UNIVERSITY OF TECHNOLOGY SYDNEY
DATA MINING
(SUBJECT: 32130)

Data Mining

Research Proposal

(Assignment 3)

TABLE OF CONTENTS

1 Introduction

This document will present a proposal for applying data mining techniques for the purpose of gaining insights into the nature of the debilitating Parkinson's disease. Microarray DNA data and clinical data, to be gathered by means of a volunteer collection programme, will serve as the source data to be analysed. A project strategy, plan and schedule will be constructed as a means to satisfy the primary objective of aiding the development of better treatments or a cure for Parkinson's disease through the identification of any genes that play a role in its development.

2 Background

Parkinson's Disease is a progressive degenerative neurological condition of the brain that controls motor functions. The disease, increasingly over time, reduces the ability of the brain to send signals to the part of the brain that controls movement (or motor functions).

> Parkinson's disease is the second most common neurodegenerative disease, after Alzheimer's disease. It is estimated that the disease currently affects at least 500,000 Americans. The disease occurs when nerve cells in an area of the brain known as the substantia nigra die or become impaired. Normally, these neurons produce an important brain chemical known as dopamine. The loss of dopamine-producing brain cells results in the four main symptoms of Parkinson's disease: tremors, rigidity of the limbs, slowness of movement, and impaired balance and coordination.(NIH 2003)

Stem cells have been used in clinical trials involving Monkeys, Rats and Mice in order to deliver added dopamine to the affected area of the brain of these animal subjects to alleviate Parkinson-like symptoms (Wade 2002). These stem cells have been converted into the type of brain cell responsible for producing dopamine in the brain. These trials have had varying levels of success, but highlight the ultimate goal of finding a treatment to halt the damage done to the dopamine-producing brain cells (substantia nigra) so they can function normally.

More recently the National Institute of Health of America (NIH), took a major in identifying the α-synuclein gene to be implicated in the development of a rare familial form of early-onset Parkinson's disease (NIH 2003). However, this study is not the end of the road – rather it's more like we are turning the corner into a whole new road. This study, while being a breakthrough, only involved subjects from a small portion of a family tree

More evidence is mounting for genetic links to Parkinson's disease. This proposal will put forward the use of data mining to discover crucial variances in genetic sequences that could be associated with the disease, hopefully leading to a cure.

3 Data Collection Programme

The data collection programme will gather Microarray DNA data and DNA signatures via DNA samples (small amount of blood taken from the subject) and clinical/environmental data via a survey.

Microarray DNA data is a measure of gene expression at any one moment. Genes on the genome are responsible for protein synthesis, so the gene expression is a (continuous value) measure of how active a particular gene is at that time in synthesising the protein (Watson 2003). This data gives us the ability to do comparisons of gene expression

The collection programme will be targeting a balanced sample of both sufferers and non-sufferers, as well as aiming for a good spread of racial backgrounds (such that there is a reasonable amount of variability in the genetic samples). As the symptoms of Parkinson's disease are not typically witnessed until after the age of 60 (Eeden et al. 2003), the data collection programme will be gathering data from subjects between the ages of 60 and 70. This measure will allow the study to have a reasonable level of certainty that subjects non showing any symptoms during the study are not likely to develop the symptoms after the study.

It should be noted that this data collection programme is not within the scope of this proposal. It has already been planned, and it is currently (at the time of this writing) in the early stages of gathering appropriate volunteers. This programme is the basis for this proposal. It is due for completion in May 2004.The project plan for the proposed data mining project will be aligned with key dates identified by the data collection programme.

4 Objectives of the study

The general or primary object of the study is:

> *To provide information that will be of vital importance to the development of treatments for Parkinson's disease, either to treat the symptoms of the disease, or as a cure for the disease.*

Another important objective of the study is to provide the information required to develop a precise diagnostic test for the disease that could be used well in advance of any symptoms being present using DNA sampling. This could prove to be very beneficial if early treatment of the disease ends up being more effective in controlling the damage caused by the disease.

Should the study fail to produce information that could directly lead to a treatment or cure, it should at least be able to feed into further research, perhaps in identifying a number of genes that are *not* implicated in the disease.

5 Possible Outcomes

- Identify genes that play a role in the malfunctioning regulation of the neurotransmitter dopamine in Parkinson's disease patients.
- Identify any environmental factors that may influence the timing and severity of the symptoms.
- Provide the basis research for possible gene therapies that may be developed to make corrective measures to patients with the defective genes before they witness any symptoms.
- Identify gene sequences to give rise to the prominence of the disease for the purposes of diagnosis by DNA sampling.

6 Potential Beneficiaries

The knowledge acquired from this study will be primarily aimed at Biochemists and Biochemical organizations with the intent and capability of developing treatments or cures for Parkinson's disease given the evidence of one or more genetic mutations that cause the condition.

As stated previously in this document, rather than producing direct evidence of a genetic link (in terms of isolation of all genes implicated), it is likely that this study will provide information that will benefit further research in the area. One more step towards the cure for Parkinson's disease. In this case, the beneficiaries will be other scientific and/or genetic researchers.

7 Success Criteria

In order to keep the project focused, the proposal includes the keeping of both measurable and judgement-based criteria for success in the form of key performance indicators (KPI) and qualitative targets and milestones. These criteria should be developed and maintained in order to support the primary objectives state earlier in this document.

8 Resources Available

Data collection of microarray data, DNA signatures and patient clinical records provide the data source for analysis. The target for the data collection programme is to obtain 5000 samples. This proposal expects that figure to be achievable in the timeframe.

9 Resources Required

Personnel: 1 Neurological expert (consultant), 3 x Data miners, Project manager, Research liaison officer and office administrator.

The Neurological expert will be brought onboard periodically to ensure the data analysis being done will be guided in the right direction and should be able to catch any invalid assumptions that may be made by the data miners alone.

The Research Liaison officer will ensure that the efforts within the team are communicated to relevant parties. This will be mostly acting out the role of the sales of the resulting IP of the research. Should the project be government funded, the role is likely to be similar, but would be more of a mediator role.

The Project Manager will ensure that the projects tasks are carried out, according to the schedule and within the budgetary constraints. Other typically Project Management responsibilities, such as risk management, would be assigned to the Project Manager.

The Office Administrator would ensure the general business requirements of the operation are upheld.

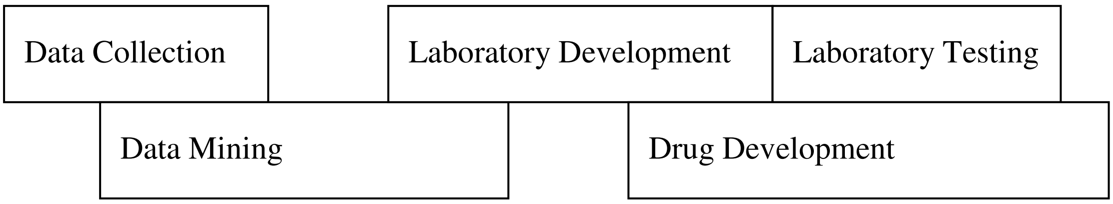
Hardware: 2 x Sun Fire V880 servers and 5 PCs.

The Sun Fire V880 servers will do the serious data crunching and model building required to undertake this project. There is the potential to set these two servers up in a grid configuration such that they can operate on the same task at the same time.

Software: Data Mining tools (technology selection subject to tool prototyping phase).

10 Context

To put the data mining project in the context of the overall scientific endeavour to find a cure for Parkinson’s disease, the following major tasks are also required.



The Data Collection Programme has already been discussed, there fore needs no further explanation.

The Laboratory Development would come from using the output of the data mining project to clone genes sequences that have been identified as contributors to the disease. This may lead to drugs that may block this type of gene if it is over active or regulate it, if it is under active. This project could also be used to develop good copies of the problem gene (or genes) for use in gene therapies (once this practice become feasible on human subjects.

Laboratory testing is always done to validate the developmental laboratory work in witnessing how they work in practice, typically on animal subjects.

The Laboratory work will be feeding into drug development and vice versa, in an iterative manner, until the final product can be produced.

11 Risks and Contingencies

This project is exploratory by nature; therefore the risk of not providing any useful or conclusive information needs to be managed.

The Project Manager will need to take a risk management stance in running the project, as the risk of failure due to an ill focussed strategy will be high. The ability to switch to a contingency plan when an objective will not be achievable within a realistic timeframe will be paramount.

Some external risks to the project are:

- Incorrect or excessively noisy data provided by the data collection programme. This will be managed by doing early validation of the data using sample data.
- Inappropriate or inflexible tools used (and unable to produce quality models). This risk will be mitigated somewhat by the prototyping of various tools and techniques on the sample data set.
- Competitor produces the results before this project is complete. In this scenario the project could slightly change its objectives to validating the other study.
- Funding is cut during the project. This risk can be managed by ensuring that there is always some level of information that would be useful for further

research efforts. This means that all results from modelling should be generated frequently.

12 Costs and Benefits

Non-exclusive rights to use the data from the data collection programme: $150,000

Hardware costs:

Sun Fire V880 x 2: $270,000

PCs x 5: $10,000

Software: estimate at $200,000

Office lease and equipment for 12 months: $20,000

Office administration: $10,000

Personnel costs:

Neurological expert: Estimated 3 months in man-hours for consultant - $100,000

3 x Data Miners for 12 months full-time: $360,000

1 x Project Manager for 12 months full-time: $150,000

1 x Research liaison officer for 12 months part-time: $45,000

1 x Office Administrator full-time: $60,000

Total costs: $1,375,000 AUD over 12 months.

Hardware re-sale value (@ 35% depreciation): $182,000 AUD

The benefits of the program could be massive, particularly if the data suggests a clear pathway to a cure or better treatment of the disease. The research Liaison officer is on the project to sell the benefits of the results of the project to researchers that are able to turn this information into a treatment.

13 Process Methodology

The process standard used for the planning and execution of the data mining project will be CRISP-DM (http://www.crisp-dm.org/). CRISP-DM is a hierarchical approach to process modelling specifically formulated for data mining projects. At the highest level of abstraction are phases. Within each phase there are a collection of tasks. For these tasks there may be a variety of tools and techniques that can be employed to met the data mining objective. The project plan will only show the level of phases, however, during the implementation of the project the more specific planing involving the generic tasks, specialized tasks and process instances will be done.

14 Technical Implementation

14.1 Why Data Mining is appropriate for the task?

The analysis of vastly complex and highly dimensional data such as genetic Mircoarray DNA data requires sophisticated data modelling techniques and learning algorithms to decipher and interpret such data for the purposes of pattern matching (or recognition) for prediction. The human genome contains approximately 30,000 genes, with an average of 3000 nucleotide bases per gene. That's a dimensionality of 90 million nucleotide bases to cover just the gene-coded regions of the genome, which accounts for just 2% of the entire genome (*Early Insights from the Human DNA Sequence* 2003). This study will only be concerned with the regions of the genome that code for proteins (i.e. genes). Non-coding or junk DNA will not be regarded in this study.

14.2 Data Mining Objectives

Identify one or more genes implicated in causing the death of the nerve cells substantia nigra. At the moment, the only way to judge if a patient has the disease is to measure the symptoms. This data will be:

- Does the subject show signs of the symptoms? (binary)
- If so, what are the severity of the symptoms (categorical and to a degree subjective)
- Time period the symptoms have been observable.

DNA Microarray data will be used to uncover the variance in gene expression. This should lead to the identification of genes that play a role in the disease.

If time is available, this information can then be used to look more closely at the structure of the genes that the offenders. This is also a data mining exercise in looking for patterns in gene structure that may lead to the abnormal destruction of the nerve cells.

The Neurological expert will be used to determine what the appropriate cut-off criteria for success is when it comes to identifying gene expression variation.

15 Data Mining Phases & Tasks

15.1 Data Understanding

15.1.1 Prototyping / Data Exploration

Using a small subset of the data that will be used in the study, the team will do some prototyping of a number of tools and techniques that could be applied to this problem. This prototyping phase will be strictly for the purposes of evaluating the appropriateness of the tools and techniques evaluated. No model built from this phase will be kept to influence the final model (or models) built.

It will be possible for this phase to overlap with the data collection programme, as we will arrange for some early access data to be available for testing.

The prototyping phase will be guided by the knowledge of similar research done in the past. For example, other researchers have been successful in using clustering techniques when using microarray data to isolate genes implicated in Multiple Sclerosis (Jiang, Tang & Zhang). This information would influence the approach taken to establish appropriate techniques during the prototyping phase.

Performance constraints imposed by the tools and techniques would also be extrapolated from the sample data during the prototyping phase to ensure feasibility within the project schedule.

15.1.2 Verify Data Quality

Given the sample data from the data collection programme, it will also be a prudent time to verify that the data is of sufficient quality. At this stage in the process, verifying the quality means that we would check that there are not excessive missing values (particularly in the genetic samples, where we don't expect any missing genetic data). Also, that the qualitative data (i.e. clinical data) is being captured diligently enough to prove useful. At this stage we cannot check at the coverage of samples taken is adequate (i.e. covering many genetic variations), as we are only dealing with an early sample.

We expect that the data collection programme will be responsible for verifying the quality of the microarray data using triangulation (i.e. using more than one machine to process the same sample, for a subset of the entire sample collection, to verify the results are consistent). This technique will ensure the machines are operating as expected.

15.2 ***Data Preparation***

15.2.1 Data Selection

Depending on the lessons learned during prototyping, there might be the opportunity to refine the strategy for what data attributes would be required to be kept in the included set of data for model building. This may be due to some redundant data attributes, or those that are not likely to have any impact final result. The decision to deselect any subset of the genetic data gathered is not likely, as the aim of the study is to find those genes that play a part in the disease. However, the clinical data may be a good target for this exercise.

15.2.2 Data Cleansing and Construction

The relatively unstructured clinical data will need to be flattened and cleaned. If it is feasible to fill in some missing attributes with default values then that will be the responsibility of this task.

15.2.3 Data Formatting

Any tools that require the data to be in a format other than the original format will need to be transformed in to the desired format without altering the semantics of the data. For example, a neural network will require categorical data to be split into binary input and output attributes. Some binning is often required to transform numeric attributes to categorical attributes.

15.3 ***Modelling***

This phase of the data mining project is where the major deliverable is produced – the model.

The modelling phase of the project will be highly specific to the technique being applied, however, any modelling technique applied will use a training, validation and test set of data to build the model. This ensures that the model will not be too specific to the training data (this is called over-fitting) such that it should be applicable to any unseen cases (perhaps for the purpose of diagnosis).

15.4 ***Evaluation***

This phase is used to evaluate that the model produced will meet the business objectives. If this stage indicates that the model does not adequately satisfy the business requirements, the model may need to be revised. In this sense, the process is iterative on this large scale, but it is also iterative within the Modelling phase, as the model is slowly refined on each new revision of the model.

15.5 ***Deployment***

In this project the deployment is not to implement a system, but rather to generate formal reports that can be used to carry on the research in the laboratories.

16 Project Plan

The following plan is a rough schedule of the various phases that will go into the data mining project. The project can be iterative in the way the various phases are scheduled, so durations are presented in elapsed time.

Start Date	**Phase**	**Duration (elapsed time)**	**Staff Involved**
Dec 2003	Business Understanding	2 months	Project Owner, Project Manager
Feb 2004	Data Understanding	4 months	All
June 2004	Data Preparation	1 month	Data Miners
June 2004	Modelling	4 months	Project Manager, Data Miners and domain expert.
August 2004	Evaluation	1 month	Project Manager, Data Miners and domain expert.
Dec 2004	Deployment	1 month	Project Manager, Data Miners.

The total duration of the project is 12 months.

17 References

Early Insights from the Human DNA Sequence [Online]. 2003, Available: http://www.ornl.gov/sci/techresources/Human_Genome/publicat/primer2001/4.shtml [Accessed 25 Oct 2003].

Eeden, S. K. V. D., Tanner, C. M., Bernstein, A. L., Fross, R. D., Leimpeter, A., Bloch, D. A. & Nelson, L. M. 2003, 'Incidence of Parkinson's Disease: Variation by Age, Gender, and Race/Ethnicity', *American Journal of Epidemiology Online*.

Jiang, S., Tang, C. & Zhang, L., *DNA Microarray Technology, Data Mining Help Researchers Differentiate among Patients with Multiple Sclerosis* [Online]. Available: http://www.globaltechnoscan.com/25thApr-2ndMay01/microarray.htm [Accessed 1 Nov 2003].

NIH 2003, *Major New Finding on Genetics of Parkinson's Disease Zeroes in on Activity of Alpha Synuclein* [Online]. Available: http://www.nih.gov/news/pr/oct2003/nia-30.htm [Accessed 3 Nov 2003].

Wade, N. 2002, 'Progress Is Reported on Parkinson's Disease', *The New York Times*, June 21, Available at: http://www.parkinsoncolorado.org/news2_files/nyt_62102.htm

Watson, J. 2003, *DNA: The Secret of Life*, William Heinemann.