaptive, evolving knowledge repository systems and methods for personalized modeling and their integration and dynamic interaction.

2.1 Inductive and transductive reasoning

The inductive reasoning approach has been widely used in all fields of science. The inductive reasoning approach is concerned with the creation of a model (a function) from all available data, representing the entire problem space. The model is then applied to new data (deduction) (Levey et al 1999; Anderson et al, 1990).The model is usually created without taking into account any information about a particular new vector. An error is measured to estimate how

well the new data fits into the model. This method has been used widely such as in the Framingham risk calculations. Figure 2.1 explains inductive reasoning in the form of block diagram.

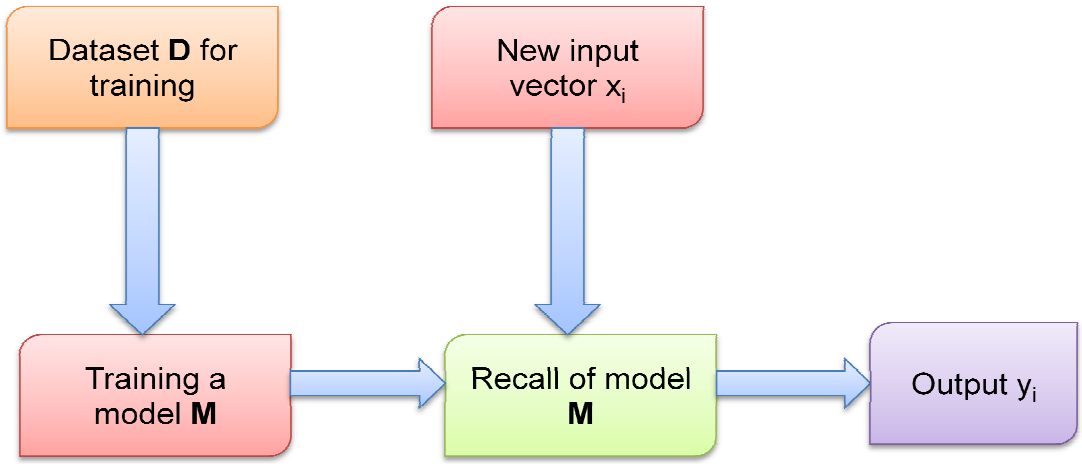


Figure 2.1. A block diagram of an inductive reasoning system. A global model M is created based on data samples from D and then recalled for every new vector x_i (From: Song and Kasabov, 2004).

On the other hand, the transductive inference, introduced by Vapnik (1998), is defined as a method used to estimate the value of a potential model (function) for only a single point of space (that is, a new data vector) by utilizing additional information related to that vector. While the inductive learning and inference approaches are useful, when a global model of the problem is needed, even in its very approximate form, the transductive approach is more appropriate for applications where the focus is not on the model, but rather on every personalized case. This approach seems to be more appropriate for clinical and medical applications where the focus needs to be centered on individual patient's conditions.

The transductive approach is related to the common sense principle (Bosnic et al, 2003) which states that to solve a given problem one should avoid solving a more general problem as an intermediate step. In the past, transductive reasoning has been implemented for a variety of classification tasks such as text classification (Chen et al, 2003; Joachims, 1999), heart disease diagnostics (Wu et al, 1999), synthetic data classification using a graph-based approach (Li and Yuen, 2001), digit and speech recognition (Joachims, 2003), promoter recognition in bioinformatics (Kasabov and Pang, 2004), image recognition (Li and Chua, 2003) and image classification (Proedrou et al, 2002), micro-array gene expression classification (Wolf and Mukherjee, 2004; West et al, 2001) and biometric tasks such as face surveillance (Li and Wechsler, 2004).

This reasoning method is also used in prediction tasks such as, predicting if a given drug binds to a target site (Weston et al, 2003), evaluating prediction reliability in regression (Bosnic et al, 2003) and providing additional measures to determine the reliability of predictions made in medical diagnosis (Kukar, 2003). Out of several research papers that utilize transductive approach, transductive support vector machines (Joachims, 1999) and semi-supervised support vector machines (Bennett and Demiriz, 1998) are often citied (Sotiriou et al, 2003).

In transductive reasoning, for every new input vector $\mathbf{x}_i$ that needs to be processed for a prognostic/classification task, the nearest neighbors N_i, which form a data subset D_i, are derived from an existing dataset D and a new model M_i is dynamically created from these samples to approximate the function in the locality of point $\mathbf{x}_i$ only. The system is then used to calculate the output

value y_i for the input vector $\mathbf{x}_i$. Figure 2.2 and 2.3 explain the transductive reasoning.

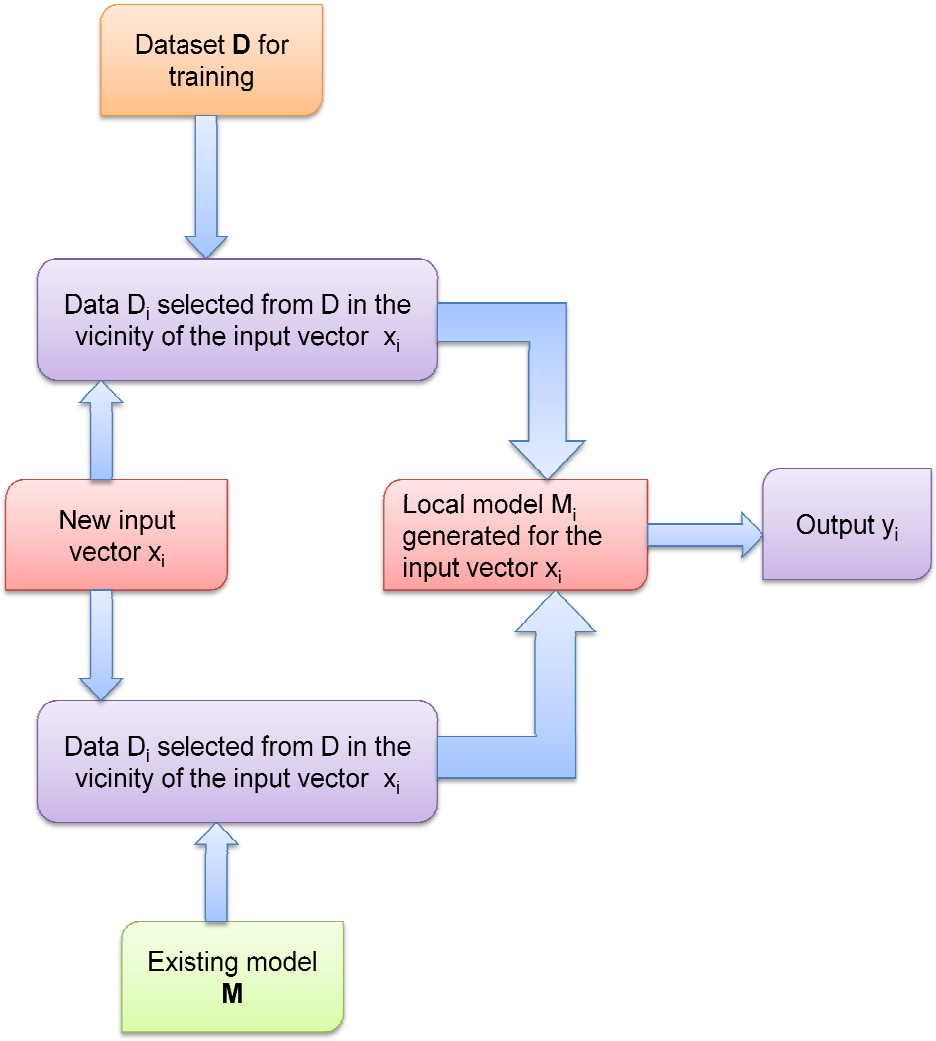
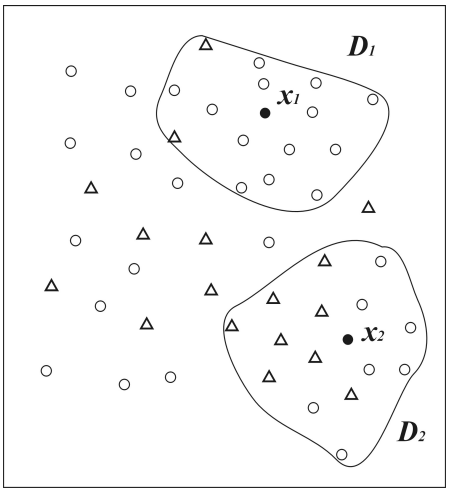


Figure 2.2. A block diagram of a transductive reasoning system. An individual model M_i is trained for every new input vector x_i with data samples D_i selected from a data set D, and data samples $D_{o,i}$ generated from an existing model (formula) M (if such a model exists). Data samples in both D_i and $D_{o,i}$ are similar to the new vector x_i according to a defined similarity criteria (From: Song and Kasabov, 2006).



● – a new data vector

○ – a sample from *D*

Δ – a sample from *M*

Figure 2.3. Example of transductive reasoning. In the centre of a transductive reasoning system is the new data vector (here illustrated with two vectors – $\mathbf{x}_1$ and $\mathbf{x}_2$), surrounded by a fixed number of nearest data samples selected from the training data D and/or generated from an existing model M (From: Song and Kasabov, 2006).

This approach has been implemented with a radial basis function (Song and Kasabov 2004) in medical decision support systems and time series prediction problems, where individual models are created for each input data vector. The results indicate that transductive inference performs better than inductive inference models mainly because it exploits the structural information of unlabeled data.

2.2 Global, local and personalized modeling

Global Modeling: In global modeling, a model is created from data that covers the whole problem space and is represented as a single function, e.g. a linear

regression formula or a support vector machine. The global model gives the big picture but not the individual profile. It has difficulty adapting to new data.

Local Modeling: In local modeling, a set of local models are created from data, each representing a sub-space, for example, a cluster of the problem, e.g. a set of rules; a set of local regressions, etc. Local models are created to evaluate the output function for only a sub-space of the problem space. Multiple local models (e.g. one for each cluster of data) together constitute the complete model of the problem over the whole problem space. Local models are often based on clustering techniques such as k-means (Mitchell et al 1997).

A cluster is a group of similar data samples, where similarity is measured predominantly as Euclidean distance in an orthogonal problem space. Clustering techniques include: k-means (Mitchell et al 1997); Self-Organizing Maps (SOM) (Kohonen 1997; DeRisi et al 1996), fuzzy clustering (Dembele and Kastner 2003; Futschik and Kasabov 2002; Bezdek 1981), hierarchical clustering (Alon et al 1999), simulated annealing (Lukashin and Fuchs 2001). In fuzzy clustering, one sample may belong to several clusters to a certain membership degree, the sum of which is 1. Local models are easier to adapt to new data and can provide a better explanation for individual cases.

Personalized modeling: In personalized modeling, a model is created for a single point (patient record) of the problem space only using transductive reasoning. A personalized model is created “on the fly” for every new input vector and this individual model is based on the closest data samples to the new samples taken from a data set. The K-nearest neighbours (K-NN) method is one example of the personalized modeling technique. In the K-NN method, for every new sample, the nearest K samples are derived from a data set using

a distance measure, usually Euclidean distance, and a voting scheme is applied to define the class label for the new sample (Vapnik 1998, Mitchell et al 1997). In the K-NN method, the output value y_i for a new vector x_i is calculated as the average of the output values of the k nearest samples from the data set D_i .In the weighted K-NN method (WKNN), the output y_i is calculated based not only on the output values (e.g. class label) y_j of the K, NN samples, but also on a weight w_j , that depends on the distance of them to x_i.

Global models capture trends in data that are valid for the whole problem space, and local models capture local patterns, valid for clusters of data. Both models contain useful information and knowledge. Local models are also adaptive to new data as new clusters and new functions that capture patterns of data in these clusters can be incrementally created. Both global and local modeling approaches usually assume a fixed set of variables but if new variables, along with new data, are introduced with time, the models are very difficult to modify to accommodate these new variables. Modification can be done in the personalized models, as they are created ‘‘on the fly’’ and can accommodate any new variables, provided that there is data for them.

All the three approaches are useful for complex modeling tasks and all of them provide complementary information and knowledge, learned from the data. For each individual data vector (e.g. a patient) an individual, local model that fits the new data is needed, rather than a global model, in which the data is matched without taking into account any specific information about the new data. A few personalized modeling approaches the weighted-weighted K nearest neighbour (WWKNN), the neuro fuzzy inference method (NFI) and the transductive neuro-

fuzzy inference system (TWNFI) have been described in detail in the next sections.

2.3 Weighted K-nearest neighbour method (WKNN)

In the K-NN method, the output value y_i for a new vector x_i is calculated as the average of the output values of the k nearest samples from the data set D_i (Kasabov, 2007(b)). In the weighted K-NN method (WKNN) the output y_i is calculated based not only on the output values (e.g. class label) y_j of the K NN samples, but also on a weight w_j, that depends on the distance of them to x_i:

$$y_i = \frac{\sum_{j=1}^{Ni} w_j y_j}{\sum_{j=1}^{Ni} w_j} \qquad (2.1)$$

where: y_j is the output value for the sample x_j from D_i and w_j are their weights calculated based on the distance from the new input vector:

$$w_j = [\max(\mathbf{d}) - (dj - \min(\mathbf{d}))] / \max(\mathbf{d}) \qquad (2.2)$$

The vector $\boldsymbol{d} = [d_1, d_2, \ldots, d_{Ni}]$ is defined as the distances between the new input vector x_i and the N_i nearest neighbours $(\boldsymbol{x}_j, y_j)$ for $j = 1$ to N_i

$$d_j = \text{sqrt}\left[\sum_{I=1 \text{ to } V} (x_{i,I} - x_{j,I})^2\right] \qquad (2.3)$$

where: V is the number of the input variables defining the dimensionality of the problem space; $x_{i,I}$ and $x_{j,I}$ are the values of variable x_i in vectors x_i and $\boldsymbol{x}_j$, respectively. The parameters max($\boldsymbol{d}$) and min($\boldsymbol{d}$) are the maximum and

minimum values in ***d*** respectively. The weights w_j have the values between min(***d***)/max(***d***) and 1; the sample with the minimum distance to the new input vector has the weight value of 1, and it has the value min(***d***)/max(***d***) in case of maximum distance. The weighted K-NN method (WKNN) is the simplest transductive method.

2.4 Weighted-weighted K nearest neighbour algorithm for transductive reasoning (WWKNN) and personalized modeling

In WWKNN (Kasabov, 2007(b)) (Appendix A) the distance between a new input vector and the neighbouring ones is weighted, and also variables are ranked according to their importance in the neighbourhood area. WWKNN is also a simple but fast transductive method and is used to solve a classification problem and two classes, represented by 0 (class 1) and 1 (class 2) output class labels, the output for a new input vector x_i has the meaning of a "*personalized probability*" that the new vector x_i will belong to class 2. In order to finally classify a vector x_i into one of the (two) classes, there has to be a probability threshold selected *Pthr*, so that if $y_{i>=}$ *Pthr*, then the sample x_i is classified as class 2. For different values of the threshold *Pthr,* the classification error is, generally, different. In the WWKNN, the calculated output for a new input vector depends not only on the number of its neighbouring vectors and their output values (class labels), but also on the distance between these vectors and the new vector which is represented as a weight vector (W). It is assumed that all V input variables are used and the distance is measured in a V-dimensional Euclidean space with all variables having the same impact on

the output variable. But when the variables are ranked in terms of their discriminative power of class samples over the whole V-dimensional space, we can see that different variables have different importance to separate samples from different classes, therefore it has a different impact on the performance of a classification model. If we measure the discriminative power of the same variables for a sub-space (local space) of the problem space, the variables may have a different ranking. Using the ranking of the variables in terms of discriminative power within the neighborhood of K vectors, when calculating the output for the new input vector, is the main idea behind the WWKNN algorithm, which includes one more weight vector to weigh the importance of the variables. The Euclidean distance d_j between a new vector x_i and a neighboring one x_j is calculated now as:

$$d_j = \text{sqr}\left[\sum_{I=1 \text{ to } V} \left(c_{i,I} \left(x_{i,I} - x_{j,I}\right)\right)^2\right] \quad (2.4)$$

where $c_{i,I}$ is the coefficient weighting variable x_i in neighborhood of x_i. It can be calculated using a Signal-to-Noise Ratio (SNR) procedure that ranks each variable across all vectors in the neighborhood set D_i of N_i vectors:

$$\boldsymbol{C_i} = (c_{i,1}, c_{i,2}, \ldots, c_{i,V}) \quad (2.5)$$

$$c_{i,I} = S_I / \sum (S_I), \quad \text{for: } I = 1,2,\ldots,V, \quad (2.6)$$

where:

$$S_1 = \text{abs}\left(M_1^{(\text{class 1})} - M_1^{(\text{class 2})}\right) / \left(\text{Std}_1^{(\text{class 1})} + \text{Std}_1^{(\text{class2})}\right) \quad (2.7)$$

Here $M_1^{(class\ 1)}$ and $Std_1^{(class\ 1)}$ are respectively the mean value and the standard deviation of variable x_i for all vectors in D_i that belong to class 1. The new distance measure, that weighs all variables according to their importance as discriminating factors in the neighborhood area D_i, is the new element in the WWKNN algorithm when compared to the WKNN. Using the WWKNN algorithm a ‘‘personalized’’ profile of the variable importance can be derived for any new input vector, that represents a new piece of ‘‘personalized’’ knowledge, but WWKNN lacks a feature selection algorithm and needs modifications in terms of feature selection.

2.5 Neuro-Fuzzy Inference Method (NFI) for personalized modeling

The Neuro-Fuzzy Inference (NFI) method is a more complex transductive and dynamic neural-fuzzy inference system with local generalization, in which, either the *Zadeh-Mamdani* type fuzzy inference engine (Zadeh, 1988; 1965) or the *Takagi-Sugeno* fuzzy inference engine (Takagi and Sugeno, 1985) can be used. The local generalization means that in a sub-space (local area) of the whole problem space, a model is created and this model performs generalization in this area. In the NFI model, *Gaussian* fuzzy membership functions are used in each fuzzy rule for both antecedent and consequent parts. A steepest descent (back-propagation) learning algorithm is used for optimizing the parameters of the fuzzy membership functions (Lin and Lee, 1996; Wang, 1994). Figure 2.4 explains NFI in the form of a block diagram. The NFI learning algorithm is detailed in Appendix B. The distance between vectors **x** and **y** is measured as a weighted normalized Euclidean distance defined as follows:

$$\|\boldsymbol{x}-\boldsymbol{y}\| = \left[\frac{1}{P}\sum_{j=1}^{P} w_j |x_j - y_j|^2\right]^{\frac{1}{2}} \tag{2.8}$$

where: $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{R}^P$; w_j are weights assigned to the input variables x_j.

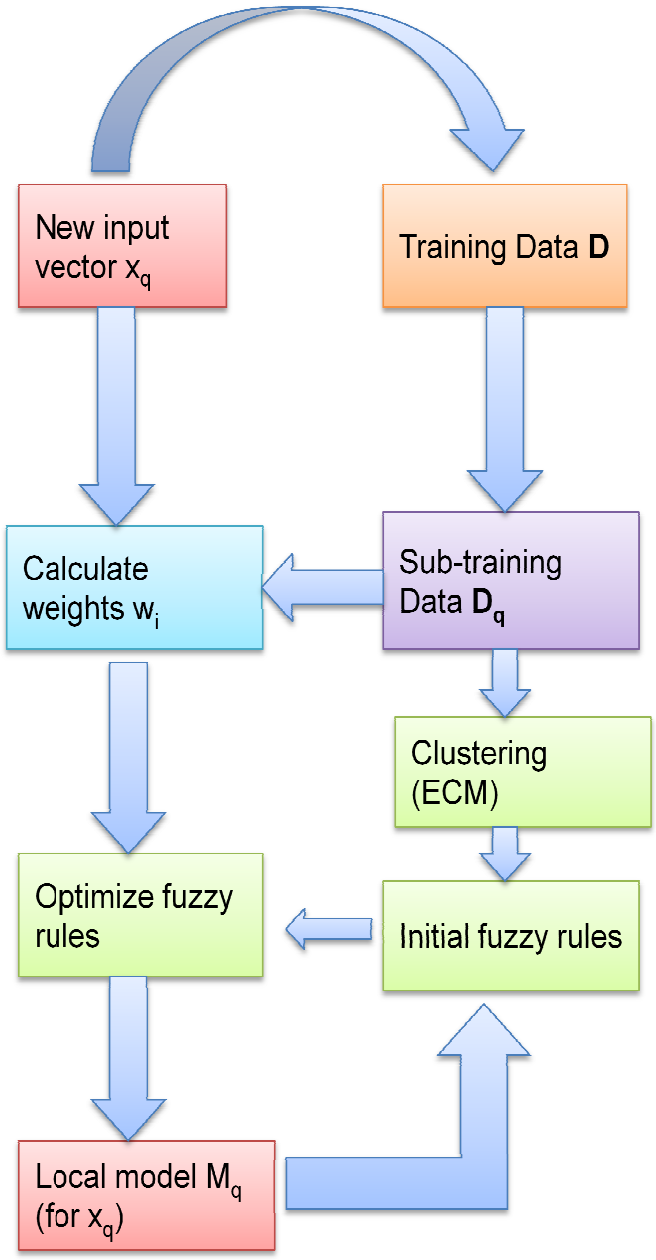


Figure 2.4. A block diagram of the NFI learning algorithm (From: Song and Kasabov, 2004).

NFI gives higher accuracy than K-NN although both are transductive methods because K-NN takes the average of outputs of the nearest neighbors while NFI uses samples to create and train local fuzzy inference system (Song and Kasabov, 2004). NFI has been applied for time series prediction, classification and personalized modeling in medical decision support system.

2.6 Transductive neuro-fuzzy inference system with weighted data normalization (TWNFI) for personalized modeling

Transductive neuro-fuzzy inference system with weighted data normalization (TWNFI) is an improved and advanced transductive method that uses the similar principles as NFI but the difference between NFI and TWNFI is that, in TWNFI, data is first normalised and then it looks for nearest samples (Appendix C). A general block diagram of the TWNFI algorithm is shown in Figure 2.5.

TWNFI has been used for many applications such as time series prediction (Song and Kasabov, 2006) and personalized renal function evaluation. The TWNFI system has been applied for modeling and predicting the future values of a chaotic time series on the Mackey-Glass (MG) data set (Farmer and Sidorowitch, 1987), which has been used as a benchmark problem in the areas of neural networks, fuzzy systems and hybrid systems (Crodder and Grossberg 1990). The TWNFI model performs better than the other models (Song and Kasabov, 2006).

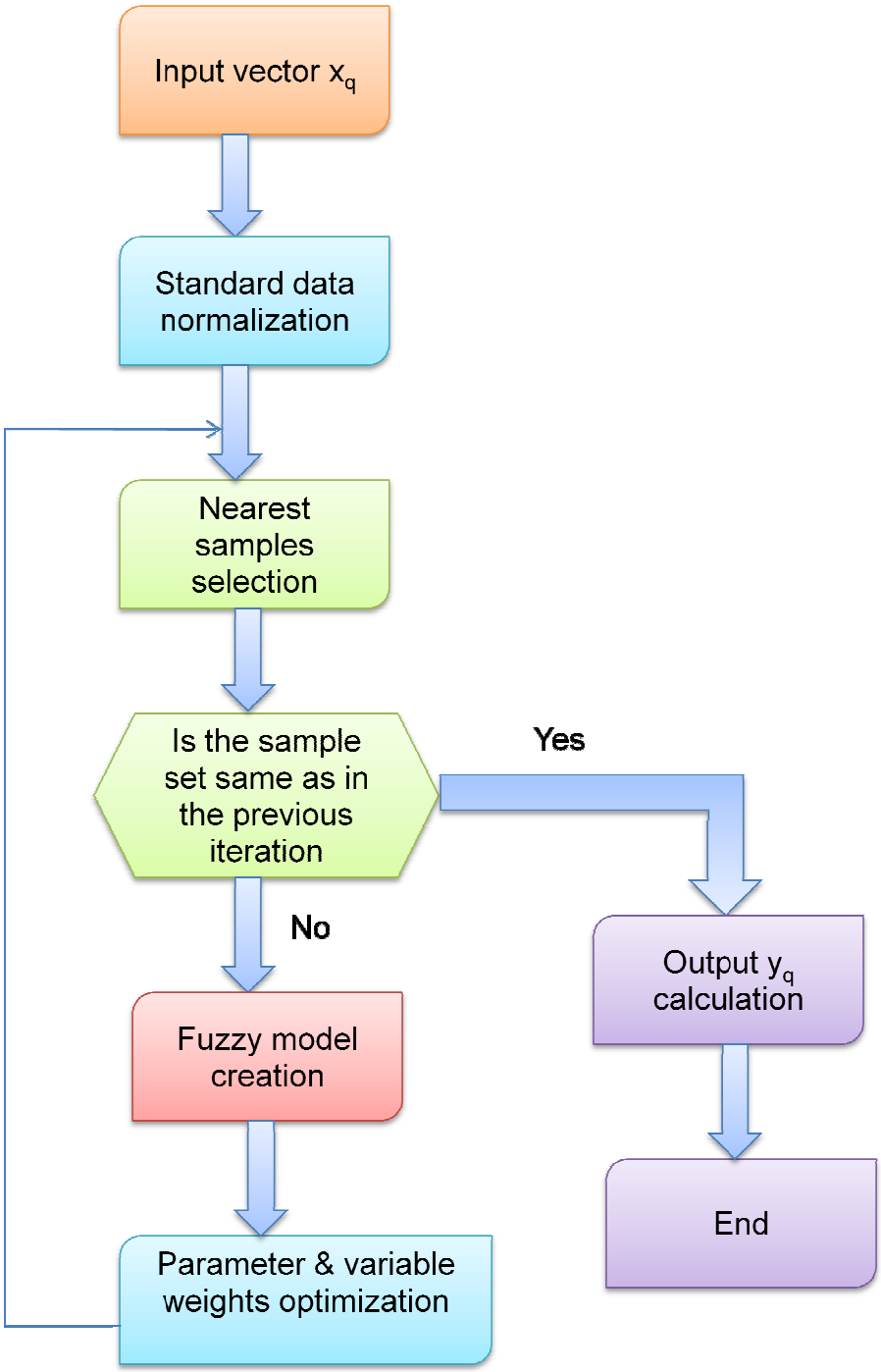


Figure 2.5. A block diagram of the TWNFI algorithm (From: Song and Kasabov, 2006).

The TWNFI method not only results in a “personalized” model with a better accuracy of prediction for a single new sample, but also depicts the most significant input variables (features) for the model that may be used for personalized medicine. Therefore, the TWNFI method can be applied for the creation of a personalized model of a patient for medical diagnosis, disease prognosis, and treatment planning.

Due to the completion of human genome sequencing, the new revolution in the field of biotechnology is personalized medicine (Walsh, 2009). Personalized medicine is very promising and a growing area of research where TWNFI, a personalized modeling method, can be broadly utilized. For medical applications, TWNFI has been used for personalized renal function evaluation and results were compared with several well known methods which have been applied to similar problems, such as the MDRD logistic regression function (Levey et al., 1999), which is widely used in the renal clinical practice, and other machine learning techniques, such as: MLP neural network (Neural network Toolbox, 2002) adaptive neural fuzzy inference system (ANFIS) (Jang, 1993), and dynamic evolving neural fuzzy inference system (DENFIS) (Kasabov and Song, 2002). It was found that the TWNFI method results in better accuracy and also depicts the average importance of the input variables, represented as calculated weights. For every patient, represented by a sample, a personalized model can be created and used to evaluate the output for the patient and also to estimate the importance of the variables for this patient for one particular data sample.

The TWNFI performs better local generalization over new data as it develops an individual model for each data vector that takes into account the new input

vector location in the space, and it is an adaptive model, in the sense that input-output pairs of data can be added to the data set continuously and immediately made available for transductive inference of local models. This type of modeling is called “personalized”, and appears to be promising for medical decision support systems. As the TWNFI creates a unique sub-model for each data sample, it usually needs more performance time than an inductive model, especially in the case of training and simulating on large data sets.

Creating a personalized model for a patient to predict an important event for this patient, such as risk of cardiovascular disease or a recurrence of cancer, could increase the accuracy of the prediction, especially when more variables are included in the prognosis, for example, clinical, genetic, environmental, social, etc. that relate to the problem, unlike variables that were used in the past to create one model for all, such as in the Anderson’s formula (Anderson et al, 1990), derived from data from the North American population to predict a cardiovascular event for everybody. A major advantage of the personalized modeling is that for a single vector (a single patient data), an individual model is created that can be used to explain the calculated output value.

The TWNFI method has several advantages when compared to the previously developed inductive inference methods when used on the same datasets:

- TWNFI performs a better local generalization over new data as it develops an individual model for each data vector that takes into account the new input vector location in the space. This type of modeling is called “personalized”, and appears to be very promising for medical decision support systems.

- TWNFI works well in a large dimensional space of many variables (for example: number of genes, clinical and nutritional variables).
- TWNFI is an adaptive model, in the sense that input-output pairs can be added to the data set continuously and immediately made available for transductive inference of local models.
- The TWNFI method can be used at different times on different number of variables (different dimensions) and over data vectors characterized by missing values. A personalized model can be created with the use of only the variables available in the new vector which is not possible when using global models.

As the TWNFI method creates a unique sub-model for each data sample, it usually needs more performance time than inductive models, especially in the case of training and simulating large data sets. If there are some simulating data samples that are same or very similar to each other, the TWNFI model will create the same or very similar models for them repeatedly. It is therefore advantageous to use both incremental, inductive reasoning (e.g. ECOS) to reveal a global model (the “big picture”), and the TWNFI reasoning for accurate personalized inference and decision making, or to store some already evolved personalized models for further use, which is one of current research topics. Time demands of the method depend mainly on the search algorithm, while searching for similar data to the new vector in a database.

2.7 Summary

This chapter has explained inductive and transductive reasoning and has described global, local and personalized modeling methods. Examples of existing transductive methods of personalized modeling have also been

explained. The local models, such as multiple linear regression method, provide information about individual samples based on local models or clusters created from data. But in the personalized models, such as WWKNN and TWNFI, a model is created for a single space based on nearest samples.

WWKNN is a simple transductive method while TWNFI is a much more advanced method and involves optimization of variables and samples. The above mentioned local and personalized methods are adaptive but local models, such as the multiple linear regression method, are difficult to adapt when new variables are added. On the contrary, personalized model TWNFI can accommodate new variables. In the present research, I have used TWNFI for cardiovascular disease risk prognosis with clinical and nutritional data in Chapter 6; type 2 diabetes risk prognosis based on general, clinical and genetic data in Chapter 7. I have also compared this method with local model such as multiple linear regression and a simple personalized method WWKNN.

Chapter 3. Ontology Systems for Knowledge Engineering: A Review

This chapter defines ontology and its applications. This chapter gives comparative analysis about the tools for developing ontology. This chapter also explains different methods for building ontology. The details about the existing ontologies are also listed in later sections of this chapter.

3.1 What is Ontology?

The American Heritage Dictionary defines ontology as "*the branch of metaphysics that deals with the nature of being*" (Swartout, 1999, p.18) or a systematic explanation of being (Corcho et al, 2003). Ontology has been inherited from philosophy. Philosophically, ontology is a systematic account of being or existence. There have been many attempts done to define ontology and definition of ontology has been evolved over time. The first definition was given by Neches and colleagues (1991) "*an ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary*" (p.42).

Wielinga and Schreiber (1993) defined ontology as "*a theory of what entities can exist in the mind of a knowledgeable agent*". Gruber (1993) defines ontology as a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. Ontology is defined as "*concise and unambiguous description of principle relevant entities with their potential, valid relations to each other*" by Uscold and Gruninger (1996, p. 94). Later on van Heijst and colleagues (1997) extended Gruber's definition of ontology as "*an explicit knowledge level specification of conceptualization, which may be affected by the particular*

domain and task it is intended for" (p. 229). Borst in 1997 modified Gruber's definition and defined ontology as a formal specification of a shared conceptualization.

Studer and colleagues (1998) merged these definitions and defined ontology as "*a formal, explicit specification of a shared conceptualization. Conceptualization refers to an abstract model of some phenomenon in the world by having indentified the relevant concepts of the phenomenon. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine readable. Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group*" (p. 185). Chandrasekaran et al (1999) defined ontology as content theories about the sorts of objects, properties of objects and relations between objects that are possible in a specified domain of knowledge.

In bioinformatics, ontology is defined as knowledge engineering and knowledge sharing repository. Ontology aims to discover new knowledge from existing knowledge. Ontology is a database to collect existing knowledge in one domain and to share and reuse for knowledge discovery.

3.2 Applications of ontology

Ontology is a hierarchical structure of concepts and their relationships built in order to extract new knowledge. Ontology in terms of bioinformatics can be interpreted as the representation of existing knowledge of life and the discovery of new knowledge. Ontology is concerned with making information and knowledge explicit; it includes descriptions of concepts and their relationships.

The main applications of ontologies are:

- Extracting and collecting knowledge;
- Sharing knowledge and reusing formally represented knowledge among systems (Yu, 2006; Fensel 2004; Chandrasekaran et al, 1999; Guarino, 1997);
- Terminology management (Yu, 2006);
- Establishing a vocabulary for representing knowledge (Chadrasekaran et al, 1999);
- Data storage, retrieval and analysis (Baker et al, 1999);
- Finding relationships between concepts;
- Knowledge repository;
- Discovering new knowledge;
- Reuse knowledge for decision support systems (Yu,2006);
- Integration, interoperability (Yu, 2006);
- Making conceptual information computationally available (Baker et al, 1999).

3.3 Tools for developing an ontology

There are a number of tools and environments available for building ontologies. These tools are aimed at providing support for the ontology development process and for the subsequent ontology usage. The very first tool for building ontology was Ontolingua server and was developed at the Knowledge Systems Laboratory at Stanford University (Farquhar et al, 1996). At the same time another tool, Ontosaurus, was developed by the Information Science Institute at the University of South California (Swartout et al, 1997). Later other tools were developed to build ontologies; namely: Tadzebao and Webonto (Domingue,

1998), Protégé 2000, WebODE (Arpirez et al, 2001), OntoEdit (Sure et al, 2002), OILEd (Bechhofer et al, 2001), DUET (Kogut et al, 2002), Chimaera (McGuinness et al, 2000). Duineveld and colleagues (1999) did comparative analysis of Ontolingua, Webonto, Protégé, OntoSaurus and WebODE. Corcho and colleagues (2003) compared different ontology building tools. Table 3.1 presents a comparison of most of tools available for building ontology.

Ontolingua was built to ease the development of Ontolingua ontologies with a frame based web application. Initially the main module in the Ontolingua server was the ontology editor, later on other modules like Webster, an equation solver, an open knowledge based connectivity (OKBC) server and Chimaera (an ontology merging tool) were included. The ontology editor also provides translators to languages such as Loom, Prolog, CORBA's IDL, CLIPS etc. An ontology in Ontoligua Server can be edited remotely (Chaudhri et al, 1998).

OntoSaurus was developed by Information Sciences Institute at the University of South California. It consists of two modules; one is an ontology server and the other is a web browser for Loom ontologies. The ontology server uses Loom as its knowledge representation system. It translates from Loom to Ontolingua, KIF, KRSS and C++. These can be accessed with the OKBC protocol.

Webonto was developed by the Knowledge Media Institute (KMI) at the Open University. Webonto is an ontology editor for OCML ontologies. It supports the editing and browsing of ontologies collaboratively, allowing synchronous and asynchronous discussions about developed ontologies over the web, using a standard web browser. The Ontolingua server, Ontosaurus and Webonto have strong relationships with specific languages (Ontolingua, LOOM and OCML, respectively).

WebODE was developed at the Artificial Intelligence Lab at the Technical University of Madrid (UPM) and is the successor of the Ontology Design Environment (ODE) (Blazquez et al, 1998). WebODE cannot be used as a standalone, as it is a web-based interface. WebODE can be used by all the services and applications plugged into the server especially by WebODE's Ontology Editor.

OntoEdit has been developed by AIFB at Karlsruhe University. It is an extensible and flexible environment based on plug-in architecture. Plugins provide functionality to browser and edit ontologies. There are two versions of OntoEdit available; one is Ontoedit Free and other one is OntoEdit Professional. OilEd was developed by University of Manchester as an ontology editor. OilEd has evolved and is now called DAML+OIL. DUET was developed by AT and T Government Solutions Advanced Systems Group. This tool is suitable for database designers and system engineers who can model their ontologies with UML and then translate them into DAML +OIL which can be applied to software systems.

Protégé 2000 was the ontology tool developed by Stanford Medical Informatics at Stanford University. It is an open source, standalone with an extensible architecture. There are many plug-ins available for storage of files, importation and exportation (FLogic, Jess, OIL, XML and Prolog) and OKBC access.

Protégé is extensible and provides a stable, robust infrastructure for more specific research in knowledge-based systems (Gennari et al, 2001). The Protégé tool has evolved and the latest version, Protégé 3, is more user-friendly and has more plug-ins, such as Datamaster for incorporating large data files inside an ontology.

Table 3.1

Comparison of ontology development tools.

	Ontololingua	**OntoSaurus**	**Webonto**	**Protege**	**WebODE**	**Onto Edit**	**OILEd**	**DUET**
Developers	KSL (Stanford university)	ISI (USC)	KMI	SMI (Stanford University)	UPM	Ontoprise	University of Manchester	ATandT
Availability	Free Web access	Open source evaluation version	Free web access license	Open source	Free web access license	Freeware and Licenses	Freeware	Freeware
Software architechure	Client/server	Client/server	Client/server	Standalone	3-tier	Standalones and client server	Standalone	Plugin
Extensibility	None	None	No	Plugins	Plugins	Plugins	No	No
Ontology storage	Files	Files	File	File DBMS (JDBC)	DBMS (JDBC)	File DBMS (v3.0)	File	No
Import from languages	Ontolingua IDL KIF	Loom IDL ONTO KIF C++	OCML	XML, RDF(S), XML Schema	XML, RDF(S), XML Schema	XML RDF(S) FLogic DAML+OIL	RDF(S), OIL, DAML+OIL	DAML +OIL
Export to languages	KIF-3.0 CLIPS CML ATP CML Rule engine Epikit IDL KSL rule engine LOOM OKBC syntax	Loom IDL ONTO KIF C++	OCML Ontolingua GXL RDF(S) OIL	XML, RDF(S), XML Schema, FLogic, CLIPS, Java, HTML	XML, RDF(S) OIL DAML +OIL CARIN FLogic prolog Jess Java	XML RDF(S) FLogic DAML+OIL SQL-3	OIL RDF(S) DAML+OIL SHIQ Dotty HTML	DAML + OIL
Built-in-interface engine	No	Yes	Yes	Yes (PAL)	Yes (Prolog)	Yes (Onto-Broker)	Yes (FaCT)	No
Graphical views	No	No	Yes	Yes	Yes	No	No	Yes
Zooms	No	No	No	Yes	No	No	No	No
Ontology libraries	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes

The Protégé system is an environment for knowledge-based systems and I have used for building the chronic disease ontology in this research. The reasons for selecting Protégé environment are:

- Protégé is a free open source environment which can be used on windows and Mac platforms;
- Protégé is the most user friendly tool to develop an ontology;
- Protégé allows graphical tools that model classes and their attributes and relationships and to manipulate ontologies;
- Protégé has various inbuilt visualization plug-ins for graphical visualization of the ontology;

- Protégé is not a web-based tool, so it can be used without internet connection once installed on a computer;
- Protégé has extendibility through plug-ins and has a library of plug-ins available;
- Protégé plug-ins provide functionality such as merging;
- Ontology from protégé can be imported into and exported from different formats including XML, UML and RDF (Resource Description Framework).

3.4 Methods for developing ontology

Before developing ontology, an aim is set; for example what has to be achieved by the ontology, what are the main goals of building the ontology. These further provide guidance and a way to proceed with building the ontology. In modern computer science, ontology is a data model that represents a set of concepts, information and data within a domain and the relationships between those concepts. Ontology is used to reason and make inferences about the objects within that domain (Gruber 1993).

Ontology is generally written as a set of definitions of a formal vocabulary of objects and relationships in the given domain. It supports the sharing and reuse of formally represented knowledge among systems (Fensel 2004; Chandrasekaran et al, 1999). In recent years, ontologies have been adopted in many business and scientific communities as a way to share, reuse and process domain knowledge (Fensel 2004). As a database technology, ontologies are commonly coded as triple stores (subject, relationship, object), where a network of objects is formed by relationship linkages, as a way of storing semantic information (Owens 2005; Berners Lee et al 2001).

There have been many methods for developing ontologies. The main methods include; the Cyc knowledge base development (Lenat and Guha, 1990), the Enterprise Ontology (Uschold and King, 1995), The Toronto Virtual Enterprise (TOVE) project ontology (Gruninger and Fox, 1995), the Espirit KACTUS project (Bernaras et al, 1996), METHONTOLOGY (Gomez-Perez et al, 1996), the SENSUS ontology (Swartout et al, 1997) and the On-To-Knowledge methodology (Staab et al, 2001).

Cyc knowledge base was one of the first computational ontologies (Lenat, 1995). Cyc development consists of three phases; the first one is to manual codification of articles and pieces of knowledge, the second and third phases consist of acquiring new, common sense knowledge using natural language or machine learning tools. In the Enterprise ontology, four steps are proposed; the first one to identify the purpose of the ontology, the second to build the ontology, the third to evaluate and the fourth is to document the ontology. This mainly consists of capturing knowledge, coding it and integrating it. This method also includes three strategies for identifying the main concepts in the ontology: a top-down approach, a bottom-up approach, and a middle-out approach. In Top-down, the most abstract concepts are identified, then refined into more specific concepts; In the bottom-up approach, the most specific concepts are identified first and then generalized into more abstract concepts and in the middle-out approach, the most important concepts are indentified first and then generalized and specialized into other concepts.

The Tronto Virtual Enterprise project (The TOVE) is a very formal method and is inspired by the development of knowledge-based systems using first order logic. It includes intuitively identifying the main scenarios i.e. main applications

of the ontology; then using a set of competency questions to determine the scope of the ontology. This method can be used as a guide to transform informal scenarios into computable models.

The method proposed in the KACTUS project includes a bottom-up strategy in which a knowledge base is built specifically for a particular application. First the goal of the ontology is set and then ontology is built. On the other hand, Sensus includes top-down approach for deriving domain-specific ontologies from huge ontologies. This method promotes sharability of knowledge as the same base ontology is used to develop ontologies for particular domains.

METHONTOLOGY is a method that enables construction of ontologies at the knowledge level for building ontologies either from scratch, reusing other ontologies as they are or by re-engineering them. It includes identification of the ontology development process, a life cycle based on evolving prototypes and particular techniques.

The On-To-Knowledge methodology includes the identification of the goals to be achieved by knowledge management tools. This method is mainly based on an analysis of usage scenarios. The main steps involved here are; kick-off (to identify the ontology's requirements and competency questions), refinement (to produce a mature and application-oriented ontology), evaluation (to check requirements and competency questions, and check the ontology) and ontology management.

Corcho and colleagues (2003) suggest that none of these approaches are fully mature and are not unified: each group applies their own approach. A lot of effort is required to create a methodology for building ontology.

3.5 Existing ontologies

Several medical ontologies (Pisanelli 2004) have been created including the Open Bio-medical Ontology (OBO) (http://www.bioontology.org/), the Gene Ontology (GO) (http://www.geneontology.org/) (Ashburner et al 2000), the Food ontology (Cantais et al, 2005), Microarray Gene Expression Data (MGED) ontology (MO) (http://mged.sourceforge.net/) (Whetzel et al, 2006(b)), Ontology for Functional Genomics Investigations (FuGo) (http://fugo.sourceforge.net/ontologyInfo/ontology.php) (Whetzel, 2006(a)), the Disease ontology.

OBO: The goal of a Biomedical Ontology is to allow scientists to create, disseminate, and manage biomedical information and knowledge in machine process-able form to accessing and use this biomedical information in research. OBO includes about 60 domain specific ontologies (Smith et al, 2007) associated with phenotypic and bio-molecular information.

GO: The Gene Ontology (GO) project provides a controlled vocabulary to describe gene and gene product attributes in any organism (Shegogue and Zheng, 2005). The GO project is an effort to address the need for consistent descriptions of gene products in different databases. The project began as collaboration between three model organism databases, FlyBase (Drosophila), the Saccharomyces Genome Database (SGD), and the Mouse Genome Database (MGD), in 1998. Since then, the GO Consortium has grown to include many databases, including several of the world's major repositories for plant, animal (mouse, rat), human, and microbial genomes. But this is not all. GO explains three structural vocabularies for genes, biological process, molecular process and cellular function of genes in one model. According to the 2007

update of the world-wide molecular database collection, there are 968 freely available gene/protein related databases (Galperin 2007). Since 2004, a total of 110-170 databases have been added each year (Galperin 2007, 2006, 2005). Therefore intelligent integration of relevant knowledge needs to be embodied in any bio-data ontology that deals with personalized decision support.

Food Ontology: The Food ontology was developed by the Personalized Information Platform for Health and Life Sciences (PIPS) project funded by the European commission. The aim behind the food ontology for diabetes control is to build health care delivery models for health and knowledge services support.

The main domains of the Food Ontology are medical, food and nutrition, patient record and data, treatments and products available. This ontology will provide nutritional advice for diabetic patients by telling how nutrition affects health and how the type of nutrition can be changed. The Food ontology has been modeled using the Protégé environment and the methodology used is the Ontology 101 development process (Noy and McGuiness, 2001). The Food ontology presents all kinds of foods available with their nutritional information and daily recommendations and comprises of 177 classes, 53 properties and 632 instances (Cantais et al, 2005)

MO: MO (MEGD ontology) has been developed by the Microarray Gene Expression Data (MGED) Society. MO is an annotation resource for microarray data and contains 229 classes, 110 properties and 658 instances. MO was further expanded to include clinical, epidemiological and biomedical imaging in FuGo.

FuGo: FuGo is an effort to provide resources for functional genomics investigations, annotations which include design, protocols, instruments, data and type of analysis (Whetzel et al, 2006(a)). FuGo is the collaborative project of the MO working group, the MGED reporting structure for Biological Investigations (RSBI), the HUPO Proteomics Standards Initiative and the Metabolomic Society. FuGo aims to provide a mechanism for annotation of functional genomics experiments which comprise different biological and technological domains (Whetzel et al, 2006(a)).

The Disease Ontology is a controlled medical vocabulary designed to facilitate the mapping of diseases and associated conditions to particular medical codes such as ICD9CM, SNOMED and others (http://diseaseontology.sourceforge.net/). The Disease Ontology can also be used to associate model organism phenotypes to human disease as well as for medical record mining.

Simultaneously with the emerging need for standardized nomenclatures and concept ontologies for biosciences, the new science of systems biology has emerged. It is needed for the grand unification of biological (and medical) knowledge for basic and applied research. Importantly, systems biology is the ultimate tool for describing metabolic and genetic networks interacting with environmental variables to produce phenotypes of all organisms, including health and disease in individuals. Systems biology knowledge is essential for both personalized medicine and molecular epidemiology studies of human diseases in stratified populations (Nicholson 2006). In such systems, biological knowledge needs to be represented, stored and analyzed in a standardized

ontological framework, so that data from different domains of biology and medicine can be properly integrated.

A standardized ontology framework makes data easily available for advanced methods of analysis, including artificial intelligence algorithms, that can tackle the multitude of large and complex datasets for clustering, classification, and rule inference for biomedical and bioinformatics applications.

3.6 Conclusion

In the present research, an effort has been made to create an ontology database for a chronic disease oriented domain with clinical, genetic and nutritional information related to three chronic diseases. The aim behind the building of the chronic disease ontology (CDO) is to share and collect all the existing knowledge related to three chronic diseases in one domain and to extract new knowledge and reuse knowledge to build a personalized model to predict risk for the respective diseases.

This chapter explained the existing tools and methods available for building ontologies. The Protégé tool has been used to build the chronic disease ontology. A mixture of all the above mentioned methods, existing ontologies and the protégé platform has been used for construction of the chronic disease ontology. The chronic disease ontology is described in detail in the next chapter.

Chapter 4. A Novel Chronic Disease Ontology (CDO) for Information Storage and Knowledge Discovery

The first section of this chapter includes a detailed explanation of the chronic disease ontology and its detailed domains and knowledge discovery that have been created through this PhD study. The aim of the ontology is to store and reuse the knowledge related to three chronic diseases particularly the genetic information and to discover new knowledge from it.

4.1 Chronic Disease Ontology (CDO)

Chronic conditions develop over the course of a life-time in the presence of a number of interrelated factors including; genetic predisposition, nutrition and lifestyle. With the development and completion of human genome sequencing, we are able to trace the genes responsible for proteins and metabolites that are linked with these diseases. A Protégé-based ontology has been developed for entering and linking concepts and data for chronic diseases. The ontological representation provides the framework into which information about individual patients; disease symptoms, gene maps, diet and life history details can be inputted, and risks, profiles, and recommendations derived.

I have created original chronic disease ontology (CDO) that is an ontology database for three chronic diseases; cardiovascular disease, type 2 diabetes and obesity. The chronic disease ontology (CDO) consists of five main domains namely; organism domain, molecular domain, medical domain, nutritional domain and a biomedical informatics map. These are the five main subclasses of the chronic disease ontology. These subclasses contain further subclasses and instances (Table 4.1). Each subclass has a set of slots which provides

information about each instance. The detailed information about each class and subclass is described in next subsections.

Table 4.1

General structure of the chronic disease ontology sub-domains.

Organism domain	**Molecular domain**	**Medical domain**	**Nutritional domain**	**Biomedical informatics map**
Human	Gene	Disease	Nutrients	Disease gene map
Group	Mutation	Clinical findings	Source	
Population group	Protein	Signs	Function	
Patient group		Symptoms		
		Laboratory tests		

4.1.1 Organism Domain

The organism domain contains information about human which is categorized into two groups; the population group and the patient group. The population group contains information about different populations, and there are fifty different population groups in the ontology, and each population is also linked with the genes and diseases by means of different mutations. Individual patient data is contained in the patient group. Large datasets can be imported with the help of the data-master plug-in and can be stored inside the ontology. Each new patient's data can be added manually in this sub class. The general structure of the organism domain is demonstrated in figure 4.1.

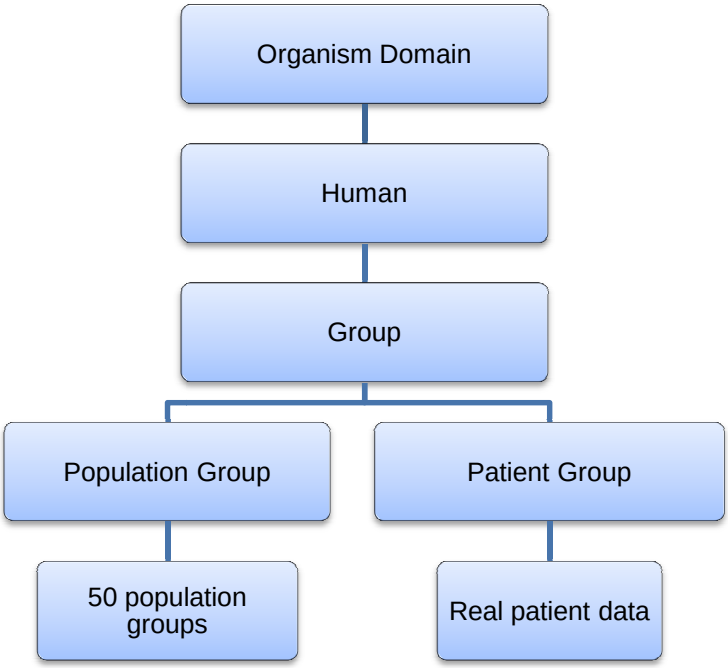


Figure 4.1. The general structure of the organism domain in the chronic disease ontology.

4.1.2 Molecular Domain

Genes are hereditary material which pass on to the next generation and also carry genetic information. If any kind of mutation occurs during the course of transfer that may cause disease. The genes involved in the present ontology, are those involved in the three diseases by different mutations such as single nucleotide polymorphism or deletions. The molecular domain contains detailed information about genes.

Genes have further subclasses; proteins and mutations. The chronic disease ontology contains information about the genes involved in cardiovascular disease, type 2 diabetes and obesity. There are 71 genes in the ontology and these are being regularly updated with newly discovered genes. Relevant information about the genes has been collected from different sources such as

the gene ontology, gene bank, gene cards, and the human gene mutation database. The genes listed in Table 4.2 in the chronic disease ontology contain a lot of information. Each gene has different slots relating to information about the gene and also has relationships with other classes (Figure 4.2).

For each gene, information such as gene symbol, synonyms, chromosomal location, annotations, function, description, proteins produced by the gene, cause of disease, the population they have been studied in so far, responsible for disease, a link to an external source for that gene, frequency in the population and the map to which it is linked. This information can be used for discovering new knowledge.

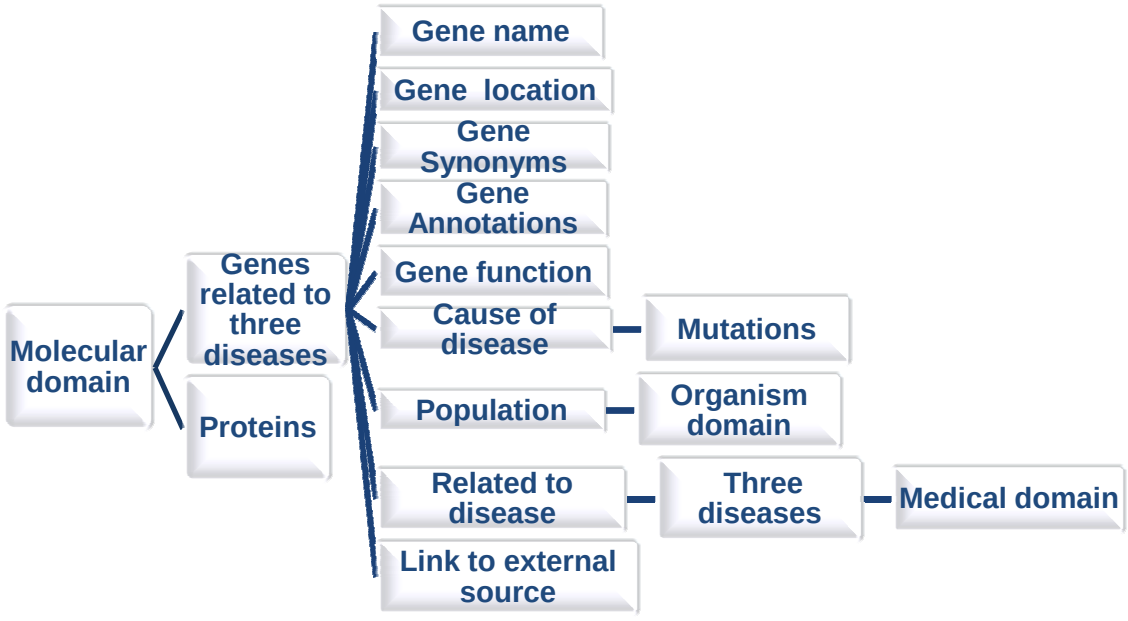


Figure 4.2. General structure of molecular domain in the chronic disease ontology.

Detailed information about all the genes in the chronic disease ontology is listed in Table 4.2.These genes have been found from various sources, literature and combined together. The genes which are common for these three diseases are included in the ontology because these three diseases have many common genes which are responsible for causing these three diseases. The genes present in the chronic disease ontology only contain common genes which are present in all three diseases. There are 71 genes in the ontology which, by means of mutations or polymorphism, can cause cardiovascular disease, type 2 diabetes and obesity. The newly discovered genes can be updated in ontology on regular basis.

Table 4.2

List of genes present in the chronic disease ontology (CDO).

	Gene Symbol	**Gene Name**	**Location**	**Function**
1	ACE	Angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	17q23	Blood pressure regulation
2	ADD1	Adducin 1 (alpha)	4p16.3	Positive regulation of protein binding
3	ADIPOQ	Adiponectin, C1Q and collagen domain containing.	3q27	Positive regulation of cholesterol transport
4	ADM	Adrenomedullin	11p15.4	Blood circulation
5	ADRB2	Adrenergic, beta-2-, receptor, surface	5q31-q32	Protein binding
6	AGT	Angiotensinogen (serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 8)	1q42-q43	Blood pressure regulation
7	AGTR1	Angiotensin II receptor, type 1	3q21-q25	Blood circulation

8	AHSG	Alpha-2-HS-glycoprotein	3q27	Negative regulation of insulin receptor signaling pathway
9	ANGPT4	Angiopoietin 4	20p13	Vascular endothelial growth factor receptor binding
10	ANGPTL3	Angiopoietin-like 3	1p31	Positive regulation of lipid catabolic process
11	APOA4	Apolipoprotein A-IV	11q23	Lipid metabolism
12	APOB	Apolipoprotein B (including Ag(x) antigen)	2p24-p23	Lipid transport
13	APOC3	Apolipoprotein C-III	11q23.1-q23.2	Negative regulation of lipoprotein lipase activity
14	APOE	Apolipoprotein E	19q13.2	Cholesterol homeostasis
15	BBS4	Bardet-Biedl syndrome 4	15q22.3-q23	Heart looping
16	CAPN10	Calpain 10	2q37.3	Positive regulation of glucose import
17	CAPN5	Calpain 5	11q14	Signal transduction
18	CCL2	Chemokine (C-C motif) ligand 2	17q11.2-q21.1	Protein binding
19	CD36	CD36 antigen (collagen type I receptor, thrombospondin receptor), probable pseudogene	7q11.2	Lipid metabolism
20	CEBPA	CCAAT/enhancer binding protein (C/EBP), alpha, probable pseudogene	19q13.1	Generation of precursor metabolites and energy
21	CETP	Cholesteryl ester transfer protein, plasma.	16q21	Cholesterol metabolism

22	CHGA	Chromogranin A (parathyroid secretory protein 1)	14q32	Regulation of blood pressure
23	CRP	C-reactive protein, pentraxin-related	1q21-q23	Low-density lipoprotein binding
24	ENPP1	Ectonucleotide pyrophosphatase/phosphodiesterase 1	6q22-q23	Negative regulation of glucose import
25	FABP2	Fatty acid binding protein 2, intestinal	4q28-q31	Fatty acid metabolism
26	FABP4	Fatty acid binding protein 4, adipocyte	8q21	Fatty acid binding
27	FGB	Fibrinogen, B beta polypeptide	4q28	Blood pressure regulation
28	FGF1	Fibroblast growth factor 1 (acidic)	5q31	Signal transduction
29	FLT1	Fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)	13q12	Vascular endothelial growth factor receptor activity
30	GHRL	Ghrelin precursor	3p26-p25	Positive regulation of body size
31	GNB3	Guanine nucleotide binding protein (G protein), beta polypeptide 3	12p13	Blood pressure regulation
32	HIF1A	Hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)	14q21-q24	Positive regulation of glycolysis
33	HNF4A	Hepatocyte nuclear factor 4, alpha	20q12-q13.1	Lipid metabolism
34	ICAM1	Intercellular adhesion molecule 1 (CD54), human rhinovirus receptor	9p13.3-p13.2	Transmembrane receptor activity
35	IGF1	Insulin-like growth factor 1 (somatomedin C)	12q22-q23	Insulin-like growth factor receptor binding

36	IGFBP1	Insulin-like growth factor binding protein 1	7p13-p12	Insulin-like growth factor receptor binding
37	IL1RN	Interleukin 1 receptor antagonist	2q14.2	Interleukin-1 receptor antagonist activity
38	IL6	Interleukin 6 (interferon, beta 2)	7p21	Cell surface receptor linked signal transduction
39	IL8	Interleukin 8	4q13-q21	Negative regulation of cell proliferation
40	INS	Insulin	11p15.5	Insulin receptor binding
41	INSR	Insulin receptor	19p13.3-p13.2	Insulin binding
42	IRS1	Insulin receptor substrate 1	2q36	Insulin-like growth factor receptor signaling pathway
43	IRS2	Insulin receptor substrate 2	13q34	Glucose metabolism
44	ITLN1	Intelectin 1 (galactofuranose binding)	1q22-q23.5	Positive regulation of glucose import
45	KCNJ11	Potassium inwardly-rectifying channel, subfamily J, member 11	11p15.1	Negative regulation of insulin secretion
46	LEP	Leptin (obesity homolog, mouse)	7q31.3	Energy reserve metabolism
47	LIPC	Lipase, hepatic	15q21-q23	Triglyceride catabolic process
48	LIPE	Lipase, hormone-sensitive	19q13.2	Protein binding
49	LIPG	Lipase, endothelial	18q21.1	Triglyceride metabolic process
50	LMNA	Lamin A/C	1q21.2-q21.3	Structural molecule activity

51	LPL	Lipoprotein lipase	8p22	Lipoprotein lipase activity
52	MMP2	Matrix metallopeptidase 2 (gelatinase A, 72kda gelatinase, 72kda type IV collagenase)	16q13-q21	Protein binding
53	MTHFR	5,10-methylenetetrahydrofolate reductase (NADPH	1p36.3	Blood circulation
54	NAIP	NLR family, apoptosis inhibitory protein	5q13.1	Positive regulation of transcription factor activity
55	NOS3	Nitric oxide synthase 3 (endothelial cell)	7q36	Amino acid metabolism
56	PLTP	Phospholipid transfer protein	20q12-q13.1	Lipid metabolism
57	PPARG	Peroxisome proliferative activated receptor, gamma	3p25	Response to nutrient
58	PRKAA2	Protein kinase, AMP-activated, alpha 2 catalytic subunit.	1p31	Signal transduction
59	PRKAG1	Protein kinase, AMP-activated, gamma 1 non-catalytic subunit	12q12-q14	Protein amino acid phosphorylation
60	RBP4	Retinol binding protein 4, plasma	10q23-q24	Cardiac muscle development
61	RETN	Resistin	19p13.2	Hormone activity
62	SCD	Stearoyl-CoA desaturase (delta-9-desaturase)	10q23-q24	Stearoyl-CoA 9-desaturase activity
63	SELE	Selectin L (lymphocyte adhesion molecule 1)	1q22-q25	Inflammatory response
64	SERPINE1	Serine (or cysteine) proteinase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1.	7q21.3-q22	Blood coagulation

65	SOD3	Superoxide dismutase 3, extracellular	4p16.3-q21	Soluble fraction
66	TCF7L2	Transcription factor 7-like 2 (T-cell specific, HMG-box)	10q25.3	Positive regulation of insulin secretion
67	TGFB1	Transforming growth factor, beta-induced, 68kda.	5q31	Negative regulation of transcription
68	TNF	Tumor necrosis factor (TNF super family, member 2)	6p21.3	Negative regulation of lipid catabolic process
69	TNFRSF1B	Tumor necrosis factor receptor super family, member 1B	1p36.3-p36.2	Protein binding
70	VEGF	Vascular endothelial growth factor	6p12	Signal transduction
71	WFS1	Wolfram syndrome 1 (wolframin)	4p16	Glucose homeostasis

As an example, the information for the gene ACE has been shown as a screenshot in Figure 4.3. From Figure 4.3 it is clear that the ACE gene is involved mainly in blood pressure regulation and therefore can cause cardiovascular disease, type 2 diabetes and obesity. Mutations such as insertion, point mutation or gross deletion in gene ACE can cause the above mentioned diseases. This information is very useful for predicting risk of cardiovascular disease and can be used to predict risk by integrating this with a personalized model. Similarly information about other genes can also be obtained and used for discovering new knowledge.

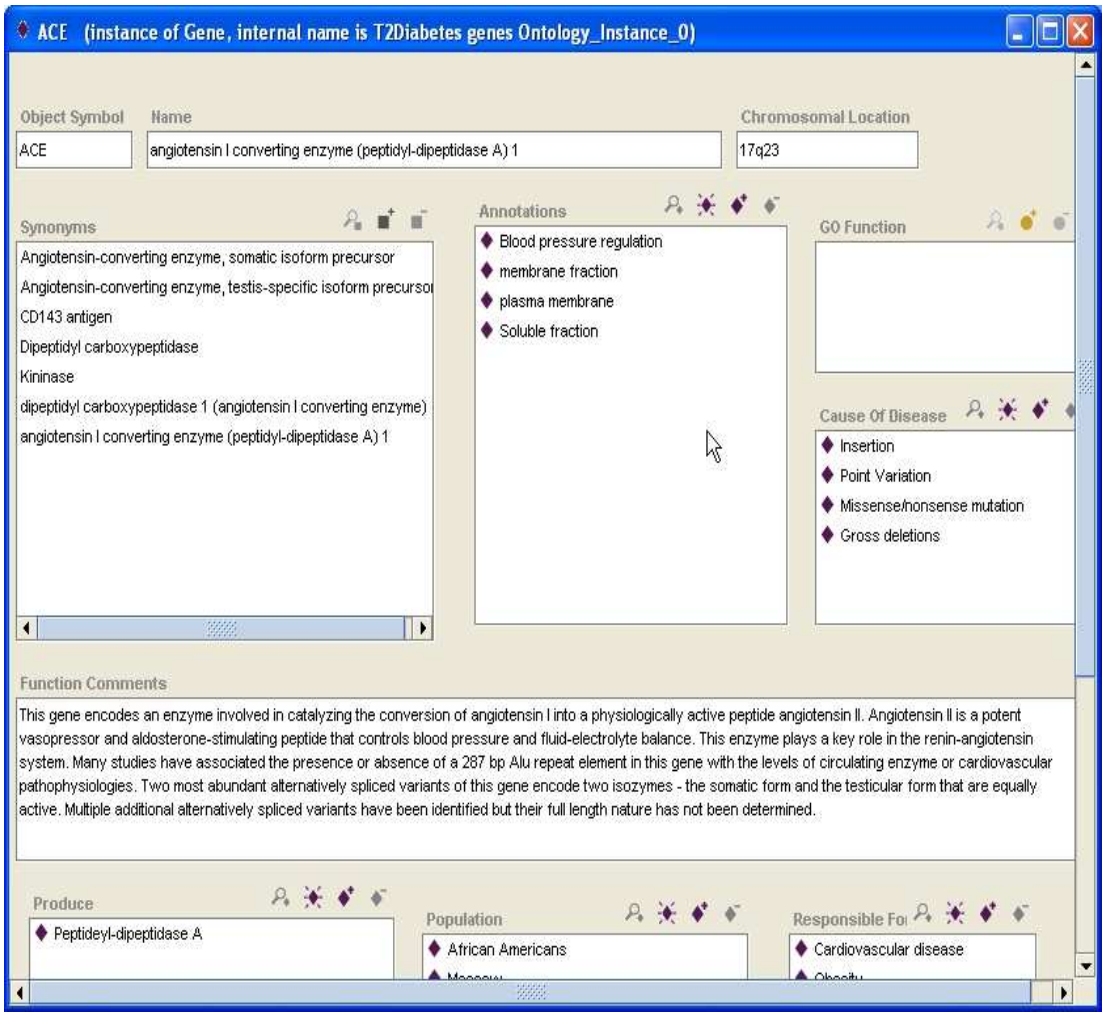


Figure 4.3. A screenshot from the chronic disease ontology showing information about the gene ACE.

The genetic knowledge in the ontology can be used to discover new knowledge and relationships between these genes in relation to mutations and diseases for better prediction and recommendation.

4.1.3 Medical Domain

The medical domain contains information about the three chronic diseases, cardiovascular disease, obesity and type 2 diabetes. This domain contains the description, signs, symptoms and clinical information for the above mentioned

diseases. For example, details about blood pressure, body mass index and many other clinical variables such as blood test reports including total cholesterol, high-density lipo-proteins etc. linked to these diseases are included in this domain explaining more about the diseases and their signs and symptoms.

4.1.4 Nutritional Domain

The nutritional domain contains information about different nutrients and their function inside the body. It also contains information about sources from which these nutrients can be obtained. This domain also includes information about the function of nutrients. The main nutrients in this domain are carbohydrates, fats, minerals, proteins and vitamins.

4.1.5 Biomedical informatics map

The clinical, genetic and nutritional information within the chronic disease ontology can be better understood by creating maps (which explain the relationship of concepts) of the diseases and the information inside the ontology. This domain of the ontology contains disease gene maps created for understanding relationships between genes and types of mutations which cause disease and uses them to discover new knowledge. Figure 4.4 is an example of 10 genes used to create a biomedical gene map of various mutations involved in causing type 2 diabetes.

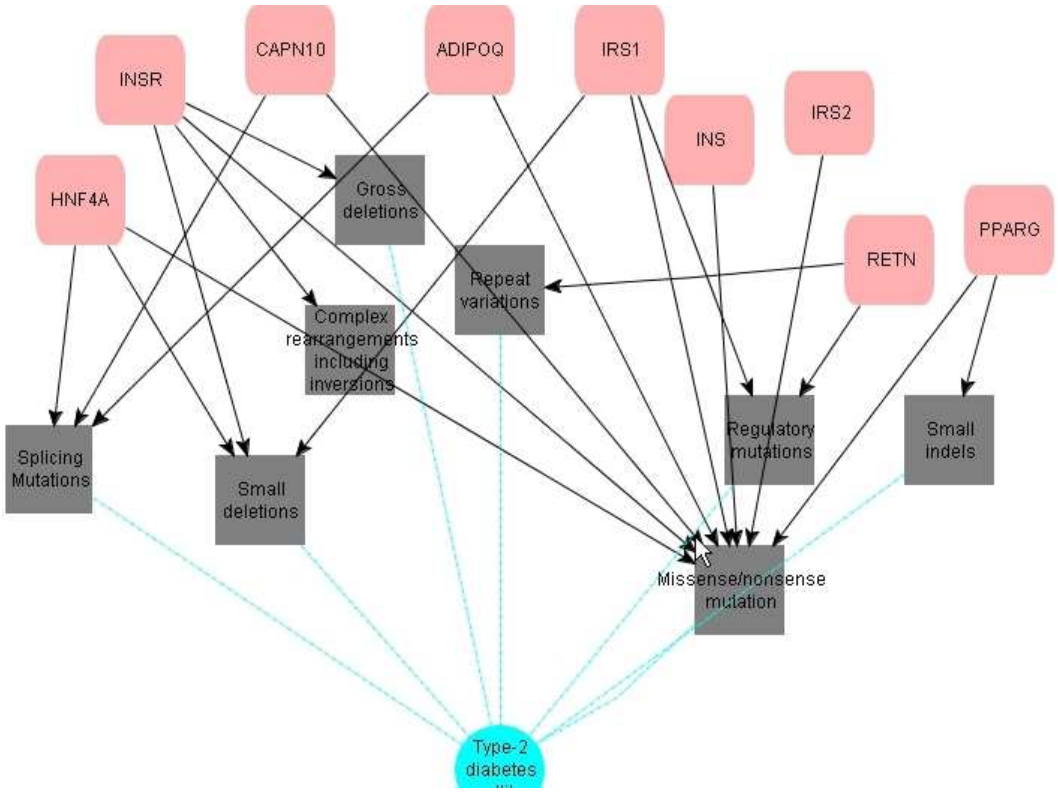


Figure 4.4. Picture of a disease gene map for type-2 diabetes showing few genes related to type 2 diabetes through various mutations.

It can be seen from the map that most of the genes cause type-2 diabetes by means of missense or nonsense mutation. Missense or nonsense mutation is defined as single base-pair substitutions in coding regions which are presented in terms of a triplet change with an additional flanking base included if the mutated base lies in either the first or third position of the triplet.

This example only shows the existence of missense mutation and comparison with other type of mutations for these 10 genes only as this has been created for illustration purpose only. So by creating map for all the genes present in the ontology, most common type of mutation causing type 2 diabetes or cardiovascular disease can be identified.

4.1.6 Information retrieval

The information stored inside the ontology can be retrieved and used for further analysis. The information can be obtained by using a query tool. This query tool looks inside the ontology from the point of view of instance slots such as gene function, gene location on the chromosome, cause of disease (types of mutations etc.), presence in population type and relation to disease. Figure 4.5 shows a screenshot of the query tool with one example.

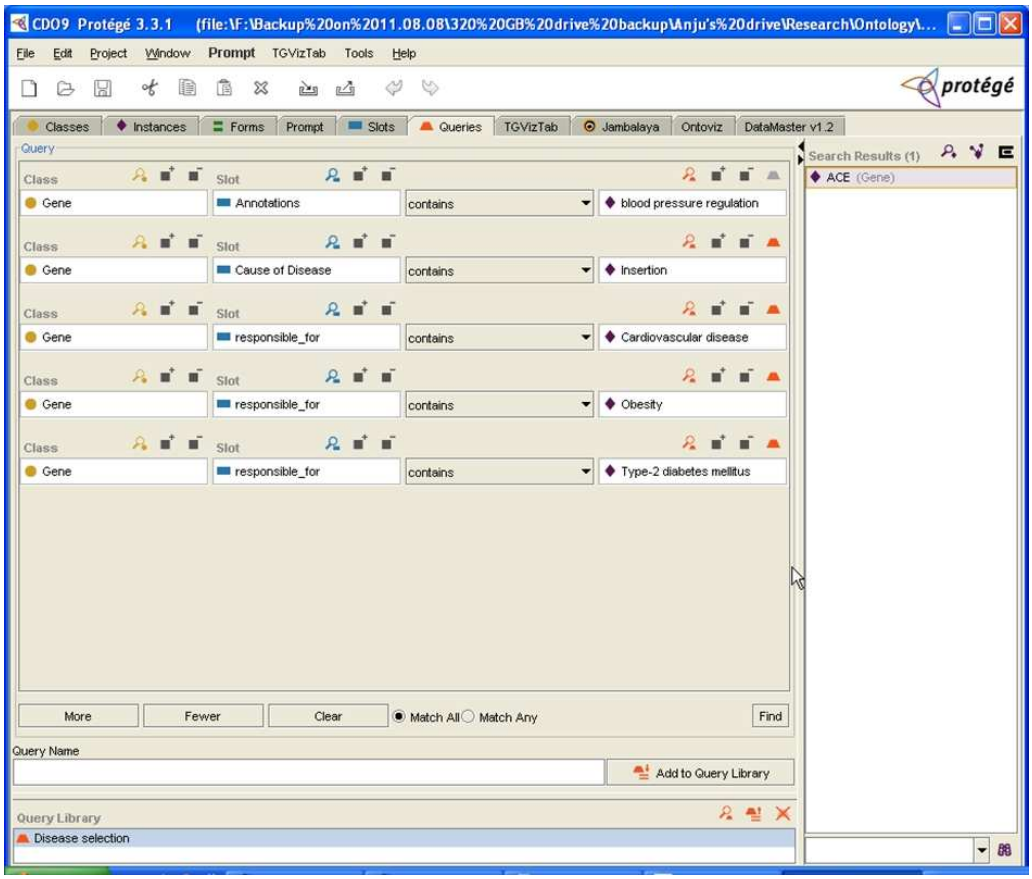


Figure 4.5. A screenshot of an example of the query tool showing a gene list responsible for the regulation of blood pressure and causing cardiovascular disease, obesity and type 2 diabetes by means of insertion (a type of mutation).

For example, if we try to find genes related to blood pressure regulation and by means of deletion, cause cardiovascular disease, the information retrieved from the ontology is a list of genes involved in blood pressure regulation; AGT, FGB, GNB3 and ACE. But the gene involved in blood pressure regulation, which by means of insertion cause the three chronic diseases is only ACE gene. Information about other genes can also be similarly retrieved.

4.1.7 Visualization of the ontology

The graphical presentation of relations in the ontology can be used to navigate, browse and visualize the information available within the chronic disease ontology.

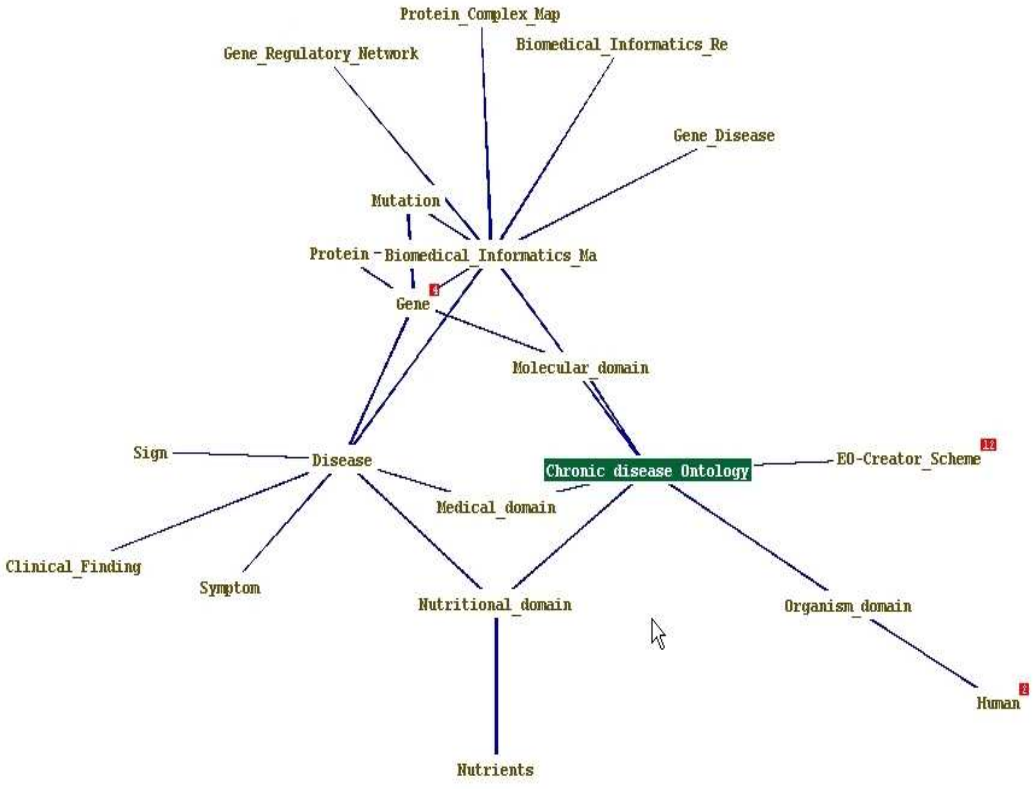


Figure 4.6. Visualization for the structure of the chronic disease ontology using TGViz plug-in.

Visualization explains the general structure of the ontology and also the concepts and relationships existing among the genes and diseases in the CDO. In Protégé, there are many plug-ins that can be used to navigate, browse and visualize existing knowledge. An example of the chronic disease ontology visualization with TGViz is illustrated in Figure 4.6.

Using TGViz plug-in, a particular instance or a particular class can be selected and displayed in the form of a hierarchical graph. The vizualisation picture reveals the links and relationship between different classes of ontology domain.

4.2 Knowledge discovery through the chronic disease ontology (CDO)

The chronic disease ontology (CDO) has been created in order to collect, store and reuse the information for new discoveries in the field of bioinformatics. The aim behind building the CDO is:

- To collect, store, share and reuse knowledge about three chronic diseases (particularly the genes related to these diseases) in one place, as the information about these chronic diseases and related genes is available in various web sources and literature. The aim is to collate all this information in one place. Combined information for these diseases and particular genes is not available at one place. The chronic disease ontology contains collated information from different sources and this information can be reused to discover new knowledge. The chronic disease ontology serves as a knowledge repository.

- To extract new knowledge and to reuse the knowledge for further discoveries. The information stored in the chronic disease ontology can be extracted and reused.
- **Example 1:** If one wants to obtain a list of genes that are present on chromosome 2, a list of about 41 genes is obtained by inserting chromosome 2 in the query slot (Figure 4.7).

  Once the list is obtained, by double clicking on each gene, information about the genes can be obtained. This information can be used by genetic companies and the cost of gene sequencing can be reduced by sequencing genes related to the three chronic diseases on the same chromosome.

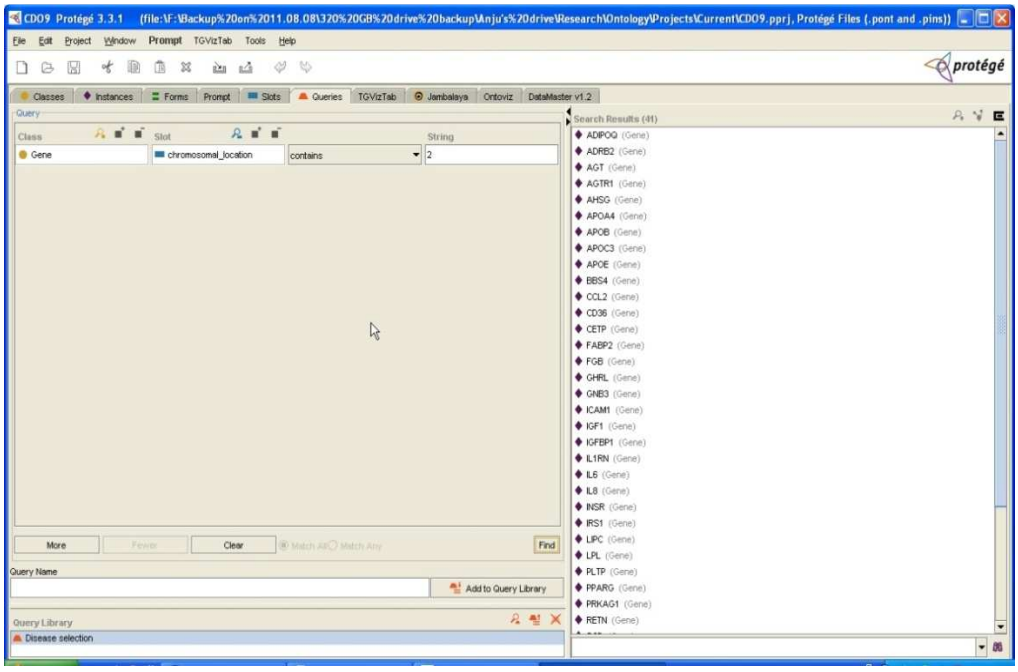


Figure 4.7. A screenshot of an example of a gene list obtained from the chronic disease ontology at chromosome 2.

Similarly, sequencing for similar kinds of mutations can also be cost effective. Studying a gene mutation for a particular disease in a particular

population can also be useful for sequencing and analysis. More particular information can be obtained by adding more queries at one time such as cause of disease, annotations or population in which these genes are present. The number of genes can be reduced by adding more options in the query tool such as kind of mutations. After obtaining a list of genes on chromosome 2, for example, if one wants to see the genes on the chromosome with dinucleotide repeat polymorphism, the list of 41 genes is reduced to only 9 genes (Figure 4.8).

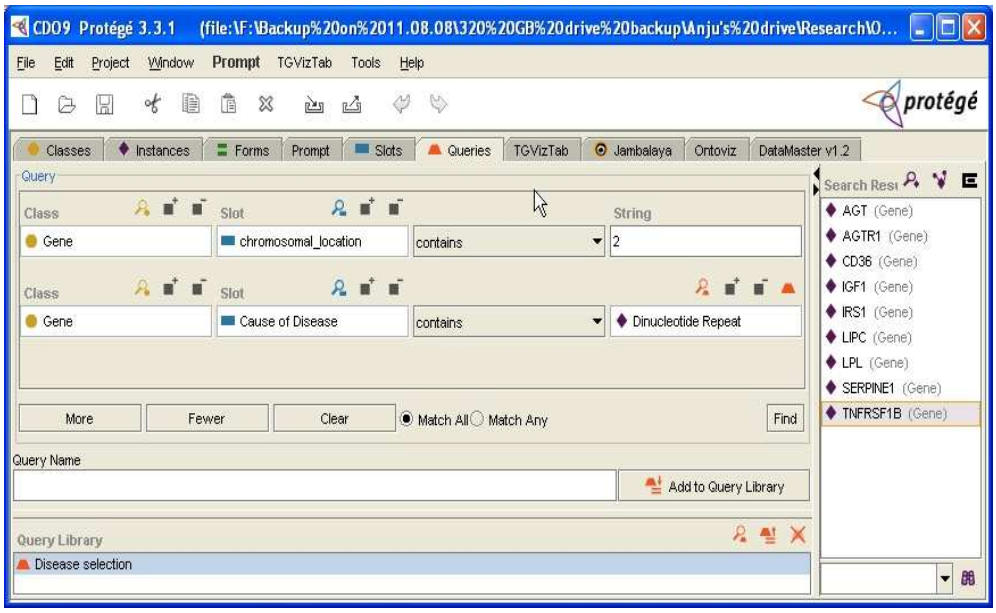


Figure 4.8. A screenshot of a list of genes present on chromosome 2 in the chronic disease ontology which cause disease by dinucleotide repeat mutation.

- **Example 2:** If one needs to obtain a list of genes involved in blood circulation, a list of seven genes ADM, AGTR1, APOA4, APOB, APOE, LPL and MTHFR is discovered (Figure 4.9).

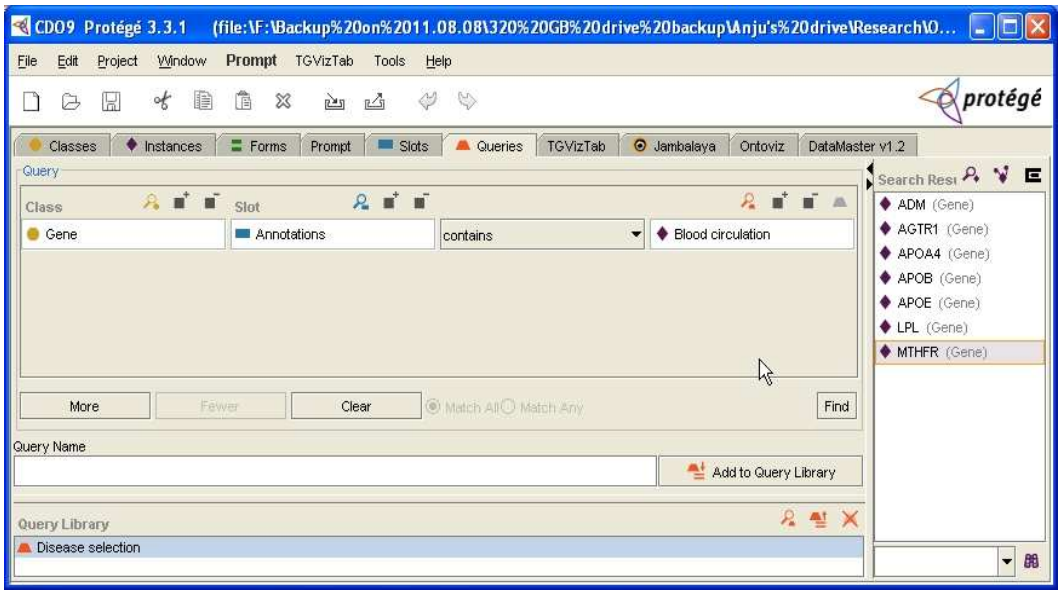


Figure 4.9. A screenshot of a list of genes involved in blood circulation obtained from the chronic disease ontology.

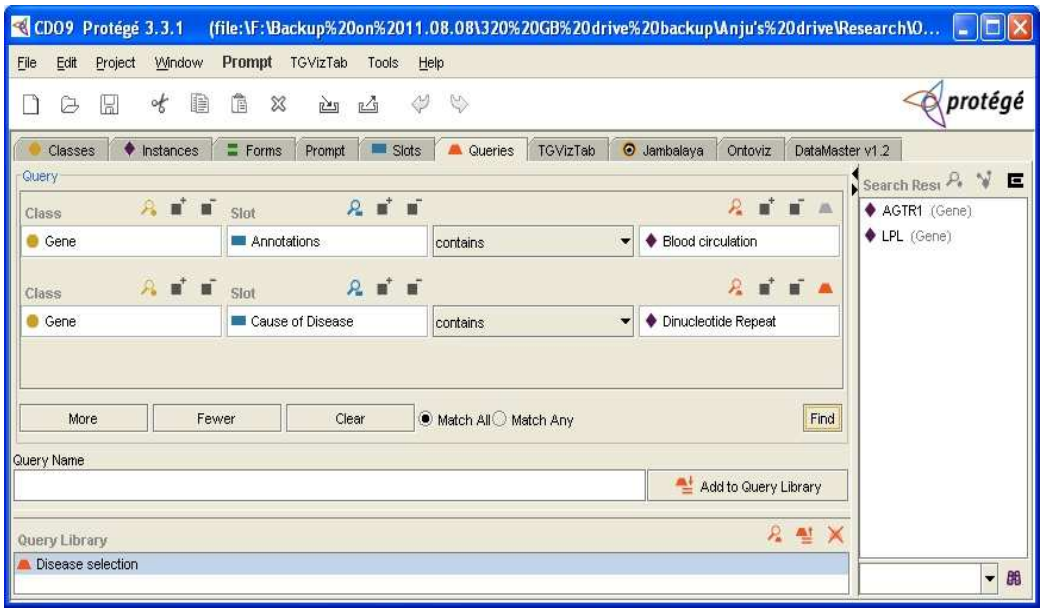


Figure 4.10. A screenshot of a list of genes (AGTR1 gene and LPL gene) involved in blood circulation that cause disease by dinucleotide repeat mutation.

Mutations in these genes are responsible for causing cardiovascular disease. To obtain information for about genes involved in cardiovascular disease through a particular mutation again another query slot can be added. It has been found that only 2 genes AGTR1 and LPL cause disease by means of dinuncleotide repeat polymorphism (Figure 4.10).

The genes AGTR1 and LPL in the normal form are involved in blood circulation. When tandem repeat occur in any of these genes, a pattern of two or more nucleotides is repeated and the repetitions are directly adjacent to each other and may cause disease.

Similarly more knowledge about genes and their relations can be obtained by selecting different slots such as population in which the mutations have been discovered or by changing different mutation types in the query and can be used for further discoveries. The information retrieved can be used in integration with personalized models for better prediction and recommendations.

4.3 Summary

This chapter explained the domains of the chronic disease ontology and gave examples of new discoveries made through the ontology (such as different genes responsible for causing cardiovascular disease or type 2 diabetes on the same chromosome or with similar kind of mutations).

The aim of the chronic disease ontology is to store, reuse knowledge and discover new knowledge and to use it in integration with a personalized model. The next chapter explains the framework for integrating the chronic disease ontology with a personalized model.

Chapter 5. An Integrated Framework of Ontology and Personalized Modelling for Knowledge Discovery

This chapter explains the framework for integrating the chronic disease ontology and a personalized risk evaluation system. The challenge is to create computational platforms that dynamically integrate the ontology and a set of efficient machine learning methods, including new methods for personalized modeling that would manifest better accuracy at a personal level and facilitate new discoveries in the field of bioinformatics.

5.1 Integration framework for ontology and personalized modeling

A novel ontology based decision support framework and a development platform is described in this chapter, which allows for the creation of global knowledge representation for local and personalized modeling and decision support. The main modules are: an ontology module, a machine learning module and an interface to import and export knowledge from and to ontology. The ontology and machine learning module evolve through continuous learning from new data. Results from the machine learning procedures can be entered back into the ontology thus enriching its knowledge base and facilitating new discoveries**.**

The framework presented here and the software platform bring together ontology knowledge repository and machine learning techniques to facilitate sophisticated adaptive data and information storage, retrieval, modeling and knowledge discovery. The framework utilizes ontology based data, as well as new knowledge inferred from the data embedded in the ontology. The platform allows for the adaptation of an existing knowledge base to new data sources and through entering results from machine learning and reasoning models. A

generic diagram of the framework is shown in Figure 5.1. It consists of three main modules: an ontology knowledge and data repository module, a machine learning module and interface. The interface module between ontology and machine learning module is specific for every application such as importing knowledge from ontology module and using it in machine learning module and importing knowledge gained from machine learning module and feeding the new knowledge to ontology module.

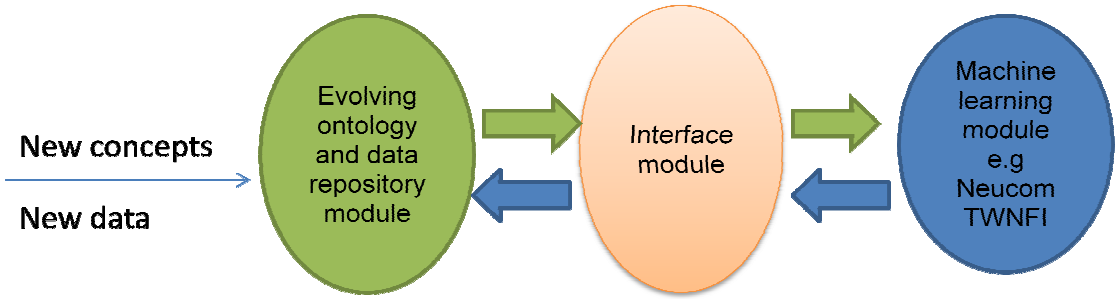


Figure 5.1. The ontology-based personalized decision support (OBPDS) framework consisting of three interconnected parts: (1) An ontology/database module; (2) Interface module; (3) A machine learning module.

The general framework from Figure 5.1 is implemented as a software platform with the following characteristics:

- Protégé ontology development environment
- Data import module to enter external multimodal data into ontology
- Data retrieval module to search and retrieve relevant data from an ontology

- Machine inference module that includes local and personalized techniques such as those described in Chapter 2, included for example in the decision support environment NeuCom (Appendix E).
- User-friendly interface modules that can be tailored to specific applications in different knowledge domains.
- A module for updating the ontology, based on classification and clustering results from the machine inference module.

A sample implementation schema of an ontology-based personalized decision support system in biomedicine is shown in Figure 5.2.

The system from Figure 5.2 is able to combine data from numerous sources to provide individualized patient/case reports and recommendations on actions/interventions to help modify the outcome in a desired direction based on previously accumulated information on input variables and outcomes in the database.

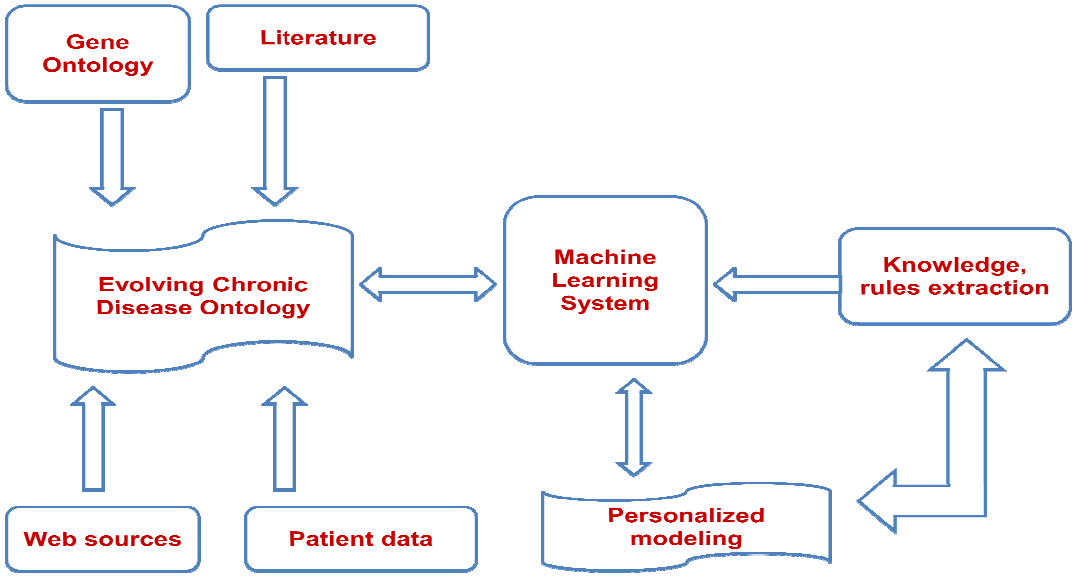


Figure 5.2. The general framework for the ontology based personalized risk evaluation system.

In keeping with the overall vision of bringing together machine learning and ontologies into a single integrated environment, we propose using the patterns produced by the machine learning module to refine the structure of the existing ontology. One way of doing this is to extract relevant features from a database that resides within an ontology to create local profiles, and then to enter the extracted features and profiles back into the ontology in order to enrich it and to discover new relationships.

Feature selection has long been known to be a key success factor in improving the accuracy of the classification/prediction process in machine learning (Kasabov, 2008; Kasabov, 2007(a), 2002; Witten and Frank, 2000). Since ontologies link related concepts together, they can be used to extract a set of related features of different kinds (e.g. clinical, genetic, cognitive, etc.) for a particular machine learning model. For example, in classifying whether patients are at high risk or low risk of contracting heart disease, an ontology such as the Chronic Disease Ontology (CDO), described earlier in Chapter 4 can be used to determine all the currently known risk factors (encompassing the clinical, genomic and demographic data types). Since the predictors used are acknowledged to be the best that are currently known, we could expect performance to improve over uninformed or adhoc methods of feature selection from only a single database.

A major challenge is using the newly discovered knowledge to further evolve existing ontologies. In general, the knowledge extracted from machine learning methods falls into three distinct categories; those that refer to:

1) concepts that already exist in the ontology

2) concepts not covered by the existing ontology

3) changes in the nature of existing concepts in the ontology

In terms of category 1, no changes need to be made to the existing ontology. Categories 2 and 3 pose significant challenges as they could represent knowledge hitherto unknown to the knowledge engineer. The naïve approach, immediately refining the ontology, may not be desirable, given that the ontology represents the collective wisdom and knowledge of world-class domain experts gained through their life experiences. A more prudent approach would be to monitor such knowledge over a period of time and only update the ontology when a clear and consistent trend emerges that shows that such knowledge persistently improves the accuracy of predictions on newly arriving data. The rank aggregation technique proposed by Domshlak et al (2007) and the knowledge pattern technique proposed by Clark et al (2004) provides us with the right tools for assessing when changes should be made to the existing ontology.

The problem of linking ontologies with machine learning systems requires the building of a specific interface. To enable a machine learning workbench to automatically obtain the right data, there should be shared contextual "understanding" between the learning system and the ontology itself, as each may have their own contextual meaning which may differ from the other. Thus the integration of these local contexts is yet another challenging issue, and, as discussed by Maamar, et al. (2006), and Satyanarayanan (2001), should address the following issues: how can changes in a concept be detected; how should the context be found and stored within the systems/data; how should the context be taken into account; how should an inference engine obtain sufficient

information to act in a context-aware manner. A further issue is the mutual trust between the system and user / data source; and whether the system retrieves accurate and relevant information.

Local and transductive inference methods only focus on a small area of data space and its relevant information (Kasabov, 2007(a, b); Song and Kasabov, 2006). Thus new incoming data will dynamically change the contextual meaning of the information within the database; especially when the new data point is being introduced near the area of interest. This can lead to changes in how the data is being clustered, or it may strengthen a particular cluster. Either way, the changes will affect the ontology, because as the data changes, the representation depicted by the ontology will need to be updated. Therefore, to accommodate the dynamics of the data, the ontology must be able to evolve. Evolving the ontology involves modifying the originally designed ontology based on the knowledge and new clustering discovered during the inference (Gottgtroy et al 2006).

In general, the ontology evolution process can be classified into conceptual changes, and explication changes (Lenzerini, et al 2004). Conceptual changes deal with and include new concepts or relationships which are emerging or flagging already existing concepts which display a diminishing level of support from new data streaming in; while explication changes focus on modifications to the description of the concepts, such as adding a new description or property to a concept. As a general guide, Uschold and Grüninger (1996) and Maedche (2002) offer good frameworks for ontology building and learning. However, here, the main interest is in evaluating and refining the frameworks; to assure

that the evolved ontology will still reflect the real world which it represents, as well as to refine the process in order to support its evolving nature.

At an early stage, ontology evolution focuses on the ontology learning process by proving from the machine learning process. In its subsequent stages, the system will grow and further evolve. This includes the ability of the machine learning module to automatically select appropriate data from the database and for the ontology to detect newly emerging concepts or relationships. For instance, the patient's health and medical data stored in the database might be stored in several separate tables, thus the ontology and context mediation system will help the machine learning module (e.g. NeuCom) to collect data from relevant columns and tables based on the information and relationships described in the ontology. For example, if the user wants to perform chronic disease analysis, a context mediation system can be used to ensure that the system will collect all of the right information about chronic diseases, but not about other functions.

After the machine learning has performed its analysis on a given data set, and identified new relationships, these new findings will be fed back into the ontology and will be noted. However, this doesn't mean that this new relationship will immediately be acknowledged as new concepts. It will be noted as possible discovery but confirmed by further evidence to establish its status.

Sometimes, when we analyze a set of data using one methodology, for example numerical prediction, it may not show any new findings, but if we combine it with the result of some other methodology, such as pattern recognition or clustering on the same data set, the combined results may reveal new insights. These new insights can then be used to update and/or evolve the

ontology. Therefore, the capability of the system to evolve is not just limited to a certain learning method, or findings.

The implementation of this technique will raise the issue of how one can be sure that the particular concepts or relationships already have enough evidence to be claimed as new findings. As we are using the rank and weighting methods to overcome this issue, we believe that by adopting the rank aggregation technique proposed by Domshlak et al (2007) will help us ensure that the ontology evolution process will not go amiss. We will also adopt the knowledge pattern technique proposed by Clark et al (2004), to help us ascertain whether the emerging concepts fit with certain knowledge patterns and are reliable new findings.

5.2 Knowledge discovery through the integration of personalized modeling tools and the chronic disease ontology (CDO)

The platform described above can be used to create ontology and simulation systems for various bioinformatics and biomedical applications, such as, chronic disease (e.g. heart disease, obesity, diabetes) and personal risk evaluation. The following examples illustrate knowledge discovery through the integration of the chronic disease ontology and a personalized model.

- Example 1: Use of knowledge from the chronic disease ontology (CDO) within personalized model. The new discoveries made through the chronic disease ontology explained in Chapter 4 can be used for disease risk evaluation. Figure 5.3 shows a generic diagram how the new knowledge discovered from the chronic disease ontology can be used in integration with machine learning methods (personalized modeling methods).

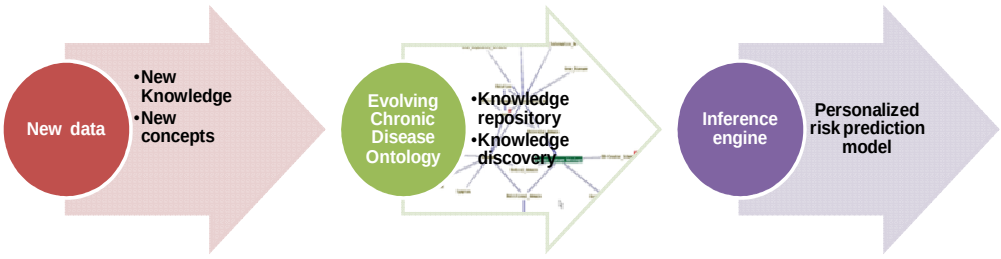


Figure 5.3. Example of framework for use of knowledge from the chronic disease ontology (CDO) to personalized model.

The ACE gene (Angiotensin I converting enzyme) controls blood pressure and fluid-electrolyte balance. Different mutations such as insertion or point mutations in the ACE gene can lead to cardiovascular disease particularly in the African-American and Moscow populations. So if the new subject comes from African-American or Moscow population for personalized risk evaluation, the system looks for the information and recommends that if the subject has the ACE gene mutation, either insertion or point mutation or gross deletion, the person is at high risk of having cardiovascular disease.

- Example 2: Use of knowledge from the personalized model for cardiovascular disease within the chronic disease ontology (CDO) and reuse with subsequent subjects. The information discovered from the personalized model, particularly TWNFI (explained in Chapter 6 for cardiovascular risk evaluation), in the form of

personalized profiles or rules is inputted into the ontology and is used when a new subject comes with a similar profile (Figure 5.4).

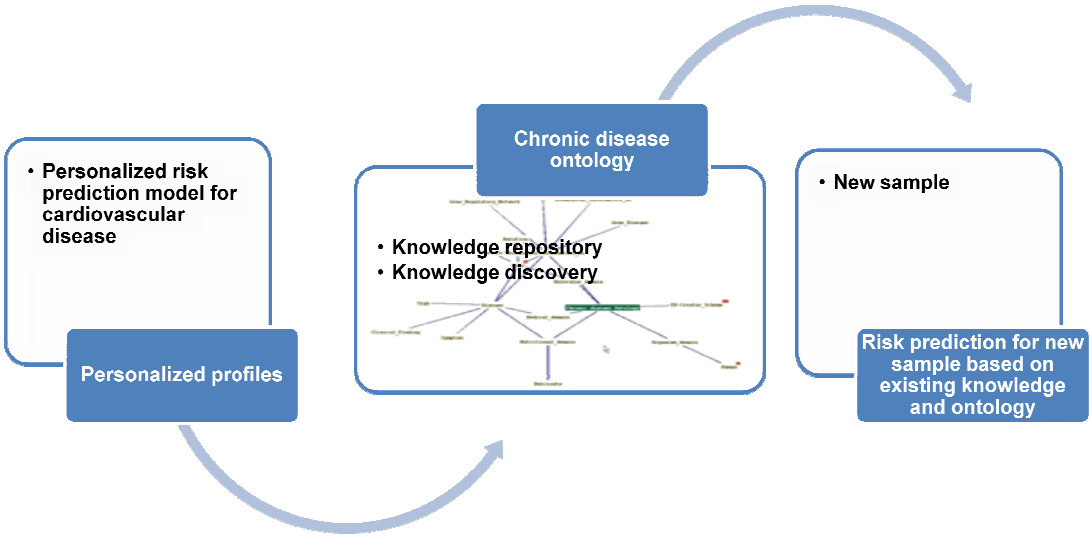


Figure 5.4. An example of utilization of knowledge from the personalized risk evaluation model for cardiovascular disease within the chronic disease ontology (CDO) and reuse for subsequent subjects.

It has been found that a high ratio of total cholesterol and HDL and waist circumference are the major risk determinant factors for cardiovascular disease. So these factors are linked with certain genes in the chronic disease ontology such as the gene ADIPOQ (Adiponectin). This gene is expressed in adipose tissue and the encoded protein from the gene ADIPOQ circulates in the plasma and is involved with metabolic and hormonal processes. ADIPOQ regulates energy homeostasis and glucose and lipid metabolism. ADIPOQ is responsible for positive regulation of cholesterol transportation and mutations, particularly single nucleotide polymorphism or substitution in the gene, can alter the regulation of cholesterol and is a direct cause of cardiovascular disease. If a

new subject with a high ratio of total cholesterol and HDL, is linked to gene ADIPOQ mutation they can be advised accordingly.

- Example 3: Use of knowledge from the personalized model for type 2 diabetes within the chronic disease ontology (CDO) and reuse for subsequent subjects (Figure 5.5). The information discovered from gene data in Chapter 7 can be inputted in the ontology and can be used for any new subject. It has been found that genes associated with male and female subjects are different in the Italian population. This information has been added to the chronic disease ontology (CDO) and can be used for any new subject with a similar profile.

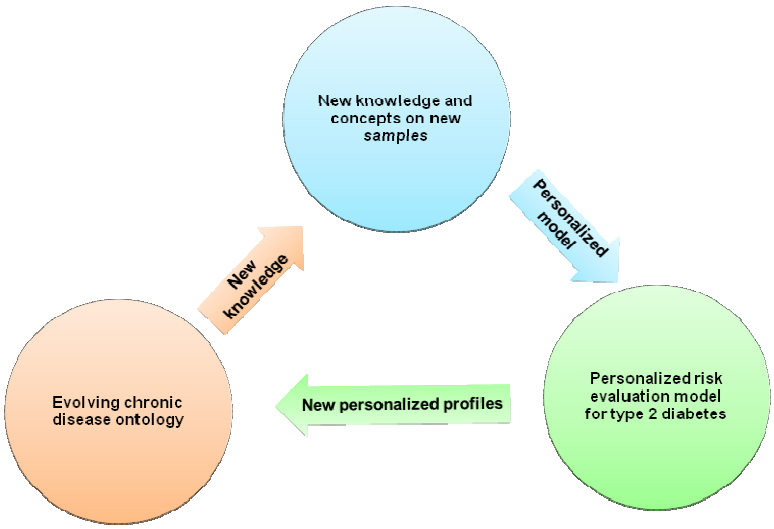


Figure 5.5. An example of use of knowledge from the personalized model for type 2 diabetes within the chronic disease ontology (CDO) and reuse for subsequent subjects.

For example, a male subject with high cholesterol or fasting blood glucose is directly linked to gene tumor necrosis factor (TNF) (responsible for regulation of

insulin secretion) and is at high risk of having type 2 diabetes in the Italian population. Mutation, particularly, point mutation in gene TNF is responsible for negative regulation of insulin secretion and is a cause of type 2 diabetes. This information is stored inside the ontology and can be used to predict type 2 diabetes risk and recommendations for new subjects with a similar profile.

It has been found that genes causing type 2 diabetes in male and female population are different. So each set of genes can be included in the ontology and can be used for integration with the personalized modeling. Furthermore, the new knowledge and profiles for different subjects can be added to the ontology and can be used for risk prediction and recommendations for type 2 diabetes for new subjects with similar profiles.

5.3 Conclusion

Chapter 5 explained the general framework for integration of the chronic disease ontology with personalized risk analysis modeling. The knowledge from the ontology and the personalized model can be used for new discoveries and better prediction. The integration of the ontology and the personalized risk prediction model has been explained with three examples. This chapter provided a general framework for the integration approach proposed in the thesis. The following two chapters explain this integration platform with two case studies for cardiovascular risk evaluation (Chapter 6) and type 2 diabetes risk evaluation (Chapter 7).

Chapter 6. Cardiovascular Disease Risk Evaluation Based on the Chronic Disease Ontology (CDO)

This chapter explains the description of the New Zealand National Nutrition Survey 1997 data, variables used and the processes used to determine the relationships between variables and their association with risk of getting cardiovascular disease. This chapter also explains the existing methods used to predict risk of cardiovascular disease and a new personalized method using demographic, clinical, anthropometric and nutritional variables.

6.1 Cardiovascular disease, prevalence and description

Cardiovascular disease is the most prevalent disease in the world and is the leading cause of mortality (Bonow et al, 2002; Yusuf et al, 2001). According to the World Health Organization in 2005 about 30% of all deaths occur due to cardiovascular disease all over the world (WHO, 2009). Also according to the World Health Organization (WHO) (2009) over 80% of these deaths occurred almost equally in men and women in low and middle income countries such as China and India. By 2015, it is estimated that almost 20 million people will die from cardiovascular disease every year. In New Zealand, cardiovascular disease is the leading cause of death, accounting for 40% of all deaths (New Zealand Guidelines Group, 2003(b)).

Cardiovascular disease represents all the disorders associated with heart and blood vessels. The most common disorders that fall under cardiovascular disease are; coronary artery disease, ischemic heart attack, heart failure, high blood pressure and stroke. A number of risk factors classified as, non-modifiable, modifiable and intermediate risk factors, are associated with cardiovascular disease. Age, sex, ethnicity and genes are non-modifiable or

irreversible factors for cardiovascular disease. The main modifiable risk factors for cardiovascular disease are unhealthy diet, smoking and lack of exercise, which in turn lead to intermediate risk factors such as being overweight, obesity, high blood pressure, raised blood glucose levels and raised blood lipids (Nesto, 2008; Cannon, 2007). According to the WHO (2009), unhealthy diet, smoking and physical inactivity (lifestyle) are associated with about 80% of the cardiovascular disease burden.

The most common feature of cardiovascular disease is the gradual clogging of blood vessels by fatty or fibrous material which forms plaque. This fatty material gradually builds up on the blood vessel walls, narrowing the arteries and reducing their elasticity. This eventually prevents vital oxygen and nutrients from reaching the cells. This condition is often referred to as hardening of the arteries or arteriosclerosis. Any artery in the body can be affected. Blocking of arteries associated with the heart, brain or kidneys, or those to the eyes and legs may cause angina (chest pain), heart attack, stroke, renal failure, blindness or claudication (pain in legs) respectively. Risk of cardiovascular disease is enhanced by unhealthy lifestyle habits such as, food high in saturated fat, sugar, salt, low levels of physical activity and tobacco use (Robitaille et al, 2007; World Health Organization, 2003). Improving diet and increasing physical activity from conception onwards can prevent the major outcomes of cardiovascular disease.

6.2 Existing methods for predicting risk of cardiovascular disease

Historically, efforts have been undertaken to improve quality of life and reduce the complications from cardiovascular disease through creating risk prediction tools that inform treatment. The Framingham heart study has been operational

for more than 50 years (Anderson et al, 1990). Using longitudinal data, several tools have subsequently been developed using “The Anderson equation” to calculate risk of 5 years or 10 years such as Modified Sheffield tables, Joint European charts, New Zealand tables, British charts, Canadian tables, Joint European guidelines. A number of tools in different formats are available including risk charts, tables, online risk calculators which are available as stand-alone or web-based applications for personal computers or as stand-alone applications for personal digital assistants. Most of these tools use input variables such as:

- Age
- Gender
- Systolic blood pressure
- Diastolic blood pressure
- Fasting total cholesterol
- Fasting high-density lipoprotein cholesterol
- Smoking status
- Diabetes status diagnosed from fasting blood glucose/ oral glucose tolerance test

QRISK2 is an example of a recently developed cardiovascular disease risk algorithm for the different ethnic groups of England and Wales. This algorithm also includes all the above mentioned variables in addition to renal disease, arterial fibrillation and rheumatoid arthritis (Hippisley-Cox et al, 2008).

The predicted outcome of these tools varies. Tools such as Sheffield tables, Joint British charts and European charts only calculate the risk of coronary heart disease; with others, such as the New Zealand tables, calculate the risk of

coronary heart disease and stroke, peripheral vascular disease is included as an outcome of the Birmingham Heartlands calculator (Sheridan et al, 2003). All the tools present the information about the prediction of risk differently, as (i) numeric or graphic (ii) supporting explanations (iii) a point estimation of risk. Others, such as the Sheffield tables, provide a range of risks or simply state whether a predefined treatment threshold to initiate therapy had been exceeded. Some tools also provide a comparison of the risk factors to an individual of the same age or sex who has either average risk or no risk. Few tools provide treatment advice or links to evidence based treatment guidelines.

For the New Zealand population, New Zealand cardiovascular risk charts have been developed using The Framingham Heart Study Cardiovascular risk-prediction equation (Milne et al, 2003; Jackson, R., 2000), and a web-based tool, PREDICT-CVD, has subsequently been developed (Bannink et al, 2006). These charts use information like age, gender, blood pressure, total cholesterol to HDL, smoking and diabetes status same as used in the Framingham equation but also include Maori, Pacific and Indian sub-continent ethnicity. These charts predict five year age specific incidence of cardiovascular disease in a cohort of New Zealand men aged 34-74 years and women aged 35-69 years. PREDICT-CVD generates CVD risk for five years and also evidence based risk management recommendations based on national guidelines (Bannink et al, 2006).

All the above-mentioned tools include similar information such as age, sex, total cholesterol, high-density lipoprotein cholesterol, smoking status, diabetes and blood pressure and some include previous medical and family history. It has been observed that the effect of dietary changes on cardiovascular disease

related risk factors such as cholesterol, blood glucose levels, obesity, and blood pressure differs significantly among individuals. Some people are hyper-responders or highly sensitive to dietary interventions while others are insensitive or hypo-responders (Ordovas and Corella, 2007).

By way of personalized recommendation tools for the New Zealand population, an online tool has been developed by a pharmacogenetics company (Theranostic Labs, 2009) to recommend warfarin dosage to cardiovascular patients based on gender, ethnicity, height, weight, age, amiodarone dosage, statins dosage, indication of any cardiac disorder or CYP2C9 and VKORC1 genotype. This tool gives recommendations based on global data but utilizes general, medical and genetic information.

A similar approach has been used to build a personalized risk evaluation in the research presented in this thesis. Nutrients have been used along with the other previously used variables to predict the risk of cardiovascular disease. The aim of the present research is to create personalized predictions of risk of cardiovascular disease and provide nutritional advice for dietary nutrients based on demographic, clinical, anthropometric and nutritional data.

6.3 Data Exploration

1997 New Zealand Nutrition Survey (NNS97) data was collected to profile the nutritional status of adult New Zealanders. Health was assessed via measurements of body size, blood pressure and fasting blood biochemistry analysis. NNS97 has been used for creating a prediction model for cardiovascular disease. This data mainly includes information on nutrient intake and nutrition related clinical measures, anthropometric measures and general information. 1997 New Zealand Nutrition Survey was conducted to gain

information about the nutritional status of the New Zealand population (Quigley and Watts, 1997). The aim of this section of the thesis is to explore the relationships between demographic, anthropometric, nutrition and clinical fasting blood measurements for risk of CVD in the NNS97 data. The outcome variable selected for risk of CVD was elevated blood pressure/hypertension.

6.3.1 Description of selected data

Data available was from 5,613 subjects living in New Zealand belonging to the age range of 15 to 97. For purposes of this analysis only 2,875 complete sets of data were used from the initial 5,613 subjects. 2,738 subjects were not included in data as these subjects had missing values, as a blood sample was not provided. Table 6.1 presents details of the missing data.

Table 6.1

Description of subjects in NNS97 data.

	Number of subjects present in compete dataset	**Number of subjects present in selected data**
Age Group		
15-25 y	895	386
26-35 y	1257	646
36-45 y	1112	608
46-55 y	769	434
56-65 y	673	367
66-75 y	563	293
76-85 y	300	128
86 y+	44	13
Gender		
Male	2310	1305

Female	3304	1570
Ethnicity		
European/Pakeha	3944	2264
Other European	341	177
New Zealand Maori	765	261
Samoan	212	65
Cook Island Maori	93	27
Tongan	61	7
Niuean	23	9
Tokelauan	7	2
Fijian	19	7
Other Pacific Islander	5	2
Southeast Asian	18	5
Chinese	43	13
Indian	30	12
Other Asian	28	12
Other ethnic groups	24	12

Based on features that are associated with a higher risk of cardiovascular disease, 2,875 sets of subject data were considered in the analysis and used to build the risk prediction model. The NNS97 data containing 2,875 subjects included 1,305 male and 1,570 female subjects with each subject having 20 input variables as listed in Table 6.2.

Table 6.2

List of variables from NNS97 data for initial experiments.

General	**Clinical and Blood Analysis**	**Anthropometric measures**	**Dietary Analysis**
Age Gender Ethnicity	Systolic Blood Pressure Diastolic Blood Pressure Pulse Total cholesterol HDL cholesterol Haemoglobin Blood pressure medication	Waist circumference Subscapular skinfold Triceps skinfold	Energy Protein Carbohydrates Sugar Total fat Total saturated fatty acids Salt

6.3.2 Rationale for selecting variables

Grouping and selection of demographic, clinical and dietary variables considered for this analysis are discussed below.

Demographic: Age is the most important factor in determining the risk of cardiovascular disease. It has been observed that blood pressure, a major risk factor for cardiovascular disease, rises progressively with age (Neal et al, 2002). Risk of cardiovascular disease begins in early life, between 35 and 44 years are when manifestations of cardiovascular disease are often detected on screening (Thom et al, 2004). The New Zealand Guidelines Group (2003(b)) recommends risk assessment for the age group of 45 or above for most asymptomatic males and the age group of 55 or above for most asymptomatic females. Different population groups have different risks for cardiovascular disease. Maori, Pacific Islanders or Indians are at a higher risk of cardiovascular disease at a younger age (Whittaker et al, 2006).

Clinical and Blood Analysis: Blood pressure is a direct measure of cardiovascular function as it measures the force of blood against the walls of blood vessels. Systolic blood pressure is the pressure as the heart contracts; diastolic blood pressure is the pressure when the heart relaxes between ventricular contractions. The normal range for blood pressure is considered to be 120 mm Hg systolic and 80 mm Hg diastolic, written as 120/80. According to American Heart Association (2009) and Medicine Net (2009) high blood pressure is referred to as 'The Silent Killer' and is considered as a warning sign for heart attack or stroke.

Raised total cholesterol is also an indicator of increased risk of heart attack. Cholesterol is a waxy, fat like substance, which is found in all cells of the body. Cholesterol is present in hormones, vitamin D, liver salts and all cell membranes. Cholesterol is transported in the blood in small packages called lipoproteins. Broadly speaking, in the blood there are two types of cholesterol, low-density and high-density lipoproteins; LDL and HDL. LDL is often called "bad cholesterol" as with other substances it may accumulate on the walls of arteries. HDL is called "good cholesterol" because high levels are associated with lowered risk of cardiovascular disease. High-density lipoprotein carries blood cholesterol from other parts of the body to the liver and the liver then removes cholesterol from the blood. So the higher the value of HDL cholesterol lowers the risk of cardiovascular disease.

The desirable level of total cholesterol is less than 4.0 millimoles per litre, above 6.0 millimoles per litre is considered very high. High-density lipoprotein cholesterol levels more than 1.0 millimoles per litre are desirable and less than 1.0 millimoles per litre is considered low and a major risk factor for cardiovascular disease (Lipid Management Guidelines, 2001). The ratio of total

cholesterol to HDL cholesterol has been used in current research for determining risk of cardiovascular disease. The ratio of total cholesterol to HDL cholesterol of more than 4.5 millimoles per litre is considered high risk of cardiovascular disease (New Zealand Guidelines Group, 2003(a)).

Anthropometric measures: Waist circumference is one of the most practical tools for assessing abdominal fat for chronic disease risk (Zhu et al, 2005; Wang and Hoy, 2004; Despres, 1990) because anatomically, the fat within the intra-abdominal cavity is closely associated with fat absorbed from the gut and blood supply of the liver. Accumulation of fat in the liver decreases the sensitivity of the liver to insulin (which is associated with the development of type 2 diabetes) and increases the production of very low-density lipoprotein cholesterol (VLDL cholesterol), which is a major risk factor for cardiovascular disease (Zhu et al, 2005).

In Europeans, men with a waist circumference of up to 94 cm is considered low risk, between 94cm and 101 cm, increased risk, greater than or equal to 102 cm, very high risk. For European women a waist circumference up to 80cm is considered low risk, between 80-87 cm, increased risk and more than or equal to 88 cm is very high risk of metabolic complications and obesity (Turley, 2008). It has been confirmed by the International Day for the Evaluation of Abdominal Obesity (IDEA) study that waist circumference is an important predictor, along with blood pressure, body mass index (BMI), cholesterol and blood glucose, of a person's risk of cardiovascular disease (Balkau et al, 2007; Wang and Hoy, 2004). Along with abdominal obesity, truncal obesity is also a major risk factor for cardiovascular disease. The best measure of truncal obesity is the ratio of sub scapular to triceps skinfold measurements (Okosun et al, 2006).

Dietary analysis: The New Zealand Ministry of Health recommends a diet high in fibre and low in fat, sugar and salt for health i.e. prevention and treatment of chronic diseases such as cardiovascular disease (Carson et al, 2004). The modifiable lifestyle factors, diet and physical inactivity directly affect the development of arteriosclerosis, which is the main cause of cardiovascular disease. Altering diet, for example, with the partial substitution of carbohydrates with either protein or monounsaturated fat can lower blood pressure, improve lipid levels, and reduce estimated cardiovascular risk (Appel et al, 2005). Over-eating can result in obesity, which in turn can lead to cardiovascular disease. An imbalance in total energy intake and macronutrients, protein, fat (including saturated fat), carbohydrates (including sugar) is related to the choice and quantity of food eaten each day.

6.3.3 Statistical Analysis

For statistical analysis, each subject has been categorized into a binary variable. There are a number of systems used to detect risk for cardiovascular disease. Two of the most commonly used are the International Diabetes Federation IDF (Alberti, Zimmet and Shaw, 2006) and the Adult Treatment Panel (ATPIII) (Cleeman, 2006). The definition for hypertension for both systems is the same

If a subject has

systolic blood pressure >=130 mmHg

or diastolic blood pressure >=85 mmHg

or treatment for hypertension (takes blood pressure medication), he or she is classified as class 1 which means high risk of developing cardiovascular

disease. All others belong to class 0 which means lower risk of developing cardiovascular disease (CVD).

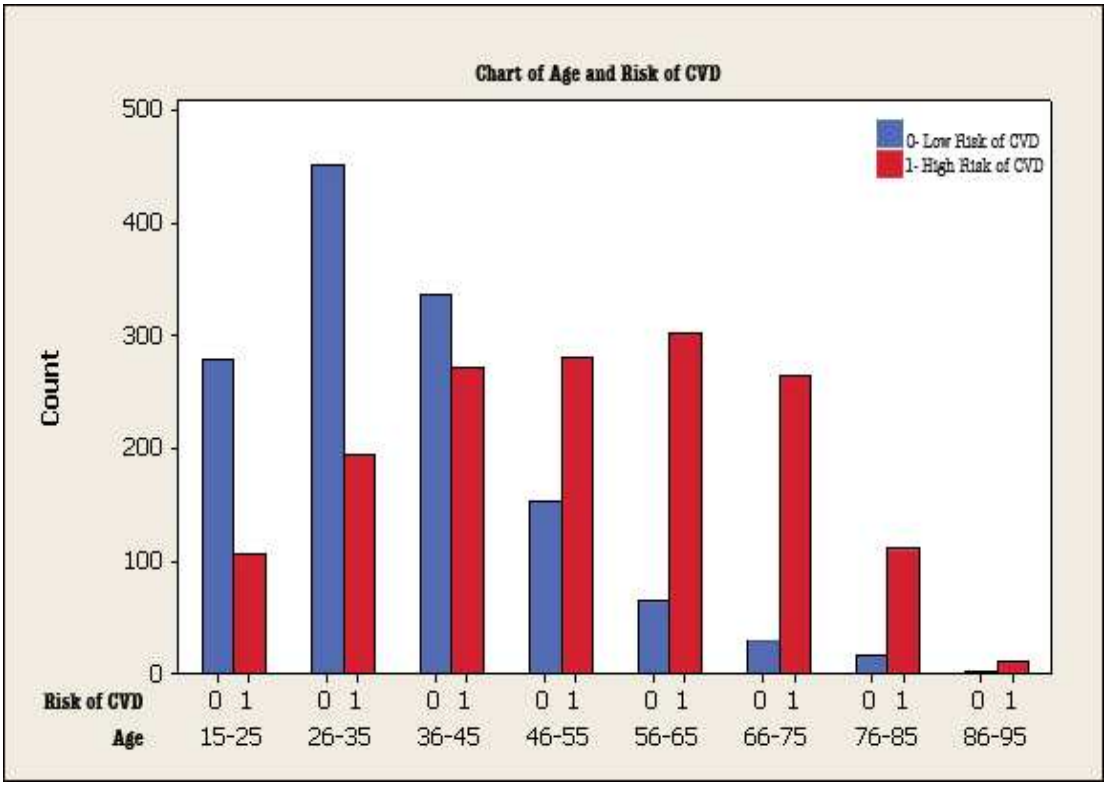


Figure 6.1. Bar graph of NNS97 data for all subjects with age and risk of cardiovascular disease (n=2,875).

The data was then analyzed using different tools. In total there were 2,875 subjects used to develop the models. Risk of CVD (class1) starts at the age of 15 but there are only a few subjects with a high risk of cardiovascular disease under the age of 36 but in most of the cases in the male, female combined data, risk of cardiovascular disease increases after 36 years of age and keeps on increasing with an increase in age (Figure 6.1).

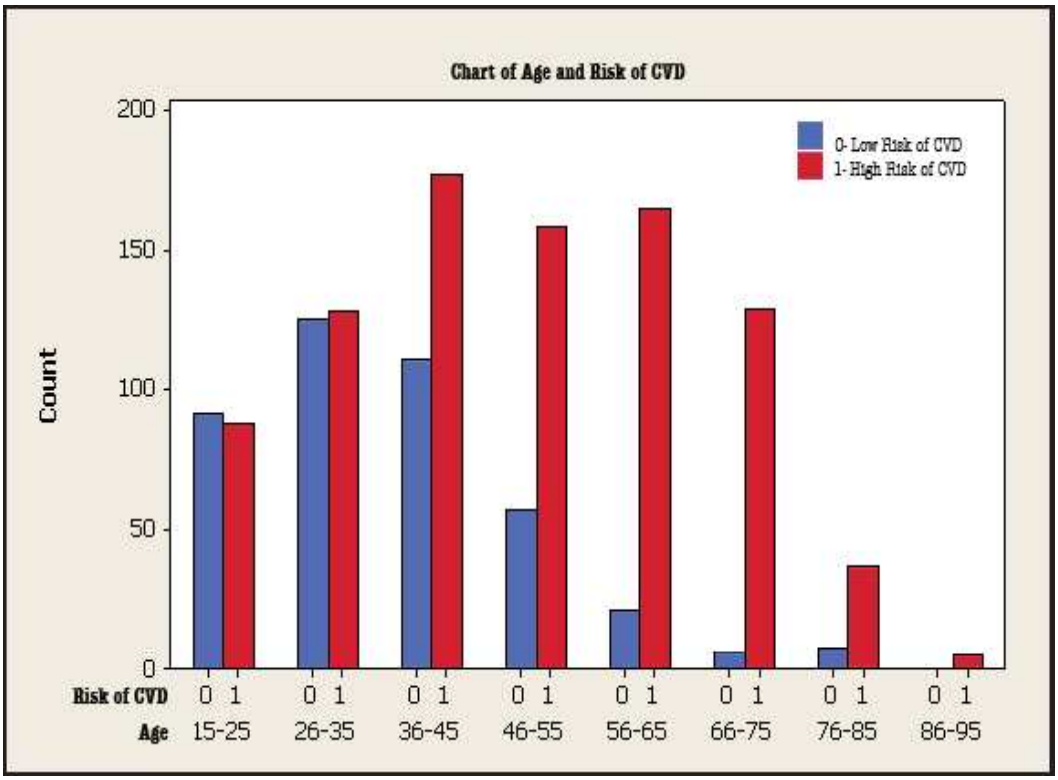


Figure 6.2. Bar graph of NNS97 male data for age and risk of cardiovascular disease (n=1,305).

It has been found from the bar graph (Figure 6.2), in male subjects the risk of cardiovascular disease is higher after 36 years of age and keeps increasing, with increase in age. Male subjects with higher risk of cardiovascular disease are more after the age of 36. While in female subjects, risk of cardiovascular disease is seen to be higher after 46 years of age (Figure 6.3). For male subjects risk of cardiovascular disease appears more after 36 years of age which is earlier than in females, as in female subjects the risk increases after 46 years of age.

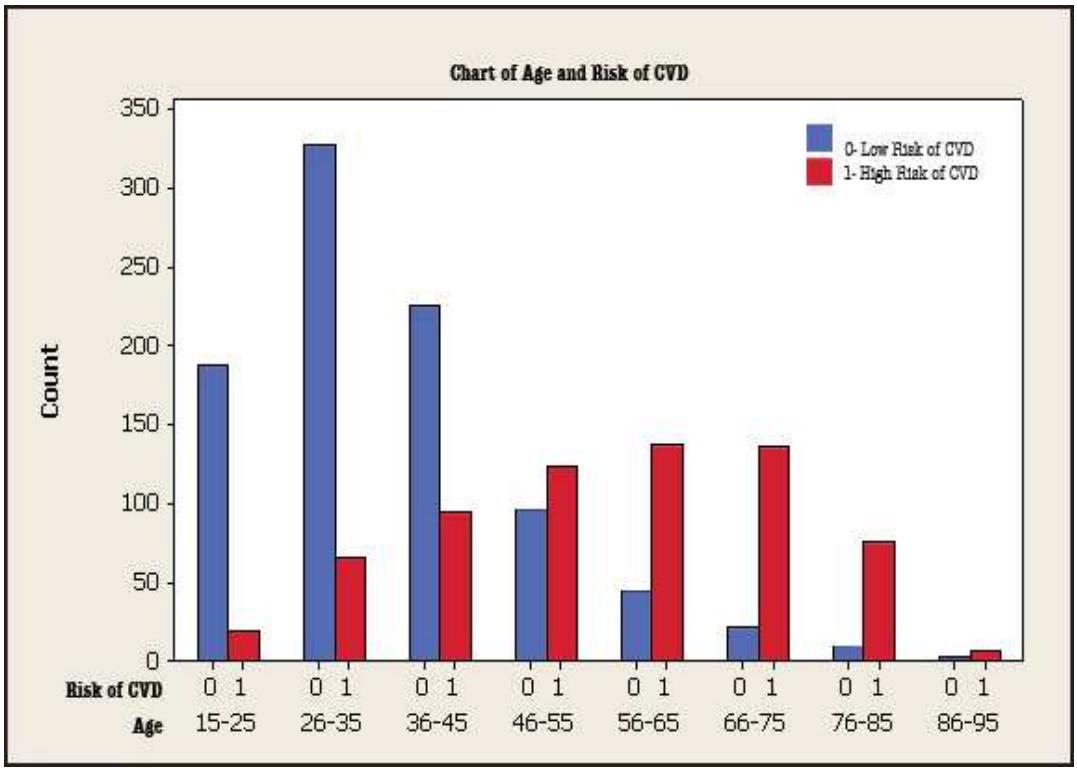


Figure 6.3. Bar graph of NNS97 female data for age and risk of cardiovascular disease (n=1,570).

It has been found that more male subjects are at high risk of cardiovascular disease than female subjects (Table 6.3). As out of total sample at risk of cardiovascular disease, about 57.5% are male subjects and 42.5% are female subjects. It is also clear from Table 6.3 that numbers of subjects at higher risk of cardiovascular disease are more after 45 years of age.

In current data more samples belong to European ethnicity. There are few samples in number of other ethnic groups. For purposes of this analysis and as numbers of non-European were relatively small self-defined ethnicity coded as European (2) and non-European (1), it is clear from the Table 6.3 that Europeans are at high risk of having cardiovascular disease than other ethnic groups.

Table 6.3

Prevalence of hypertension (risk factor for cardiovascular disease) in 2,875 subjects from the National Nutrition Survey 1997.

Age range	Total subjects(n)	Total subjects at high risk %(n)	Total male subjects (out of total subjects at high risk) at high risk % (n)	Total female subjects (out of total subjects at high risk) at high risk % (n)
15-25 y	386	27.7% (107)	88.2% (88)	17.8% (19)
26-35 y	646	30% (194)	66% (128)	34% (66)
36-45 y	608	44.5% (271)	65.3% (177)	34.7% (94)
46-55 y	434	64.8% (281)	56.2% (158)	43.8% (123)
56-65 y	367	82.3% (302)	54.6% (165)	45.4% (137)
66-75 y	293	90.4% (265)	48.7% (129)	51.3% (136)
76-85 y	128	87.5% (112)	33% (37)	67% (75)
86 y+	13	84.6% (11)	45.5% (5)	54.5% (6)
European	2441	54.5% (1330)	56.8% (756)	43.2% (574)
Non-European	434	49% (213)	61.5% (131)	38.5% (82)
Total	2875	53.7% (1543)	57.5% (887)	42.5% (656)

The average values for each variable for all subjects and maximum values for male and female subjects are presented in Table 6.4. The blood pressure of the subjects has been taken three times for each subject and the average value of the three values was taken. Summary descriptive statistics for anthropometric measurements, biochemical analysis of fasting blood and dietary analysis are presented in Table 6.4. On average, males had higher blood pressure, haemoglobin and total cholesterol, and lower HDL cholesterol than females.

Table 6.4

Average, maximum and minimum values of the selected variables in whole, male and female population.

Sr. No.	Variables	Average value	Male			Female		
			Minimum	Maximum	Average	Minimum	Maximum	Average
1.	Age (years)	44	15	97	44	15	93	44
2.	Systolic Blood Pressure	133	98	231	139	85	234	128
3.	Diastolic Blood Pressure	80.30	54	122	83	40	121	77
4.	Pulse (beats/min.)	69	37	109	68	43	113	70
5.	Waist circumference(cm)	87.7	65.4	160.4	93.5	59	150	82.9
6.	Sub scapular/triceps skinfold	1.2	0.2	4.8	1.8	0.2	7.5	1.2
7.	Ratio of total cholesterol/HDL	4.7	1.6	16.4	5.2	1.9	14.9	4.4
8.	Haemoglobin(g/L)	142	105	186	151	82	170	135
9.	% Protein intake	15.8	1.4	51.9	15.6	3.2	50.6	15.9
10.	% Carbohydrate intake	35.1	4.6	64.7	35.3	4.1	71.5	34.9
11.	% Sugar intake	18.5	0.2	242.3	15.9	0.5	184.4	21.3
12.	% Total fat intake	46.1	12.6	90.1	45.1	4.8	89.8	46.9
13.	% Total saturated fat intake	20.9	2.3	79.7	19.9	1.6	89.2	21.9
14.	Salt intake(mg)	3002	123	20845	3666.1	80	13548	2449.4

The ratio of total cholesterol to HDL cholesterol has been calculated by dividing total cholesterol by HDL cholesterol. The average value of ratio of total cholesterol to HDL cholesterol is 4.7 with the highest value of 16.4 among males and lowest of 1.6. In females the highest value of ratio of total cholesterol to HDL cholesterol is 14.8 and the lowest value of 1.86.

For macronutrients (nutrients that provide energy) percentage values relative to total energy intake have been calculated. The formula used to calculate percentage energy provided by protein, sugar and carbohydrates is listed in

appendix D. The average percentage protein intake was 15.8; it has been found that maximum percentage consumption of protein was 51.8 for males and 50.6 for females and the minimum percentage consumption is 1.4 for males and 3.2 for females.

The average percentage intake for total fat is 35.1; ranging from 4.5 to 64.7 in males and 4.1 to 71.5 in females. The average for percentage consumption of saturated fat is 18.5 with maximum of 242.3 in males and 184.4 in females; minimum 0.2 in males and 0.5 consumed in females. Average salt intake in whole data is 3001.7 milligrams. Minimum intake of salt in males has been found 123 milligrams and maximum in males 20845 milligrams. Minimum intake of salt in females is 80 milligrams and maximum is 13548 milligrams.

Software used:

NeuCom (KEDRI, Auckland, New Zealand) was used for detailed analysis of the data. NeuCom is a self-programmable, learning and reasoning computer environment based on connectionist (Neuro-computing) modules. NeuCom learns from data using connectionist modules. The modules can adapt to new incoming data in an on-line, incremental, life-long learning mode, and can extract meaningful rules that help users discover new knowledge in their respective fields (www.theNeuCom.com). Different modules of NeuCom have been used for the data analysis. Details about the various modules in NeuCom are listed in Appendix E.

Signal to noise ratio ranking of variables:

For data analysis, the first step was the signal to noise ratio method to get a quantitative measure for each variable within the whole data. This is a method

that extracts the signals from each variable by comparison with the outcome (in this case hypertension) and ranks the variables according to the strength of association.

Signal to noise ratio gives a quantitative measure of how much each variable in a given data set for classification discriminates one class (considered as the “signal”) from the other class (classes) (considered as “noise”). Variables that have higher values for an output class versus other classes are ranked higher as they have a higher SNR. Inputs to this function are: the data set, and the number of variables to rank. Signal to noise ratio is a well known method used for feature selection (Goh, 2005).

Signal to noise ratio for the case of two-class problem for a variable x is calculated as absolute difference between the mean value (M1x) of the variable for class 1 and the mean value (M2x) of this variable for class 2, divided to the sum of the respective standard deviations. The formula to calculate signal to noise ratio is shown in equation 6.1 (Kasabov, 2007(a)).

$$\mathrm{SNR} = \mathrm{abs}\,(\mathrm{M1x} - \mathrm{M2x})/(\mathrm{Std1x} + \mathrm{Std2x}) \qquad (6.1)$$

This method performs very well for continuous variables (variables with range of values). This method does not perform well for binary (variables which define presence or absence such as yes or no) or categorical or nominal variables (variables which have two categories). Signal to noise ratio gives the ranking of variables, higher the value of SNR, higher the importance of the variable. The ranking of variables can be visualized in the form of bar chart by using NeuCom.

In this section, signal to noise ratio has been used to understand the ranking of already selected variables. Figure 6.4 shows the ranking of all variables based

on the signal to noise ratio for the whole data. According to the signal to noise ratio, age and waist circumference were the most important factors for determining the risk of cardiovascular disease. Ratio of total blood cholesterol to HDL cholesterol is ranked after waist circumference. Gender and ethnicity are included in the data but does not have any significance in signal to noise ratio being categorical variables.

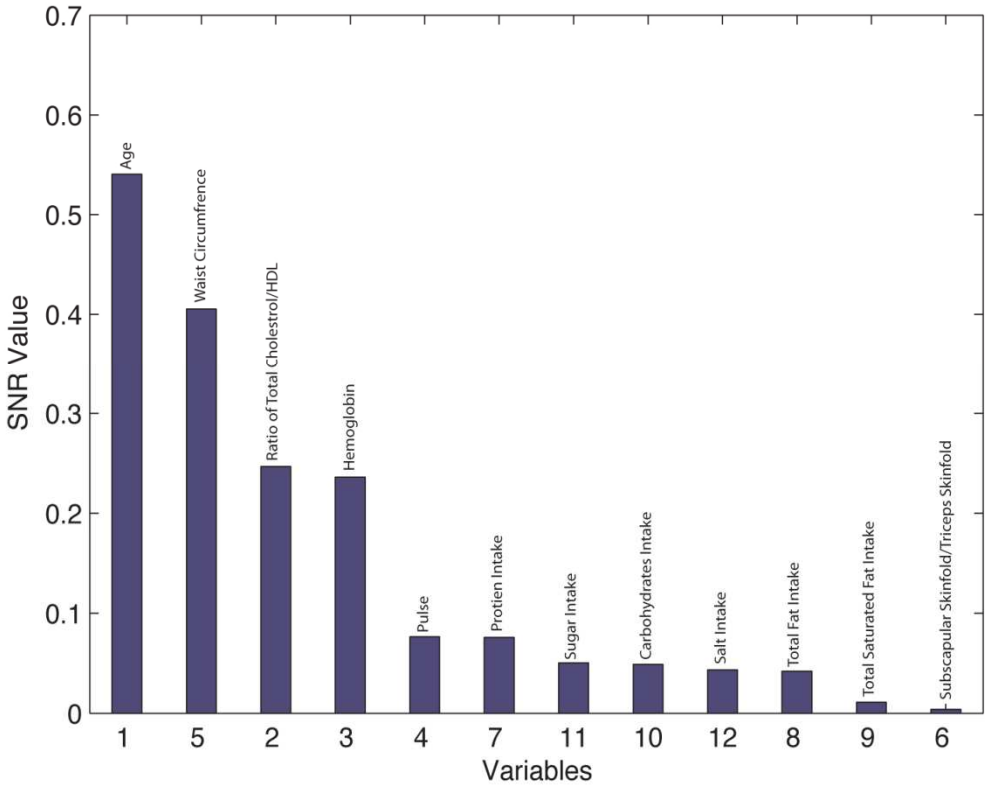


Figure 6.4. Bar graph showing variables ranked (highest to lowest) according to signal to noise ratio for the whole data.

Total saturated fat intake and ratio of sub scapular skinfold and triceps skinfold are the least important factors for the risk of cardiovascular disease. As described in section 6.3.2, males have different risk determinant factors than

females; such as male subjects have high risk for cardiovascular disease at the age of 45 or above and for female subjects is 55 or above, ranking of variables using signal to noise ratio has been done separately for male and female subjects. Figure 6.5 shows the ranking of all variables for male subjects and Figure 6.6 shows the ranking of variables for female subjects.

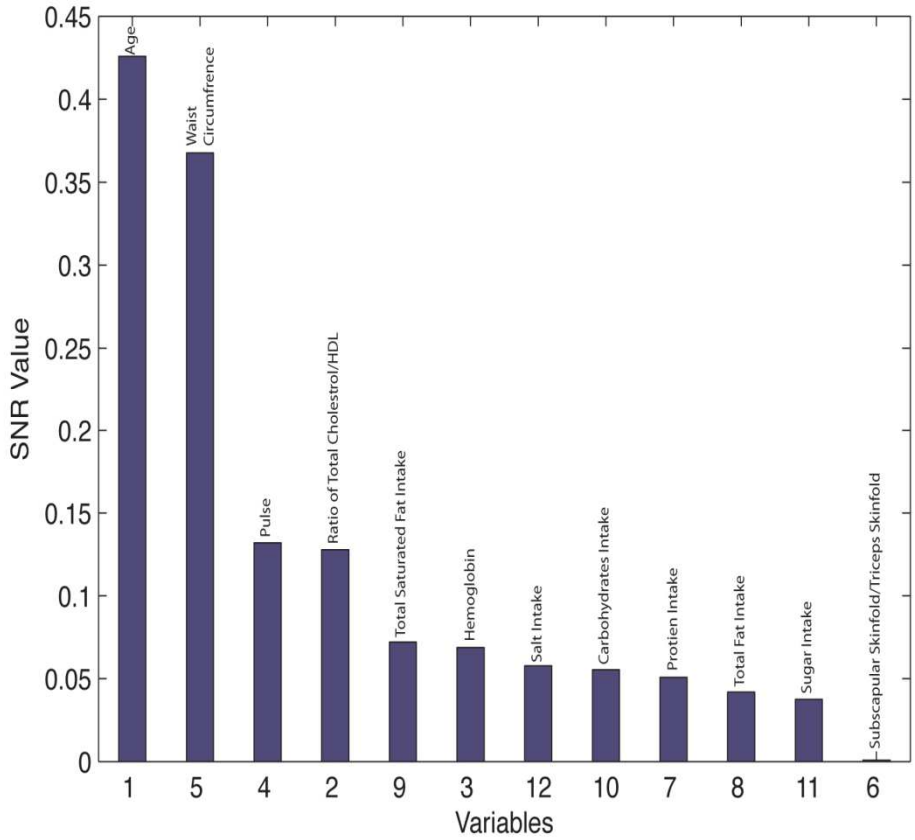


Figure 6.5. Bar graph showing variables ranked (highest to lowest) according to signal to noise ratio for male subjects only.

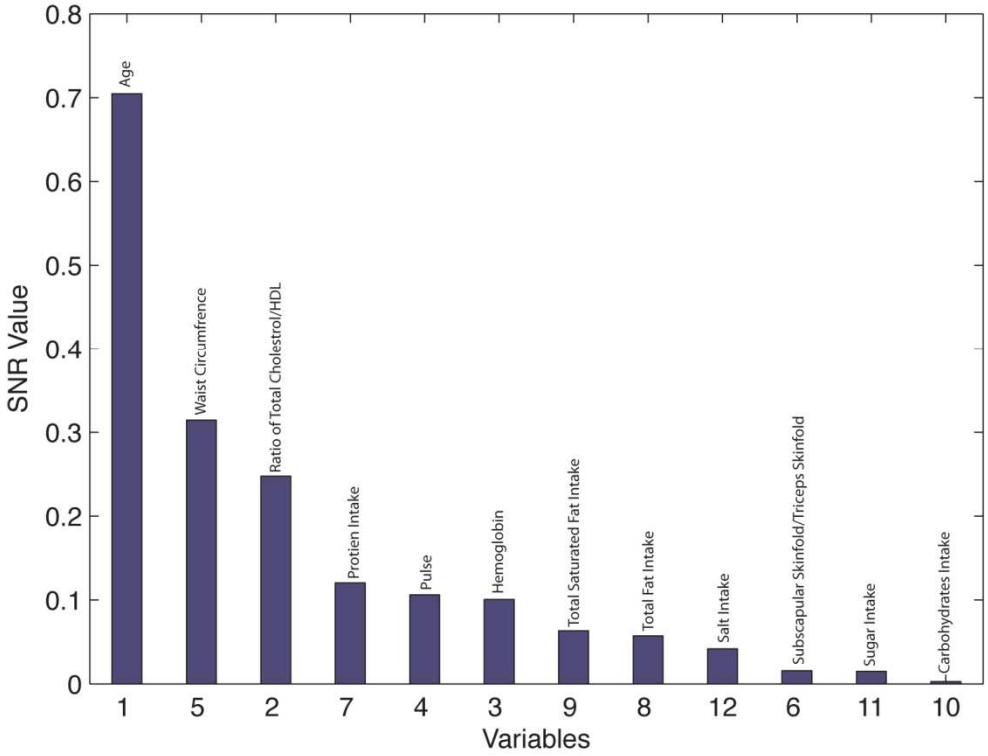


Figure 6.6. Bar graph showing variables ranked (highest to lowest) according to signal to noise ratio for female subjects only.

Age and waist circumference has been found to be equally important in both male and female subjects, but pulse is found to be more important in male subjects than in female subjects. Ratio of total cholesterol to HDL cholesterol and protein intake is found to be more important than pulse in female subjects.

Protein intake is not found to be very important in male subjects. Sugar intake is not found to be very important in either subject group but has a high ranking when looking at the whole data. Ethnicity has been used for male and female subjects but does not give significance as ethnicity is categorical variables whose value is either 1 or 2. Ratio of sub-scapular skin-fold to triceps skinfold is

found to have little importance in female subjects but shows no importance for male subjects and for the whole data.

Correlation analysis of variables:

The next step was to analyze data to derive a correlation coefficient. The correlation coefficient provides a measure of the linear relationships among the variables of the entire dataset (Figure 6.7). Later the data was stratified by gender and the analysis was repeated (Figures 6.8 and 6.9). Table 6.5 represents the correlation results of the variables for male and female datasets.

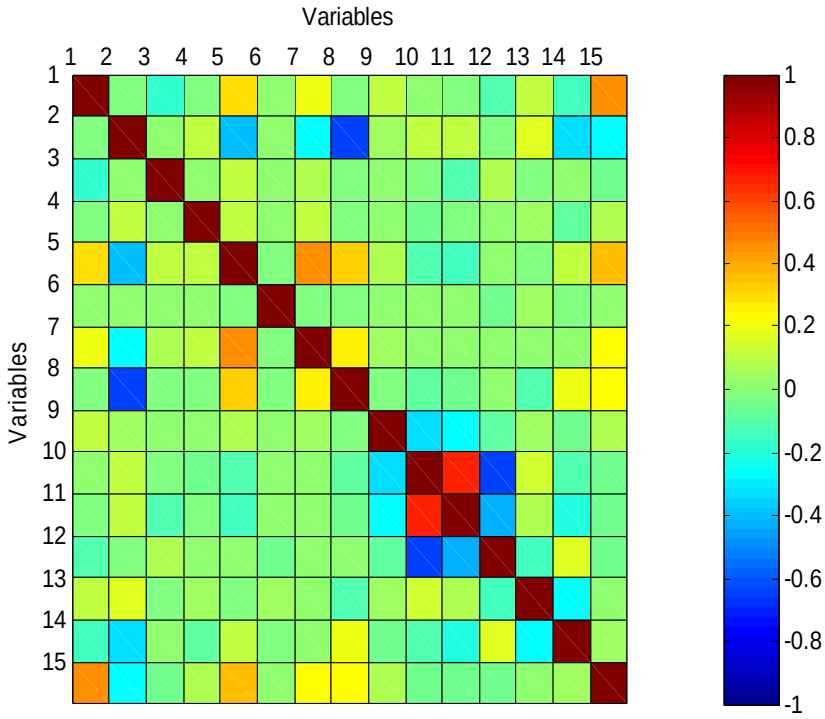


Figure 6.7. Linear relationship between the variables (listed below) using a correlation coefficient for the whole data.

1-Age, 2-gender, 3-ethnicity, 4-pulse, 5-waist circumference, 6-ratio of subscapular skinfold to triceps skinfold, 7-ratio of total cholesterol to HDL, 8-

haemoglobin, 9-%protein intake, 10-% carbohydrate intake, 11-%sugar intake, 12-%fat intake, 13-%saturated fat intake, 14-salt intake, 15-risk of CVD.

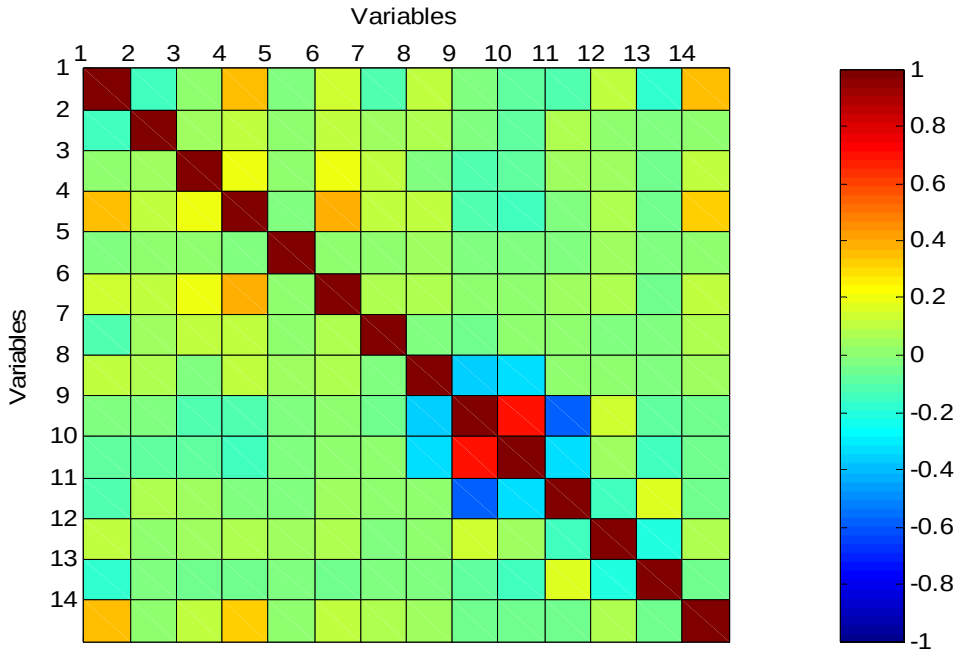


Figure 6.8. Linear relationship between the variables (listed below) using a correlation coefficient for male subjects.

1-Age, 2-ethnicity, 3-pulse, 4-waist circumference, 5-ratio of subscapular skinfold to triceps skinfold, 6-ratio of total cholesterol to HDL, 7-haemoglobin, 8-%protein intake, 9-% carbohydrate intake, 10-%sugar intake, 11-%fat intake, 12-%saturated fat intake, 13-salt intake, 14-risk of CVD.

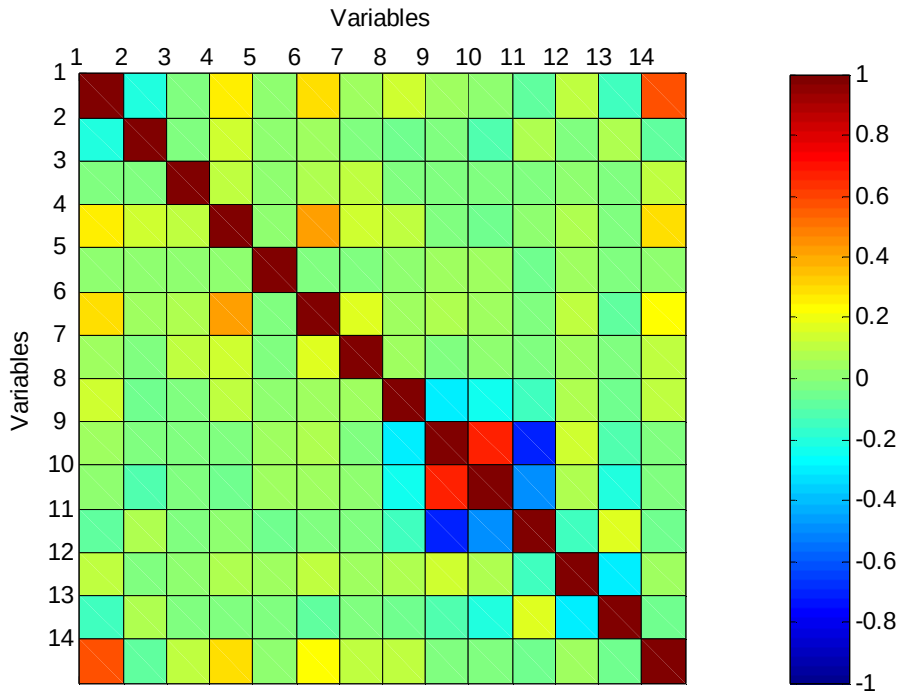


Figure 6.9. Linear relationship between the variables (listed below) using a correlation coefficient for female subjects.

1-Age, 2-ethnicity, 3-pulse, 4-waist circumference, 5-ratio of subscapular skinfold to triceps skinfold, 6-ratio of total cholesterol to HDL, 7-haemoglobin, 8-%protein intake, 9-% carbohydrate intake, 10-%sugar intake, 11-%fat intake, 12-%saturated fat intake, 13-salt intake, 14-risk of CVD.

For male data the positive correlations were shown between age and waist circumference and risk of cardiovascular disease and for age, negative correlations were found with ethnicity, haemoglobin, total fat intake and salt intake. Pulse had a positive correlation with waist circumference and ratio of total cholesterol to HDL and a negative correlation with carbohydrates and sugar intake. Waist circumference had a positive correlation with ratio of total cholesterol to HDL and risk of cardiovascular disease and pulse, waist

circumference, protein intake and total fat intake had a negative correlation with carbohydrates and sugar intake.

Carbohydrate intake had a positive correlation with sugar and saturated fat intake; carbohydrates intake and sugar intake have negative correlation with total fat intake. Total fat intake was found to have a positive correlation with salt intake. Saturated fat intake was found to have a negative correlation with total fat intake and salt intake.

For female data, age was shown to have a positive correlation with waist circumference, ratio of total cholesterol and HDL, risk of cardiovascular disease and a negative correlation with ethnicity and salt intake. Ethnicity had negative correlation with sugar intake and risk of cardiovascular disease. Waist circumference had a positive correlation with ratio of total cholesterol/HDL and risk of cardiovascular disease. Ratio of total cholesterol had a positive correlation with haemoglobin. Carbohydrate intake showed a positive correlation with sugar intake and saturated fat intake. Carbohydrate intake and sugar intake showed a negative correlation with protein intake, total fat and salt intake. Protein intake showed a negative correlation with total fat intake. Total fat intake had a positive correlation with salt intake. Saturated fat intake showed a negative correlation with total fat intake and salt intake.

Table 6.5

Results of correlation coefficient for male subjects, female subjects and the whole dataset.

Variables	Male subjects		Female subjects		Whole data	
	Positive relationship	Negative relationship	Positive relationship	Negative relationship	Positive relationship	Negative relationship
Age (years)	Waist circumference Risk of CVD	Ethnicity Haemoglobin Total fat intake Salt intake	Waist circumference Ratio of total cholesterol/HDL Risk of CVD	Ethnicity Salt intake	Waist circumference Ratio of total cholesterol/HDL Risk of CVD	Ethnicity Salt intake
Gender	-	-	-	-	Saturated fat	Haemoglobin Waist circumference Ratio of total cholesterol/HDL Carbohydrate intake Sugar intake Salt intake Risk of CVD
Ethnicity (European or non European)	-	Age	-	Age Sugar intake Risk of CVD	-	Age
Pulse (beats/min.)	Waist circumference Ratio of total cholesterol/HDL	Carbohydrate intake Sugar intake	-	-	-	-
Waist circumference (cm)	Age Pulse Ratio of total cholesterol/HDL Risk of CVD	Carbohydrate intake Sugar intake	Age Ratio of total cholesterol/HDL Risk of CVD	-	Age Ratio of total cholesterol/HDL Haemoglobin Risk of CVD	Gender Carbohydrate intake Sugar intake
Sub scapular /triceps skinfold	-	-	-	-	-	-
Ratio of total cholesterol/ HDL cholesterol	Waist circumference Pulse	-	Age Waist circumference Haemoglobin Risk of CVD	-	Age Waist circumference Haemoglobin Risk of CVD	Gender
Haemoglobin (gm/lt)	-	Age	Ratio of total cholesterol/HDL	-	Waist circumference Ratio of total cholesterol/HDL Salt intake Risk of CVD	Gender
% Protein intake	-	Carbohydrate intake Sugar intake	-	Carbohydrate Sugar intake Total fat intake	-	Carbohydrate intake Sugar intake
% Carbohydrate intake	Sugar intake Saturated fat intake	Pulse Waist circumference Protein intake Total fat intake	Sugar intake Saturated fat intake	Protein intake Total fat intake Salt intake	Sugar intake Saturated fat intake	Gender Waist circumference Protein intake Total fat intake
% Sugar intake	Carbohydrate intake	Pulse Waist circumference Protein intake Total fat intake	Carbohydrate intake	Ethnicity Protein intake Total fat intake Salt intake	Carbohydrates intake	Gender Waist circumference Protein intake Total fat intake Salt intake
% Fat intake	Salt intake	Carbohydrate Sugar Saturated fat	Salt intake	Protein intake Carbohydrates intake	Salt intake	Carbohydrates intake Sugar intake

		intake		Sugar intake Saturated fat intake		Saturated fat intake
% Saturated fat intake	Carbohydrates intake	Age Total fat intake Salt intake	Carbohydrates intake	Total fat intake Salt intake	Gender Carbohydrates intake	Total fat intake Salt intake
Salt intake (gm)	Total fat intake	Age Saturated fat intake	Total fat intake	Age Carbohydrates intake Sugar intake Saturated fat intake	Total fat intake Haemoglobin	Age Gender Saturated fat intake Sugar intake
Risk of CVD	Age Waist circumference	-	Age Waist circumference Ratio of total cholesterol/HDL	Ethnicity	Age Waist circumference Ratio of total cholesterol/HDL Haemoglobin	Gender

For the whole dataset, it was found that age had a positive correlation with waist circumference, ratio of total cholesterol and HDL, risk of cardiovascular disease and a negative correlation with ethnicity and salt intake. Gender had a positive correlation with saturated fat intake and a negative correlation with haemoglobin, waist circumference, ratio of total cholesterol/HDL, carbohydrates intake, sugar intake, salt intake and risk of cardiovascular disease.

Waist circumference had a positive correlation with ratio of total cholesterol/HDL and risk of cardiovascular disease and a negative correlation with carbohydrates and sugar intake. Haemoglobin showed a positive correlation with waist circumference, ratio of total cholesterol/HDL, salt intake and risk of cardiovascular disease.

Carbohydrate intake showed a positive correlation with sugar intake and saturated fat intake. Carbohydrates intake and sugar intake showed a negative correlation with protein intake and total fat intake. Sugar intake also showed a negative correlation with salt intake. Total fat intake had a positive correlation with salt intake and a negative correlation with saturated fat intake. Saturated fat intake showed a negative correlation with salt intake.

From the correlation analysis (Table 6.5), it can be concluded that major factors affecting the risk of cardiovascular disease are age, waist circumference and ratio of total cholesterol/HDL. Age is negatively correlated to ethnicity and salt intake which means that with an increase in age, salt intake is decreased. Waist circumference, protein intake and salt intake have a negative correlation with carbohydrate intake and sugar intake. Haemoglobin has been found to have a positive correlation with the ratio of total cholesterol to HDL which means that with an increase in haemoglobin, total cholesterol to HDL decreases, leading to reduced risk of cardiovascular disease.

6.4 Risk prediction and knowledge discovery with the ontology based personalized decision support (OBPDS)

All the methods currently known for predicting risk of cardiovascular disease only use clinical variables, but as each person is different and responds to each nutrient and medicine differently (Barton, 2008), it is important to build a tool that can predict personalized or individualized risk of cardiovascular disease. Different methods have been used for predicting risk of cardiovascular disease for example, global, local and personalized. Initially the multiple linear regression method in NeuCom was applied, followed by application of the weighted-weighted K nearest neighbor algorithm and finally the transductive neuro-fuzzy inference system was applied to predict the risk of cardiovascular disease.

Before a model is selected for personalized risk evaluation, model has to be assessed how the results of the model will generalize to an independent test. The method of assessing how the results of a statistical analysis will generalize to an independent test is called cross-validation. Cross-validation is

done to see how accurately the risk prediction model will perform for real data. There are different methods of cross validation (Kasabov, 2007(a)).

Train-test split cross-validation: In train-test split method randomly splits the dataset into training and test data (validation data). For the each set of split data the accuracy is checked by using test data.

K-fold cross-validation: This method involves splitting dataset into K subsets. K can be 3, 5, 10. In *K*-fold cross-validation, the original sample is randomly partitioned into *K* subsamples. Of the *K* subsets, a single subset is considered as the test data (validation data) for testing the model and the remaining K-1subsets are used as training data. The cross-validation process is then repeated *K* times (the *folds*), with each of the *K* subsamples used exactly once as the validation data.

Leave-one-out cross-validation: Leave-one-out cross-validation method involves using a single sample from the original data as the validation data, and the remaining data as the training data. This process is repeated for all the data so that each sample in the data is used once as the test data (validation data).

The advantage of train-test split method (over *k*-fold cross validation) is that the proportion of the data split into test and train is not dependent on the number of folds such as in k-fold cross-validation. The disadvantage of train-test split method is that test data may overlap as few may be selected even more than once but some samples may never be selected in the validation subsample. In train-test split method, results will vary if the process is repeated with different random splits. The advantage of k-fold cross-validation method over train-test split method is that, in k-fold cross-validation all data is used for

both training and testing and in train-test split method each observation is used for testing only once. The disadvantage of leave-one-out cross-validation is that it is very time consuming while process larger datasets because in case of large dataset process is repeated a number of times.

Cross-validation is useful only if the test data (validation data) and test data are taken from the same population. If carried out properly, and if the test data and training data are from the same population, cross-validation is nearly unbiased. In the present case, dataset is very large so the data is randomly split into two equal parts and one half is used as train data and other half as test data.

The results obtained from each method were compared and the accuracy of each parameter was checked. The experimental details and output of each model are described in detail in the following paragraphs.

Multiple Linear regression: The multiple linear regression method was used in NeuCom. The data was split into two equal parts randomly. One part was used to train and other set was used to test and the outcome was stored accurately stored. The accuracy was then checked in Matlab with different threshold values. It was observed that the highest accuracy of both classes could be achieved with a threshold value of 0.5. Results from the multiple linear regression were then compared with the other methods and this is described in the following pages.

Weighted-weighted K nearest Neighbor Algorithm (Kasabov, 2007(b)): The weighted-weighted K-nearest neighbor method was also used to predict personalized risk of disease. The data has been split randomly into two equal parts. The system was trained with first half of the data and the other half of the

data was used as test subjects. This method was used with different parameters such as different numbers of neighbors and best set of parameters was selected based on accuracy.

Transductive neuro-fuzzy inference system with weighted data normalization (TWNFI) (Song and Kasabov, 2006): TWNFI was used to predict personalized risk of cardiovascular disease. The data was split into two equal parts. Initially one part of data was used as test subjects and the other 50% of the data was used as training subjects. The experiments have been repeated by randomly splitting data into two equal parts. The data was then tested using all 13 variables (age, ethnicity, pulse, waist circumference, ratio of subscapular and triceps skinfold, ratio of total cholesterol and HDL, haemoglobin, percentage protein intake, percentage carbohydrates intake, percentage sugar intake, percentage fat intake and salt intake) by using a different number of parameters each time. The parameters which had been changed to train the system were (i) number of neighbors, (ii) threshold value, (iii) number of learning epochs and (iv) number of rules based on clusters to be extracted. The system was tuned with the best values of parameters to do the experiments, 18 nearest neighbors, threshold 0.5, 100 learning epochs and 6 to 8 rules to be extracted. The detailed description of all the experiments is as follows:

Personalized modeling: All 13 variables (age, ethnicity, pulse, waist circumference, ratio of subscapular and triceps skinfold, ratio of total cholesterol and HDL, haemoglobin, percentage protein intake, percentage carbohydrates intake, percentage sugar intake, percentage fat intake and salt intake) were used. The data was split into two parts; male data and female data. For the first

experiment male data was used to create a personalized model. Male data was divided randomly into two equal parts. The first 50% of the data was used to test subjects and the second half was used for training. The output was stored and the same process was repeated by using the second half to test and first half to training data. The results were stored and tested for accuracy by checking with the actual output. Accuracy was compared with the other methods as well. Tables 6.6 and 6.7 represent the comparative results of accuracy of the different methods for male data and female data respectively. Table 6.8 gives the accuracy results and comparison of accuracy with other models for the whole data. Data was split 50-50 twice. Each time the first half was used as training and second half as test data (independent data), accuracy results are shown in Table 6.6, 6.7, 6.8.

For male subjects (Table 6.6), it was found that for all the above mentioned methods, comparable accuracy was achieved at threshold value 0.6. It was found that multiple linear regression and TWNFI give better accuracy than WWKNN at threshold 0.6. Although the multiple linear regression method gives nearly equal accuracy at same threshold for male subjects, TWNFI, which is a personalized model, gives more knowledge about the variables by ranking them and generates personalized profiles or set of rules for each subject which can be used for better prediction than global models.

Table 6.6

Accuracy (%) results comparison of NNS 97 data using 13 variables for male data.

Threshold		MLR (class1) (High risk)	MLR (class0) (Low risk)	WWKNN (class1) (High risk)	WWKNN (class0) (Low risk)	TWNFI (class1) (High risk)	TWNFI (class0) (Low risk)
0.9	1st expt.	21.2	97.13	16.35	88.52	33.93	91.87
	2nd expt.	21.46	98	4	98	31.86	93
0.8	1st expt.	40.36	90.67	36.3	75.36	51.86	83.25
	2nd expt.	39.82	91	19	88	45.58	89.5
0.7	1st expt.	80.14	58.4	55.24	55.5	65.16	68.66
	2nd expt.	55.97	80	43.36	68.5	59.07	71
0.6	1st expt.	74.52	62.68	71.93	39	76.55	53.83
	2nd expt.	73.23	66.5	71.02	45	71.24	60
0.5	1st expt.	86.58	37.32	83.99	26.32	83.99	39
	2nd expt.	85.84	40.5	89.38	20	80.75	45
0.4	1st expt.	95.94	18.42	92	15.55	90.98	25.12
	2nd expt.	92.7	20.5	98.23	1.5	86.73	30
0.3	1st expt.	99.44	4.78	95.83	6.7	95.72	11.96
	2nd expt.	97.79	8	99.78	0	92.92	16.5

Table 6.7

Accuracy (%) results comparison of NNS 97 data using 13 variables for female data.

Threshold		MLR (class1) (High risk)	MLR (class0) (Low risk)	WWKNN (class1) (High risk)	WWKNN (class0) (Low risk)	TWNFI (class1) (High risk)	TWNFI (class0) (Low risk)
0.9	1st expt.	16.16	98.58	6.25	98.58	22.1	96.94
	2nd expt.	15.81	98.68	2.13	99.78	20.8	93.3
0.8	1st expt.	29.57	97.7	16.92	96.83	37.35	94.97
	2nd expt.	29.79	97.59	6.99	98.68	45.3	92.6
0.7	1st expt.	43.75	95.4	27.9	93.76	50.15	91.47
	2nd expt.	44.68	96.05	20.97	96.49	57.4	89.6
0.6	1st expt.	57.93	92.23	40.09	87.75	58.54	87.64
	2nd expt.	56.84	92.32	32.52	92.54	60.7	83.4
0.5	1st expt.	69.21	86.11	53.05	78.88	65.24	84.14
	2nd expt.	68.39	84.65	45.59	83.55	68.9	77.3
0.4	1st expt.	79.42	76.59	65.4	67.61	71.04	79.1
	2nd expt.	78.42	74.34	63.53	70.39	77.27	70
0.3	1st expt.	87.96	61.6	74.85	53.83	76.37	71.12
	2nd expt.	88.75	58.77	79.94	48.03	83.8	66.9

Table 6.8

Accuracy (%) results comparison of NNS97 data using 13 variables for the whole dataset.

Threshold		MLR (class1) (High risk)	MLR (class0) (Low risk)	WWKNN (class1) (High risk)	WWKNN (class0) (Low risk)	TWNFI (class1) (High risk)	TWNFI (class0) (Low risk)
0.9	1st expt.	19.05	98.12	12.05	95.42	28.9	95.35
	2nd expt.	22.73	98.11	4.28	99.27	30.48	93.61
0.8	1st expt.	35.5	95.5	28.06	90.09	45.69	91.29
	2nd expt.	36.9	95.5	16.31	95.94	45.05	89.7
0.7	1st expt.	52.17	90.62	43.62	81.76	58.78	84.31
	2nd expt.	50	91.58	38.24	88.68	55.61	83.6
0.6	1st expt.	67.47	82.96	58.39	72.45	68.89	77.03
	2nd expt.	64.84	84.76	58.29	77.21	66.58	76.05
0.5	1st expt.	79.2	70.8	70.84	62.39	76.02	69.97
	2nd expt.	75.4	71.7	74.06	59.22	73.8	68.94
0.4	1st expt.	88.92	58.33	80.69	51.28	82.5	62.16
	2nd expt.	85.7	56.6	85.7	40.78	80.88	60.81
0.3	1st expt.	94.56	43.77	86.91	39.04	87.49	52.55
	2nd expt.	92.51	40.20	92.91	25.54	86.63	49.93

For female subjects (Table 6.7), it was found that comparable accuracy for both classes have been achieved at threshold value 0.5. As with male subjects, multiple linear regression and TWNFI give almost equal accuracy for female subjects at threshold value 0.5. However, multiple linear regression gives only the outcome for each subject but TWNFI gives a lot more than simply predicting risk outcome. TWNFI, along with risk prediction, also ranks the variables and generates profiles for each subject based on nearest neighbours.

It was observed that the multiple linear regression method (local model) and TWNFI (personalized model) give very comparable accuracy for both the class and both give better accuracy than WWKNN. The reason WWKNN didn't perform better than TWNFI is that WWKNN first ranks variables based on signal to noise ratio. As ethnicity is also included in data, signal to noise ratio cannot perform well for binomial variables.

When comparing the overall accuracy for whole dataset, TWNFI gives higher accuracy than multiple linear regression for class 1 at threshold value 0.6 and multiple linear regression gives higher accuracy than TWNFI for class 0. The multiple linear regression method uses a set of local models which are created from data and each model represents a sub-space. The TWNFI method is a much more advanced method and optimizes parameters and variable weights. TWNFI along with a better accuracy of personalized risk prediction and estimating importance for the variables for a subject also generates a set of profiles or rules based on nearest neighbors for personalized prediction and recommendations. TWNFI results in better accuracy as it develops a

personalized model for each subject and is adaptive as it performs well when new data is added.

Examples of personalized model for male subjects: To create a personalized model for an individual using TWNFI, two different persons represented by Subject 1 and 2 were used to build a personalized model and their individual weights for all variables were compared with global weights and outcomes were compared with other methods such as multiple linear regression and WWKNN. For every new subject a personalized model can be created and can be used to predict the person's risk of developing cardiovascular disease. In the first example, subject one is high risk male subject and subject two is low risk male. Table 6.9 shows the experimental results for the first example.

It was found that the predicted outcome by TWNFI was very near to the desired output. In comparison to the other methods, like multiple linear regression and WWKNN, TWNFI results in better accuracy. TWNFI also shows the importance of variables which results in more efficient personalized prediction. Therefore, we are able to understand more about each variable and can discover new relationships between variables. For male Subject 1, salt intake has been found to be the most important factor. After salt intake, haemoglobin, sugar intake and protein intake which were found to be the most important factors for determining risk of cardiovascular disease.

Table 6.9

Examples of TWNFI personalized models for two different male subjects; high risk and low risk male subjects; showing different weights for the same variables with global weights representing importance of variables.

	Subject 1 (High risk male)		**Subject 2 (Low risk male)**		
Input Variables	**Values of input**	**Weights of input variables**	**Values of input**	**Weights of input variables**	**Global weights**
Age (years)	62	0.5246	17	0.9542	0.96995
Ethnicity	1	0.6384	1	0.8796	0.02035
Pulse (beats/min.)	54	0.772	53	0.9231	0.18915
Waist circumference(cm)	113.1	0.6754	73.3	1	0.8325
Sub scapular/triceps skinfold	1.7	0.8199	0.7	0.9051	0.869
Total cholesterol/HDL	6.9	0.574	3.1	0.9809	0.94245
Haemoglobin(gm/L)	155	0.999	147	0.9972	0.5412
% Protein intake	19.3	0.9446	18.4	0.8545	0.08865
% Carbohydrate intake	47.0	0.7942	48.5	0.804	0.33
% Sugar intake	15.9	0.9813	18.4	0.8678	0.00665
% Total fat intake	34.1	0.8145	33.7	0.8812	0.2117
% Total saturated fat intake	14.6	0.741	9.2	0.8851	0.2863
Salt intake(mg)	4545	1	3833	0.871	0.1056
Desired Output	1		0		
Predicted output with Multiple linear regression		0.8997		0.2277	
Predicted output with WWKNN		0.4722		0.6566	
Predicted output with TWNFI		1.0031		0.1361	

On the other hand, for male Subject 2, waist circumference was found to be the most important factor and this put him at low risk of cardiovascular disease. After waist circumference, haemoglobin, ratio of total cholesterol to HDL and age were found to be the important risk determining factors for male Subject 2. When comparing these with the global weights of variables, which are totally different. Age was found to be the most important factor for determining risk of cardiovascular disease.

Along with the weights of variables and the outcome, in TWNFI, fuzzy rules are generated depending on nearest neighbors. More knowledge is stored inside these fuzzy rules. Zadeh-Mamdani type is used in TWNFI to generate rules which consist of “if” and “then” parts. Rules or profiles are generated on the basis of closest neighbors. For example, if age is an input variable, it gives an input fuzzy membership function and then an output. Each fuzzy rule explains a part of the particular space depending on distance. The fuzzy rules are generated on the basis of Gaussian membership functions. The optimization of membership functions leads to the search for the overall optimal solution for the parameter spaces. Each rule covers a cluster of subjects which are similar to a defined degree among variables and outcomes as well (Figure 6.10). Each rule represents the centre of each cluster. The risk of cardiovascular disease is predicted between 0 and 1. In the given examples, risk of cardiovascular disease is considered high if predicted outcome is more than 0.5 and risk of cardiovascular disease is low if predicted outcome is less than 0.5 (a threshold value of 0.5).

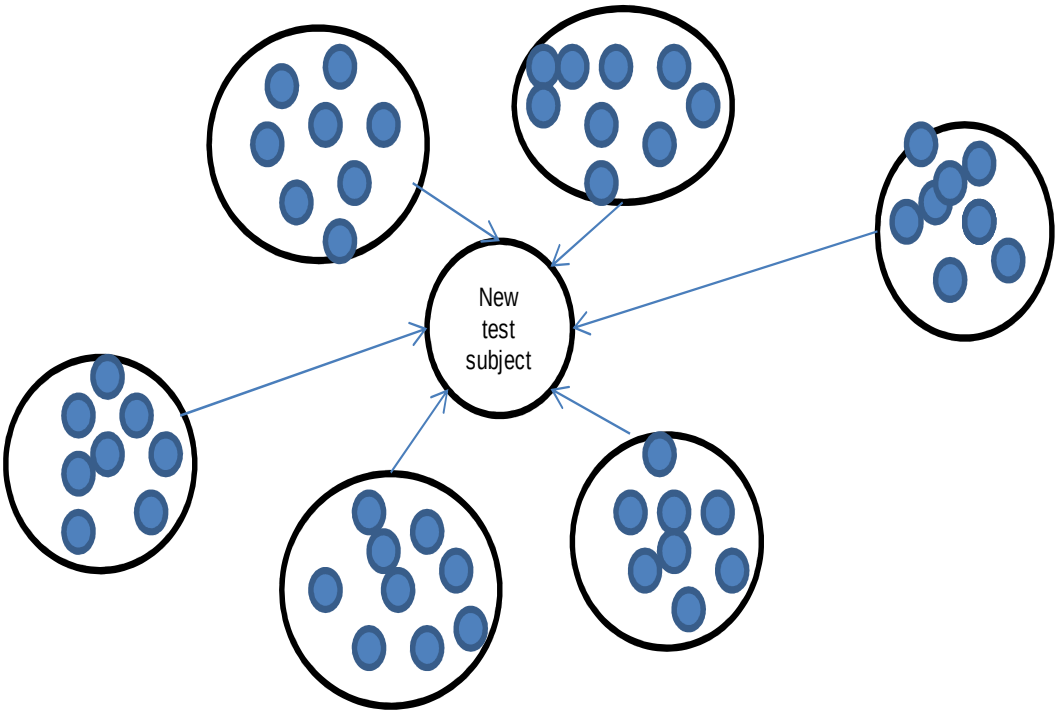
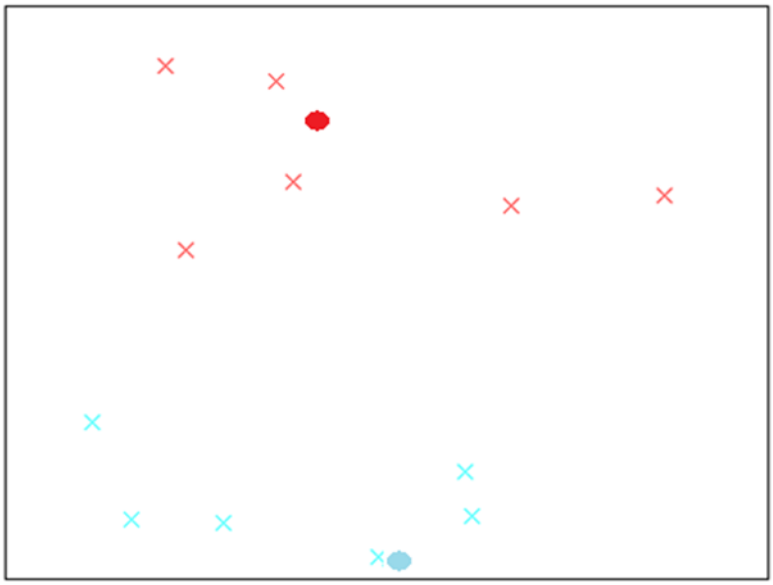


Figure 6.10. Illustration of rules extraction from clusters based on nearest subjects.

The system was tuned with different numbers of fuzzy rules and the best number of rules is between 6 and 8, and the best number of nearest neighbors for training is 18. Sets of 6 rules have been generated for each subject that is based on nearest subjects. The rules are based on cluster radius and cluster centre. Figure 6.11 illustrates male Subjects 1(red) and 2 (blue) with their nearest neighbors. Sets of rules or profiles are generated on the basis of these neighbors.



Cluster center (class1 : high risk of cardiovascular disease)

Cluster center (class 0: low risk of cardiovascular disease)

Male Subject 1 (belong to class 1)

Male Subject 2 (belong to class 2)

Figure 6.11. Example of male Subjects 1 and 2 with cluster centers based on nearest neighbors using principal component analysis (PCA).

The set of rules for male Subject 1 based on gaussian membership function are shown below.

Rule 1:

If	Age (years)	is about	60
	Pulse (beats/min.)	is about	52
	Waist circumference (cm)	is about	112.46
	Sub-scapular/triceps skinfold	is about	1.67
	Ratio of total cholesterol to HDL	is about	8.31
	Haemoglobin (gm/L)	is about	154.35

	Protein intake	is about	22.36
	Carbohydrate intake	is about	42.76
	Sugar intake	is about	17.90
	Total fat intake	is about	36.08
	Total saturated fat intake	is about	5.21
	Salt intake (mg)	is about	2162.67
Then	Risk of cardiovascular disease	is	**High**

Rule 1 explains that higher age is a major risk factor for determining risk of cardiovascular disease. With increase in age risk of cardiovascular disease increases or if the ratio of cholesterol to HDL is greater and higher waist circumference with higher intake of fat and sugar, then the risk of having cardiovascular disease increases.

Rule 2:

If	Age (years)	is about	62
	Pulse (beats/min.)	is about	66
	Waist circumference (cm)	is about	100.62
	Sub-scapular/triceps skinfold	is about	2.23
	Ratio of total cholesterol to HDL	is about	6.11
	Haemoglobin (gm/L)	is about	155.34
	Protein intake	is about	16.48
	Carbohydrate intake	is about	41.78
	Sugar intake	is about	15.23
	Total fat intake	is about	42.93
	Total saturated fat intake	is about	7.95
	Salt intake (mg)	is about	5289.24
Then	Risk of cardiovascular disease	is	**High**

Rule 2 explains that risk of cardiovascular disease increases with increase in age, waist circumference or ratio of total cholesterol to HDL. Also higher

intakes of carbohydrates, fat, sugar and salt also lead to cardiovascular disease.

Rule 3:

If	Age (years)	is about	56
	Pulse (beats/min.)	is about	50.16
	Waist circumference (cm)	is about	98.68
	Sub-scapular/triceps skinfold	is about	0.89
	Ratio of total cholesterol to HDL	is about	5.74
	Haemoglobin (gm/L)	is about	161.61
	Protein intake	is about	13.87
	Carbohydrate intake	is about	48.20
	Sugar intake	is about	15.79
	Total fat intake	is about	34.57
	Total saturated fat intake	is about	12.38
	Salt intake (mg)	is about	5220.04
Then	Risk of cardiovascular disease	is	**High**

The above mentioned profile explains that both nutritional and clinical variables are important risk determinant factors. For recommendations, reducing the intake of carbohydrate, sugar, fat and salt can reduce the risk of cardiovascular disease.

Rule 4:

If	Age (years)	is about	60
	Pulse (beats/min.)	is about	62.93
	Waist circumference (cm)	is about	114.60
	Sub-scapular/triceps skinfold (mm)	is about	1.39
	Ratio of total cholesterol to HDL (mmol/L)	is about	7.28
	Haemoglobin (gm/L)	is about	155.17

	Protein intake	is about	10.67
	Carbohydrate intake	is about	48.28
	Sugar intake	is about	31.08
	Total fat intake	is about	31.59
	Total saturated fat intake	is about	19.15
	Salt intake (mg)	is about	3014.08
Then	Risk of cardiovascular disease	is	**High**

The above mentioned rule explains that risk of cardiovascular disease increases with increase in age but if, with increasing age, the intake of carbohydrates, fat, sugar and salt is decreased the risk of cardiovascular disease may be reduced.

Rule 5:

If	Age (years)	is about	72
	Pulse (beats/min.)	is about	49
	Waist circumference (cm)	is about	99.67
	Sub-scapular/triceps skinfold	is about	2.07
	Ratio of total cholesterol to HDL	is about	7.74
	Haemoglobin (gm/L)	is about	135.72
	Protein intake	is about	17.08
	Carbohydrate intake	is about	48.59
	Sugar intake	is about	19.37
	Total fat intake	is about	32.63
	Total saturated fat intake	is about	13.91
	Salt intake (mg)	is about	4430.59
Then	Risk of cardiovascular disease	is	**High**

Rule 5 explains that risk of cardiovascular disease increases with higher intake of carbohydrates, fat, sugar and salt. Higher intake of fat leads to higher levels

of ratio of total cholesterol to HDL which is a major risk factor for cardiovascular disease.

Rule 6:

If	Age (years)	is about	49
	Pulse (beats/min.)	is about	56
	Waist circumference (cm)	is about	105.59
	Sub-scapular/triceps skinfold	is about	1.60
	Ratio of total cholesterol to HDL	is about	4.91
	Haemoglobin (gm/L)	is about	146.10
	Protein intake	is about	14.59
	Carbohydrate intake	is about	39.59
	Sugar intake	is about	10.02
	Total fat intake	is about	32.47
	Total saturated fat intake	is about	0.91
	Salt intake (mg)	is about	1806.05
Then	Risk of cardiovascular disease	is	**Low**

Rule 6 explains that having a normal ratio of total cholesterol to HDL, lower intakes of carbohydrate, fat, sugar and salt reduces the risk of cardiovascular disease.

The above mentioned rules for male Subject 1 can be used for personalized risk evaluation and recommendation. These profiles or rules have been generated on the basis of nearest neighbors. So from all the rules explained above for male subject 1, it can be recommended that with increase in age risk of cardiovascular disease increases or if the ratio of cholesterol to HDL is higher and there is greater waist circumference, higher intake of fat and sugar, then the risk of having cardiovascular disease increases. Also these explained that age, waist circumference, ratio of total cholesterol to HDL, intake of

carbohydrates, fat, sugar and salt are the important disease risk determinant factors. Although age is an irreversible risk factor, the rules imply with increased intake of carbohydrates, fat, sugar and salt and increased age there is increased risk of cardiovascular disease for this subject and reduction in dietary intake may reduce this risk.

Similarly a set of rules or profiles were generated for male Subject 2, which are based on nearest neighbors and can be used for better prediction. The set of rules generated for male subject 2 are listed below:

Rule 1:

If	Age (years)	is about	24
	Pulse (beats/min.)	is about	53
	Waist circumference (cm)	is about	78.50
	Sub-scapular/triceps skinfold	is about	1.07
	Ratio of total cholesterol to HDL	is about	3.34
	Haemoglobin (gm/L)	is about	144.43
	Protein intake	is about	12.32
	Carbohydrate intake	is about	51.43
	Sugar intake	is about	22.59
	Total fat intake	is about	36.76
	Total saturated fat intake	is about	7.41
	Salt intake (mg)	is about	5643.79
Then	Risk of cardiovascular disease	is	**Low**

Rule 1 explains that age has always been an important determinant factor for cardiovascular disease. Male subjects of lower age are at the least risk of having cardiovascular disease. Lower levels of ratio of total cholesterol to HDL also reduce the risk of cardiovascular disease.

Rule 2:

If	Age (years)	is about	16
	Pulse (beats/min.)	is about	64
	Waist circumference (cm)	is about	68.78
	Sub-scapular/triceps skinfold	is about	0.75
	Ratio of total cholesterol to HDL	is about	3.89
	Haemoglobin (gm/L)	is about	155.17
	Protein intake	is about	15.02
	Carbohydrate intake	is about	53.10
	Sugar intake	is about	27.70
	Total fat intake	is about	32.44
	Total saturated fat intake	is about	22.58
	Salt intake (mg)	is about	2477.09
Then	Risk of cardiovascular disease	is	**Low**

The above mentioned rule also explains that there is reduced risk of cardiovascular disease at a younger age and with a lower ratio of total cholesterol to HDL.

Rule 3:

If	Age (years)	is about	30
	Pulse (beats/min.)	is about	56
	Waist circumference (cm)	is about	89.40
	Sub-scapular/triceps skinfold	is about	0.59
	Ratio of total cholesterol to HDL	is about	2.10
	Haemoglobin (gm/L)	is about	150.79
	Protein intake	is about	18.66
	Carbohydrate intake	is about	45.22
	Sugar intake	is about	16.13
	Total fat intake	is about	36.41

	Total saturated fat intake	is about	6.09
	Salt intake (mg)	is about	2668.95
Then	Risk of cardiovascular disease	is	**Low**

According to rule 3, lower levels of ratio of total cholesterol and HDL, smaller waist circumference, lower intake of carbohydrates, fat, sugar and salt reduce the risk of cardiovascular disease.

Rule 4:

If	Age (years)	is about	16
	Pulse (beats/min.)	is about	63
	Waist circumference (cm)	is about	82.09
	Sub-scapular/triceps skinfold	is about	0.44
	Ratio of total cholesterol to HDL	is about	2.31
	Haemoglobin (gm/L)	is about	152.91
	Protein intake	is about	14.89
	Carbohydrate intake	is about	45.60
	Sugar intake	is about	16.93
	Total fat intake	is about	39.81
	Total saturated fat intake	is about	3.36
	Salt intake (mg)	is about	7050.46
Then	Risk of cardiovascular disease	is	**Low**

This above mentioned rule confirms the explanation given for the other rules which state that a lower intake of carbohydrates, fat, sugar and salt reduces the risk of cardiovascular risk.

Rule 5:

If	Age (years)	is about	21
	Pulse (beats/min.)	is about	59
	Waist circumference (cm)	is about	82.16

	Sub-scapular/triceps skinfold	is about	1.53
	Ratio of total cholesterol to HDL	is about	3.63
	Haemoglobin (gm/L)	is about	154.64
	Protein intake	is about	17.01
	Carbohydrate intake	is about	43.25
	Sugar intake	is about	23.43
	Total fat intake	is about	39.83
	Total saturated fat intake	is about	6.87
	Salt intake (mg)	is about	3650.04
Then	Risk of cardiovascular disease	is	**High**

The rule 5 states that a lower ratio of total cholesterol to HDL, reduced waist circumference or lower intake of carbohydrates, fat, sugar and salt reduces the risk of cardiovascular disease.

Rule 6:

If	Age (years)	is about	17
	Pulse (beats/min.)	is about	62
	Waist circumference (cm)	is about	67.78
	Sub-scapular/triceps skinfold	is about	1.06
	Ratio of total cholesterol to HDL	is about	3.84
	Haemoglobin (gm/L)	is about	153.06
	Protein intake	is about	19.48
	Carbohydrates intake	is about	57.74
	Sugar intake	is about	14.65
	Total fat intake	is about	23.13
	Total saturated fat intake	is about	19.81
	Salt intake (mg)	is about	3495.92
Then	Risk of cardiovascular disease	is	**Low**

Rule 6 also explains a similar pattern to the other rules for male Subject 2. Reduced waist circumference and ratio of total cholesterol to HDL and lower intake of carbohydrates, fat, sugar and salt decreases the risk of cardiovascular disease.

From all the above mentioned rules generated for male Subject 2, it can be concluded that being younger and having a low ratio of total cholesterol to HDL as well as a lower intake of fat and sugar, the subject is most likely to be at very low risk of having cardiovascular disease. By maintaining normal levels of the major risk factors such as ratio of total cholesterol to HDL, waist circumference and lower intake of carbohydrates, fat, sugar and protein the risk of cardiovascular disease may be reduced. As age is an irreversible risk factor and risk of cardiovascular disease is lower at younger ages the model implies that if the intake of carbohydrates, fat, sugar and salt is reduced with increased age, risk of cardiovascular disease may also be reduced.

Examples of personalized models for female subjects: The second personalized model example has been created for female subjects. Two subjects from the female dataset, one being at high risk and the other at low risk have been taken and the weights for variables have been compared. Table 6.10 shows the results for the two female subjects.

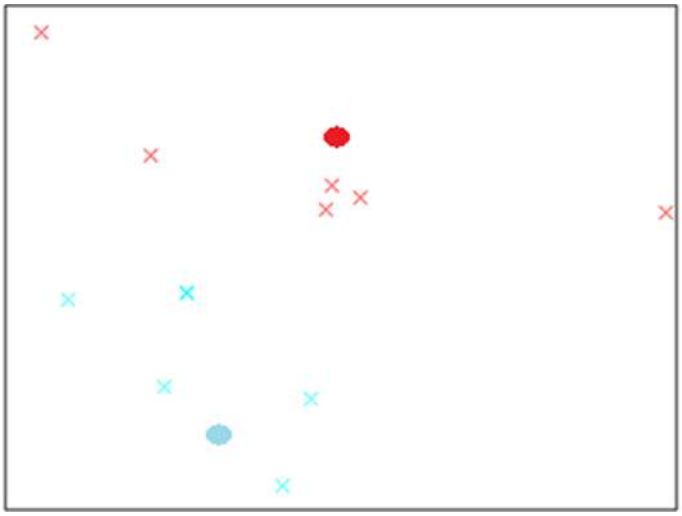
Female Subjects 1 and 2 have been found to have similar risk outcome for cardiovascular disease to the desired outcome by using TWNFI. The outcome predicted by TWNFI was compared with multiple linear regression and WWKNN and it was found that TWNFI gave higher accuracy than the other methods and more information can be obtained about variables in TWNFI.

Table 6.10

Example of TWNFI personalized models for two female subjects; high risk and low risk; showing different weights for the same variables with global weights representing the importance of variables.

	Subject 1 (High risk female)		**Subject 2 (Low risk female)**		
Input Variables	**Values of input variables**	**Weights of input variables**	**Values of input variables**	**Weights of input variables**	**Global weights**
Age (years)	71	0.8882	30	0.95	1
Ethnicity	1	0.737	2	0.546	0.01315
Pulse (beats/min.)	67	0.7617	57	0.5447	0.29535
Waist circumference(cm)	72.2	0.8732	83.6	0.9469	0.38165
Sub scapular/triceps skinfold	0.6	0.779	1.7	1	0.35945
Total cholesterol/HDL	5.4	1	3.8	0.5752	0.76515
Haemoglobin(g/L))	127	0.8751	133	0.5717	0.2465
% Protein intake	11.1	0.7792	15.2	0.8682	0.5216
% Carbohydrate intake	43.6	0.796	46.9	0.5825	0.6064
% Sugar intake	4.9	0.784	13.8	0.4766	0.0277
% Total fat intake	45.6	0.727	38.0	0.5369	0.085
% Total saturated fat intake	21.0	0.8456	13.9	0.802	0.2652
Salt intake(mg)	2335	0.7557	3207	0.5769	0.25995
Desired Output	1		0		
Predicted output with Multiple linear regression		0.6956		0.1549	
Predicted output with WWKNN		0.5672		0.1747	
Predicted output with TWNFI		0.8983		0.0090	

Female Subject 1 was found to be at high risk of cardiovascular disease and ratio of total cholesterol to HDL was the most important risk factor. Age, haemoglobin and waist circumference were ranked after ratio of total cholesterol to HDL. On the other hand, for female Subject 2, ratio of sub scapular to triceps skinfold was the most important factor. But when compared with the global weights of the female dataset, which were very different it can be seen that each individual needs personalized prediction and advice. Along with the importance of variables and high accuracy, set of rules or profiles have been generated for each female subject.



× Cluster center (class1: high risk of cardiovascular disease)

× Cluster center (class 0: low risk of cardiovascular disease)

● Female Subject 1(belong to class 1)

● Female Subject 2 (belong to class 2)

Figure 6.12. Example of female Subjects 1 and 2 with cluster centers based on nearest neighbors using principal component analysis (PCA).

As with the male subjects, rules or profiles were generated for the female subjects on the basis of nearest samples. Figure 6.12 illustrates the centre of each neighbor cluster with the two female test samples belonging to two different classes. These rules or profiles can be used for knowledge extraction and knowledge discovery.

Rule 1:

If	Age (years)	is about	75
	Pulse (beats/min.)	is about	66
	Waist circumference (cm)	is about	80.31
	Sub-scapular/triceps skinfold	is about	1.39
	Ratio of total cholesterol to HDL	is about	4.20
	Haemoglobin (gm/L)	is about	128.86
	Protein intake	is about	13.87
	Carbohydrate intake	is about	45.28
	Sugar intake	is about	10.64
	Total fat intake	is about	41.17
	Total saturated fat intake	is about	9.28
	Salt intake (mg)	is about	2071.27
Then	Risk of cardiovascular disease	is	**High**

Rule 1 explains that older subjects with increased ratio of total cholesterol to HDL are at risk of cardiovascular disease. Higher intake of carbohydrate, fat, sugar and salt may also increase the risk of cardiovascular disease.

Rule 2:

If	Age (years)	is about	68
	Pulse (beats/min.)	is about	73
	Waist circumference (cm)	is about	78.18
	Sub-scapular/triceps skinfold	is about	0.90

	Ratio of total cholesterol to HDL	is about	6.42
	Haemoglobin (gm/L)	is about	132.98
	Protein intake	is about	14.40
	Carbohydrate intake	is about	46.48
	Sugar intake	is about	18.08
	Total fat intake	is about	39.32
	Total saturated fat intake	is about	21.23
	Salt intake (mg)	is about	3058.25
Then	Risk of cardiovascular disease	is	**High**

Rule 2 explains that higher levels of ratio of total cholesterol to HDL at older age leads to increased risk of cardiovascular disease in the presence of higher intake of carbohydrates, fat, sugar and salt.

Rule 3:

If	Age (years)	is about	57
	Pulse (beats/min.)	is about	69
	Waist circumference (cm)	is about	74.82
	Sub-scapular/triceps skinfold	is about	0.77
	Ratio of total cholesterol to HDL	is about	3.77
	Haemoglobin (gm/L)	is about	138.21
	Protein intake	is about	12.16
	Carbohydrate intake	is about	35.93
	Sugar intake	is about	11.44
	Total fat intake	is about	42.13
	Total saturated fat intake	is about	20.54
	Salt intake (mg)	is about	2221.08
Then	Risk of cardiovascular disease	is	**High**

The major risk determinant factors for cardiovascular disease on the basis of Rule 3 are waist circumference, high levels of haemoglobin and increased intake of carbohydrates, fat, sugar and salt.

Rule 4:

If	Age (years)	is about	66
	Pulse (beats/min.)	is about	58
	Waist circumference (cm)	is about	70.41
	Sub-scapular/triceps skinfold	is about	1.47
	Ratio of total cholesterol to HDL	is about	3.71
	Haemoglobin (gm/L)	is about	128.85
	Protein intake	is about	12.48
	Carbohydrate intake	is about	41.44
	Sugar intake	is about	22.64
	Total fat intake	is about	46.18
	Total saturated fat intake	is about	23.81
	Salt intake (mg)	is about	1424.55
Then	Risk of cardiovascular disease	is	**High**

Rule 4 explains that the factors contributing to higher risk of cardiovascular disease are mainly, older age and increased intake of carbohydrates, fat, sugar and salt.

Rule 5:

If	Age (years)	is about	75.10
	Pulse (beats/min.)	is about	73.97
	Waist circumference (cm)	is about	78.01
	Sub-scapular/triceps skinfold	is about	0.81
	Ratio of total cholesterol to HDL	is about	3.99
	Haemoglobin (gm/L)	is about	105.86

	Protein intake	is about	12.21
	Carbohydrate intake	is about	45.20
	Sugar intake	is about	16.23
	Total fat intake	is about	42.67
	Total saturated fat intake	is about	14.49
	Salt intake (mg)	is about	2867.26
Then	Risk of cardiovascular disease	is	**High**

Rule 5 explains that risk of cardiovascular disease increases with an increase in age, waist circumference and higher intake of carbohydrates, fat, sugar and salt.

Rule 6:

If	Age (years)	is about	55
	Pulse (beats/min.)	is about	71
	Waist circumference (cm)	is about	77.17
	Sub-scapular/triceps skinfold	is about	0.60
	Ratio of total cholesterol to HDL	is about	6.08
	Haemoglobin (gm/L)	is about	131.99
	Protein intake	is about	19.46
	Carbohydrate intake	is about	33.27
	Sugar intake	is about	12.61
	Total fat intake	is about	41.24
	Total saturated fat intake	is about	10.39
	Salt intake (mg)	is about	2219.09
Then	Risk of cardiovascular disease	is	**High**

According to Rule 6, the risk of cardiovascular disease increases with increased waist circumference, higher levels of ratio of total cholesterol to HDL. Higher intake of dietary nutrients such as carbohydrates, fat, sugar and salt leads to increased risk of cardiovascular disease.

From above mentioned rules for female Subject 1, it can concluded that higher age with higher waist circumference and higher ratio of total cholesterol to HDL along with higher intake of fat, saturated fat, sugar leads to a high risk of getting cardiovascular disease. Age has been an important risk factor for cardiovascular disease even though age cannot be reversed, and ingesting large amounts of dietary nutrients such as carbohydrates, fat, sugar and salt always increases the risk of cardiovascular disease.

A similar set of rules for female Subject 2 were generated from TWNFI which can be used further for better prediction and recommendations. The set of rules generated for female Subject 2 are listed below:

Rule 1:

If	Age (years)	is about	36
	Pulse (beats/min.)	is about	63
	Waist circumference (cm)	is about	81.50
	Sub-scapular/triceps skinfold	is about	0.59
	Ratio of total cholesterol to HDL	is about	4.15
	Haemoglobin (gm/L)	is about	141.21
	Protein intake	is about	13.22
	Carbohydrate intake	is about	50.72
	Sugar intake	is about	15.49
	Total fat intake	is about	36.61
	Total saturated fat intake	is about	13.03
	Salt intake (mg)	is about	3194.51
Then	Risk of cardiovascular disease	is	**Low**

Rule 1 explains that risk of cardiovascular disease is much lower at younger ages if the intake of carbohydrates, fat, sugar and salt is low.

Rule 2:

If	Age (years)	is about	37
	Pulse (beats/min.)	is about	62
	Waist circumference (cm)	is about	93.36
	Sub-scapular/triceps skinfold	is about	2.07
	Ratio of total cholesterol to HDL	is about	3.45
	Haemoglobin (gm/L)	is about	127.79
	Protein intake	is about	15.44
	Carbohydrate intake	is about	40.13
	Sugar intake	is about	15.78
	Total fat intake	is about	40.26
	Total saturated fat intake	is about	8.68
	Salt intake (mg)	is about	3155.04
Then	Risk of cardiovascular disease	is	**Low**

Rule 2 explains that risk of cardiovascular disease is reduced with decreased intake of carbohydrates, fat, sugar and salt. Ratio of total cholesterol to HDL is also an important factor for determining risk of cardiovascular disease.

Rule 3:

If	Age (years)	is about	28
	Pulse (beats/min.)	is about	51
	Waist circumference (cm)	is about	84.12
	Sub-scapular/triceps skinfold	is about	1.65
	Ratio of total cholesterol to HDL	is about	3.81
	Haemoglobin (gm/L)	is about	138.04
	Protein intake	is about	8.32
	Carbohydrate intake	is about	48.75
	Sugar intake	is about	17.80
	Total fat intake	is about	43.07

	Total saturated fat intake	is about	9.97
	Salt intake (mg)	is about	1243.41
Then	Risk of cardiovascular disease	is	**Low**

Rule 3 explained that maintaining the levels of total cholesterol to HDL to a normal range and a decreased intake of carbohydrates, fat, sugar and salt reduces the risk of cardiovascular disease at a young age.

Rule 4:

If	Age (years)	is about	37
	Pulse (beats/min.)	is about	54
	Waist circumference (cm)	is about	72.62
	Sub-scapular/triceps skinfold	is about	2.12
	Ratio of total cholesterol to HDL	is about	3.82
	Haemoglobin (gm/L)	is about	140.52
	Protein intake	is about	19.02
	Carbohydrate intake	is about	36.65
	Sugar intake	is about	11.51
	Total fat intake	is about	44.5
	Total saturated fat intake	is about	6.62
	Salt intake (mg)	is about	3385.8
Then	Risk of cardiovascular disease	is	**Low**

Rule 4 suggested that maintaining the levels of total cholesterol to HDL to a normal range and a decreased intake of carbohydrates, fat, sugar and salt reduces the risk of cardiovascular disease at a young age.

Rule 5:

If	Age (years)	is about	23
	Pulse (beats/min.)	is about	60
	Waist circumference (cm)	is about	76.20

	Sub-scapular/triceps skinfold	is about	1.65
	Ratio of total cholesterol to HDL	is about	4.04
	Haemoglobin (gm/L)	is about	139.05
	Protein intake	is about	13.26
	Carbohydrate intake	is about	51.69
	Sugar intake	is about	26.74
	Total fat intake	is about	35.40
	Total saturated fat intake	is about	17.31
	Salt intake (mg)	is about	5436.92
Then	Risk of cardiovascular disease	is	**Low**

Age is a very important disease risk determinant factor. Rule 5 states that reduced intake of carbohydrates, fat, sugar and salt is also very important in reducing the risk of cardiovascular disease.

Rule 6:

If	Age (years)	is about	38
	Pulse (beats/min.)	is about	59
	Waist circumference (cm)	is about	76.73
	Sub-scapular/triceps skinfold	is about	2.98
	Ratio of total cholesterol to HDL	is about	3.41
	Haemoglobin (gm/L)	is about	133.89
	Protein intake	is about	10.23
	Carbohydrate intake	is about	49.43
	Sugar intake	is about	19.32
	Total fat intake	is about	40.63
	Total saturated fat intake	is about	32.70
	Salt intake (mg)	is about	1976.80
Then	Risk of cardiovascular disease	is	**Low**

Rule 6 suggests that the risk of cardiovascular disease is almost nil, if the ratio of total cholesterol to HDL is less and intake of dietary nutrients such as carbohydrates, fat, sugar and salt is decreased at the same time.

From the above mentioned rules for female subject 2, it can be concluded that major disease risk factors are age, waist circumference, and ratio of total cholesterol to HDL, intake of dietary nutrients such as carbohydrates, fat, sugar and salt.

At younger ages risk of cardiovascular disease is lower if the waist circumference is lower and the ratio of total cholesterol to HDL is also within the normal range along with a lower intake of calories from carbohydrates, fat and sugar and salt intake is reduced.

6.5 Integrated framework of ontology based personalized cardiovascular disease risk analysis

The chronic disease ontology (Chapter 4) will be used with the personalized risk evaluation system. The framework presented here with a software platform bring together the chronic disease ontology (CDO) and machine learning techniques to facilitate sophisticated adaptive data and information storage for chronic diseases and information retrieval, personalized modeling and knowledge discovery. Figure 6.13 presents the framework model example for cardiovascular disease risk analysis.

The framework utilizes the chronic disease ontology based data, as well as new knowledge inferred from the data embedded in the ontology. The platform allows for the adaptation of an existing knowledge base to new data sources and through entering results from machine learning and reasoning models.

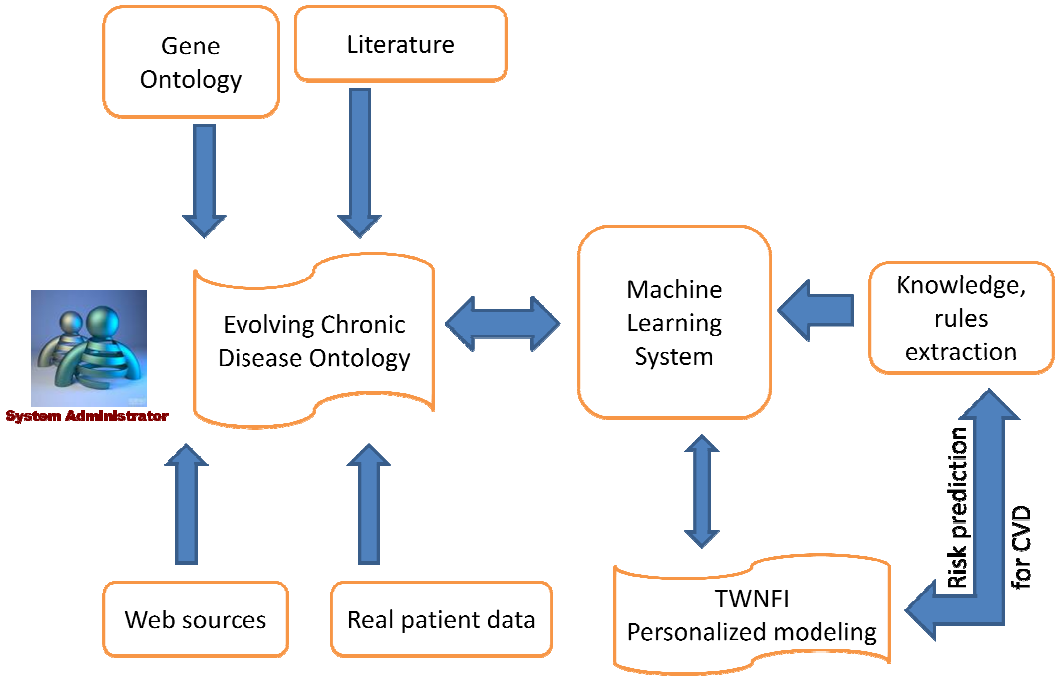


Figure 6.13. Integrated framework of ontology based personalized cardiovascular disease risk analysis.

The first module is the chronic disease ontology (CDO) developed in Protégé and is a knowledge and data repository module; the second module is a machine learning module such as a personalized method (TWNFI). The third module is an interface module between the two modules, which is specific for every application.

The system is able to combine data from numerous sources to provide individualized person/case reports and recommendations on actions/interventions to help modify the outcome in a desired direction based on previously accumulated information on input variables and outcomes in the database. This framework is explained here with few examples.

6.6 Examples of integration of the chronic disease ontology and the personalized risk evaluation system for cardiovascular disease

There is one module, the chronic disease ontology which contains all the knowledge and data for three chronic diseases including the genetic information explained in Chapter 4. The interface system between the CDO and the personalized model is able to extract all the information about cardiovascular disease and pass the information on to the personalized risk evaluation system.

Example1: The personalized risk evaluation system (TWNFI) predicts risk of developing cardiovascular disease on the basis of the real patient data available and the information retrieved from the ontology. If a person has high blood pressure, the interface looks for a gene responsible for cardiovascular disease by means of high blood pressure, for example, gene Adductin (ADD1) is a gene responsible for high blood pressure (Pedrinelli et al, 2006). Then, taking that gene into consideration, the person might have a mutation of that gene or may be linked to a mutation of that gene. Further recommendations such as genetic tests, risk of cardiovascular disease can be derived. The gene risk of hypertension or other associated symptoms of cardiovascular disease can be used for risk analysis and personalized advice regarding disease.

Example2: Similarly the system can find other genes related to other manifestations of cardiovascular disease. For example, gene Apolipoprotein E (APOE) is associated with lipid metabolism (Lahoz et al, 2001); a person with a high ratio of cholesterol to HDL must have allelic variations in the gene as allelic variation in APOE is associated with high levels of cholesterol to HDL.

Example3: Another example is polymorphism in gene FABP2 which is associated with fasting blood glucose and high cholesterol and hence with

higher risk of cardiovascular disease in the Argentinean population (Gomez et al, 2007). If a test subject is from the Argentinean population and has high levels of cholesterol or a high intake of fats, the system will look for the genes associated with high levels of cholesterol in the Argentinean population and then will discover that person might have polymorphism in gene FABP2.The model will identify new patterns in the new data and then will update the ontology with new knowledge and can use this new knowledge for subsequent subjects.

Example4: Similarly, the MYBPC3 gene, which is responsible for cardiac disease by means of mutation in Asian population, particularly Indian populations (BBC News, 2009) has been updated in the ontology and can be used when a subject with from Asian population comes, the system will recommend genetic test for MYBPC3 gene for better risk prediction for cardiac disease. The framework presented here is able to use existing information for new subject, to retain new information discovered from new subject and to reuse for new subject.

6.7 Conclusion

The National Nutrition Survey data 1997 (NNS97) from New Zealand population has been described. NNS97 data has been analyzed for prediction of risk of cardiovascular disease. From the data analysis I have found that:

- Age, waist circumference and ratio of total cholesterol to HDL are the most important risk determinant factors (Nesto, 2008; Despres et al, 1990).

- Total saturated fat intake and ratio of sub scapular skinfold to triceps skinfold are the least important factors for the risk of cardiovascular disease in the current data.
- Age has a positive correlation with waist circumference, ratio of total cholesterol to HDL, risk of cardiovascular disease and a negative correlation with ethnicity and salt intake.
- Gender has a positive correlation with saturated fat intake and a negative correlation with haemoglobin, waist circumference and ratio of total cholesterol to HDL, carbohydrate intake, sugar intake, salt intake and risk of cardiovascular disease.
- Waist circumference has a positive correlation with the ratio of total cholesterol to HDL and risk of cardiovascular disease and a negative correlation with carbohydrate and sugar intake.
- Haemoglobin shows a positive correlation with waist circumference, ratio of total cholesterol to HDL, salt intake and risk of cardiovascular disease. Carbohydrate intake shows a positive correlation with sugar intake and saturated fat intake. Carbohydrate intake and sugar intake show a negative correlation with protein intake and total fat intake. Sugar intake also shows a negative correlation with salt intake. Total fat intake has a positive correlation with salt intake and a negative correlation with saturated fat intake. Saturated fat intake shows a negative correlation with salt intake.
- Among dietary variables, protein intake has a negative correlation with carbohydrate and sugar intake showing that carbohydrate or sugar rich diet leads to risk of cardiovascular disease but substituting carbohydrates or sugar with protein will decrease the risk of

cardiovascular disease which was also explained earlier by Appel and colleagues in 2005.

To predict risk of CVD, several methods have been developed so far including risk evaluation charts, web based interfaces which include clinical variables and smoking status but no method includes nutritional variables for prediction of risk of cardiovascular disease. The novelty of the method explained in this chapter is that it includes nutritional variables along with clinical and anthropometric variables.

In the present study, the TWNFI method has been found to have better accuracy and to be more informative than other methods. This method provides a personalized approach for risk analysis. In TWNFI, along with risk of CVD, ranking of variables and sets of rules for those variables can also be obtained which results in more personalized and accurate recommendations.

The NNS97 data used in the present study provides information about dietary variables but the main limitation of the data is lack of information about physical activity and smoking status which are also very important factors for determining risk of cardiovascular disease. Another limitation is that this data is cross sectional rather than longitudinal i.e. older people in the National Nutrition Survey have a different history to younger ones, so it cannot be assumed that increased waist circumference and age will apply to every individual. TWNFI can be used on other data sets with more new variables as with more input variables more knowledge can be derived for risk evaluation for cardiovascular disease.

Chapter 7. Type 2 Diabetes and Obesity Risk Evaluation and Knowledge Discovery Based on the Chronic Disease Ontology (CDO)

This chapter explores type 2 diabetes gene data from Italian population living in Italy. This chapter explains type 2 diabetes, existing methods to predict the risk of type 2 diabetes. This chapter also introduces a new personalized type 2 diabetes risk evaluation approach using clinical and genetic variables. The last section of this chapter explains the integration framework for personalized model for type 2 diabetes risk evaluation and chronic disease ontology.

7.1 Type 2 diabetes, prevalence and description

Type 2 diabetes mellitus (T2DM) is one of the most common chronic “lifestyle” diseases with a high prevalence throughout the world (The FIELD Study Investigators, 2004; Wild et al, 2004; Lindstrom and Tuomilehto, 2003; Zimmet et al, 2001; King et al, 1998; King and Rewers, 1993). Type 2 diabetes is directly responsible for 5% of all deaths globally. According to the World Health Organization, in 2005, approximately 1.1 million people died from type 2 diabetes. The number of adults with T2DM in the world is estimated to increase by 122% from 135 million in 1995 to 300 million in 2025, which will reflect an increase of about 42% in developed countries and 170% in developing countries (Kiberstis, 2005; King et al, 1998).

Type 2 diabetes is a major health problem in New Zealand (Cheng, 2006; Joshy and Simmons, 2006; New Zealand Guidelines Group, 2003(a); Moore and Lunt, 2000; Simmons, 1996 (a, b)). Prevalence of type 2 diabetes is very high in New Zealand and been called ‘The Quiet Killer’ (Laugesen, 2006). About 200,000

people are affected with type 2 diabetes and nearly 4000 deaths are caused by type 2 diabetes every year in New Zealand (Cheng, 2006). In the 2000 New Zealand Health Strategy one of thirteen population health objectives is to "reduce the incidence and impact of diabetes" and prevention of type 2 diabetes is one of three disease priority areas identified (King, 2000).

There are two main types of diabetes mellitus; type-1 and type-2. Both result in glucose being present in urine which is where the name diabetes mellitus comes from. The translation is freely flowing (diabetes) and tasting like honey (mellitus). Type-1 diabetes is an auto-immune disease in which insulin producing pancreatic beta cells are destroyed, so insulin production in the body is stopped abruptly resulting in increased blood sugar.

Type 2 diabetes, characterized by persistent abnormally high blood glucose concentrations, is a serious disease which progressively develops throughout a life-time. The underlying defects are related to insulin action, insulin secretion or both. The disease progresses to the point that the beta cells of the endocrine pancreas can no longer produce insulin and exogenous insulin must be injected for survival.

Type 2 diabetes cannot be cured, but careful control of blood glucose by diet, exercise and/or drugs can prevent or delay the complications of this disease. When insulin is absent or ineffective, blood glucose concentration increases resulting in abnormal function of all cells of the body. Type 2 diabetes is the most common type of diabetes and develops more slowly. Globally about 90% of all cases of diabetes are type 2 diabetes (New Zealand Guidelines Group, 2003(a); Zimmet et al, 2001; Astrup and Finer, 2000).

The non-modifiable factors which cause type 2 diabetes are genetic factors, ethnicity and age (Rose et al, 2004). Being over 45 years of age raises the risk of developing type 2 diabetes but the age limit varies among different ethnic groups and is becoming lower as obesity increases. The main risk factors include:

- having a first-degree relative (such as a parent, brother, or sister) with diabetes; namely a genetic or epigenetic factor;
- for women, having had gestational diabetes, or giving birth to at least one overweight baby ;
- being male; men have higher risk at an earlier age than women;
- having blood pressure of 140/90 or higher;
- having abnormal cholesterol levels or triglyceride concentrations in fasting blood;
- lifestyle (being inactive or doing no physical exercise);
- being overweight, obese or having a high waist measurement;
- being Maori, of Pacific or Indian ethnic origins.

7.2 Obesity, prevalence and description

Major risk factors of type 2 diabetes are also those that are associated with the accumulation of excess body fat. Accumulation of excess fat is termed “obesity”. An obesogenic environment (Berkeley and Lunt, 2006) and obesity are risk factors for many other chronic diseases including cardiovascular disease and type 2 diabetes (Rigby and James, 2003; Wilson et al, 2001; Kopelman, 2000; Must et al, 1999). Obesity is also referred to as the “New World Syndrome” (Nammi et al, 2004).

The global prevalence of obesity has increased substantially over the last fifteen years (Christakis and Fowler, 2007; James, 2004; Caterson and Gill, 2002; Wilson et al, 2001). The World Health Organization declared obesity as a 'global epidemic' in 1997 (World Health Organization, 2000). According to the International Obesity taskforce (2009) about 300 million people around the world are estimated to be obese.

The prevalence of overweight and obesity in New Zealand continues to increase (Johnston, 2009; Wilson et al, 2001). About 26.5% of adults in New Zealand are obese (Johnston, 2009). As per the Social report 2007, of the Ministry of Social Development, New Zealand, the number of obese people has doubled from 10 percent to 20 percent in the male population and from 13 percent to 22 percent in the female population. Obesity in New Zealand is more prevalent among Maori and Pacific Islanders than other ethnic groups (Social report, 2007). Obesity is associated with significant health problems, health cost and increased risk of early death in the New Zealand population (Wilson et al, 2001; Wright, 2001).

Obesity is the accumulation of excess adipose tissue to the extent that it impairs both physical and psychosocial health and well being. In clinical practice, obesity is crudely measured by using a formula which combines weight and height called body mass index (BMI) (James, 2004; Kopelman, 2000). The body mass index is calculated as weight in kilograms divided by square of height in meters. According to the WHO classification of BMI for obesity, a BMI 30kg/m^2 or above is considered obese (Table 7.1).

Table 7.1

WHO classification of BMI for obesity (World Health Organization, 2000).

Classification	Body Mass Index (BMI)	Risk of co morbidities
Underweight	< 18.5 kg/m^2	Low (but risk of other clinical problems increased)
Normal range	18.5–24.9 kg/m^2	Average
Overweight	≥25 kg/m^2	
Pre-obese	25.0–29.9 kg/m^2	Increased
Obese class 1	30.0–34.9 kg/m^2	Moderate
Obese class 2	35.0–39.9 kg/m^2	Severe
Obese class 3	≥ 40.0 kg/m^2	Very severe

Obesity is a heterogeneous group of conditions with multiple causes. Obesity is determined by interaction of genetic, environmental and psychosocial factors through the physiological mediators of energy intake and expenditure (Kopelman, 2000). The treatment for obesity in primary care has mainly focused on weight loss. Obesity is particularly associated with an increased risk of developing type 2 diabetes or non-insulin dependent diabetes (James and Rigby, 2004; Hu et al, 2001; Wing et al, 2001; Colditz et al, 1995; 1990). Obesity is the main risk factor associated with about 80-90 percent of known cases of type 2 diabetes (Astrup and Finer, 2000). Since type 2 diabetes is closely associated with obesity and is the main aetiological cause of type 2 diabetes, Astrup and Finer in 2000 proposed the term ‘diaobesity’ as it reflects both an etiology and clinical presentation.

7.3 Diabetes prediction models

It has become a major task for public health administrators to identify individuals and groups with high risk of type 2 diabetes and apply targeted interventions. There have been several models, namely, 'The global diabetes model' (Brown et al, 2000(a, b)), 'The diabetes risk score' (Lindstrom and Tuomilehto, 2003), the 'Archimedes diabetes model' (Eddy and Schlessinger, 2003(a, b)), the Diabetes risk score in Oman (Al-Lawati and Tuomilehto, 2007), the 'Genetic Risk Score' (Cornelis, et al, 2009) (Table 7.2). All these models predict risk of future complications associated with type 2 diabetes in people with diagnosed type 2 diabetes.

The global diabetes model (GDM) is a continuous, stochastic micro simulation (individual by individual approach) model of type 2 diabetes. The GDM is a computer program and predicts longevity, quality of life, medical events and expenditures for groups and individuals with type 2 diabetes. The GDM calculates rates and probabilities of the medical events in diabetic individuals (Brown et al, 2000 (a, b)).

The diabetes risk score was developed in 2003 by Lindstrom and Tuomilehto to predict future risk of type 2 diabetes. The Archemides model (Eddy and Schlessinger, 2003(a, b)) is a mathematical model with a person by person simulation. This model predicts diabetes and its complications, coronary artery disease, congestive heart failure and asthma. This model uses Markov and Monte Carlo models to describe the progress of disease (Herman, 2003).

The 'German diabetes risk score' is the only tool which predicts risk of having type 2 diabetes and looks at how to lower the risk. It is valid for people between 35 to 65 years of age. The German diabetes risk score, a publically

available tool (Schulze et al, 2007) requires information about age, waist circumference, height, history of hypertension, physical activity, smoking, and consumption of red meat, whole grain bread, coffee and alcohol.

Table 7.2

Comparison of exiting methods to predict risk of type-2 diabetes.

Prediction Model	Year	Population	Method	Variables used	Outcome
Genetic Risk Score	2009	European	Logistic regression	Age, gender, BMI, life style, family history, genetic variants	Risk for type 2 diabetes
German Risk score	2007	German	Cox regression models with forward selection	Age, Waist circumference, height, history of hypertension, physical activity, smoking, consumption of red meat, whole grain bread, coffee, alcohol	Risk for type-2 diabetes or undiagnosed diabetes
Diabetes risk score	2007	Oman	Backward stepwise logistic regression	Age, gender, waist circumference, BMI, Systolic blood pressure, diastolic blood pressure, family history of diabetes, diabetes test (OGTT)	Risk for diabetes
Archiemedes	2003			Age, sex, weight, Family history of diabetes/ heart disease, weight, Blood pressure, Cholesterol, fasting glucose, A1C, health history (smoking), medication related to diabetes, blood pressure, cholesterol	Prediction of risk of diabetes, treatment and complications
Diabetes risk score	2003	Finland	Multivariate logistic regression method	Age, BMI, Waist circumference, use of BP medication, history of blood glucose, physical activity, daily consumption of vegetables, fruits and berries	Future risk for diabetes
Global diabetes model	2000		Monte Carlo method	Age, gender, race/ethnicity, duration of diabetes, SBP, HDL, LDL, triglycerides, HbA1C, smoking, use of aspirin	Prevalence of diabetes, retinopathy, nephropathy, neuropathy

It has been reported that from the existing methods for predicting risk of type 2 diabetes, The Archimedes Model predicts the risk with better sensitivity and specificity than other models (Stern et al, 2008). Recently, the 'Genetic Risk Score' has been developed which uses multiple genetic as well as conventional risk factors (Cornelis et al, 2009). Because these methods calculate risk of type 2 diabetes globally and they are not the same as the proposed methodology in this thesis, personalized approach for risk prediction of type 2 diabetes is described in this chapter. In the previous chapter, a model was created to predict risk of cardiovascular disease using clinical and nutritional variables. In this chapter a personalized model has been created for type 2 diabetes risk prediction using clinical and genetic variables.

The aim of the research reported in this chapter was to create a personalized model for predicting risk of type 2 diabetes. In this chapter genetic variables have been used along with clinical variables to create a personalized model to predict risk of type 2 diabetes. The next section of this chapter describes the data and methods used for creating a diabetes risk model using genetic markers along with clinical variables.

7.4 Data Exploration

The data used for creating a personalized model for predicting risk of type 2 diabetes used sample data collected and derived from the Italian people living in Italy. The description and process of analysis of the data and variables is described in the following section.

7.4.1 Description of selected data

The dataset was for a total of 74 subjects, in which there are 48 male subjects and 26 female subjects and 93 variables. The variables include 87 type 2

diabetes genes, age, gender, haemoglobin and fasting total cholesterol, triglycerides and glucose and whether a diagnosis of type 2 diabetes had been made. The list of all 87 genes and the 6 other variables is described in table 7.4. This data has two classes, class 0 are subjects without diagnosed type 2 diabetes and class 1 are subjects who have been diagnosed with type 2 diabetes. Prevalence of type 2 diabetes in this group was high (Table 7.3). More (seven out of ten) of the male subjects had been diagnosed with type 2 diabetes, than female (six out of ten). Age ranges were the same for both sexes, for men the range was 39 to 67 years and for women 40 to 63 years

Table 7.3

Distribution of male and female subjects classified as diagnosed without or with type 2 diabetes.

	Class 0 (without T2DM)	**Class 1 (with T2DM)**	**Total**
Male	15 (58%)	33(69%)	48
Female	11(42%)	15 (31%)	26
	26	48	74

7.4.2 Rationale for selecting variables

Data was divided into three categories; general variables such as age, gender, clinical variables such as triglycerides and cholesterol and genetic variables and these clusters were used to build a personalized risk prediction model for type 2 diabetes.

General and clinical variables: The previous name for type 2 diabetes was “adult onset diabetes” because in the past, type 2 diabetes rarely occurred in young people (McKinlay and Marceau, 2000). Men have been shown to have an earlier age of onset of type 2 diabetes than women (Meisinger et al, 2002).

So gender is an important consideration and separate models need to be developed for each gender. Raised concentrations of cholesterol and triglycerides in the blood are risk factors for type 2 diabetes. According to the New Zealand Guidelines Group (2003b), total cholesterol higher than 4 mmol/L or triglycerides higher than 1.7 mmol/L are considered as major risk factors for type 2 diabetes and cardiovascular disease (New Zealand Guidelines Group, 2003(a)).

Genes related to type 2 diabetes: Genes are very important risk factors for type 2 diabetes (Diamond, 2003; Zimmet, 1997; 1992; Neel, 1982; 1962) and many have been identified. Neel (1982; 1962) proposed the theory of thrifty genes responsible for causing type 2 diabetes. According to this theory there are genes that are associated with metabolically thrifty traits such as accumulation of body fat. The natural selection would ensure survival in times of famine and starvation. However when there is an abundance of high calorie foods, the expression of genes would continue to efficiently regulate the metabolism of macronutrients and promote the rapid storage of fat. This may lead to obesity and type 2 diabetes (Neel, 1962). The present data comprises 87 genes which have previously been shown to be directly or indirectly responsible for type 2 diabetes (Table 7.4). Out of these 87 genes only six genes were selected on the basis of signal to noise ratio as described in next section.

Table 7.4

List of clinical variables and genes used for personalized risk evaluation and knowledge discovery.

No	Variables	Description
1	Gender	Male, female
2	Age	years
3	Haemoglobin	g/L
4	Fasting blood glucose	mmol/L
5	Cholesterol	mmol/L
6	Triglycerides	mmol/L
7	ANG	Angiogenin, ribonuclease, RNase A family, 5
8	ANGPT1	Angiopoietin 1
9	ANGPT2	Angiopoietin 2
10	ANGPTL2	Angiopoietin-like 2
11	ANGPTL3	Angiopoietin-like 3
12	CEACAM1	Carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein)
13	COL15A1	Collagen, type XV, alpha 1
14	COL4A2	Collagen, type IV, alpha 2
15	CTGF	Connective tissue growth factor
16	CXCL10	Chemokine (C-X-C motif) ligand 10
17	EDIL3	EGF-like repeats and discoidin I-like domains 3
18	EPHB2	EPH receptor B2
19	FBLN5	Fibulin 5
20	FGA	Fibrinogen alpha chain
21	FGF1	Fibroblast growth factor 1 (acidic)
22	FGF2	Fibroblast growth factor 2 (basic)
23	FLT1	Fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)
24	HEY1	Hairy/enhancer-of-split related with YRPW motif 1
25	IFNB1	Interferon, beta 1, fibroblast
26	IFNG	Interferon, gamma
27	IL12A	Interleukin 12A (natural killer cell stimulatory factor 1, cytotoxic lymphocyte maturation factor 1, p35)
28	ITGAV	Integrin, alpha V (vitronectin receptor, alpha polypeptide, antigen CD51)
29	MMP2	Matrix metallopeptidase 2
30	PECAM1	Patelet/endothelial cell adhesion molecule

31	PGK1	Phosphoglycerate kinase 1
32	PLG	Plasminogen
33	PRL	Prolactin
34	SEMA3F	Sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3F
35	SERPINB5	Serpin peptidase inhibitor, clade B (ovalbumin), member 5
36	SERPINC1	Serpin peptidase inhibitor, clade C (antithrombin), member 1
37	TEK	TEK tyrosine kinase, endothelial
38	TGFA	Transforming growth factor, alpha
39	TGFB1	Transforming growth factor, beta 1
40	THBS1	Thrombospondin 1
41	TIE1	Tyrosine kinase with immunoglobulin-like and EGF-like domains 1
42	TIMP2	TIMP metallopeptidase inhibitor 2
43	TIMP3	TIMP metallopeptidase inhibitor 3
44	TNF	Tumor necrosis factor (TNF superfamily, member 2)
45	TNFSF15	Tumor necrosis factor (ligand) super family, member 15
46	TNMD	Tenomodulin
47	VEGFA	Vascular endothelial growth factor A
48	VEGFB	Vascular endothelial growth factor B
49	VEGFC	Vascular endothelial growth factor C
50	ACHE	Acetylcholinesterase (Yt blood group)
51	AKT1	v-akt murine thymoma viral oncogene homolog 1
52	ANGPT4	Angiopoietin 4
53	BAI1	Brain-specific angiogenesis inhibitor 1
54	BMP2	Bone morphogenetic protein 2
55	CDKN1A	Cyclin-dependent kinase inhibitor 1A (p21, Cip1)
56	CHGA	Chromogranin A (parathyroid secretory protein 1)
57	COL18A1	Collagen, type XVIII, alpha 1
58	COL4A3	Collagen, type IV, alpha 3 (Goodpasture antigen)
59	CSF3	Colony stimulating factor 3 (granulocyte)
60	CXCL12	Chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)
61	DLL4	Delta-like 4 (Drosophila)
62	EFNB2	Ephrin-B2
63	F2	Coagulation factor II (thrombin)
64	FLT3	Fms-related tyrosine kinase 3
65	FN1	Fibronectin 1
66	FST	Follistatin

67	GJB1	Gap junction protein, beta 1, 32kDa
68	GRN	Granulin
69	HIF1A	Hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)
70	HSPG2	Heparan sulfate proteoglycan 2
71	IL17A	Interleukin 17A
72	IL21	Interleukin 21
73	IL6	Interleukin 6 (interferon, beta 2)
74	IL7	Interleukin 7
75	IL8	Interleukin 8
76	ITGA5	Integrin, alpha 5 (fibronectin receptor, alpha polypeptide)
77	ITGAV	Integrin, alpha V (vitronectin receptor, alpha polypeptide, antigen CD51)
78	ITGB3	Integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61)
79	KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
80	MMP3	Matrix metallopeptidase 3 (stromelysin 1, progelatinase)
81	MMRN1	Multimerin 1
82	NOS3	Nitric oxide synthase 3 (endothelial cell)
83	NT5C3	5'-nucleotidase, cytosolic III
84	PF4	Platelet factor 4
85	POSTN	Periostin, osteoblast specific factor
86	PRKAB1	Protein kinase, AMP-activated, beta 1 non-catalytic subunit
87	PTGS2	Prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)
88	SELL	Selectin L
89	TGFB1	Transforming growth factor, beta 1
90	THBS2	Thrombospondin 2
91	TIMP1	TIMP metallopeptidase inhibitor 1
92	TNNI1	Troponin I type 1 (skeletal, slow)
93	VASH1	Vasohibin 1

For gene expression data, total RNA was extracted from whole peripheral blood samples using an RNA easy Mini Kit (Qiagen) according to the manufacturer's instructions. RNA was quantified by spectrometry while the quality was confirmed by gel electrophoresis. Then reverse transcription of the purified RNA was reverse transcribed using Superscript III reverse

transcriptase (Invitrogen). Gene expression assays was carried out with TaqMan® Low Density Array Fluidic card (TaqMan® Human Angiogenesis Array, registered trademarks) based on Applied Biosystems PRISM® 7900HT comparative dd CT (delta delta or 2 delta) method. The ddCt method is one of the first methods used to calculate gene expression results. ddCt is an approximation method and is easy to implement. However, it reduces lot of experiment effort by making these assumptions and is easy to implement, and in many cases they return results similarly to other non-approximation methods (Livak and Schmittgen, 2001). In present data for normal subjects the value of genes is 1. Value 1 of reach gene for normal subjects results from ratio that machine sets in default by using following formula:

ddCT = [(ct 18S) - (ct sample patient gene)] / [(ct 18S) - ct of the normal patient gene)].

The above mentioned formula has been used to calculate gene expression value for subjects with diabetes and on the basis of an assumption that all normal subjects have gene expression value 1.

7.4.3 Statistical Analysis

During data analysis, it has been found that more male subjects have type 2 diabetes in comparison to female subjects. The males with type 2 diabetes were older than those without type 2 diabetes. The male subjects age ranges between 39 years to 67 years and female subjects age ranges from 40 years to 63 years. Most of the male subjects in diseased class belong to higher ages. The mean, highest and lowest values of the variables for whole data, male and female subjects are listed in Table 7.5.

Table 7.5

Comparison of minimum, maximum and average values of clinical variables among male and female subjects.

	Male subjects			Female subjects		
	Minimum	**Maximum**	**Average**	**Minimum**	**Maximum**	**Average**
Age	39	67		40	63	
Haemoglobin (g/L)	12.0	17.6	15.0	11.9	15.5	13.6
Glucose (mmol/L)	4.0	20.0	7.6	4.2	14.5	6.4
Cholesterol (mmol/L)	3.3	7.9	5.2	3.7	7.6	5.3
Triglycerides (mmol/L)	0.5	4.6	1.9	0.4	3.5	1.4

From table 7.5, it is found that the average values of fasting blood glucose and triglycerides for males are higher than the recommended concentrations (mentioned in section 7.4.2). Females had lower haemoglobin and triglyceride concentrations than males.

The data has been sequentially analysed with different methods including signal to noise ratio, t-test and correlation analysis for feature selection before building a model to predict risk of type 2 diabetes. This analysis was done using NeuCom and Siftware.

Signal to noise ratio of variables: The data was analyzed with the signal to noise ratio method to get a quantitative measure for the influence of each variable in the whole dataset on the outcome measure. This method extracts the signals from each variable, and then ranks the variables according to the strength of association when compared with the outcome. Higher values of

signal to noise ratio for a variable explains that that variable is more important. Figure 7.1 shows the ranking of first 25 variables based on the signal to noise ratio for all subjects. Signal to noise ratio was used to select the genes with the highest signal to noise ratio from the total of 87 genes. For feature selection student's t-test has been performed by using Siftware (Appendix F). Results achieved from signal to noise ratio are exactly similar to student's t-test.

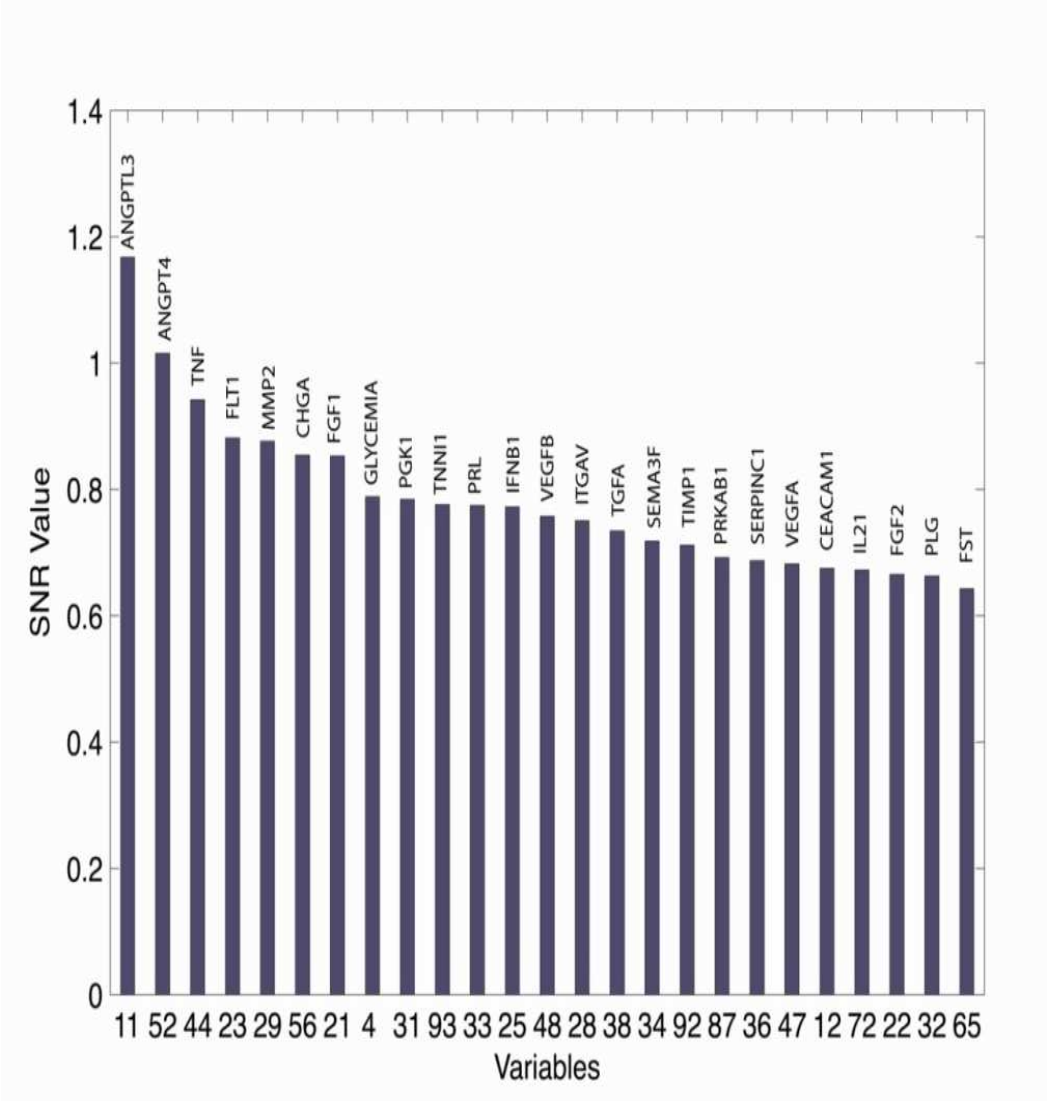


Figure 7.1. Bar graph showing ranked variables (highest to lowest) for whole data using signal to noise ratio for prediction of type 2 diabetes by gene markers. ANGPTL3, AGPT4, TNF genes are ranked at high position.

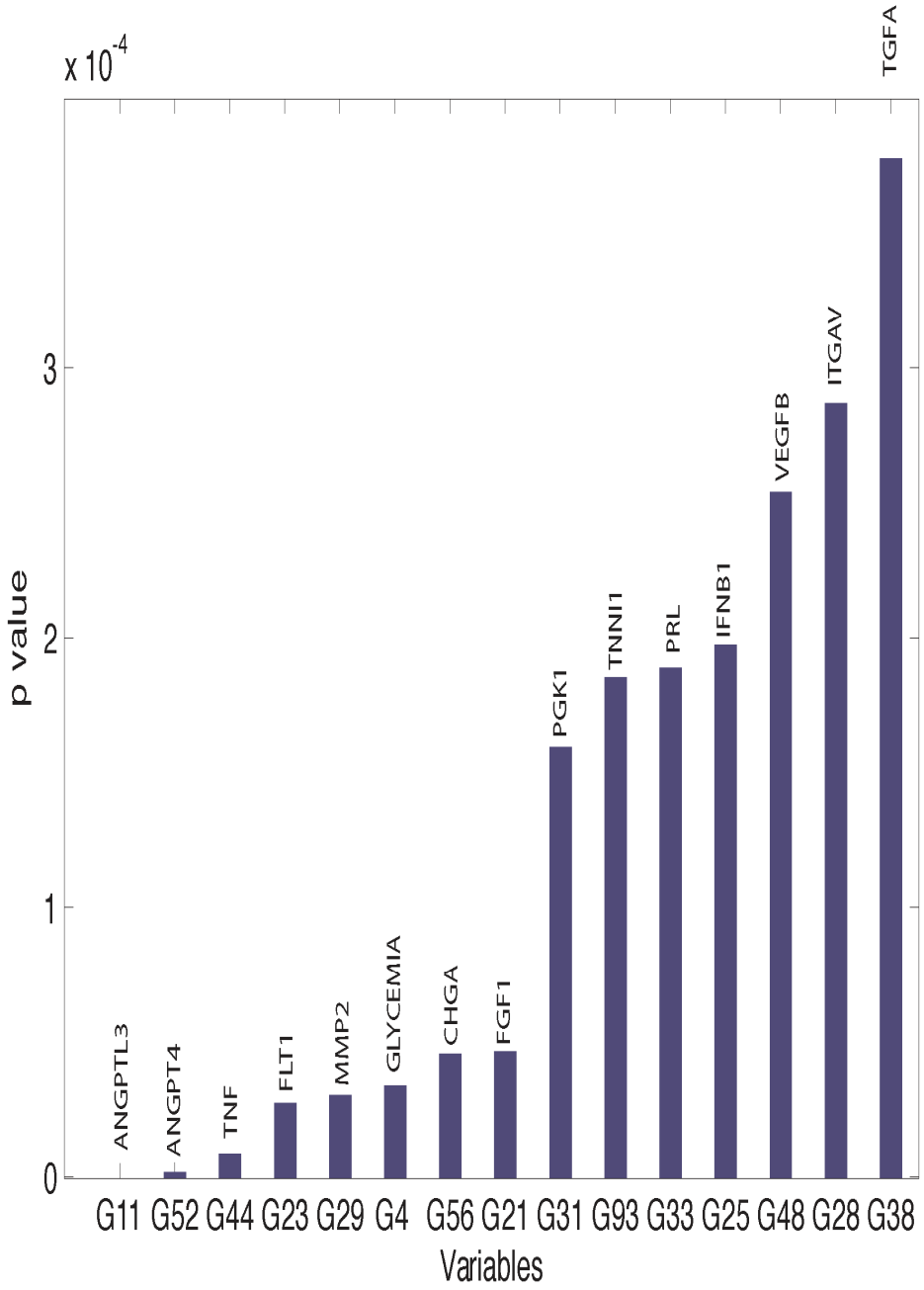


Figure 7.2. Bar graph showing ranked variables (highest to lowest) for whole data for prediction of type 2 diabetes by gene markers using p-value derived from t-test. The lowest p-value explains the most important gene.

According to signal to noise ratio and t-test for the combined male and female subjects, genes ANGPTL3, ANGPT4, TNF, FLT1, MMP2 and CHGA are ranked

highest (Figures 7.1 and 7.2). Interestingly, gene CHGA has not been ranked at same high position for male and female subjects separately.

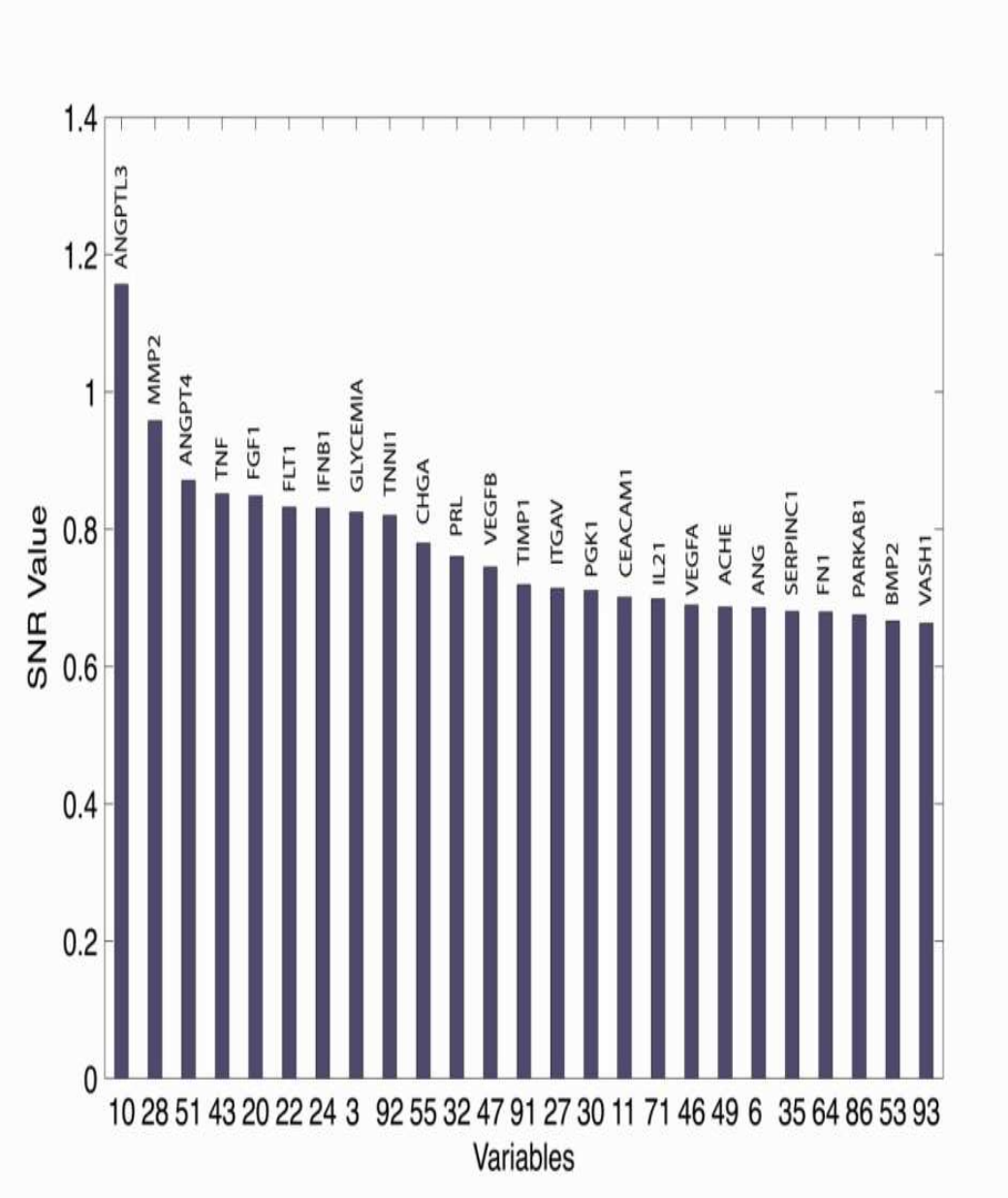


Figure 7.3. Bar graph showing ranked variables (highest to lowest) for male subjects using signal to noise ratio for prediction of type 2 diabetes by gene markers. ANGPTL3 and MMP2 are the most important genes for male subjects and are ranked at highest position.

It has been found that genes HIF1A and TNFSF15 have been ranked in first six genes but both of these genes are least important for male subjects or while compared with whole data, as these genes do not even rank in first 25 genes (Figure 7.3). Similar results have been found for male subjects by using another method (t-test) by using Siftware (Figure 7.4). Figures 7.5 and 7.6 show the results of signal to noise ratio and t-test for female subjects.

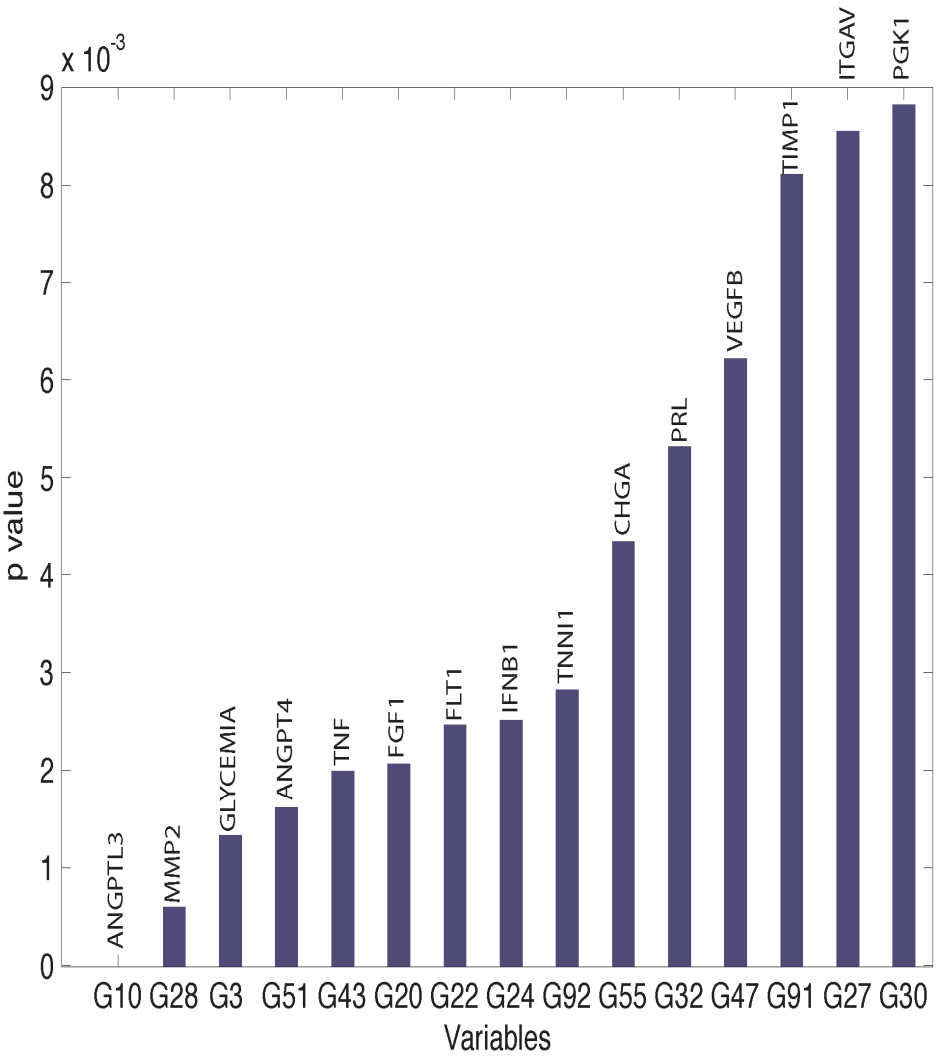


Figure 7.4. Bar graph showing ranked variables (highest to lowest) for male subjects for prediction of type 2 diabetes by gene markers using p-value derived from t-test. ANGPTL3 and MMP2 are most important genes.

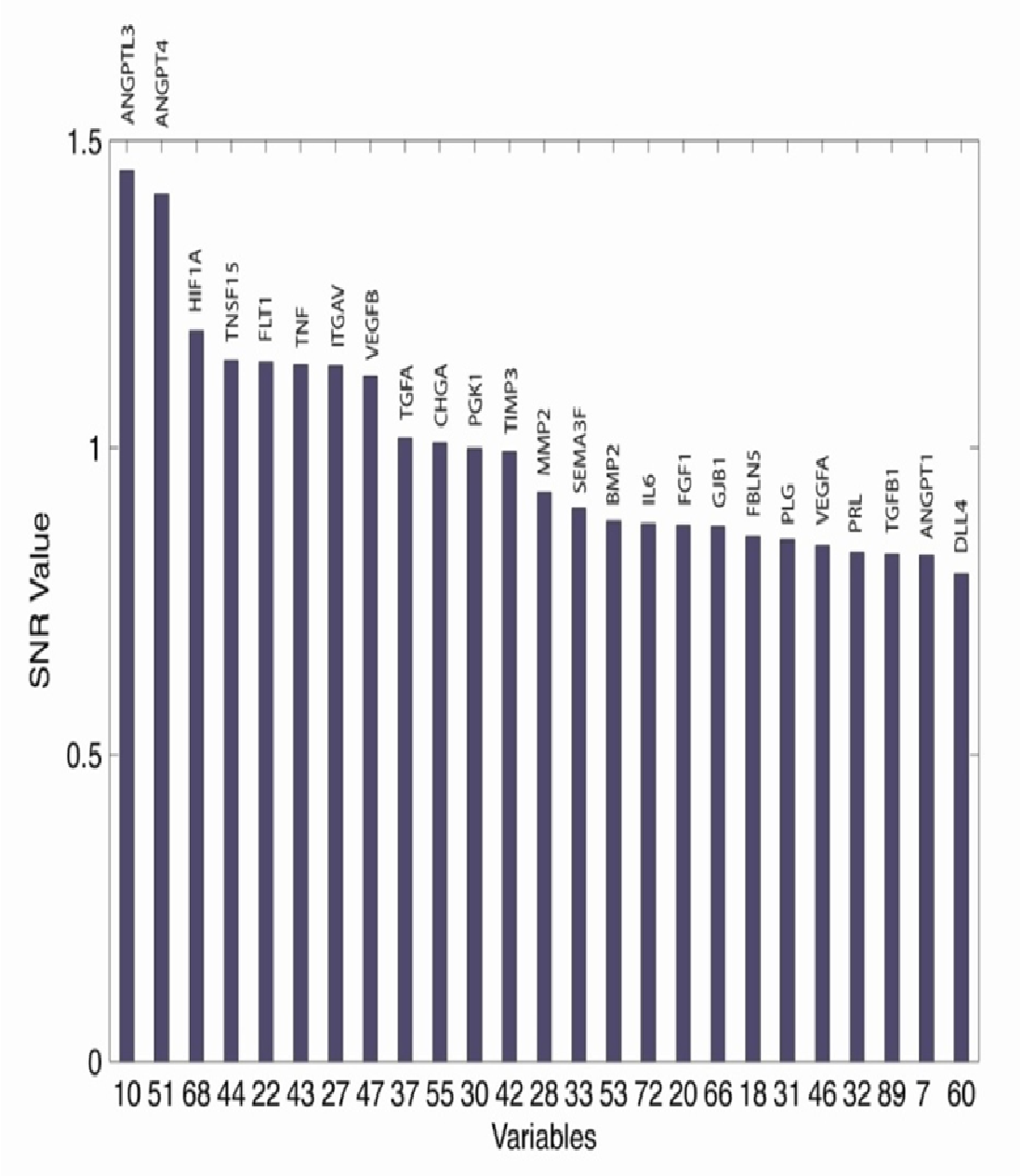


Figure 7.5. Bar graph showing ranked variables (highest to lowest) for female subjects using signal to noise ratio for prediction of type 2 diabetes by gene markers. ANGPTL3 and ANGPT 4 are the most important genes for female subjects.

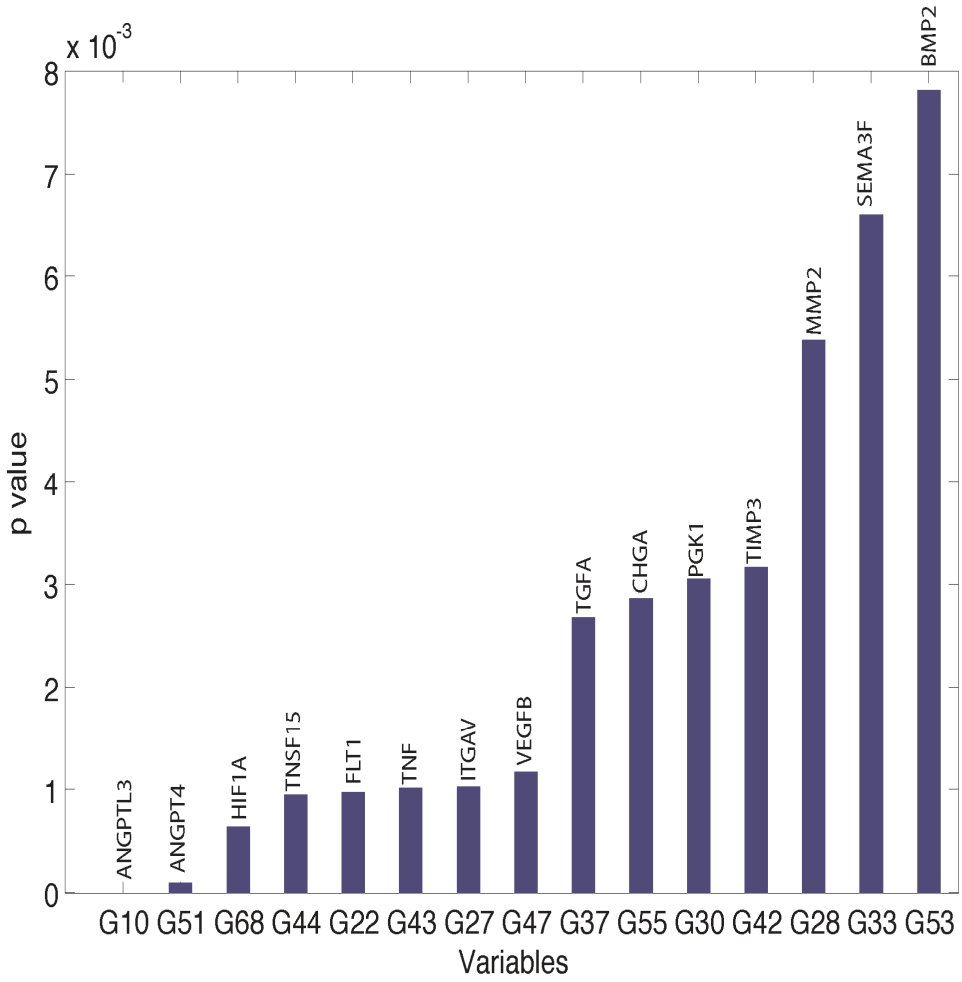


Figure 7.6. Bar graph showing ranked variables (highest to lowest) for female subjects for prediction of type 2 diabetes by gene markers using p-value derived from t-test. ANGPTL3 and ANGPT4 are the most important genes for female subjects.

On the basis of signal to noise ratio and student's t-test for male and female subjects, it was found that male and female subjects show different patterns of ranking of genes. For male subjects, genes ANGPTL3, MMP2, ANGPT4, TNF, FGF1 and FLT1 has been found the most important genes and genes ANGPTL3, ANGPT4, HIF1A, TNFSF1S, FLT1 and TNF have been found the most important genes for female subjects.

Genes MMP2 and FGF1 have been ranked at position 2 and 5 respectively in male subjects while these genes are ranked at position 13 and 17 respectively in female subjects. On the other hand genes HIF1A and TNSF1S which are ranked on position 3, 4 respectively in female subjects, same genes have not been even ranked in first 25 genes in male subjects.

The first six genes of highest importance for male and female subjects were selected for further analysis and to build a personalized sex-specific risk prediction model. As genes are ranked differently as per signal to noise ratio for male and female subjects, different genes have been selected for personalized modeling. Table 7.6 shows the names of selected genes by sex and for the data combined.

Table 7.6

List of first six genes for whole data and male, female subjects according to signal to noise ratio and t-test.

	Male subjects	**Female subjects**	**Whole data**
1	ANGPTL3	ANGPTL3	ANGPTL3
2	MMP2	ANGPT4	ANGPT4
3	ANGPT4	HIF1A	TNF
4	TNF	TNSF1S	FLT1
5	FGF1	FLT1	MMP2
6	FLT1	TNF	CHGA

The gene CHGA which was ranked at sixth position for whole data but the same gene was ranked at position 10 for male and female subjects. Table 7.7 and table 7.8 show the genes and their description for male and female subjects respectively.

Table 7.7

List of genes selected for personalized modeling for male subjects with their description.

	Gene symbol	Gene name	Gene Function
1	**ANGPTL3**	Angiopoietin-like 3	o Cholesterol homeostasis o Cholesterol metabolic process o Fatty acid metabolic process o Glycerol metabolic process o Positive regulation of lipid catabolic process
2	**MMP2**	Matrix metalloproteinase-2	o Protein binding o Metalloendopeptidase activity o Regulation of enamel mineralization o Proteolysis
3	**ANGPT4**	Angiopoietin 4	o Vascular endothelial growth factor receptor binding
4	**TNF**	Tumor necrosis factor	o Chronic inflammatory response to antigenic stimulus o Inflammatory response o Negative regulation of lipid catabolic process o Regulation of insulin secretion
5	**FGF1**	fibroblast growth factor 1	o Growth factor activity o Protein binding o Multicellular organismal development o Signal transduction o Fibroblast growth factor receptor signaling pathway
6	**FLT1**	fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)	o Vascular endothelial growth factor receptor activity o Growth factor binding o Transmembrane receptor protein tyrosine kinase signaling pathway o Positive regulation of vascular endothelial growth factor receptor signaling pathway o Positive regulation of cell proliferation

Table 7.8

List of genes selected for personalized modeling for male subjects with their description.

	Gene symbol	Gene name	Gene Function
1	**ANGPTL3**	Angiopoietin-like 3	o Cholesterol homeostasis o Cholesterol metabolic process o Fatty acid metabolic process o Glycerol metabolic process o Positive regulation of lipid catabolic process
2	**ANGPT4**	Angiopoietin 4	o Vascular endothelial growth factor receptor binding
3	**HIF1A**	Hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)	o Transcription factor binding o RNA polymerase II transcription factor activity, enhancer binding o Regulation of transforming growth factor-beta production o Positive regulation of vascular endothelial growth factor production o Positive regulation of glycolysis
4	**TNFSF15**	Tumor necrosis factor (ligand) super family, member 15	o Signal transduction o Negative regulation of endothelial cell proliferation o Activity of NF-kappaB-inducing kinase activity o Cytokine metabolic process
5	**FLT1**	fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)	o Vascular endothelial growth factor receptor activity o Growth factor binding o Transmembrane receptor protein tyrosine kinase signaling pathway o Positive regulation of vascular endothelial growth factor receptor signaling pathway
6	**TNF**	Tumor necrosis factor	o Chronic inflammatory response to antigenic stimulus o Inflammatory response o Negative regulation of lipid catabolic process o Regulation of insulin secretion

Correlation analysis of variables: The genes selected from signal to noise ratio and t-test were subsequently used to examine associations with the general and clinical variables using correlation analysis. Different genes have been used for each datasets i.e. for all subjects, male only and female only. Figure 7.7 shows the correlation analysis of all data with general, clinical and six genes. Figures 7.8, 7.9 show the linear relationship between variables for male and female subjects respectively.

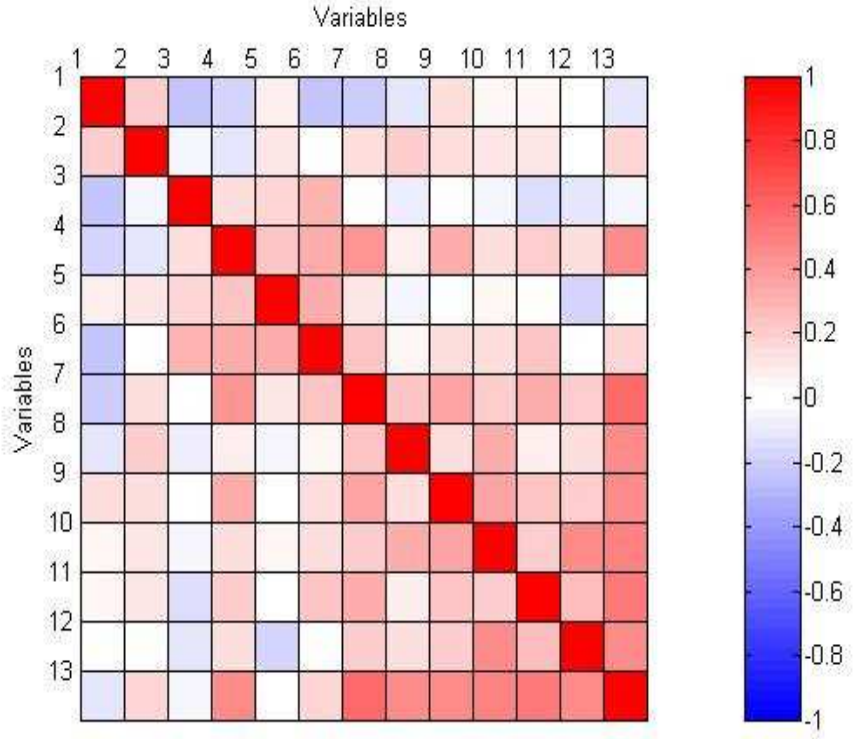


Figure 7.7. Linear relationships between general, clinical and genetic variables (listed below) for whole data using correlation coefficient (Red colour: high positive correlation).

1-age, 2-gender, 3- haemoglobin, 4-glucose, 5-cholesterol, 6-triglycerides, 7-ANGPTL3, 8-ANGPT4, 9-TNF, 10- FLT1, 11-MMP2, 12-CHGA, 13- Disease

It was found that all the genes were highly associated with each other and genes provided the best explanation for associations with a diagnosis of type 2 diabetes. Gene ANGPTL3 was the most important gene and was positively correlated with type 2 diabetes.

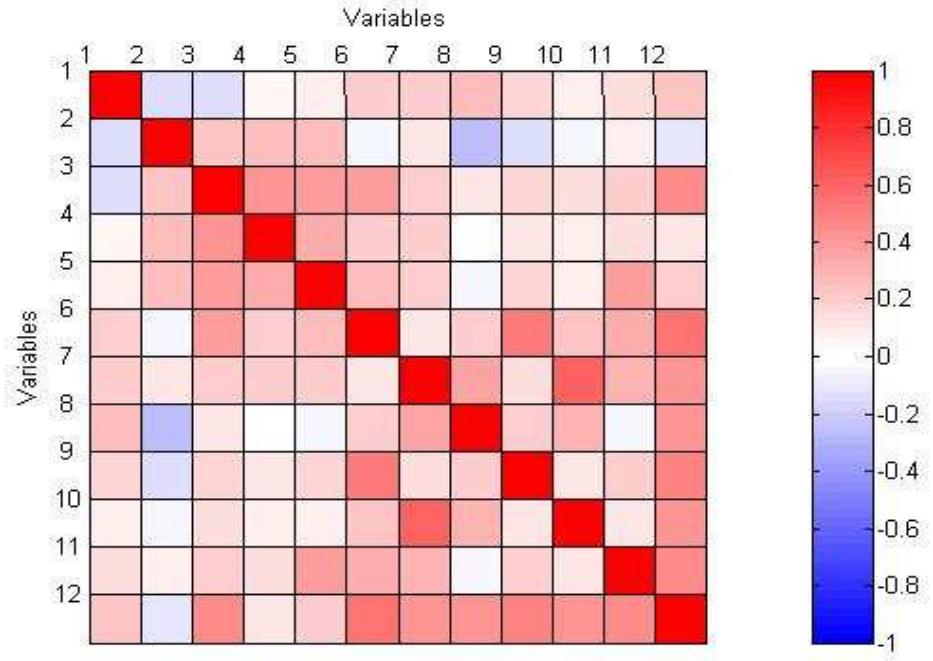


Figure 7.8. Linear relationships between the general, clinical and genetic variables (listed below) for male subjects using correlation coefficient. (Red colour: high positive correlation).

1-age, 2-haemoglobin, 3-glucose, 4-cholesterol, 5-triglycerides, 6-ANGPTL3, 7-MMP2, 8-ANGPT4, 9- TNF, 10-FGF1, 11-FLT1, 12- Disease

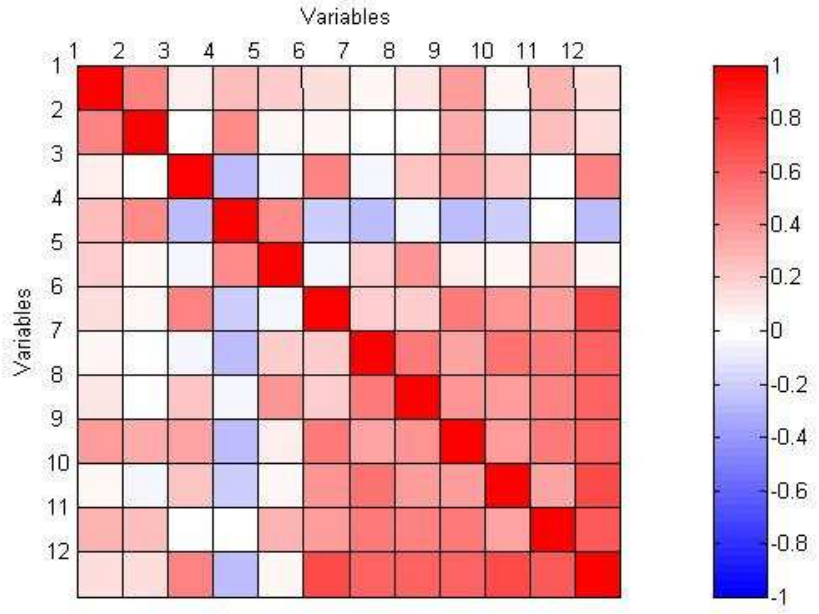


Figure 7.9. Linear relationships between the general, clinical and genetic variables (listed below) for female subjects using correlation coefficient. (Red colour: high positive correlation).

1-age, 2-haemoglobin, 3-glucose, 4-cholesterol, 5-triglycerides, 6-ANGPTL3, 7-ANGPT4, 8-HIF1A, 9- TNSF1S, 10-FLT1, 11-TNF, 12- Disease

From the correlation analysis of variables of male and female subjects seprately, similar patterns were found as for all subjects combined. The presence of ANGPTL3 genes had a positive correlation with type 2 diabetes and FGF1 gene was highly correlated to disease for female subjects.

7.5 Risk prediction method and knowledge discovery

The existing methods to predict risk of type 2 diabetes include general and clinical variables only (Table 7.2). An effort was made to create a model to predict risk of type 2 diabetes by using general, clinical and genetic variables.

According to the signal to noise ratio, the six different genes for male and female subjects most associated with type 2 diabetes were selected and separate models for male and female subjects were built. Different methods were then used for type 2 diabetes risk prediction methods such as multiple linear regression using NeuCom (global, inductive method), WWKNN and TWNFI (personalized methods) by using different sets of genes and it was found that highest accuracy was achieved with six set of genes and further experiments were carried out with six genes and clinical variables.

For the purpose of cross-validation, leave-one-out cross-validation method has been used. This method is used because the dataset is very small. In leave one out cross validation method one subject is taken out as the test subject and the rest of the subjects are used as training subjects and this process is repeated for all the subjects. As dataset is very small, leave one out cross validation method can be carried out in less time. Results obtained by each risk prediction method were compared for accuracy and details and results of each method are explained below.

Multiple linear regression (MLR): NeuCom was used to apply this method. Because the dataset is small, the leave one out cross validation method has been applied. The accuracy of the results obtained by multiple linear regression method was compared with other risk prediction methods at different threshold values.

Weighted-weighted K nearest Neighbor Algorithm (WWKNN) (Kasabov, 2007(b)): The first method for building personalized method to predict risk of type 2 diabetes was weighted-weighted K nearest neighbor method (explained in Chapter 2). Leave one out method has been used for cross-validation. The

system was tuned by changing the number of nearest neighbors to achieve high accuracy and the best number of neighbors is 4 as data subject is very small.

Transductive neuro-fuzzy inference system with weighted data normalization (TWNFI) (Song and Kasabov, 2006): TWNFI was used to build personalized model to predict risk of type 2 diabetes. Because the dataset is small in comparison to the dataset used in previous chapter to predict risk of cardiovascular disease, leave one out cross validation method has also been used here.

Table 7.9

Accuracy (%) comparison of diabetes data using clinical and genetic variables for male subjects.

Threshold	MLR (class1) (High risk)	MLR (Class0) (Low risk)	WWKNN (class1) (High risk)	WWKNN* (Class0) (Low risk)	TWNFI (class1) (High risk)	TWNFI* (Class0) (Low risk)
0.9	57.58	100	57.58	100	72.73	100
0.8	57.58	100	66.67	100	72.73	100
0.7	63.64	100	72.73	100	75.76	100
0.6	72.73	100	72.73	100	78.79	100
0.5	75.76	100	78.79	100	78.79	100
0.4	84.85	93.33	81.82	100	81.82	93.33
0.3	84.85	80.00	84.85	100	84.85	93.33

*See the description after Table 7.10.

The methods have been tested on different set of parameters such as (i) number of neighbors, (ii) threshold value, (iii) number of learning epochs and (iv) number of rules based on clusters. It has been found that the highest

accuracy has been achieved at 4 nearest neighbors (because dataset is very small).

The above mentioned three methods were used to predict risk of type 2 diabetes and accuracy of each method was checked. Because male and female subjects were modeled separately, results have also been checked separately for accuracy. Tables 7.9, 7.10 give the comparison of accuracy results of male subjects and female subjects respectively at different threshold values.

Table 7.10

Accuracy (%) comparison of diabetes data using clinical and genetic variables for female subjects.

Threshold	MLR (class1) (High risk)	MLR (Class0) (Low risk)	WWKNN (class1) (High risk)	WWKNN* (Class0) (Low risk)	TWNFI (class1) (High risk)	TWNFI* (Class0) (Low risk)
0.9	46.67	90.91	33.33	100	86.67	100
0.8	60.00	90.91	33.33	100	86.67	100
0.7	73.33	90.91	46.67	100	86.67	100
0.6	80.00	81.82	53.33	100	86.67	100
0.5	80.00	72.73	66.67	100	93.33	100
0.4	86.67	72.73	66.67	100	93.33	100
0.3	93.33	72.73	73.33	100	93.33	100

*See the description below.

It was found that for male subjects, multiple linear regression gave highest accuracy for class 1(high risk) at 0.4 threshold and for class 0 (low risk) at 0.5 threshold accuracy decreases. With WWKNN, class 0 gives 100% accuracy and highest accuracy for class 1 is achieved at 0.3 threshold. In TWNFI, accuracy for class 1 is achieved at 0.3 threshold and accuracy of class 0 is

100% until threshold 0.5 but at threshold value 0.4 the accuracy for class 0 decreases which is quite similar to multiple linear regression.

The accuracy for class 0 is shown as 100 percent by using WWKNN and TWNFI. This is because the gene expression values of selected genes for normal subject is set to 1 and for subjects with type 2 diabetes is calculated by using ddct method explained earlier in last section of 7.4.2. That is why both methods give one hundred percent accuracy for class 0 even though this is only for the sake of the explanation of the given experimental data rather than for real patient application. The class 1 accuracy is realistic and can be implemented for predicting risk for real patients.

Personalized modeling: As every person has a different genetic admixture, therefore personalized prediction and treatment is required for each person. Different models were required for male and female subjects. Table 7.11 shows results from example of personalized modeling for two male subjects. Subject 1 belongs to class 1 (with type 2 diabetes) and subject 2 belongs to class 0 (without type 2 diabetes).

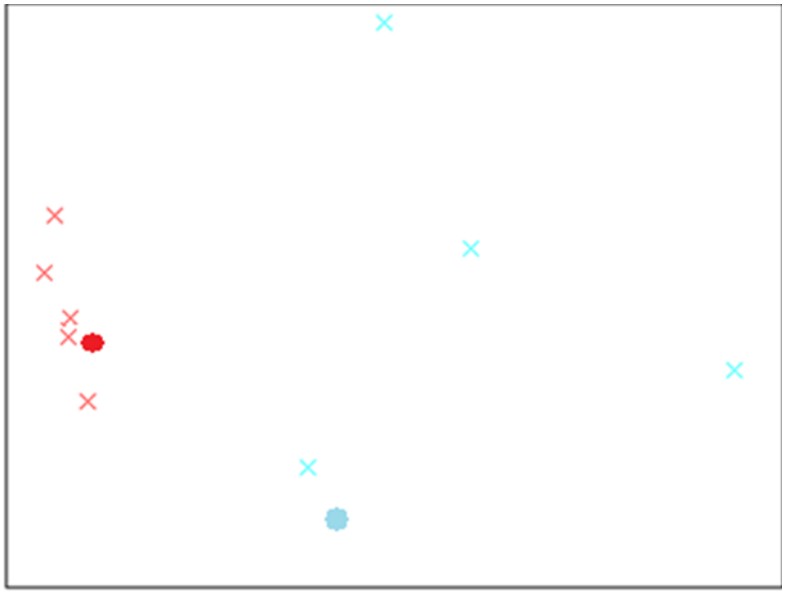
It was found that highest accuracy was achieved with the TWNFI method. TWNFI not only gives highest accuracy, also gives weights of variables as per their importance for risk of disease. For each subject in present example, separate weight of each variable has been presented and compared with global weights of variables for male subjects. It is very interesting that male subject 1 and 2 both have higher values of fasting blood glucose, cholesterol and triglycerides, the genes were more important factors to predict the risk of type 2 diabetes for male subject 2.

Table 7.11

Examples of TWNFI personalized models for two different male subjects; high risk and low risk; with weight of variables and genes with global weights representing importance of the variables.

	Subject 1 (High risk male)		Subject 2 (Low risk male)		
Input Variables	**Values of input**	**Weights of input variables**	**Values of input**	**Weights of input variables**	**Global weights/ importance (male)**
Age (years)	58	0.7729	42	0.9625	0.8393
Haemoglobin (g/L)	12	0.8521	15.4	0.7847	0.8429
Fasting blood glucose (mmol/L)	5.6111	0.7507	5.1111	0.9352	0.8769
Cholesterol (mmol/L)	4.7582	0.7478	5.3013	0.752	0.8104
Triglycerides (mmol/L)	2.1225	0.6961	2.6983	0.7413	0.8327
ANGPTL3	32.14	0.7617	1	0.9269	0.9254
FGF1	17.9446	0.7295	1	0.641	0.8228
FLT1	24.059	0.651	1	0.7059	0.8096
MMP2	4.4584	0.6797	1	0.8802	0.9009
TNF	3.3048	1	1	0.8495	0.8699
ANGPT4	14.1165	0.6705	1	1	0.904
Actual output	1		0		
Predicted output with Multiple linear regression		0.7963		0.1378	
Predicted output with WWKNN		1.127		0	
Predicted output with TWNFI		1.002		0	

By comparing weights for each variable of each subject, it was found that for male subject 1, gene TNF was found to be the most important gene associated with type 2 diabetes while for male subject 2, ANGPT4 gene has been weighted the highest, while for all the male subjects the ANGPTL3 gene has been found most important factor for type 2 diabetes.



Cluster center (class1: high risk of type 2 diabetes)

Cluster center (class 0: low risk of type 2 diabetes)

Male Subject 1 (belong to class 1)

Male Subject 2 (belong to class 0)

Figure 7.10. Example of male Subjects 1 and 2 with cluster centers based on nearest neighbors using principal component analysis (PCA).

TWNFI along with high accuracy and importance of variables also provides set of rules based on the clusters formed based on nearest neighbors. Each rule

contains a lot of information. Rules or profiles for male subjects were generated on the basis of nearest samples. Figure 7.10 illustrates the cluster centers for two test male subjects; one at high risk and one at low risk of type 2 diabetes. The risk of type 2 diabetes is predicted between 0 and 1. In the given examples, risk of type 2 diabetes is considered high if predicted outcome is more than 0.5 and risk of type 2 diabetes is low if predicted outcome is less than 0.5 (a threshold value 0.5 is set to predict risk for type 2 diabetes).

The set of rules for male subject one is listed below:

Rule 1:

If	Age (years)	is about	67
	Haemoglobin (g/L)	is about	13.19
	Fasting blood glucose (mmol/L)	is about	5.66
	Cholesterol (mmol/L)	is about	4.53
	Triglycerides (mmol/L)	is about	1.20
	ANGPTL3 gene	is about	12.71
	FGF1 gene	is about	13.71
	FLT1 gene	is about	15.87
	MMP2 gene	is about	9.94
	TNF gene	is about	8.19
	ANGPT4 gene	is about	7.47
Then	Risk of type 2 diabetes	is	**High**

The rule 1 explains that with increase in age risk of having type 2 diabetes increases. Higher levels of cholesterol, triglycerides and fasting blood glucose are directly associated with or can explain risk of type 2 diabetes.

Rule 2:

If	Age (years)	is about	56
	Haemoglobin (g/L)	is about	14.90
	Fasting blood glucose (mmol/L)	is about	10.78
	Cholesterol (mmol/L)	is about	5.79
	Triglycerides (mmol/L)	is about	2.40
	ANGPTL3 gene	is about	48.51
	FGF1 gene	is about	19.77
	FLT1 gene	is about	2.35
	MMP2 gene	is about	22.10
	TNF gene	is about	2.47
	ANGPT4 gene	is about	40.86
Then	risk of type 2 diabetes	is	**High**

The rule 2 can be explained as increase in age, increases the risk of having type 2 diabetes. Higher levels of cholesterol, triglycerides and fasting blood glucose are important type 2 diabetes risk factors. Gene ANGPTL3 and gene ANGPT4 seems to be the most important disease risk determinant factors.

Rule 3:

If	Age (years)	is about	43
	Haemoglobin (g/L)	is about	12.80
	Fasting blood glucose (mmol/L)	is about	6.00
	Cholesterol (mmol/L)	is about	3.34
	Triglycerides (mmol/L)	is about	2.96
	ANGPTL3 gene	is about	13.41
	FGF1 gene	is about	22.17
	FLT1 gene	is about	7.75

	MMP2 gene	is about	20.39
	TNF gene	is about	19.16
	ANGPT4 gene	is about	30.45
Then	risk of type 2 diabetes	is	**High**

Rule 3 explains that ANGPT4 gene and MMP2 genes are major risk determinant factors for type 2 diabetes.

Rule 4:

If	Age (years)	is about	47
	Haemoglobin (g/L)	is about	13.40
	Fasting blood glucose (mmol/L)	is about	10.33
	Cholesterol (mmol/L)	is about	4.77
	Triglycerides (mmol/L)	is about	1.10
	ANGPTL3 gene	is about	24.78
	FGF1 gene	is about	3.98
	FLT1 gene	is about	1.17
	MMP2 gene	is about	37.30
	TNF gene	is about	14.77
	ANGPT4 gene	is about	0.43
Then	risk of type 2 diabetes	is	**High**

From the above mentioned rule 4, it can be explained if age and cholesterol levels are higher; the risk of type 2 diabetes is also high. The higher the values of genetic variables especially here MMP2 gene, higher are the risk of type 2 diabetes.

The information in the above four rules for subject 1 can be used for prediction and recommendations. Furthermore, these four rules explain that, with growing age and higher concentrations of total cholesterol, the risk of type 2 diabetes

increases. The major risk determinant factors for type 2 diabetes were age, cholesterol, triglycerides and glucose.

Similarly a set of rules is generated for male subject 2 for better prediction, recommendation and knowledge discovery. The set of rules generated for male subject 2 are:

Rule 1:

If	Age (years)	is about	59
	Haemoglobin (g/L)	is about	15.78
	Fasting blood glucose (mmol/L)	is about	5.49
	Cholesterol (mmol/L)	is about	6.47
	Triglycerides (mmol/L)	is about	1.99
	ANGPTL3 gene	is about	0.84
	FGF1 gene	is about	1.00
	FLT1 gene	is about	1.00
	MMP2 gene	is about	1.03
	TNF gene	is about	1.00
	ANGPT4 gene	is about	0.66
Then	risk of type 2 diabetes	is	**Low**

For male subject 2 rule 1 explains that if the fasting blood glucose, triglycerides and cholesterol are in normal range then there is almost no risk of getting type 2 diabetes. Also if the genes FLT1, FGF1 and TNF are 1, then risk of type 2 diabetes is decreased.

Rule 2:

If	Age (years)	is about	46
	Haemoglobin (g/L)	is about	15.48

	Fasting blood glucose (mmol/L)	is about	5.31
	Cholesterol (mmol/L)	is about	5.41
	Triglycerides (mmol/L)	is about	2.59
	ANGPTL3 gene	is about	0.92
	FGF1 gene	is about	1.00
	FLT1 gene	is about	0.99
	MMP2 gene	is about	0.90
	TNF gene	is about	0.96
	ANGPT4 gene	is about	0.80
Then	risk of type 2 diabetes	is	**Low**

Rule 2 explains that low levels of fasting blood glucose, cholesterol and triglycerides reduce the risk of type 2 diabetes.

Rule 3:

If	Age (years)	is about	43
	Haemoglobin (g/L)	is about	14.16
	Fasting blood glucose (mmol/L)	is about	5.10
	Cholesterol (mmol/L)	is about	4.83
	Triglycerides (mmol/L)	is about	0.86
	ANGPTL3 gene	is about	1.03
	FGF1 gene	is about	1.02
	FLT1 gene	is about	0.98
	MMP2 gene	is about	0.21
	TNF gene	is about	0.74
	ANGPT4 gene	is about	0.96
Then	risk of type 2 diabetes	is	**Low**

The above mentioned rule also explains the similar advice for reduced risk of type 2 diabetes with reduced fasting blood glucose, cholesterol and triglycerides.

Rule 4:

If	Age (years)	is about	47
	Haemoglobin (g/L)	is about	16.59
	Fasting blood glucose (mmol/L)	is about	5.48
	Cholesterol (mmol/L)	is about	3.92
	Triglycerides (mmol/L)	is about	0.70
	ANGPTL3 gene	is about	0.79
	FGF1 gene	is about	1.00
	FLT1 gene	is about	1.00
	MMP2 gene	is about	1.02
	TNF gene	is about	0.99
	ANGPT4 gene	is about	0.68
Then	risk of type 2 diabetes	is	**Low**

Rule 4 explains that risk of having type 2 diabetes is influenced by genes and clinical variables. The risk of type 2 diabetes is reduced, if the values of triglycerides and cholesterol are low. Also higher values of fasting blood glucose are directly associated with type 2 diabetes. But the gene variables have values more than 1 which can increase the risk of getting type 2 diabetes. The genes FGF1, FLT1 are the important genes for reducing the risk of tpe-2 diabetes.

From the above mentioned set of rules, it can be concluded that if cholesterol or triglycerides level is low or the values of genetic variables are 1, then risk of

type 2 diabetes is very low. The genetic variables are very important factors for determining risk of type 2 diabetes.

Examples of personalized model for female subjects: Another personalized model has been created for two female subjects; one being normal and the other diseased. Table 7.12 shows the results of two female examples with their variable weights and comparison with global weights.

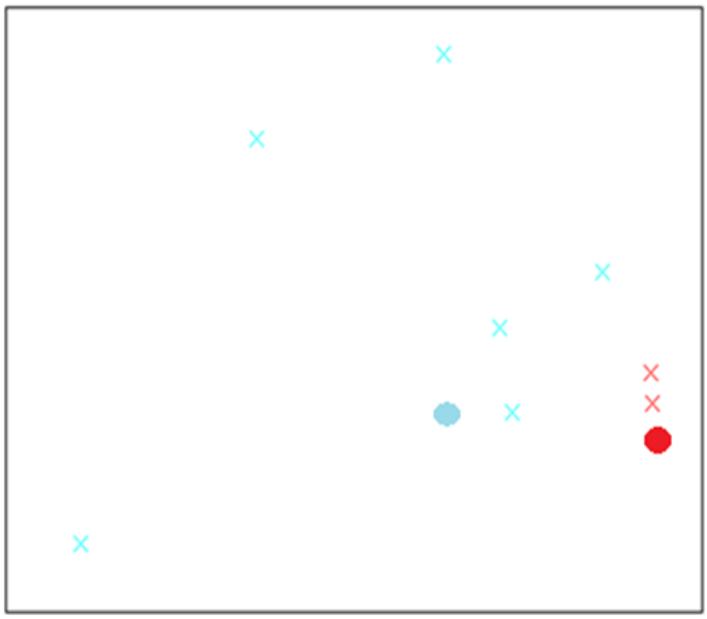
TWNFI along with evaluating risk of type 2 diabetes puts weights on variables for better understanding the importance of variables. According to weights, it has been found from both female examples that both subjects have different genes as risk determinant factors. For female subject one, gene ANGPTL3 is the most important risk determinant gene while for female subject 2, FLT1 gene is the most important risk factor.

Table 7.12

Examples of TWNFI personalized models for two different female subjects; high risk and low risk; with weights of variables and genes with global weights representing importance of variables.

	Subject 1 (High risk female)		**Subject 2 (Low risk female)**		
Input Variables	**Values Of input**	**Weights of input variables**	**Values of input**	**Weights of input variables**	**Global weights/ importance (female)**
Age (years)	47	0.9919	62	0.9888	0.9938
Haemoglobin (g/L)	11.9	0.9857	13.9	0.9977	0.9924
Fasting blood glucose (mmol/L)	14.22	0.991	5.55	0.9958	0.9959
Cholesterol (mmol/L)	3.74	0.9893	5.92	0.9948	0.9952
Triglycerides (mmol/L)	0.98	0.9924	1.08	0.9963	0.9958
ANGPTL3	21.1889	1	1	0.9973	0.9946
FLT1	0.385	0.9912	1	1	0.9991
TNF	0.1882	0.991	1	0.9972	0.9969
TNFSF15	3.9508	0.9911	1	0.9956	0.9958
ANGPT4	22.6942	0.9911	1	0.9963	0.9962
HIF1A	1.005	0.9934	1	0.9962	0.9954
Output	1		0		
Predicted output with Multiple linear regression		0.7594		0.1040	
Predicted output with WWKNN		1.34		0	
Predicted output with TWNFI		1.00		0	

For male subject 1 after gene ANGPTL3, gene HIF1A is also an important disease risk determinant factor. Along with weights of each variable, set of fuzzy rules or profiles are generated for each female subject that explains more about the variables and can be used for more specific recommendation. Illustration for the center of cluster is shown in figure 7.11.



Cluster center (class1: high risk of type 2 diabetes)

Cluster center (class 0: low risk of type 2 diabetes)

Female Subject 1 (belong to class 1)

Female Subject 2 (belong to class 0)

Figure 7.11. Example of female Subjects 1 and 2 with cluster centers based on nearest neighbors using principal component analysis (PCA).

Set of rules or profiles are generated for each female subject based on cluster centers of nearest subjects. The rules for female subject are listed as follow:

Rule 1:

If	Age (years)	is about	59
	Haemoglobin (g/L)	is about	14.80
	Fasting blood glucose (mmol/L)	is about	14.50
	Cholesterol (mmol/L)	is about	5.51
	Triglycerides (mmol/L)	is about	1.41
	ANGPTL3 gene	is about	412.87
	FLT1 gene	is about	2.35
	TNF gene	is about	12.87
	TNFSF15 gene	is about	6.36
	ANGPT4 gene	is about	21.19
	HIF1A gene	is about	0.66
Then	risk of type 2 diabetes	is	**High**

Rule 1 explains that the most important risk determinant factors for type 2 diabetes are fasting blood glucose, cholesterol and triglycerides along with genes. Higher levels of fasting blood glucose, triglycerides and cholesterol lead to higher risk of type 2 diabetes.

Rule 2:

If	Age (years)	is about	54
	Haemoglobin (g/L)	is about	12.40
	Fasting blood glucose (mmol/L)	is about	7.17
	Cholesterol (mmol/L)	is about	5.22
	Triglycerides (mmol/L)	is about	1.41
	ANGPTL3 gene	is about	22.79
	FLT1 gene	is about	1.26
	TNF gene	is about	9.77

	TNFSF15 gene	is about	4.45
	ANGPT4 gene	is about	57.99
	HIF1A gene	is about	3.60
Then	risk of type 2 diabetes	is	**High**

The above mentioned rule explains that gene ANGPTL3, ANGPT4 are very important for determining risk of type 2 diabetes along with higher levels of fasting blood glucose and cholesterol.

Rule 3:

If	Age (years)	is about	59
	Haemoglobin (g/L)	is about	14.60
	Fasting blood glucose (mmol/L)	is about	8.44
	Cholesterol (mmol/L)	is about	5.66
	Triglycerides (mmol/L)	is about	1.21
	ANGPTL3 gene	is about	24.94
	FLT1 gene	is about	4.59
	TNF gene	is about	3.30
	TNFSF15 gene	is about	1.85
	ANGPT4 gene	is about	15.12
	HIF1A gene	is about	6.12
Then	risk of type 2 diabetes	is	**High**

Rule 3 explains that fasting blood glucose is the major risk determinant factor for type 2 diabetes. Also higher values of genes ANGPLT3 and ANGPT4 are responsible for getting type 2 diabetes.

Rule 4:

If	Age (years)	is about	49
	Haemoglobin (g/L)	is about	13.00

	Fasting blood glucose (mmol/L)	is about	5.44
	Cholesterol (mmol/L)	is about	4.14
	Triglycerides (mmol/L)	is about	0.54
	ANGPTL3 gene	is about	5.15
	FLT1 gene	is about	12.90
	TNF gene	is about	13.31
	TNFSF15 gene	is about	4.85
	ANGPT4 gene	is about	37.20
	HIF1A gene	is about	0.52
Then	risk of type 2 diabetes	is	**High**

The above mentioned rule explains that major risk factors for type 2 diabetes are genes ANGPLT3, ANGPT4, FLT1 and TNF.

The above mentioned set of rules, it can be concluded that higher values of genes are the main cause of type 2 diabetes. Also with the increase in age risk of type 2 diabetes increases, as well as higher values of cholesterol and triglycerides also lead to type 2 diabetes.

Similarly, set of rules are also generated for female Subject 2 based on nearest subjects within the data. The best set of rules generated from female subject 2 are listed and explained below:

Rule 1:

If	Age (years)	is about	58
	Haemoglobin (g/L)	is about	13
	Fasting blood glucose (mmol/L)	is about	4.28
	Cholesterol (mmol/L)	is about	5.87
	Triglycerides (mmol/L)	is about	0.93

	ANGPTL3 gene	is about	1.00
	FLT1 gene	is about	1.00
	TNF gene	is about	1.00
	TNFSF15 gene	is about	1.00
	ANGPT4 gene	is about	1.00
	HIF1A gene	is about	1.00
Then	Risk of type 2 diabetes	is	**Low**

Rule 1 explains that low levels of fasting blood glucose reduce the risk of type 2 diabetes. The genetic variables are also important factors for causing type 2 diabetes. If the value of genetic variables is 1, it reduces the chances of getting type 2 diabetes.

Rule 2:

If	Age (years)	is about	55
	Haemoglobin (g/L)	is about	14.40
	Fasting blood glucose (mmol/L)	is about	4.50
	Cholesterol (mmol/L)	is about	7.50
	Triglycerides (mmol/L)	is about	3.57
	ANGPTL3 gene	is about	1.00
	FLT1 gene	is about	1.00
	TNF gene	is about	1.00
	TNFSF15 gene	is about	1.00
	ANGPT4 gene	is about	1.00
	HIF1A gene	is about	1.00
Then	Risk of type 2 diabetes	is	**Low**

The risk of type 2 diabetes is reduced fasting blood glucose and genetic variables. Fasting blood glucose, genes variables have direct relationship with the risk of type 2 diabetes.

Rule 3:

If	Age (years)	is about	46
	Haemoglobin (g/L)	is about	15.50
	Fasting blood glucose (mmol/L)	is about	5.22
	Cholesterol (mmol/L)	is about	5.25
	Triglycerides (mmol/L)	is about	0.46
	ANGPTL3 gene	is about	14.70
	FLT1 gene	is about	2.57
	TNF gene	is about	0.53
	TNFSF15 gene	is about	1.72
	ANGPT4 gene	is about	5.15
	HIF1A gene	is about	3.49
Then	Risk of type 2 diabetes	is	**High**

Rule 3 explains that increase in fasting blood glucose, triglycerides and cholesterol increases the risk of type 2 diabetes along with gene variables.

Rule 4:

If	Age (years)	is about	49
	Haemoglobin (g/L)	is about	13
	Fasting blood glucose (mmol/L)	is about	6.56
	Cholesterol (mmol/L)	is about	5.09
	Triglycerides (mmol/L)	is about	1.67
	ANGPTL3 gene	is about	4.66

	FLT1 gene	is about	13.00
	TNF gene	is about	12.71
	TNFSF15 gene	is about	0.77
	ANGPT4 gene	is about	18.29
	HIF1A gene	is about	2.06
Then	Risk of type 2 diabetes	is	**High**

For female Subject 2, from the above mentioned set of rules, it has been found that lower value of fasting blood glucose, cholesterol and triglycerides are important factors for reducing the risk of type 2 diabetes. All the genes are also important risk determinant factors as if the value of gene variables is 1; it reduces the risk of type 2 diabetes.

7.6 Integration framework of ontology and personalized diabetes risk analysis and knowledge discovery

The chronic disease ontology that was described in chapter 4 will be used to integrate personalized modeling and ontology. The current chronic disease ontology contains most of genes which are common for three chronic interrelated diseases (cardiovascular disease, type-2 diabetes and obesity). In the present Italian dataset used for predicting risk of type 2 diabetes, clinical and genetic variables were included.

It was found that for male and female subjects different combinations of genes are more predictive of type 2 diabetes. So these genes were updated in the chronic disease ontology and the missing genes and information related to these genes was also added in to the chronic disease ontology. Similarly, any other information derived from personalized model can be added to the chronic disease ontology and the new relationships and discoveries within the chronic disease ontology can be used to improve personalized risk evaluation system.

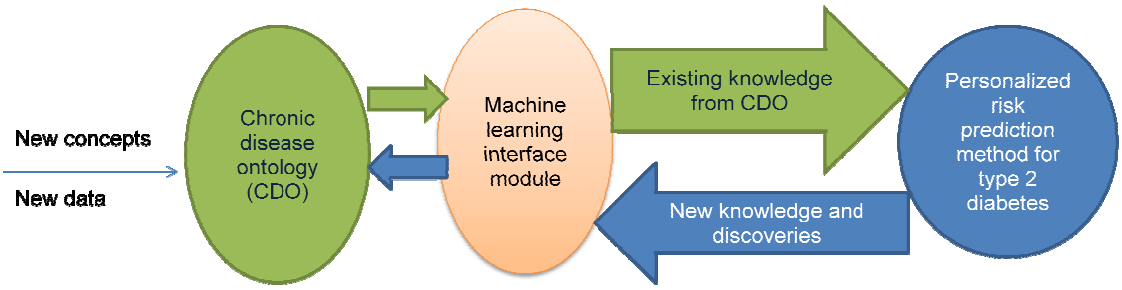


Figure 7.12. Integration framework for chronic disease ontology and personalized risk evaluation of type 2 diabetes.

The integration system for type 2 diabetes and the chronic disease ontology utilized the same modules as described in the chapter 5. The framework uses the chronic disease ontology based data and knowledge embedded in the ontology (Figure 7.12).

It also allows the adaptation of new knowledge by entering the results of the machine learning system to ontology. The first module for integration is protégé based the chronic disease ontology which is knowledge and data repository module, second module is TWNFI; a personalized modeling technique and the third module is interface between both the modules.

7.7 Examples for integration of the chronic disease ontology and personalized diabetes risk analysis model

As shown in chapter 6, integration of the chronic disease ontology and personalized model can be done for better personalized risk prediction for type 2 diabetes. This integration framework has been illustrated with the help of few examples.

Example1: The information obtained from personalized model for type 2 diabetes in the Italian dataset, such as the gene matrix metalloproteinase (MMP2), responsible for protein binding in normal person and mutated form is responsible for high risk of type 2 diabetes in the Italian male population can be added to the ontology and if a new male subject comes which is from Italian population, the same information can be for a new male subject belonging to Italian population.

Example2: It has been found that the gene hypoxia inducible factor 1 (HIF1A) acts as a normal transcription binding factor but mutation in gene is related to type 2 diabetes in females in Italian population. This information can be added to ontology and can be applied to the analysis for the next new subject from a similar population and with similar clinical features for risk prediction. Similar process can be applied for predicting risk of obesity.

Example3: It has been found that gene FTO in its inactivated state protects from risk of obesity (Fischer et al, 2009). Polymorphism in FTO gene is strongly and positively correlated to body mass index which is common measure of obesity. This knowledge has been updated in the chronic disease ontology and the system is able to use this knowledge if a similar subject with high body mass index comes, it can identify that FTO gene is active and the person may have a predisposition to obesity if dietary intake exceeds physical activity.

7.8 Conclusion

Diabetes, its global prevalence and description are detailed in chapter 7. Obesity and its prevalence are also described in section 7.2. There have been many methods developed to predict risk of type 2 diabetes. All the methods

developed so far use general and clinical variables only. None of the methods so far published have combined clinical and genetic variables together. This chapter describes how a model was built using clinical and genetic variables. In particular, I have found that

- Male subjects have high values of cholesterol and triglycerides and are more prone to type 2 diabetes.
- For male and female subjects different combinations of genes have association with type 2 diabetes.
- For male subjects, genes ANGPTL3, MMP2, ANGPT4, TNF, FGF1 and FLT1 appear to be the most important genes associated with risk of type 2 diabetes.
- For female subjects, genes ANGPTL3, ANGPTL4. HIF1A, TNSF15, FLT1 and TNF appear to be the most important factors for determining risk of type 2 diabetes.

For personalized modeling, different methods such as WWKNN and TWNFI were used and compared. I found that TWNFI gives highest accuracy along with importance of each gene and variable by optimizing each variable and weight which can be used for better prediction and recommendations.

The main limitation of this small data is lack of medical information such as whether subjects were treated for type 2 diabetes such as use of medicines e.g. metformin or statins. Secondly diabetes dataset is relatively very small, there is an increased likelihood of type 1 and type 2 errors. With a larger dataset or more contributions to the learning function and ontology more reliable predictions will be derived. The analysis presented here is a start and acts as a “proof of principle”. Another limitation of the data is, that, it is

preliminary data and not been previously used for statistical analysis. These are the first experiments carried on this dataset. This dataset is very small so leave one out cross validation method has been used. Another limitation is that there was no similar data available to perform independent test. Also gene expression data was calculated based on ddct method on the assumption that all normal subjects have gene expression value 1 which resulted in 100 percent accuracy for that class when personalized model was built.

The personalized risk evaluation system explained in this chapter utilizes 6 genes at this stage and it can be further extended with more genes and more set of clinical and general variables. The risk evaluation system can also be extended if more clinical and nutritional information is available along with genetic data. Still a better prediction system can be developed, if nutritional information and other environmental variables are known (e.g. exposure to sun for vitamin D) along with clinical and genetic variables are available. Similar approach can be used for building obesity risk prediction model by using clinical, nutritional and genetic variables.

Chapter 8. Conclusions, Discussion and Directions for Future Research

The last chapter of the thesis includes a summary of the thesis and the main contribution of this study in the field of bioinformatics. The first chapter of the thesis explained the motivation for this study and its main goals. The second chapter of the thesis introduced inductive and transductive methods of reasoning and personalized methods such as WWKNN and TWNFI. Third chapter defined ontology, different tools of constructing ontology and methods to build ontology. Fourth chapter explained the chronic disease ontology (CDO) and discoveries through ontology. Fifth chapter introduced the framework for integration of the chronic disease ontology and personalized modeling. Sixth chapter of thesis explained cardiovascular disease as an example of chronic disease, its prevalence in world and in New Zealand. Later sections of chapter 6 described the existing tools for predicting risk of cardiovascular disease. The data used to build personalized model for cardiovascular disease was also described in detail here. Different methods for personalized risk evaluation along with personalized examples were also explained. Second last chapter described type 2 diabetes, as another example of chronic diseases and also had been used to create personalized model for type 2 diabetes risk evaluation. Already existing methods for predicting risk for type 2 diabetes were also described in this chapter.

8.1 Achievements

The present study has contributed to the field of bioinformatics in terms of knowledge representation and knowledge discovery. I have created the chronic disease ontology (CDO) for knowledge collection and new discoveries

about the genes related to three chronic diseases. The major areas of achievements have been listed as follow:

(1) The chronic disease ontology is the first achievement of this study. The chronic disease ontology is a database which contains genetic, clinical and nutritional information related to three chronic diseases. The aim to build ontology is to collect all the information about these diseases from literature, World Wide Web at one place in order to reuse this information to discover new knowledge.

The genes related to these three diseases have been included in ontology from different sources. There are 71 genes in the latest version of CDO. These genes are common genes which are involved in three diseases and play important role in chronic diseases.

(2) The national nutrition health data for New Zealand population have been studied in detail. It can been concluded from the NNS 97 data analysis that

(2.1) Age and waist circumference are the most important factors for determining the risk of cardiovascular disease. Gender and ratio of total blood cholesterol to HDL cholesterol are important after waist circumference.

(2.2) Total saturated fat intake and ratio of sub scapular skinfold and triceps skinfold are the least important factors for the risk of cardiovascular disease.

(2.3) Age has positive correlation with waist circumference, ratio of total cholesterol and HDL, risk of cardiovascular disease and negative correlation with ethnicity and salt intake.

(2.4) Gender has positive correlation with saturated fat intake and negative correlation with haemoglobin, waist circumference, ratio of total cholesterol/HDL, carbohydrates intake, sugar intake, salt intake and risk of cardiovascular disease.

(2.5) Waist circumference has positive correlation with ratio of total cholesterol/HDL and risk of cardiovascular disease and negative correlation with carbohydrates and sugar intake.

(2.6) Haemoglobin shows positive correlation with waist circumference, ratio of total cholesterol/HDL, salt intake and risk of cardiovascular disease. Carbohydrate intake shows positive correlation with sugar intake and saturated fat intake. Carbohydrates intake and sugar intake show negative correlation with protein intake and total fat intake. Sugar intake also shows negative correlation with salt intake. Total fat intake has positive correlation with salt intake and negative correlation with saturated fat intake. Saturated fat intake shows negative correlation with salt intake.

(3) Another contribution is that for personalized modeling for NNS97 data, I have found that TWNFI method gives highest accuracy in comparison to multiple linear regression method and WWKNN. Also along with higher accuracy TWNFI also gives set of rules or profiles based on nearest samples which can be used further for better personalized recommendations.

The novelty of current model is using personalized approach for risk prediction for cardiovascular disease using NNS97 data with clinical, anthropometric and nutritional variables together which has not been used in existing methods so far. Along with predicting personalized risk of cardiovascular disease TWNFI generates rules or profiles for each subject

which explains the relationship between variables and can be used for better personalized disease risk prediction and further recommendations.

(4) As fourth contribution I have built a personalized advice system for diabetes risk prediction with diabetes data from Italian population. This data contains clinical and genetic variables. From the data analysis I have found that

(4.1) It has been found from the diabetes dataset that male subjects are at high risk of having type 2 diabetes than female subjects. It has been also found that male subjects in the current dataset have higher values of cholesterol and triglycerides which lead to high risk of having type 2 diabetes.

(4.2) From the analysis of this data it has been found that male and female subjects have different importance of genes responsible for causing type 2 diabetes.

(4.3) It has been found that for male population ANGPTL3, MMP2, ANGPT4, TNF, FGF1 FLT1 genes are most important genes.

(4.4) For female population ANGPTL3, ANGPT4, HIF1A, TNFSF15, FLT1 and TNF genes are the most important genes.

(4.5) For personalized different methods have been used to create personalized methods such as multiple linear regression, WWKNN and TWNFI. The results have been compared to check accuracy and it has been found that TWNFI gives highest accuracy along with importance of each variable for each subject and set of fuzzy rules or profiles for better prediction and recommendation.

(5) I have designed a framework for the integration of the ontology and the personalized modeling techniques which illustrates the integration of personalized method and ontology database for better recommendations and

advice. This framework explains how existing knowledge and new knowledge can be used together for better life style, risk evaluation and recommendations.

8.2 Further developments

The present research included many different aspects of knowledge discovery and personalized risk prediction through existing methods and data along with building knowledge repository (Chronic disease ontology) for three chronic diseases. This research can be extended in different ways.

- **Extension of The chronic disease ontology:** The chronic disease ontology is evolving and can be further extended with the new knowledge and information. The chronic disease ontology can be extended in following ways:

- New data, genes and information can be inputted from time to time and can be updated. As new genes are being found related to chronic diseases by different mutations which can be further added to existing ontology.
- At present the chronic disease ontology contains information about three chronic disease (cardiovascular disease, type 2 diabetes and obesity), but this ontology can be further developed to add more genetic information about these diseases individually instead of combined genes of these diseases.
- There are many other chronic diseases with very high prevalence such as, arthritis, cancer (breast cancer, colon cancer) etc. which can also be added to the chronic disease ontology. Also clinical, genetic and nutritional information about other chronic diseases can also be included to the chronic disease ontology.

- The chronic disease ontology can be extended in other fields e.g. in terms of medical information such as anatomical and physiological information about the organs involved in the respective chronic diseases.
- The chronic disease ontology is evolving and vast project and can be carried out for years. The only limitation of evolving the chronic disease ontology is that the new information has to be added manually, at present there is no such tool which can automatically update the existing information without duplicating or removing the existing knowledge in ontology.

- **Personalized cardiovascular disease risk prediction methods:** During the course of study national nutrition health data from New Zealand population has been used to create personalized model for cardiovascular disease risk analysis. The limitation of the data is, it does not include information about smoking status, alcohol consumption, physical activity, previous diabetes history which are very important factors in determining risk of cardiovascular disease. Also there was no genetic data available for these subjects.

  The personalized model can be further improved and tested with a different data with more variables of interest. Also there was no real patient genetic information available, if that information is available, the model can be used to predict more accurate personalized risk and recommendations. This method can be extended if more clinical and genetic data can be obtained. Another limitation is that this data is cross sectional rather than longitudinal. E.g. older people in the National Nutrition Survey have a different history to younger ones so it cannot be assumed that increased waist with time will apply to every individual. The national nutrition survey data contains

information about individual food intake can be used for more personalized food recommendations such as how many fruits or bread a day is advisable for individual. The personalized method TWNFI can also be modified to develop more efficient risk prediction along with visualizations with varied number of neighbours and more input variables.

- **Personalized type 2 diabetes risk prediction methods:** To predict risk of type 2 diabetes clinical and genetic data has been used. For building personalized model six genes has been used from male and female subjects. This model can be further extended by using more number of genes for even better prediction. This data lacks nutritional information, if nutritional information could be available along with clinical and genetic variables, it can used for disease based better recommendations. The main limitation of diabetes data is lack of medical information such as whether subjects have been treated for type 2 diabetes, taking medication (metformin or statins) as this information can change the outcome. The data size of diabetes data is very small and more information can be obtained if a large dataset is available.
- **Integration of The chronic disease ontology and personalized model:** The integration of the chronic disease ontology and personalized model can be further developed to build an efficient system for better advice and recommendations which can be used for man-kind to improve health and life style.

  An effort has been made to develop an ontology database (knowledge repository) for collecting information about three chronic diseases to discover new knowledge from existing information and also a personalized

risk evaluation system for risk analysis and recommendations which can be further extended as individual piece of work.

- **Pharmacogenomics:** The similar approach can be used for personalized recommendations for medicine. In current research, the main proposal is “personalized” risk prediction and recommendations which can be extended for “personalized” medicine. There are many web based tools for medicine prescription such as Warfarin dose recommendations in New Zealand has been developed by using genetic markers, previous medical information etc. Similarly, the integrative approach described in this thesis can be used for personalized nutritional and medicine recommendations along with risk prediction.

References

Alberti, K. G. M. M., Zimmet, P. and Shaw, J. (2006). Metabolic syndrome—a new world-wide definition. A Consensus Statement from the International Diabetes Federation. *Diabetic medicine, 23*, 469-480.

Al-Lawati, J. A. and Tuomilehto, J. (2007). Diabetes risk score in Oman: A tool to identify prevalent type 2 diabetes among Arabs of the Middle East. Diabetes Research and Clinical Practice, 77, 438-444.

Alon, U., Barkai, N., et al (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America, 96*(12), 6745-6750.

Amari, S. (1990). *Mathematical foundations of neuro-computing.* IEEE 78.

American Heart Asssociation (2009). *Blood Pressure.* Retrieved on 10 April, 2009, from: www.americanheart.org/presenter.jhtml?identifier=2114.

Anderson, K. M., Odell, P. M., Wilson, P. W. F. and Kannel, W. B. (1990). Cardiovascular disease risk profiles. *American Heart Journal, 121*(1), 293-298.

Appel, L. J., Sacks, F. M., Carey, V. J., Obarzanek, E., Swain, J. F., Miller, E. R., et al. (2005). Effects of protein, mono saturated fat, and carbohydrates intake on blood pressure and serum lipids. *The Journal of the American Medical Association, 294*(19), 2455-2464.

Arpirez, J. C., Corcho, O., Fernandez-Lopez, M. and Gomez-Perez, A. (2001). *WebODE: a scalable ontological engineering workbench.* Paper presented at the First International Conference on Knowledge Capture (KACP'01), Victoria.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H. and Cherry, J. M. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics, 25*, 25:29.

Astrup, A. and Finer, N. (2000). Redefining type 2 diabetes: 'Diabesity' or 'Obesity Dependent Diabetes Mellitus'? *Obesity Reviews, 1*, 57-59.

Atwater, W. O., and Bryant, A. P. (1900). The availability and fuel values of food materials. *Conn Agr Expt Sta 12th Ann Rpt*, 73-110.

Baker, P. G., Boble, C. A., Bechhofer, S., Paton, N. W., Stevens, R. and Brass, A. (1999). An ontology for bioinformatics applications. *Bioinformatics, 15*(6), 510-520.

Balkau, B., Deanfield, J. E., Despres, J. P., Bassand, J.-P., Fox, K. A. A., Smith, S. C., et al. (2007). International Day for the Evaluation of Abdominal Obesity (IDEA): A study of waist circumference, cardiovascular disease, and Diabetes Mellitus in 168,000 Primary care patients in 63 countries. *Circulation, 116*, 1942-1951.

Bannink, L., Wells, S., Broad, J., Riddell, T. and Jackson, R. (2006). Web-based assessment of cardiovascular disease risk in routine primary care practice in New Zealand: the first 18,000 patients (PREDICT CVD-1). *The New Zealand Medical Journal, 119*(1245).

Barton, C. (26 January, 2008). Gene Genie*. The New Zealand Herald,* pp. B1, B4, B5.

BBC News (2009). *Asian heart disease gene found*. Retrieved on 19 January, 2009, from: http://news.bbc.co.uk/2/hi/health/7833753.stm.

Bechhofer, S., Horrocks, I., Goble, C.and Stevens, R. (2001). OilEd: a reason-able ontology editor for the semantic web. *Lecture Notes in Artificial Intelligence, 2174*.

Bennett, K. P.and Demiriz, A. (1998). *Semi-supervised support vector machines.* Paper presented at the Conference on Advances in Neural Information Processing systems II, Cambridge, MA, USA.

Berkeley, J. and Lunt, H. (2006). Diabetes epidemiology in New Zealand- does the whole picture differ from the sum of its parts? *The New Zealand Medical Journal, 119*(1235).

Bernaras, A., Laresgoiti, I. and Corera, J. (1996). *Building and reusing ontologies for electrical network applications.* Paper presented at the European Conference on Artificial Intelligence (ECAI'96), Budapest, Hungary.

Berners-Lee, T., Hendler, J.and Lassila, O. (2001). The Semantic Web. *Scientific American*, May.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*: Plenum press, New York.

Blazquez, M., Fernandez-Lopez, M., Garcia-Pinat, J. M.and Gomez-Perez, A. (1998). *Building ontologies at the knowledge level using the ontology design environment.* Paper presented at the 11th International Workshop on Knowledge Acquisition, Modeling and Management (KAW'98), Banff.

Bonow, R. O., Smaha, L. A., Smith, S. C., Mensah, G. A. and Lenfant, C. (2002). World Heart Day 2002. The international Burden of cardiovascular disease: Responding to the emerging global epidemic. *Circulation, 106*, 1602-1605.

Borst, W. N. (1997). *Construction of Engineering Ontologies.* University of Tweenty, Enschede.

Bosnic, Z., Kononenko, I., et al (2003). Evaluation of prediction reliability in regression using the transduction principle. *EUROCON, 2003, The IEEE Region 8*(2), 99-103.

Brown, J. B., Palmer, A. J., Bisgaard, P., Chan, W., Pedula, K. and Russell, A. (2000(a)). The Mt. Hood challenge: cross-testing two diabetes simulation models. *Diabetes Research and Clinical Practice, 50*(3), S57-S64.

Brown, J. B., Russell, A., Chan, W., Pedula, K. and Aickin, M. (2000(b)). The global diabetes model: user friendly version 3.0. *Diabetes Research and Clinical Practice, 50*(3), S15-S46.

Cannon, C. P. (2007). Cardiovascular disease and modifiable cardio-metabolic risk factors. *Clinical Cornerstone, 8*(3), 11-28.

Cantais, J., Dominguez, D., Gigante, V., Laera, L. and Tamma, V. (2005). *An example of food ontology for diabetes control.* Paper presented at the ISWC 2005: 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6 - 10, 2005, Proceedings, eds., V. Richard Benjamins Yolanda Gil, Enrico Motta and Mark Musen, New York, USA.

Carson, S. J. A., Burke, F. M. and Hark, L. A. (2004). *Cardiovascular Nutrition. Disease Management and Prevention*: American Dietetic Association.

Caterson, I. D. and Gill, T. P. (2002). Obesity: epidemiology and possible prevention. *Best Practice and Research Clinical Endocrinology and Metabolism, 16*(4), 595-610.

Chanderasekaran, B., Josephson, J. R. and Benjamins V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems, 14*(1), 20-26.

Chaudhri, V. K., Farquhar, A., Fikes, R., Karp, P. D. and Rice, J. P. (1998). Open knowledge base connectivity 2.0.3. *Technical report*.

Chen, Y., Wang, G., et al (2003). Learning with progressive transductive support vector machine. *Pattern Recognition Letters, 24*(12), 1845-1855.

Cheng, D. (4 October, 2006). Diabetes discovery excites NZ team. *The NZ Herald,* p. A2.

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine, 357*(4), 370-379.

Clark, P., Thompson, J. and Porter, B. (2004). Knowledge Patterns. In S. Staab and R. Studer (Eds.), *Handbook on Ontologies* (pp. 191-208). Berlin: Springer-Verlag.

Cleeman, J. I. (2006). *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)*.

Colditz, G., Willett, W. C., Rotnizky, A., and Manson, J. E. (1995). Weight gain as a risk factor for clinical diabetes mellitus in women. *Annals of Internal Medicine, 122*, 481-486.

Colditz, G., Willett, W. C., Stampfer, M. J., et al (1990). Weight as a risk factor for clinical diabetes in women. *American Journal of Epidemiology, 132*, 501-513.

Corcho, O., Fernandez-Lopez, M. and Gomez-Perez, A. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data and Knowledge Engineering, 46*, 41-64.

Cornelis, M., Qi, L., Zhang, C., Kraft, P., Manson, J. A., Cai, T., et al. (2009). Joint effects of common genetic variants on the risk of type-2 diabetes

in U. S. men and women of European ancestry. *Annals of Internal Medicine, 150*, 541-550.

Crodder, G. and Grossberg, S. (1990). *Predicting the Mackey-Glass time series with cascade-correlation learning:* 1990 Connectionist models summer school, Carnegie Mellon University.

Dembele, D. and Kastner, P. (2003). Fuzzy C-means method for clustering microarray data. *Bioinformatics, 19*(8), 973-980.

DeRisi, J., Penland, L., et al. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics, 14*(4), 457-460.

Despres, J. P., Moorjani, S., Lupien, P. J., Tremblay, A., Nadeau, A. and Bouchard, C. (1990). Regional distribution of body fat, plasma lipoproteins, and cardiovascular disease. *Arteriosclerosis, Thrombosis, and Vascular Biology, 10*, 497-511.

Diamond, J. (2003). The double puzzle of diabetes. *Nature, 423*, 599-602.

Domingue, J. (1998). *Tadzebao and Webonto: Discussing, browsing and editing ontologies on the web.* Paper presented at the *11th Knowledge Acquisition Workshop (KAW98), Banff*.

Domshlak, C., Gal, A. and Roitman, H. (2007). *Rank Aggregation for Automatic Schema Matching*. *IEEE Transactions on Knowledge and Data Engineering, 19*(4), 538-553.

Duineveld, A., Studer, R., Weiden, M., Kenepa, B. and Benjamis, R. (1999). *Wondertools? A comparative study of ontological engineering tools.* Paper presented at the 12th Knowledge Acquistion Workshop (KAW99), Banff.

Eddy, D. M. and Schlessinger, L. (2003(a)). Archimedes. A trial-validated model of diabetes. *Diabetes Care, 26*(11), 3093-3101.

Eddy, D. M. and Schlessinger, L. (2003(b)). Validation of the Archimedes diabetes model. *Diabetes Care, 26*(11), 3102-3110.

Farmer, J. D. and Sidorowitch, J. J. (1987). Predicting chaotic time series. *Physical Review Letters, 59*(7), 845-848.

Farquhar, A., Fikes, R. and Rice, J. (1996). *The Ontolingua Server: A tool for collaborative ontology construction.* Paper presented at the 10th Knowledge Acquisition for Knowledge-based Systems Workshop (KAW96), Banff.

Fensel, D. (2004). *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Heidelberg.

Fischer, J., Koch, L., Emmerling, C., Vierkotten, J., Peters, T., Bruning, J. C., et al. (2009). Inactivation of the Fto gene protects from obesity. *Nature, 458*, 894-899.

Futschik, M. E. K. and N., K. (2002). *Fuzzy clustering of gene expression data.* Paper presented at the FUZZ-IEEE'02.

Galperin, M. Y. (2005). The Molecular Biology Database Collection: 2005 Update. *Nucleic Acids Research, 33*, D5-D24.

Galperin, M. Y. (2006). The Molecular Biology Database Collection: 2006 Update. *Nucleic Acids Research, 34*, D3-D5.

Galperin, M. Y. (2007). The Molecular Biology Database Collection: 2007 Update. *Nucleic Acids Research, 35*, D3-D4.

Gennari, J. H., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubezy, M., Eriksson, H., et al. (2001). The evolution of Protégé: An environment for knowledge-based systems development.

Goh, L. (2005). *Computational methods for microarray gene expression analysis through integration and knowledge discovery.*, Auckland University of Technology.

Gomez, L. C., Real, S. M., Ojeda, M. S., Gimenez, S., Mayorga, L. S. and Roque, M. (2007). Polymorphism of the FABP2 gene: a population frequency analysis and an association study with cardiovascular risk markers in Argentina. *BMC Medical Genetics, 8*(39), doi:10.1186/1471-2350-8-39.

Gomez-Perez, A., Fernandez-Lopez, M. and de Vicente, A. (1996). *Towards a method to conceptualize domain ontologies.* Paper presented at the European Conference on Artificial Intelligence (ECAI'96), Budapest, Hungary.

Gottgtroy, P., Kasabov, N. and Macdonell, S. (2006). Evolving ontologies for intelligent decision support. *Fuzzy Logic and the Semantic Web* (pp. 415-439): Elsevier.

Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*, 199-220.

Gruninger, M. and Fox, M. S. (1995). *Methodology for the design and evaluation of ontologies.* Paper presented at the Workshop on Basic Ontological Issues in Knowledge Engineering, Montreal.

Guarino, N. (1997). Understanding, building and using ontologies. *International Journal of Human and Computer Studies, 46*, 293-310.

Herman, W. H. (2003). Diabetes Modelling. *Diabetes Care, 26*(11), 3182.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., Minhas, R., Sheikh, A., et al. (2008). Predicting cardiovascular risk in England and

Wales: prospective derivation and validation of QRISK2. *British Medical Journal, 336*.

Hu, F. B., Manson, J. E., Stampfer. M. J., Colditz, G., Liu, S., Solomon, G. C., et al. (2001). Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *The New England Journal of Medicine, 345*(11), 790-797.

International Obesity Taskforce. (2009). *The global epidemic*. Retrieved on 10 March, 2009, from: http://www.iotf.org/globalepidemic.asp.

Jackson, R. (2000). Updated New Zealand cardiovascular disease risk-benefit prediction guide. *British Medical Journal, 320*, 709-710.

James, P. and Rigby, N. (2004). The challenge to movers and shakers: broad strategies to prevent obesity and diabetes. *Diabetes Voice, 49*(2).

James, P. T. (2004). Obesity: The worldwide epidemic. *Clinics in Dermatology, 22*, 276-280.

Jang, R. (1993). ANFIS: Adaptive network based fuzzy inference system *IEEE Transactions on Systems, Man, and Cybernetics, 23*(3), 665-685.

Joachims, T. (1999). *Transductive inference on text classification using support vector machines.* Paper presented at the Sixteenth International Confernce on Machine Learning, San Francisco, CA, USA.

Joachims, T. (2003). *Transductive learning via spectral graph partitioning.* Paper presented at the Twentieth International conference on Machine Learning, Washington, DC.

Johnston, M. (9 March, 2009). NZ scientist's gene research could be the answer to the obesity epidemic. *NZ Herald*.

Joshy, G. and Simmons, D. (2006). Epidemiology of diabetes in New Zealand: revisit to a changing landscape. *The New Zealand Medical Journal, 119* (1235).

Kaput, J. (2004). Diet-disease gene interactions. *Nutrition, 20*(1), 26-31.

Kaput, J. and Rodriguez, R. L. (2004). Nutritional genomics: the next frontier in the post-genomic era. *Physiological Genomics, 16*, 166-177.

Kasabov, N. (2002). *Evolving Connectionist Systems: Methods and Applications in Bioinformatics, Brain Study and Intelligent machines.* London: Springer.

Kasabov, N. (2007a). *Evolving Connectionist Systems: The Knowledge Engineering Approach (Second edition).* London: Springer.

Kasabov, N. (2007b). Global, local and personalized modeling and profile discovery in bioinformatics: An integrated approach. *Pattern Recognition Letters, 28*(6), 673-685.

Kasabov, N. (2008). Adaptive modeling and discovery in Bioinformatics: The evolving connectionist approach*. International Journal of Intelligent Systems,* 23, 545-555.

Kasabov, N. and Pang, S. (2004). Transductive support vector machines and applications in bioinformatics for promoter recognition. *Neural Information Processing Letters Review, 3*(2), 31-38.

Kasabov, N. and Song, Q. (2002). DENFIS: Dynamic, evolving neural-fuzzy inference system and its application for time-series prediction. *IEEE Transactions on Fuzzy Systems, 10*, 144-154.

Kiberstis, P. A. (2005). A surfeit of suspects. *Science, 307*, 369.

King, A. (2000). *The New Zealand Health Strategy*. Wellington: Ministry of Health.

King, H., Aubert, R. E. and Herman, W. H. (1998). Global burden of diabetes, 1995-2025. *Diabetes Care, 21*(9), 1414-1431.

King, H. and Rewers, M. (1993). Global estimates for prevalence of diabetes mellitus and impaired glucose tolerance in adults. WHO Ad Hoc Diabetes Reporting Group. *Diabetes Care, 16*(1), 157-177.

Kogut, P., Cranefielf, S., Hart, L., Dutra, M., Baclawski, K., Kokar, M., et al. (2002). UML for ontology development. *The Knowledge Engineering Review, 17*(1), 61-64.

Kohonen, T. (1997). *Self-organizing maps*. Springer Verlag.

Kopelman, P. G. (2000). Obesity as a medical problem. *Nature, 404*, 635-643.

Kukar, M. (2003). Transductive reliability estimation for medical diagnosis. *Artificial Intelligence in Medicine, 29*, 81-106.

Lahoz, C., Schaefer, E. J., Cupples, A., Wilson, W. F., Levy, D., Osgood, D., et al. (2001). Apolipoprotein E genotype and cardiovascular disease in Framingham heart study. *Atherosclerosis*, 529-537.

Lambrix, P., Habbouche, M. and Perez, M. (2003). Evaluation of ontology development tools for bioinformatics. *Bioinformatics, 19*(12), 1564-1571.

Laugesen, R. (23 April, 2006). Diabetes- the quiet killer*. The NZ Herald,* pp. C3-C4.

Lenat, D. B. (1995). Cyc: A large-scale investment in Knowledge Infrastructure. *Communications of the ACM, 38*, 33-48.

Lenat, D. B. and Guha, R. V. (1990). *Building Large Knowledge-based Systems: Representation and Inference in the Cyc Project.* Boston: Addison-Wesley.

Lenzerini, M., Milano, D. and Poggi, A. (2004). *State of the art and state of the practice including initial possible research orientations (InterOP Report).* Roma, Italy: UniRoma.

Levey, A. S., Bosch, J. P., Lewis, J. B., Greene, T., Rogers, N. and Roth, D. (1999). A more accurate method to estimate glomerular filteration rate from serum creatinine: A new prediction equation. *Annals of Internal Medicine, 130*, 461-470. For the modification of diet in renal disease study group.

Li, C. H. and Yuen, P. C. (2001). *Transductive learning: Learning Iris data with two labeled data ICANN 2001*: Springer Verlag, Heidelberg, Berlin.

Li, F. and Wechsler, H. (2004). Watch list face surveillance using transductive inference. *Lecture Notes in Computer science, 3072*, 23-29.

Li, J. and Chua, C. S. (2003). *Transductive inferences for color-based particle filter tracking.* Paper presented at the International Conference on Image Processing, 2003, Nanyang Technological University, Singapore.

Lin, C. T. and Lee, C. S. G. (1996). *Neuro fuzzy systems*: Prentice Hall.

Lindstrom, J. and Tuomilehto, J. (2003). The diabetes risk score. A practical tool to predict type 2 diabetes risk. *Diabetes Care, 26*(3), 725-731.

Livak and Schmittgen. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2-[Delta][Delta]CT method. *Methods, 25*(4), 402-408.

Lipid Management Guidelines 2001. (2001).*The Medical Journal of Australia, 175*, S57-S88.

Lukashin, A. V. and Fuchs, R. (2001). Analysis of temporal gene expression profiles: Clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics, 17*, 405-414.

Maamar, Z., Benslimane, D.and Narendra, N. C. (2006). What can context do for web services? *Communications of the ACM, 49*(12), 98-103.

Maedche, A. (2002). *Ontology learning for the semantic web.* Boston / Dordrecht / London: Kluwer Academic Publishers.

McGuinness, D. L., Fikes, R., Rice, J. and Wilder, S. (2000). *The Chimaera ontology environment.* Paper presented at the 17th National Conference on Artificial Intelligence (AAAI'00), Austin.

McKinlay, J. and Marceau, L. (2000). *US public health and the 21st century:* diabetes mellitus. *The Lancet, 356*, 757-761.

MedicineNet (2009). *High blood pressure - a silent killer*. Retrieved on 10 April, 2009, from: www.medicinenet.com/script/main/art.asp?articlekey=13118.

Meisinger, C., Thorand, B., Schneider, A., Stieber, J., Doring, A., and Lowel, H. (2002). Sex differences in risk factors for incident type 2 diabetes mellitus. *Archives of Internal Medicine, 162*, 82-89.

Mendel, J. M. (2001). *Uncertain rule-based fuzzy logic systems: Introduction and new directions.* Englewood Cliffs, New Jersey: Prentice Hall PTR.

Milne, R., Gamble, G., Whitlock, G. and Jackson, R. (2003). Framingham Heart study risk equation predicts first cardiovascular event rates in New Zealanders at the population level. *The New Zealand Medical Journal, 116* (1185).

Mitchell, M. T., Keller, R., et al (1997). Explanation-based generalization: A unified view. *Machine Learning, 1*(1), 47-80.

Moore, M. P. and Lunt, H. (2000). Diabetes in New Zealand. *Diabetes Research and Clinical Practice, 50*(Suppl.2), S65-S71.

Must, A., Spadano, J., Coakley, E. H., Field, A. E., Colditz, G. and Dietz, W. H. (1999). The disease burden associated with overweight and obesity. *The Journal of the American Medical Association, 282*(16), 1523-1529.

Nammi, S., Koka, S., Chinnala, K. M. and Boini, K. M. (2004). Obesity: An overview on its current perspectives and treatment options. *Nutrition Journal, 3*(3).

Neal, B., Chapman, N. and Patel, A. (2002). Managing the global burden of cardiovascular disease. *European Heart Journal Supplements, 4*(Supplement F), F2-F6.

Neches, R., Fikes, R. E., Finin, T., Gruber, T., Senator, T. and Swartout, W. (1991). Enabling technology for knowledge sharing. *Artificial Intelligence Magazine, 12*(3), 36-56.

Neel, J. (1962). Diabetes Mellitus: A "Thrifty" genotype rendered detrimental by "Progress"? *American Journal of Human Genetics, 14*, 353-362.

Neel, J. (1982). *The Genetics of Diabetes Mellitus*, New York: Academic.

Nesto, R. W. (2008). Comprehensive clinical assessment of modifiable cardiometabolic risk factors. *Clinical Cornerstone, 9*(Suppl1), S9-S19.

Neural Network Toolbox User's Guide, version 4 (2002). The Math Works Inc., 3 Apple Hill Drive, Natrick, Massachusetts.

New Zealand Guidelines Group (2003(a)). *Management of diabetes*. New Zealand Guidelines Group, Wellington. Retrived from : http://www.nzgg.org.nz/guidelines/dsp_guideline_popup.cfm?guidelineID=36

New Zealand Guidelines Group (2003(b)). The assessment and management of cardiovascular risk. New Zealand Guidelines Group, Wellington. Retrieved on 10 December, 2008, from: www.nzgg.org.nz/guidelines/dsp_guideline_popup.cfm?guidelineID=35

Nicholson, J. K. (2006). Global systems biology, personalized medicine and molecular epidemiology. *Molecular Systems Biology, 52*(2).

Noy, N. F. and McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology *Medical Informatics Technical Report No. SMI-2001-0880: Stanford Knowledge Systems Laboratory*.

Okosun, I. S., Boltri, J. M., Hepburn, V. A., Eriksen, M. P., & Davis-Smith, M. (2006). Regional fat localizations and racial/ethnic variations in odds of hypertension in at-risk American adults. *Journal of Human Hypertension, 20*, 362-371.

Ordovas, J. M. and Corella, D. (2007). Nutrition, Genomics, and Cardiovascular Disease Risk. In F. Kok, L. Bouwmanand F. Desiere (Eds.). *Personalised Nutrition* (pp. 49-60): CRC Press.

Owens, A. (2005). *Semantic Storage: Overview and Assessment.* Amsterdam.

Pedrinelli, R., Dell'Omo, G., Penno, G., et al. (2006). Alpha- Adducin and angiotensin-converting enzyme polymorphism in hypertension: evidence for a joint influence on albuminuria. *Journal of Hypertension, 24*, 931-937.

Pisanelli, D. M. (2004). *Ontologies in Medicine*. Amsterdam: IOS Press.

Proedrou, K., Nouretdinov, I., et al (2002). *Transductive confidence machine for pattern recognition.* Paper presented at the 13th European Conference on Machine Learning.

Quigley, R. and Watts, C. (1997). *Food comes first: Methodologies for the National Nutrition Survey of New Zealand*.

Rigby, N. and James, P. (2003). The obesity campaign view of diabetes prevention. *Diabetes Voice, 48*.

Robitaille, J., Perusse, L., Bouchard, C. and Vohl, M.C. (2007). Genes, fat intake, and cardiovascular disease risk factors in the Quebec family study. *Obesity, 15*(9), 2336-2347.

Rose, S., Lawton, B., Dowell, A. and Fenton. (2004). Risk factors for type 2 diabetes in postmenopausal New Zealand women: a cross-sectional study. *The New Zealand Medical Journal, 117*(1207).

Ruden, D. M., Luca, M. D., Garfinkel, M. D., Bynum, K. L. and Lu, X. (2005). Drosophila nutrigenomics can provide clues to human gene-nutrient interactions. *Annual Review of Nutrition, 25*, 21.21-21.24.

Satyanarayanan, M. (2001). Pervasive computing: vision and challenges. *IEEE Personal Communications, 8*(4), 10-17.

Schulze, M. B., Hoffmann, K., Boeing, H., Linseisen, J., Rohrmann, S., Mohlig, M., et al. (2007). An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. *Diabetes Care, 30*(3), 510-515.

Shegogue, D. and Zheng, W. J. (2005). Integration of the Gene Ontology into an object-oriented architecture. *Bioinformatics, 6*(113), 1-14.

Sheridan, S., Pignone, M. and Mulrow, C. (2003). Framingham-based tools to calculate the global risk of coronary heart disease. *Journal of General Internal Medicine, 18*, 1039-1052.

Simmons, D. (1996(a)). Diabetes and its complications in New Zealand: an epidemiological perspective. *New Zealand Medical Journal, 109*, 245-247.

Simmons, D. (1996 (b)).The epidemiology of diabetes and its complication in New Zealand. *Diabetic medicine, 13*, 371-375.

Smith, B., Ashburner, M., Rosse, C., et al (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology, 25*(11), 1251-1255.

Social Report (2007). *Obesity.* Retrieved on 10 March, 2009, from: www.socialreport.msd.govt.nz/2007/health/obesity.html.

Song, Q. and Kasabov, N. (2001). *A novel online, evolving clustering method and its applications.* Paper presented at the fifth biannual conference on artificial neural networks and expert systems.

Song, Q., and Kasabov, N. (2004). NFI: A Neuro-Fuzzy Inference Method for Transductive Reasoning. *IEEE Transactions on Fuzzy Systems*.

Song, Q. and Kasabov, N. (2006). TWNFI - a transductive neuro-fuzzy inference system with weighted data normalization for personalized modeling. *Neural Networks, 19*(10), 1591-1596.

Sotiriou, C., Neo, S. Y., et al (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America, 100*(18), 10393-10398.

Staab, S., Schnurr, H. P., Studer, S. and Sure, Y. (2001). Knowledge processes and ontologies. *IEEE Intelligent Systems, 16*(1), 26-34.

Stern, M., Williams, K., Eddy, D. and Kahn, R. (2008). Validation of prediction of diabetes by the Archimedes Model and comparison with other prediction models. *Diabetes Care, 31*(8), 1670-1671.

Stover, P. J. (2004). Nutritional Genomics. *Physiological Genomics, 16*, 161-165.

Studer, S., Schnurr, H. P. and Fensel, D. (1998). Knowledge engineering: principles and methods. *Data and Knowledge Engineering, 25*, 161-197.

Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R. and Wenke, D. (2002). OntoEdit: collaborative ontology engineering for the semantic web. *Lecture Notes in Computer Science, 2342*.

Swartout, W. (1999). Ontologies. *IEEE Intelligent Systems, 14*(1), 18-19.

Swartout, W., Ramesh, P., Knight, K. and Russ, T. (1997). *Toward distributed use of large scale ontologies.* Paper presented at the AAAI Symposium on Ontological Engineering, Stanford.

Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics, 15*, 116-132.

The FIELD Study Investigators. (2004). The need for a large-scale trial of fibrate therapy in diabetes: the rationale and design of the Fenofibrate Intervention and Event Lowering in Diabetes (FIELD) study. ISRCTN64783481. *Cardiovascular Diabetology*. 2004; 3:9.

Theranostic Labs (2009). *Warfarin dosage factors.* Retrieved on 4 April, 2009, from: http://theranostics.co.nz/cgi-bin/warfarin.cgi.

Thom, J. T., Kannel, W. B., Chobanian, A. and D'Agostino, R. B. (2004). Cardiovascular disease: prevalence to prevention. *Cardiovascular Nutrition. Disease Management and prevention* (pp. 3-16): American Dietetic Association.

Turley, M. (2008). *Body Size Technical Report: Measurements and classifications in the 2006/07 New Zealand Health Survey.* Wellington: Ministry of Health.

Uschold, M. and King, M. (1995). *Towards a methodology for building ontologies.* Paper presented at the IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal.

Uschold, M. F. and Grüninger, M. (1996). Ontologies: Principles, Methods and Applications. *The Knowledge Engineering Review, 11*(2), 93-155.

van Heijst, G., Schreiber, A. T. and Wielinga, B. J. (1997). Using explicit ontologies in KBS development. *International Journal of Human and Computer Studies, 46*, 183-292.

Vapnik, V. N. (1998). *Statistical Learning Theory*: Wiley Inter-Science.

Verma, A., Fiasche, M., Cuzzola, M., Iacopino, P., Morabito, F.C, Kasabov, N.(2009). Ontology Based Personalized Modeling for Type 2 Diabetes Risk Analysis: An Integrated Approach. International Conference on Neural Information Processing (ICONIP). 1-5 December, 2009. Paper accepted.

Verma, A., Gottgtroy, P., Havukkala, I., and Kasabov, N. (2005) An Ontological Representation of Nutrigenomics Knowledge about Type 2 diabetes. New Zealand's Annual Biotechnology Conference Sky City Convention Centre, Auckland, NZ. 27-28 February 2006.

Verma, A., Gottgtroy, P., Havukkala, I., and Kasabov, N. (2005) An Ontological Representation of Nutrigenomics Knowledge about Type 2 diabetes. ILSI's First International conference on Nutrigenomics Opportunities in Asia 2005.

Verma, A., Gottgtroy, P., Havukkala, I., and Kasabov, N. (2005) Understanding the Molecular Basis of Type 2 diabetes by Means of Evolving Ontologies and Intelligent Modeling. The Queenstown Molecular Biology Meeting 2005.

Verma, A., Kasabov, N., Rush, E., Song, Q. (2009) Knowledge discovery through Integration of Ontology Based Personalized Modeling for

Chronic Disease Risk Analysis. Paper accepted for Australian Journal of Intelligent Information Processing Systems.

Verma, A., Kasabov, N., Rush, E., Song, Q. (2008) Ontology Based Personalized Modeling for Chronic Disease Risk Analysis: An Integrated Approach. International Conference on Neural Information Processing (ICONIP). 24 -28 November, 2008, Springer, LNCS No 5506/07, 2009.

Kasabov, N., Song, Q., Benuskova, L., Gottgtroy, P., Jain, V., Verma, A., Havukkala, I., Rush, E., Pears, R., Tjahjana, A., Hu, R., MacDonell, S. (2008): Integrating Local and Personalized Modeling with Global Ontology Knowledge Bases for Biomedical and Bioinformatics Decision Support, chapter in: Smolin et al (eds) Computational Intelligence in Bioinformatics, Springer, 2008.

Verma, A, Song, Q & Kasabov, N., (2006). Developing “Evolving Ontology” for Personalised Risk Evaluation for Type 2 diabetes Patients. At 6th International Conference on Hybrid Intelligence, Auckland, New Zealand.

Verma, A., Song, Q. & Kasabov, N. (2008) Ontology based personalized modeling for chronic disease and risk evaluation in medical decision support system. At The 3rd Asia Pacific Nutrigenomics Conference 2008: Diet-Gene Interaction in Human Health and Disease is being held in Melbourne, Australia, from May 6-9.

Verma, A., Song, Q. & Kasabov, N. (2008) Ontology based personalized modeling for chronic disease risk evaluation in medical decision support systems. At New Zealand’s Annual Biotechnology Conference, Sky City Convention Centre, Auckland, NZ. 31 March-2 April 2008.

Verma, A., Song, Q. & Kasabov, N. (2006) Developing “Evolving Ontology” for Personalized Risk Evaluation for Type 2 diabetes Patients. At 6th

International Conference on Hybrid Intelligent Systems (HIS'06) and 4th Conference on Neuro-Computing and Evolving Intelligence (NCEI'06) 13December -15 December, 2006, Auckland, New Zealand.

Verma, A., Song, Q. & Kasabov, N.(2006) Developing "Evolving Ontology" for Nutritional Advice for Type 2 diabetes Patients Poster at Queenstown Molecular Biology Meeting 29th August – 1 September 2006. Queenstown, New Zealand.

Walsh, F. (2009). *Era of personalized medicine awaits.* Retrieved on 8 April, 2009. http://news.bbc.co.uk/2/hi/health/7954968.stm.

Wang, L. X. (1994). *Adaptive fuzzy systems and control: Design and stability analysis*: Englewood Cliffs, N J: Prentice Hall.

Wang, Z. and Hoy, W. E. (2004). Waist circumference, body mass index, hip circumference and waist-to-hip ratio as predictors of cardiovascular disease in aboriginal people. *European Journal of Clinical Nutrition, 58*, 888-893.

West, M., Blanchette, C., et al (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America, 98*(20), 11462-11467.

Weston, J., Perez-Cruz, F., et al (2003). Feature selection and transduction for prediction of molecular bioactivity for drug design. *Bioinformatics, 19*(6), 764-771.

Whetzel, P. L., Brinkman, R. R. and Causton, H. C. (2006a). Development of FuGO: ontology for functional genomics investigations. *OMICS A Journal of Integrative Biology, 10*, 199-204.

Whetzel, P. L., Brinkman, R. R. and Causton, H. C. (2006b). The MGED ontology; a resource for semantics-based description of microarray experiments. *Bioinformatics, 22*, 866-873.

Whittaker, R., Bramley, D., Wells, S., Stewart, A., Selak, V., Furness, S., et al. (2006). Will a web-based cardiovascular disease (CVD) risk assessment program increase the assessment of CVD risk factors for Maori? *The New Zealand Medical Journal, 119*(1238).

Wielinga, B. J. and Schreiber, A. T. (1993). *Reusable and sharable knowledge bases: a European perspective.* Paper presented at the First international conference on Building and Sharing of Very Large-Scaled Knowledge Bases., Tokyo, Japan.

Wild, S., Roglic, G., Green, A., Sicree, R. and King, H. (2004). Global prevalence of Diabetes. Estimates for the year 2000 and projections for 2030. *Diabetes Care, 27*(5), 1047-1053.

Wilson, B. D., Wilson, N. C. and Russell, D. G. (2001). Obesity and body fat distribution in the New Zealand population. *New Zealand Medical Journal, 114*, 127-130.

Wing, R., Goldstein, M. G., Acton, K. J., et al (2001). Lifestyle changes related to obesity, eating behaviour, and physical activity. *Diabetes Care, 24*, 114-123.

Witten, I. H. and Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*: Morgan Kaufmann.

Wolf, L. and Mukherjee, S. (2004). *Transductive Learning via Model Selection*. Massachusetts Institute of Technology, Cambridge, MA, the Center for Biological and Computational Learning.

World Health Organization (2009). *Cardiovascular diseases.* Retrieved on 10 April, 2009, from: www.who.int/mediacentre/factsheets/fs317/en/index.html.

World Health Organization (2003). Diet, nutrition and the prevention of chronic diseases. *WHO Technical Report Series 916*, Geneva, Switzerland.

World Health Organization (WHO) (2000). Obesity: Preventing and managing the global epidemic. *WHO Technical report series 2000*, no. 894. Geneva, Switzerland.

Wright, H. (2001). Draft DHB toolkit: Obesity. Wellington, Ministry of Health*, 33*. Retrieved from: http://www.newhealth.govt.nz/toolkits/diabetes.htm.2001.

Wu, D., Cristianini, N., et al (1999). *Large margin trees for induction and transduction.* Paper presented at the 16th International conference of machine learning. Morgan Kaufmann, Bled, Slovenia.

Yu, A. C. (2006). Methods in biomedical ontology. *Journal of Biomedical Informatics, 39*, 252-266.

Yusuf, S., Reddy, S., Ounpuu, S. and Anand, S. (2001). Global burden of cardiovascular diseases. Part II: Variations in Cardiovascular disease by specific ethnic groups and geographic regions and prevention strategies. *Circulation, 104*, 2855-2864.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338-353.

Zadeh, L. A. (1988). Fuzzy logic. *IEEE Computer, 21*, 83-93.

Zhu, S., Heymsfield, S. B., Toyoshima, H., Wang, Z., Pietrobelli, A.and Heshka, S. (2005). Race-ethnicity-specific waist circumference cutoffs for identifying cardiovascular disease risk factors. *American Journal of Clinical Nutrition, 81*, 409-415.

Zimmet, P. (1992). Challenges in diabetes epidemiology- from west to the rest. *Diabetes Care, 15*, 232-252.

Zimmet, P. (1997). *The Medical Challenge: Complex Traits*: Munich: Piper.

Zimmet, P., Alberti, K. G. M. M. and Shaw, J. (2001). Global and societal implications of the diabetes epidemic. *Nature, 414*, 782-787.

Appendix A WWKNN Algorithm

Using the ranking of the variables in terms of a discriminative power within the neighborhood of K vectors, when calculating the output for the new input vector, is the main idea behind the WWKNN algorithm (Kasabov, 2007(a,b)), which includes one more weight vector to weigh the importance of the variables.

The Euclidean distance d_j between a new vector $\boldsymbol{x}_i$ and a neighboring one $\boldsymbol{x}_j$ is calculated now as:

$$dj = sqr [\sum l_{= 1 \text{ to } v} (ci,l (xi,l - xj,l))2] \quad (A1)$$

where: **ci,l** is the coefficient weighing variable **xl** for in neighbourhood of ***x*i** . It can be calculated using a Signal-to-Noise Ratio (SNR) procedure that ranks each variable across all vectors in the neighborhood set Di of Ni vectors:

$$\mathbf{C}i= (c_{i,1}, c_{i,2},\ldots..,c_{i,v}) \quad (A2)$$

$c_{i,l} = S_l /sum (S_l)$, for: l =1,2,…,v, where:

$$S_l = abs (M_l^{(class\ 1)} – M_l^{(class\ 2)}) / (Std_l^{(class\ 1)} + Std_l^{(class2)}) \quad (A3)$$

Here $M_l^{(class\ 1)}$ and $Std_l^{(class\ 1)}$ are respectively the mean value and the standard deviation of variable x_l for all vectors in Di that belong to class 1.

The new distance measure, that weighs all variables according to their importance as discriminating factors in the neighborhood area Di, is the new element in the WWKNN algorithm when compared to the WKNN.

Using the WWKNN algorithm, a “personalized” profile of the variable importance can be derived for any new input vector that represents a new piece of personalized’ information

Appendix B NFI Learning Algorithm

NFI Learning Algorithm (From Song and Kasabov, 2004)

Suppose that data have been normalized (the values are between 0 and 1) and, for each new data vector $\boldsymbol{x}_q$, the NFI performs the following learning algorithm:

(1) Search in the training data set in the input space to find N_q training examples that are closest to x_q. The value for N_q can be pre-defined based on experience, or – optimized through the application of an optimization procedure. Here we assume the former approach.

(2) Calculate the distances d_i, i = 1, 2, …, N_q, between each of these data subjects and x_q. And calculate the weights $w_i = 1 - (d_i - \min(d))$, i = 1, 2, …, N_q, min(d) is the minimum value in the distance vector $d = [d_1, d_2, \dots, d_{Nq}]$.

(3) Use the *ECM* clustering algorithm to cluster and partition the input sub-space that consists of N_q selected training subjects.

(4) Create fuzzy rules and set their initial parameter values according to the ECM clustering procedure results; for each cluster, the cluster centre is taken as the centre of a fuzzy membership function (*Gaussian* function) and the cluster radius is taken as the width.

(5) Apply the steepest descent method (back-propagation) to optimize the parameters of the fuzzy rules in the local model M_q following equations (B1 – B 20).

(6) Calculate the output value y_q for the input vector x_q applying fuzzy inference over the set of fuzzy rules that constitute the local model M_q.

(7) End of the procedure.

The parameter optimization procedure is described below:

Consider the system having P inputs, one output and M fuzzy rules defined initially through the ECM clustering procedure, the l-th rule has the form of:

R_l :If x_1 is F_{l1} and x_2 is F_{l2} and … x_P is F_{lP}, then y is G_l

(*Zadeh-Mamdani* type) (B1)

or:

R_l :If x_1 is F_{l1} and x_2 is F_{l2} and … x_P is F_{lP}, then y is n_l.

(*Takagi-Sugeno* type) (B2)

Here, F_{lj} are fuzzy sets defined by the following *Gaussian* type membership function:

$$GaussianMF = \alpha \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right] \quad \text{(B3)}$$

and G_l are of a similar type as F_{lj} and are defined as:

$$GaussianMF = \exp\left[-\frac{(y-n)^2}{2\delta^2}\right]$$

(for *Zadeh-Mamdani* type) (B4)

or:

$$n_l = b_{l0} + b_{l1}x_1 + b_{l2}x_2 + \ldots + b_{lP}x_P :$$

(for *Takagi-Sugeno* type) (B5)

Using the *ModifiedCentre Average defuzzification* procedure (B5) the output value of the system can be calculated for an input vector $x_i = [x_1, x_2, \ldots, x_P]$ as follows:

$$f(\boldsymbol{x_i}) = \frac{\sum_{l=1}^{M} \frac{G^l}{\delta_l^2} \prod_{j=1}^{P} \alpha_{lj} \exp\left[-\frac{(x_{ij} - m_{lj})^2}{2\sigma_{lj}^2}\right]}{\sum_{l=1}^{M} \frac{1}{\delta_l^2} \prod_{j=1}^{P} \alpha_{lj} \exp\left[-\frac{(x_{ij} - m_{lj})^2}{2\sigma_{lj}^2}\right]}$$

(for *Zadeh-Mamdani* type) (B6)

or:

$$f(\boldsymbol{x_i}) = \frac{\sum_{l=1}^{M} n_l \prod_{j=1}^{P} \alpha_{lj} \exp\left[-\frac{(x_{ij} - m_{lj})^2}{2\sigma_{lj}^2}\right]}{\sum_{l=1}^{M} \prod_{j=1}^{P} \alpha_{lj} \exp\left[-\frac{(x_{ij} - m_{lj})^2}{2\sigma_{lj}^2}\right]}$$

(for *Takagi-Sugeno* type) (B7)

Suppose the NFI is given a training input-output data pair $[x_i, t_i]$, the system minimizes the following objective function (a weighted error function):

$$E = \frac{1}{2} w_i \left[f(x_i) - t_i \right]^2$$ (w_i are defined in step 2) (B8)

The steepest descent algorithm (BP) (Amari, 1990) is used then to obtain the formulas for the optimization of the parameters G_l, δ_l, α_{lj}, m_{lj} and σ_{lj} of *Zadeh-Mamdani* type NFI such that the value of *E* from equation (B8) is minimized:

$$G_l(k+1) = G_l(k) - \frac{\eta_G}{\delta_l^2(k)} w_i \Phi(\boldsymbol{x}_i)\left[f^{(k)}(\boldsymbol{x}_i) - t_i\right] \tag{B9}$$

$$\delta_l(k+1) = \delta_l(k) - \frac{\eta_\delta}{\delta_l^3(k)} w_i \Phi(\boldsymbol{x}_i)\left[f^{(k)}(\boldsymbol{x}_i) - t_i\right]\left[f^{(k)}(\boldsymbol{x}_i) - G_l(k)\right] \tag{B10}$$

$$\alpha_{lj}(k+1) = \alpha_{lj}(k) - \frac{\eta_\alpha}{\delta_l^2(k)\,\alpha_{lj}(k)} w_i \Phi(\boldsymbol{x}_i)\left[f^{(k)}(\boldsymbol{x}_i) - t_i\right]\left[G_l(k) - f^{(k)}(\boldsymbol{x}_i)\right] \tag{B11}$$

$$m_{lj}(k+1) = m_{lj}(k) - \frac{\eta_m}{\delta_l^2(k)\,\sigma_{lj}^2(k)} w_i \Phi(\boldsymbol{x}_i)\left[f^{(k)}(\boldsymbol{x}_i) - t_i\right]\left[G_l(k) - f^{(k)}(\boldsymbol{x}_i)\right]\left[x_{ij} - m_{lj}(k)\right] \tag{B12}$$

$$\sigma_{lj}(k+1) = \sigma_{lj}(k) - \frac{\eta_\sigma}{\delta_l^2(k)\,\sigma_{lj}^3(k)} w_i \Phi(\boldsymbol{x}_i)\left[f^{(k)}(\boldsymbol{x}_i) - t_i\right]\left[G_l(k) - f^{(k)}(\boldsymbol{x}_i)\right]\left[x_{ij} - m_{lj}(k)\right]^2 \tag{B13}$$

here,

$$\Phi(\boldsymbol{x}_i) = \frac{\prod_{j=1}^{\mathrm{P}} \alpha_{lj} \exp\left\{-\frac{[x_{ij}(k) - m_{lj}(k)]^2}{2\sigma_{lj}^2(k)}\right\}}{\sum_{l=1}^{\mathrm{M}} \frac{1}{\delta_l^2} \prod_{j=1}^{\mathrm{P}} \alpha_{lj} \exp\left\{-\frac{[x_{ij}(k) - m_{lj}(k)]^2}{2\sigma_{lj}^2(k)}\right\}} \tag{B14}$$

The steepest descent algorithm (BP) is also used to obtain the formulas for the optimization of the parameters $\boldsymbol{b}_l$, α_{lj}, m_{lj} and σ_{lj} of the *Takagi-Sugeno* type NFI such that the value of E from equation (B8) is minimized:

$$b_{l0}(k+1) = b_{l0}(k) - \eta_b w_i \Phi(\boldsymbol{x}_i)\left[f^{(k)}(\boldsymbol{x}_i) - t_i\right] \tag{B15}$$

$$b_{lj}(k+1) = b_{lj}(k) - \eta_b x_{ij} w_i \Phi(\boldsymbol{x}_i)\left[f^{(k)}(\boldsymbol{x}_i) - t_i\right] \tag{B16}$$

$$\alpha_{lj}(k+1) = \alpha_{lj}(k) - \frac{\eta_\alpha}{\alpha_{lj}(k)} w_i \Phi(\boldsymbol{x}_i)\left[f^{(k)}(\boldsymbol{x}_i) - t_i\right]\left[n_l(k) - f^{(k)}(\boldsymbol{x}_i)\right] \tag{B17}$$

$$m_{lj}(k+1) = m_{lj}(k) - \frac{\eta_m}{\sigma_{lj}^2(k)} w_i \Phi(\boldsymbol{x}_i)\left[f^{(k)}(\boldsymbol{x}_i) - t_i\right]\left[n_l(k) - f^{(k)}(\boldsymbol{x}_i)\right]\left[x_{ij} - m_{lj}(k)\right] \tag{B18}$$

$$\sigma_{lj}(k+1) = \sigma_{lj}(k) - \frac{\eta_\sigma}{\sigma_{lj}^3(k)} w_i \Phi(\boldsymbol{x}_i)\left[f^{(k)}(\boldsymbol{x}_i) - t_i\right]\left[n_l(k) - f^{(k)}(\boldsymbol{x}_i)\right]\left[x_{ij} - m_{lj}(k)\right]^2 \tag{B19}$$

here,

$$\Phi(\boldsymbol{x}_i) = \frac{\prod_{j=1}^{P} \alpha_{lj} \exp\left\{-\frac{[x_{ij}(k) - m_{lj}(k)]^2}{2\sigma_{lj}^2(k)}\right\}}{\sum_{l=1}^{M}\prod_{j=1}^{P} \alpha_{lj} \exp\left\{-\frac{[x_{ij}(k) - m_{lj}(k)]^2}{2\sigma_{lj}^2(k)}\right\}} \quad \text{(B20)}$$

where: η_G , η_δ , η_b , η_α , η_m and η_σ are learning rates for updating the parameters G_l, δ_l, $\boldsymbol{b}_j$, α_{lj}, m_{lj} and σ_{lj} respectively.

In the NFI training algorithm, the following indexes are used:

- Training data subjects: i = 1, 2, … , N;
- Input variables: j = 1, 2, … , P;
- Fuzzy rules: l = 1, 2, …, M;
- Learning epochs: k = 1, 2,..

Appendix C TWNFI Learning Algorithm

TWNFI Learning Algorithm (From Song and Kasabov, 2006)

For each new data vector $\boldsymbol{x}_q$, an individual model is created with the application of the following steps:

(1) Normalize the training data set and the new data vector $\boldsymbol{x}_q$ (the values are between 0 and 1); set the initial value for every input variable's weight to 1.

(2) Search in the training data set in the input space to find N_q training subjects that are the closest to $\boldsymbol{x}_q$ using a *weighted normalized Euclidean distance*, defined as equation C1. The value of N_q can be pre-defined, based on a preliminary analysis of the problem and the data available, or - optimized through the application of an optimization procedure. Here we assume the former approach.

(3) Calculate the distances d_i, i = 1, 2, …, N_q, using equation C1,between each data subjects and $\boldsymbol{x}_q$; and calculate the weights for each subject, $v_i = 1 - (d_i - \min(\boldsymbol{d}))$, i = 1, 2, …, N_q, $\min(\boldsymbol{d})$ is the minimum element in the distance vector $\boldsymbol{d} = [d_1, d_2, \dots, d_{Nq}]$.

(4) Use a clustering algorithm, for example *ECM* (Kasabov and Song, 2002; Song and Kasabov, 2001), to cluster and partition the input sub-space that consists of N_q selected training subjects.

(5) Create fuzzy rules and set their initial parameter values according to the clustering procedure results. Each fuzzy rule is created based on a cluster: the cluster centre is taken as the centre of the fuzzy membership function (*Gaussian* function) and the cluster radius is taken as the width.

(6) Apply the steepest descent method (*back-propagation*) to optimize the weights and the parameters of the fuzzy rules in the local model M_q following Eq. (C6 – C13).

(7) Search in the training data set to find a new set N_q of the closest to $\boldsymbol{x}_q$ subjects (Step 2); if the same subjects are found as in the last search, the algorithm goes to Step 8; otherwise, it goes to Step 3.

(8) Calculate the output value y_q for the input vector $\boldsymbol{x}_q$ applying fuzzy inference over the set of fuzzy rules that constitute the local model M_q.

(9) End of the procedure.

The weight and parameter optimization procedure is described below:

Consider the system having P inputs, one output and M fuzzy rules defined initially through the clustering procedure, the l-th rule has the form of:

R_l : If $x_1$is F_{l1} and x_2 is F_{l2} and … x_P is F_{lP}, then y is G_l . (C1)

Here, F_{lj} are fuzzy sets defined by the following *Gaussian* type membership function:

$$GaussianMF = \alpha \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right] \quad \text{(C2)}$$

and G_l are of a similar type as F_{lj} and are defined as:

$$GaussianMF = \exp\left[-\frac{(y-n)^2}{2\delta^2}\right] \quad \text{(C3)}$$

Using the *Modified Centre Average* defuzzification procedure the output value of the system can be calculated for an input vector $x_i = [x_1, x_2, \ldots, x_P]$ as follows:

$$f(\boldsymbol{x}_i) = \frac{\sum_{l=1}^{M} \frac{n_l}{\delta_l^2} \prod_{j=1}^{P} \alpha_{lj} \exp\left[-\frac{w_j^2 (x_{ij} - m_{lj})^2}{2\sigma_{lj}^2} \right]}{\sum_{l=1}^{M} \frac{1}{\delta_l^2} \prod_{j=1}^{P} \alpha_{lj} \exp\left[-\frac{w_j^2 (x_{ij} - m_{lj})^2}{2\sigma_{lj}^2} \right]} \qquad \text{(C4)}$$

Here, w_j are the current weights of the input variables and n_l is the point having maximum membership value in the lth output set.

Suppose the TWNFI is given a training input-output data pair $[x_i, t_i]$, the system minimizes the following objective function (a weighted error function):

$$\mathrm{E} = \frac{1}{2} v_i \left[f(x_i) - t_i \right]^2 \qquad \text{(C5)}$$

(v_i are defined in Step 3)

The steepest descent algorithm is used then to obtain the formulas (C6- C12) for the optimization of the parameters n_l, δ_l, α_{lj}, m_{lj}, σ_{lj} and w_j such that the value of E from Eq. (C6) is minimized:

$$n_l(k+1) = n_l(k) - \frac{\eta_n}{\delta_l^2(k)} v_i \Phi(\boldsymbol{x}_i) \left[f^{(k)}(\boldsymbol{x}_i) - t_i \right] \qquad \text{(C6)}$$

$$\delta_l(k+1) = \delta_l(k) - \frac{\eta_\delta v_i \Phi(\boldsymbol{x}_i)}{\delta_/^3(k)} \left[f^{(k)}(\boldsymbol{x}_i) - t_i \right] \left[f^{(k)}(\boldsymbol{x}_i) - n_l(k) \right] \tag{C7}$$

$$\alpha_{lj}(k+1) = \alpha_{lj}(k) - \frac{\eta_\alpha v_i \Phi(\boldsymbol{x}_i)}{\delta_l^2(k)\,\alpha_{lj}(k)} \left[f^{(k)}(\boldsymbol{x}_i) - t_i \right] \left[n_l(k) - f^{(k)}(\boldsymbol{x}_i) \right] \tag{C8}$$

$$m_{lj}(k+1) = m_{lj}(k) - \frac{\eta_m w_j^2(k) v_i \Phi(\boldsymbol{x}_i)}{\delta_l^2(k)\,\sigma_{/j}^2(k)} \left[f^{(k)}(\boldsymbol{x}_i) - t_i \right] \left[n_l(k) - f^{(k)}(\boldsymbol{x}_i) \right] \left[x_{ij} - m_{lj}(k) \right] \tag{C9}$$

$$\sigma_{lj}(k+1) = \sigma_{lj}(k) - \frac{\eta_\sigma w_j^2(k) v_i \Phi(\boldsymbol{x}_i)}{\delta_l^2(k)\ \sigma_{/j}^3(k)} \left[f^{(k)}(\boldsymbol{x}_i) - t_i \right] \left[n_l(k) - f^{(k)}(\boldsymbol{x}_i) \right] \left[x_{ij} - m_{lj}(k) \right]^2 \tag{C10}$$

$$w_j(k+1) = w_j(k) - \frac{\eta_w w_j(k) v_i \Phi(\boldsymbol{x}_i)}{\delta_l^2(k)\,\sigma_{/j}^2(k)} \left[f^{(k)}(\boldsymbol{x}_i) - t_i \right] \left[f^{(k)}(\boldsymbol{x}_i) - n_l(k) \right] \left[x_{ij} - m_{lj}(k) \right]^2 \tag{C11}$$

if $w_j(k+1) > 1$ then $w_j(k+1) = 1$; if $w_j(k+1) < 0$ then $w_j(k+1) = 0$

here,

$$\Phi(\boldsymbol{x}_i) = \frac{\prod_{j=1}^{P} \alpha_{lj} \exp\left\{-\frac{w_j^2(k)\left[x_{ij} - m_{lj}(k)\right]^2}{2\sigma_{lj}^2(k)}\right\}}{\sum_{l=1}^{M} \frac{1}{\delta_l^2} \prod_{j=1}^{P} \alpha_{lj} \exp\left\{-\frac{w_j^2(k)\left[x_{ij} - m_{lj}(k)\right]^2}{2\sigma_{lj}^2(k)}\right\}} \qquad \text{(C12)}$$

where: η_n, η_δ, η_α, η_m, η_σ and η_w are learning rates for updating the parameters n_l, δ_l, α_{lj}, m_{lj}, σ_{lj} and w_j respectively.

In the TWNFI learning algorithm, the following indexes are used:

- Training data subjects: $i = 1, 2, \ldots, N$;
- Input variables: $j = 1, 2, \ldots, P$;
- Fuzzy rules: $l = 1, 2, \ldots, M$;
- Training iterations (epochs): $k = 1, 2, \ldots$;

In this learning procedure, we use a clustering method called *ECM* (*Evolving Clustering Method*) for clustering and use a steepest descent algorithm for parameter optimization (Mendel, 2001; Lin and Lee, 1996; Wang, 1994). Although some other clustering methods can be used such as *K-means*, *Fuzzy C-means* or the *Subtractive* clustering method (Neural Network toolbox, 2002),

ECM is more appropriate because it is a fast one-pass algorithm and produces well-distributed clusters. The number of clusters, M, depends on the data distribution in the input space and it can be set up by experience, probing search or optimization methods (e.g. a *genetic algorithm*).

For generalization and simplicity, a general steepest descent method has been used in the TWNFI learning algorithm. The *Levenberg-Marquardt*, *one-step*

second back-propagation algorithm, *least-squares* method, *SVD-QR* method or some others (Mendel, 2001; Wang, 1994) may be applied in the TWNFI for parameter optimization instead of the steepest descent algorithm.

Appendix D Formulas used to calculate percentages for nutrient variables (Atwater and Bryant, 1900)

For percentage of protein, sugar and carbohydrates following formula has been used to:

$$\text{Protein\%} = \frac{\text{Protein intake(grams)} \times 16.72\text{kilojoules} \times 100}{\text{Total energy intake (kilojoules)}}$$

$$\text{Carbohydrates \%} = \frac{\text{Carbohydrates intake(grams)} \times 16.72\text{kilojoules} \times 100}{\text{Total energy intake (kilojoules)}}$$

$$\text{Sugar\%} = \frac{\text{Sugar intake(grams)} \times 16.72\text{kilojoules} \times 100}{\text{Total energy intake (kilojoules)}}$$

For percentage of saturated fat and total fat following formulas are used:

$$\text{Total fat \%} = \frac{\text{Total fat intake(grams)} \times 37.6\text{ kilojoules} \times 100}{\text{Total energy intake (kilojoules)}}$$

$$\text{Total saturated fat \%} = \frac{\text{Total saturated fat intake(grams)} \times 37.6\text{ kilojoules} \times 100}{\text{Total energy intake (kilojoules)}}$$

Appendix E NeuCom

NeuCom is a computer environment based on connectionist (Neuro-computing) modules. NeuCom is self programmable, learning and reasoning tool. NeuCom environment can be used for data analysis, modeling and knowledge discovery. NeuCom has been developed at Knowledge Engineering and Discovery Research Institute (KEDRI, http://www.kedri.info). NeuCom consists of about 60 various techniques that include different methods of machine learning and knowledge engineering methods. Figure E.1 shows the screenshot of NeuCom environment.

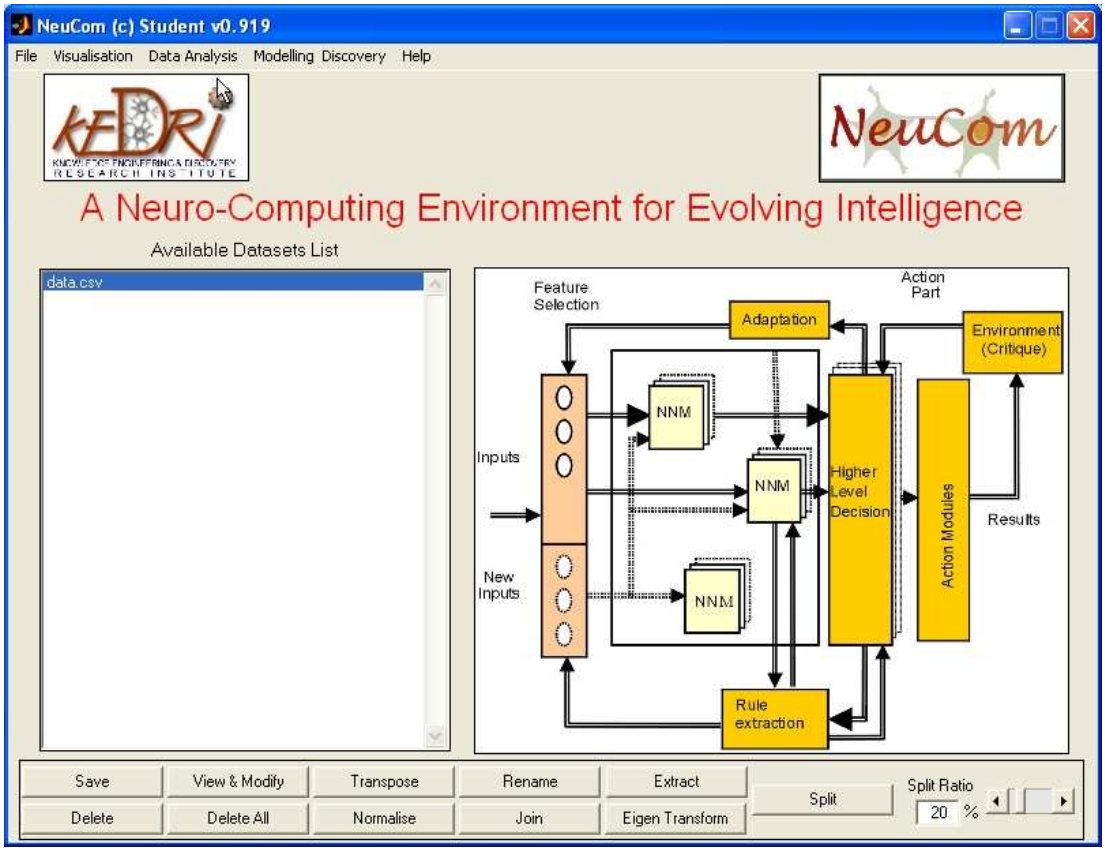


Figure E.1. Screenshot of the NeuCom environment.

The main modules of NeuCom are normalization, visualization, data analysis and modeling discovery. In NeuCom, data can be visualized by different ways such as 2D visualization, 3D visualization, class distribution, surface plot, principal component analysis visualization (PCA) and linear discriminant analysis (LDA). The 3D visualization module allows plotting the datasets in a 3D space and for classification dataset, the data points can be plotted in different colours. In this, the visualization can be dragged and rotated freely. In class distribution module, the distribution of subjects in each class for each variable, displays up to 4 variables at a time. Surface plot module can be used to view all variables and all subjects in one figure. The range of variables for plotting can be selected in this module. In principal component analysis visualization (PCA) module, variance in a dataset can be captured in terms of principal components. It can be used to reduce the dataset's dimension to summarise the most important parts of the dataset. The principal components are displayed with its relative percentage captured variance. PCA can be performed on all kind of datasets but linear discriminant analysis (LDA), only works with classification datasets.

For data analysis and feature selection, there are different modules such as signal to noise ratio, correlation coefficient and clustering methods. Signal to noise ratio(SNR) module gives a quantitative measure of how much each variable in a given data set for classification discriminates one class (considered as the "signal") from the other class (classes) (considered as "noise"). SNR can be used as a feature extraction tool for multivariate data. Variables that have higher values for an output class versus other classes are ranked higher as they have a higher SNR. Inputs to this function are: the data set, and the number of variables to rank. The clustering techniques present in

NeuCom are k-means clustering, evolving clustering method (ECM) and biclustering. K-Means clustering is one of the traditional clustering methods. In this clustering module, user can specify the number of clusters required. Bi-clustering technique is used to identify the relationship between subjects and variables. Evolving clustering method (ECM) has been developed at KEDRI.

For modeling discovery, there are several modules such as classification, prediction, and optimization and cross validation. For classification there are several sub modules which include statistical methods (multiple linear regression (MLR), support vector machine), neural networks methods (multi-layer perceptrons, radial basis function for classification) and evolving connectionist methods (evolving classification function (ECF), evolving clustering method for classification (ECMC)). Multiple linear regression module performs a least squares fit on the given multivariate data for each class. Multi-layer perceptrons (MLP) are standard neural network models for learning from data a non-linear function that discriminates (or approximates) data according to output labels (values). ECF is a clustering based classification system. ECMC is a clustering based classification algorithm which is used in this context as a classifier. For prediction, statistical methods include multiple linear regression (MLR), logarithmic regression for prediction and K Nearest Neighbour (KNN); neural networks methods include multi-layer perceptrons, radial basis function for classification and evolving connectionist methods include dynamic evolving neuro-fuzzy inference system (DENFIS) and evolving fuzzy neural networks (EFuNN). Optimization module includes general genetic algorithm, general evolutionary strategy, genetic algorithm for offline ECF optimization, evolutionary strategy for offline ECF optimization. Genetic Algorithm (GA) is presented for use on user-defined problems.

Evolutionary Strategy (ES) can be used on user-defined problems and can provide and optimize a set of parameters for that function. In Genetic Algorithm for offline ECF optimization module, a Genetic Algorithm (GA) is used to optimise the ECF network and perform feature extraction on the dataset on which the network is trained and tested. In evolutionary strategy for offline ECF optimization module, an Evolutionary Strategy (ES) is used to optimise the ECF network and perform feature extraction on the dataset on which the network is trained and tested.

NeuCom is applicable to many disciplines such as bioinformatics, neuro-informatics, medicine, health informatics, business, banking and finance, engineering, arts and design, horticulture and agriculture, geography, social and environmental sciences and in any field where problem data is available and it needs to be analysed, models need to be created, and rules or profiles need to be discovered. NeuCom is also a development environment where new intelligent systems for decision support and data analysis can be created across disciplines.

Appendix F Siftware

Siftware is an environment for analysis, modeling and profiling of gene expression data. Siftware has been developed at Knowledge Engineering and Discovery Research Institute (KEDRI, http://www.kedri.info). Figure F.1 shows the screenshot of Siftware environment.

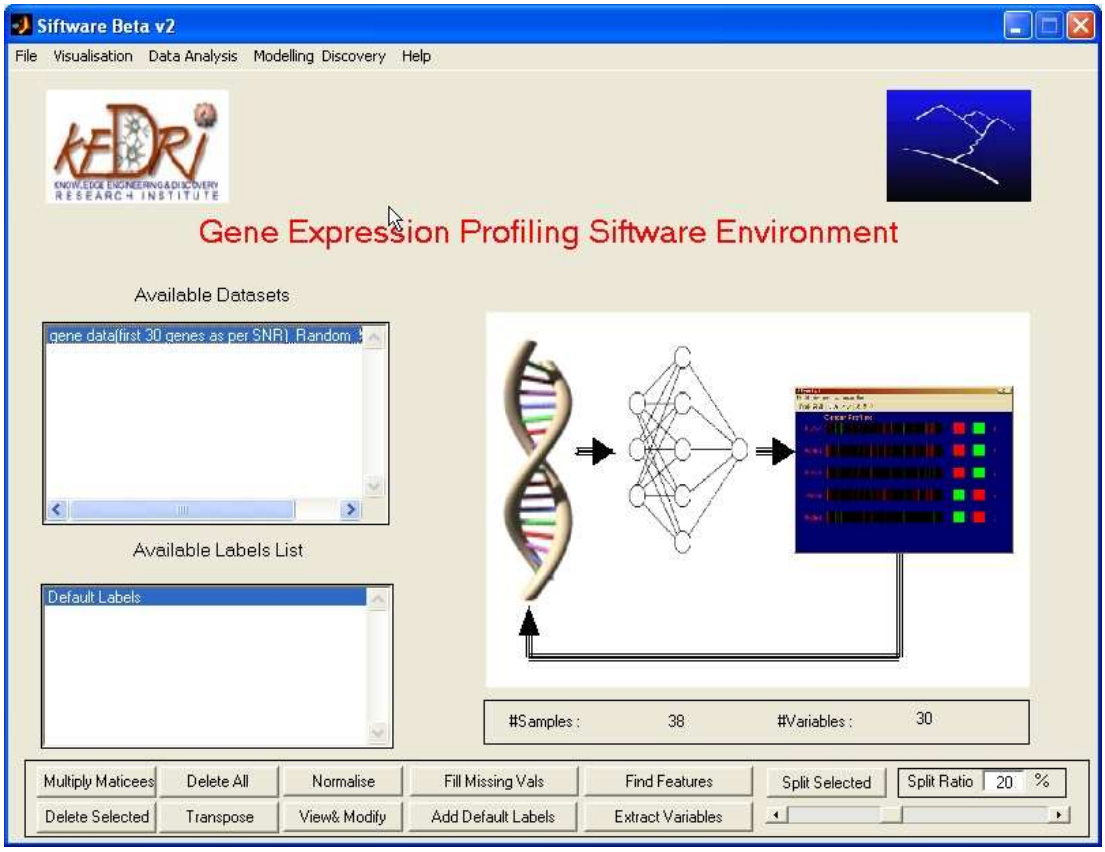


Figure F.1. Screenshot of the Siftware environment.

Siftware contains visualization, data analysis and modeling discovery modules. Visualization module includes 3D visualization, principal component analysis visualization (PCA) and linear discriminant analysis (LDA). Data analysis and feature selection in Siftware can be done by different methods such as correlation coefficient, signal to noise ratio and t-test. Clustering methods

include k-means clustering and hierarchical clustering. There are different modules for modeling and discovery in Siftware. Modeling methods include multiple linear regression and evolving classification function (ECF). For optimization of model, there are two modules genetic algorithm for ECF and genetic algorithm for multiple linear regression. In genetic algorithm for ECF and multiple linear regression, genetic algorithm is used to optimize both methods and performs feature extraction on the dataset on which the network is trained and tested.