UPTEC X 07 063
Examensarbete 20 p
December 2007

Evaluation of prediction models for biomarkers

The role of rooting models on literature networks

Sten Blomstrand

UPPSALA UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 07 063	**Date of issue Oct 2007**

Author

Sten Blomstrand

Title (English)

Evaluation of prediction models for biomarkers - the role of rooting models on literature networks

Title (Swedish)

Abstract

PLS models based on a purely mathematical approach were compared to PLS models with literature references. Microarray data analyzed was based on human cells treated with a GSK3β inhibitor substance.

Keywords

Bioinformatics, prediction modelling, PLS, GSK3β, Ingenuity, Alzheimer

Supervisors

Hugh Salter, Kerstin Nilsson
AstraZeneca

Scientific reviewer

Mikael Thollesson
EBC, Uppsala Universitet

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **31**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Evaluation of prediction models for biomarkers - the role of rooting models on literature networks

Sten Blomstrand

Sammanfattning

Att finna läkemedel mot kroniska sjukdomar som Alzheimers är en av forskarvärldens största utmaningar. För att kunna framställa dessa läkemedel måste man ha förståelse för vilka proteiner som ger upphov till sjukdomen och hur de fungerar och interagerar.

Genom att använda mikromatriser kan man mäta koncentrationer av tiotusentals gener samtidigt. Dessa mätvärden ger uppfattningar om hur proteinnivåer förändras på grund av till exempel ett läkemedel, och man kan därigenom identifiera viktiga proteiner som blir mer eller mindre påverkade. Det svåraste steget i detta tillvägagångssätt är just identifieringen av dessa proteiner.

För att underlätta detta steg brukar man använda sig av matematiska prediktionsmodeller, men även dessa kan stöta på problem. Dessa modeller kräver ofta en stor samling testdata, ofta flera hundra mikromatrissvar, och då nya läkemedel inledningsvis oftast testas på endast ett fåtal patienter ger modellerna därför inte tillförlitliga resultat. Ytterliggare problem som kan uppstå är att endast ett fåtal av de tiotusentals generna som mäts med en mikromatris är påverkade av läkemedlet. Då man försöker applicera matematiska prediktionsmodeller på sådana data försvinner dessa proteiner i mängden och man får inte heller då tillförlitliga resultat.

Genom litterära referenser kan man dock ofta hitta proteiner länkade till läkemedlets målprotein och därigenom få en uppfattning om hur dessa bör påverkas. Detta leder till att man behöver färre mikromatrissvar samt mindre beståndsdelar av dessa mikromatrissvar för att bygga matematiska prediktionsmodeller och ändå få informativa resultat.

Målet med detta examensarbete var därför att bygga rent matematiska prediktionsmodeller och jämföra dessa med matematiska prediktionsmodeller baserade på litterära referenser.

Slutsatser som drogs var att modeller baserade på litterära referenser kan ge bättre prediktionsförmåga än rent matematiska modeller. Vidare identifierades också ett antal proteiner som är kopplade till sjukdomsprocessen i Alzheimers.

Examensarbete 20p i Bioinformatik
Uppsala Universitet December 2007

Sammanfattning

Dagens mikromatriser mäter koncentrationer på tiotusentals gener samtidigt. Detta vållar dock problem då man ämnar bygga matematiska prediktionsmodeller på behandlingen av ett läkemedel med endast en eller ett fåtal mål. Problemet som uppstår är att mikromatrissvaret innehåller mestadels brus, då endast ett fåtal gener är påverkade av behandlingen. Genom litteratur kan man dock ofta hitta proteiner länkade till läkemedlets målprotein och därigenom få en uppfattning om hur dess omgivning bör påverkas. Detta leder till att man endast behöver undersöka en liten beståndsdel av mikromatrissvaret för att få informativa resultat. Målet med detta examensarbete var därför att bygga prediktionsmodeller på rent matematiska metoder och jämföra dessa med matematiska modeller med litterära referenser.

De slutsatser som drogs är att man kan öka prediktionsförmågan då man använder sig av litterära referenser, givet att dessa referenser är tillräckligt informativa. Vidare identifierades också ett fåtal proteiner som möjligtvis kan användas som biomarkörer för GSK3β inhibitorer, ett protein med starka associationer till Alzheimers sjukdom.

Contents

Abbreviations

Aβ	beta amyloid
AD	Alzheimer's Disease
AZ	AstraZeneca
FN	False Negative
FP	False Positive
GSK3β	Glycogen synthase kinase 3-β
IPA	Ingenuity Pathways Analysis
MAPT	Microtubule Associated Protein Tau
NFTs	NeuroFibrillary Tangles
PLS	Projection to Latent Structures by means of Partial Least Squares
PP	Prediction Performance
PSEN1	Presenilin 1
TN	True Negative
TP	True Positive
VIP	Variable Influence on Projection

1 Introduction and Aim

As techniques such as microarrays enable us to generate immense amounts of data, the need for analysis and interpretation increases. One challenge that accompanies this is the use of prior data to classify new samples (prediction modelling). Tools such as PLS, Neural Networks and Random Forests are widely used in bioinformatics to create these models [3, 19, 15]. Most methods are purely mathematical and thus do not take information from literature into account. Even though mathematical models may produce accurate predictions, the interpretation of how, and why, may be lost. One way to explain these questions is to look into literature.

As new findings in protein pathways constantly emerge, the vast networks that constitute the basics of protein relationships are discovered. As papers of these findings are ceaselessly published, the literature information available constantly increases.

One major issue related to these articles are the ontologies used. The most well-known organisation trying to address this problem is the Gene Ontology project (http://www.geneontology.org/), a project attempting to standardise the names of function and associations of gene products. Searchable databases such as PubMed (http://www.ncbi.nlm.nih.gov/sites/entrez) do not take different ontologies into account, and therefore the resulting findings may be more or less unrelated to the initial query. This makes automatic data-mining difficult due to the possibility of ambiguous results, and manual data-mining troublesome and time consuming for the inexperienced.

Another type of literature searchable database is provided in the knowledgebase by Ingenuity Pathways Analysis (http://www.ingenuity.com). This is a kind of combination of both article databases such as PubMed and the Gene Ontology database providing literature references to articles, biological functions and protein pathways networks.

Using the information provided by Ingenuity Pathways Analysis, the aim of this thesis project was to build and evaluate prediction models based on a purely mathematical approach and compare these to prediction models with literature references.

1.1 Biomarkers

Part of this thesis is, as prediction models are evaluated, to examine the results in search of potential biomarkers. Biomarkers are "biometric measurements that convey information about the biological condition of the subject being tested. These measurements might be a quantitative readout of a specific analyte, sophisticated image studies, or measurement of multiple analytes combined into mathematical models" [11].

The definition of a biomarker in this thesis is a single or a group of genes. The reason for this is that even though the models built are purely math-

ematical, the results will be analysed slightly further through a biological point of view.

1.2 Alzheimer's Disease

Alzheimer's Disease (AD), the cause of a common and severe type of dementia, was first described by Alois Alzheimer in 1911 as a neuropsychiatric disorder affecting the elderly. Disease symptoms caused by neurodegeneration such as loss in memory, language, object recognition and learning function now affects more than 24 million people worldwide. Today the neuropathological features of AD are considered to be neurofibrillary tangles (NFTs) and amyloid plaques [12, 10, 2, 18].

The data analysed is the numerical outcome of microarray runs on a substance being evaluated at AstraZeneca. This substance inhibits Glycogen synthase kinase 3-β (GSK3β), a protein regarded as highly involved in the process of Alzheimer's Disease. One aim for this project was to find genes linked to GSK3β (and expectantly AD as well), that can be used as biomarkers.

1.3 Mechanisms and Causes of Alzheimer's Disease

1.3.1 NFTs and amyloid plaques

NFTs are aggregates of primarily hyperphosphorylated tau (microtubule associated protein tau - MAPT) in neurons. Hyperphosphorylation of MAPT leads to structural and conformational changes in the protein, which in turn allows the protein to self-aggregate and form a compact filamentous network. The function of MAPT - stabilising microtubules and bridging these polymers with other filaments - is impaired due to the hyperphosphorylation and thus affects the stability of the cytoskeletal network [12].

Amyloid plaques are aggregates of beta amyloid (Aβ), a protein derived from proteolysis of the amyloid precursor protein (APP). There are two variants of Aβ - Aβ_{1-40} and Aβ_{1-42} - where the latter is the most aggressive in producing amyloid plaques in the human brain. Furthermore, presenilin-1 (PSEN1) is involved in the normal APP processing and it is believed that mutations in this gene are responsible for the accumulation of Aβ_{1-42} in familial Alzheimer's Disease (FAD) [12].

1.3.2 GSK3β in Alzheimer's Disease

GSK3β has been linked to both NFTs and amyloid plaques and is thus a major candidate of investigation for the understanding of AD causes. GSK3β has been associated with paired helical filaments (PHF) which are lead components in NFTs and, as well as having interactions with amyloid, tau and

presenilin-1, GSK3β is involved in neuronal apoptosis, all features of AD [9, 2].

A partial aim of this project was to incorporate such information as presented above into mathematical prediction modelling. The approach of finding this information will though not be through article databases such as PubMed or OMIM (http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM), but through a knowledge-base named Ingenuity Pathways Analysis (more on IPA in Section 2.1). The information found here will be used to extract data from the dataset (see Section 2.4) in order to build mathematical prediction models based on literature references. How Ingenuity Pathways Analysis and these mathematical methods are used will be presented in the following sections.

2 Materials and Methods

2.1 Ingenuity Pathways Analysis

Ingenuity Pathways Analysis (IPA) is a commercial product based on literature findings and has now reached version 5.1. It can be seen as a database based on manual data-mining presented in an interactive user interface. It provides extensive information on biological networks and relationships between proteins, genes, complexes, cells, tissues, drugs, and diseases as well as some mathematical significance tests to analyse expression data.

To date, 485 publications have cited the use of IPA (http://www.ingenuity.com). Even though many of these cite the use of the mathematical methods used in IPA, the main idea behind using IPA in this project is not the mathematical methods it provides, but the information it can present on protein relationships. Foremost, the simplicity of how data can be extracted and incorporated into mathematical models from Ingenuity Pathways Analysis makes this thesis project possible to perform.

The functions used in IPA for this project are protein pathways, searches, and biomarker filters.

2.1.1 Biomarker filter in IPA

Certain proteins are found in specific fluids and tissues, are present in certain species, and related to various functions and diseases. IPA provides a function of filtering datasets for proteins by these criteria. A screenshot of the interface of this filter can be seen in Figure 1.

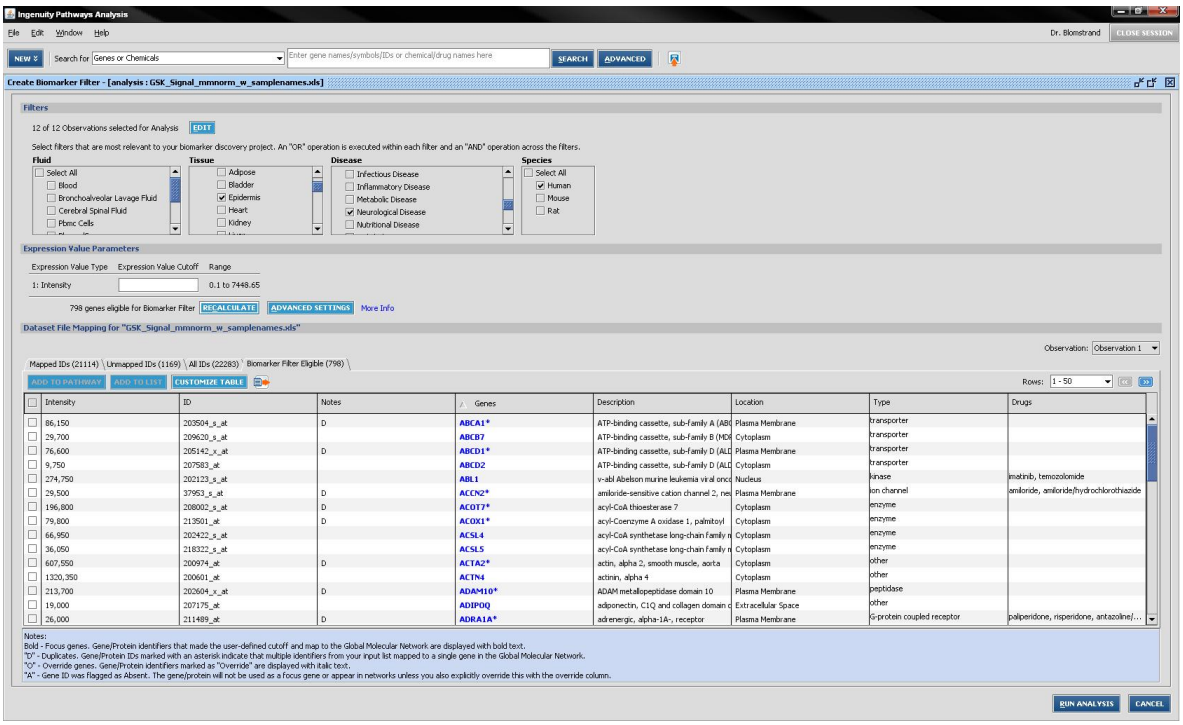

Figure 1: A screenshot of the Biomarker filter interface of IPA. Here several options are available to filter the uploaded datset for specific criteria. These criteria include if the proteins are located in specific fluids or tissues, involved in certain diseases and present in human, mouse and/or rat. The genes eligible for the set criteria show up in the lower part of the window.

2.1.2 Searches in IPA

Searches in IPA can be conducted by naming a protein, chemical or drug name. Genes can also be found by their association to diseases or functions, their type (enzyme, kinase, ion channel *etc.*) as well as their subcellular location.

2.1.3 Protein Pathways in IPA

The results of both the biomarker filter and the search queries can be added to illustrative pathways. Here proteins are presented as nodes and connected according to function. An example of a protein relationship pathway from IPA can be seen in Figure 2.

The information extracted from these functions in IPA will be incorporated into the mathematical prediction model described in the next section.

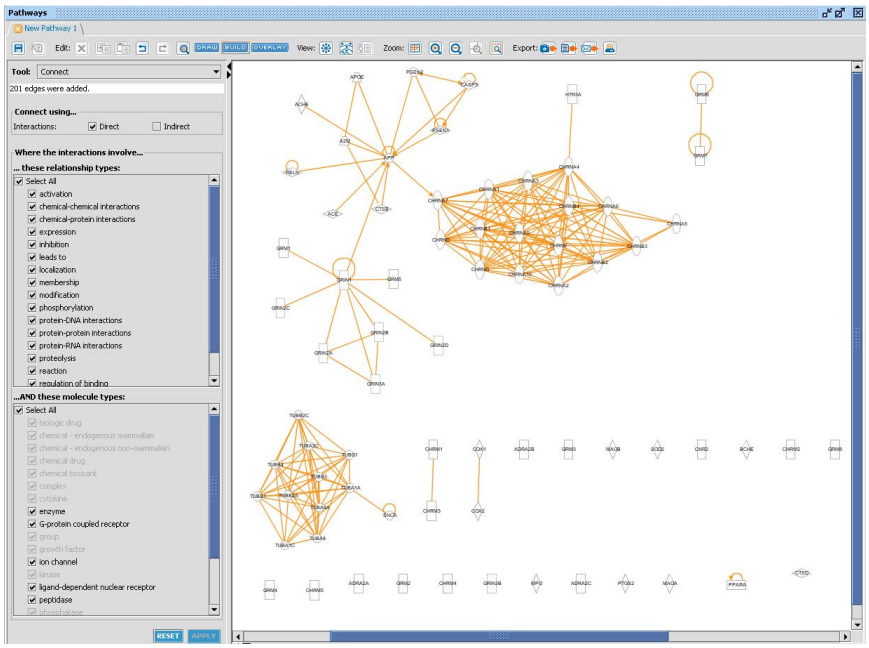

Figure 2: A biological pathway showing part of the result from a search for proteins related to Alzheimer's Disease in IPA. The most interesting part of this function in IPA is that the connections between proteins are easily visualised. This function of IPA will mostly be used to validate the resulting biomarkers to see what connections, if any, they have. Each connection is supported by at least one reference in literature (http://www.ingenuity.com).

2.2 PLS

Projections to Latent Structures by means of Partial Least Squares (anacronymed backronym PLS) is a commonly used type of prediction model. PLS is a regression model used to relate two data matrices to each other, $\mathbf{X}$ - the observed variables and $\mathbf{Y}$ - the response variables, by a linear multivariate model. Even though PLS can take several response variables (columns of Y) into account, the case where there is only one response variable will be

discussed here since this is the case for the datasets used.

The easiest and most intuitive way of presenting how PLS works is by geometry. If all columns of matrix $\mathbf{X}$ (size N-by-K) represent an axis in a K dimensional space, and each row (N) correspond to a point in this space, a line can be fitted to these using a partial least squares approach. It is important to note that in PLS, each row of $\mathbf{X}$ represents a point in the $\mathbf{Y}$-dimensional space (here size N-by-1) [6].

The first line fitted by partial least squares represents the first component of the PLS model. This line is in the direction that defines the maximum co-variance in the dataset. As one component (or Latent Variable (LV)) is calculated, the part of $\mathbf{X}$ described by this component is subtracted from the original dataset. As this procedure is repeated, more of the $\mathbf{X}$-dimensional space is taken into account by the model, leaving less and less information in the original $\mathbf{X}$ matrix.

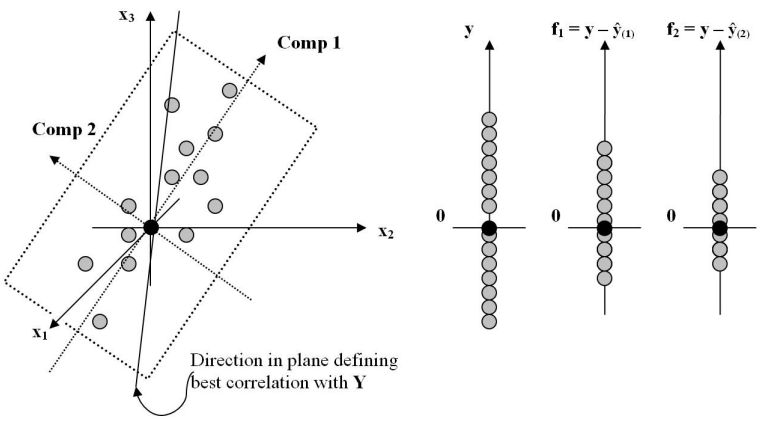

Figure 3: An illustration of a 3-dimensional variable space ($\mathbf{X}$) and a one-dimensional response space ($\mathbf{Y}$), as well as the residual of the response variable after calculating first two components. This illustration has been modified from Eriksson *et al.* [6].

As more components are used, the residual (the part that is not described by the model) of the $\mathbf{X}$ matrix is reduced, until it only contains noise. Since PLS produces a model where both $\mathbf{X}$ and $\mathbf{Y}$ are taken into account, the residual of $\mathbf{Y}$ also decreases with increasing number of components. In Figure 3 an illustration of a 3-dimensional variable space ($\mathbf{X}$) and a one-dimensional response space ($\mathbf{Y}$), as well as the residual of the response variable after calculating the first two components, can be seen.

The basic mathematics of PLS is to find the relationship between the matrices $\mathbf{X}$ and $\mathbf{Y}$ expressed as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}.$$

where $\mathbf{Y}$ is the response, $\mathbf{X}$ is the matrix containing the observed variables, $\mathbf{B}$ contains the regression coefficients, and $\mathbf{E}$ is the error.

Weighted combinations of the original $\mathbf{X}$-variables can be constructed as $t_a = \mathbf{X}w_a$, where t_a are called scores and w_a are called weights.
Similarly $u_a = \mathbf{Y}c_a$, are the weighted combinations of the $\mathbf{Y}$-variables.
These can then be re-written into

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{TC}' + \mathbf{F}$$

This method has two objectives: To approximate the $\mathbf{X}$ and $\mathbf{Y}$ spaces and to maximise the correlation between $\mathbf{X}$ and $\mathbf{Y}$ [5]. Basically, PLS models both $\mathbf{X}$ and $\mathbf{Y}$ and predicts unknown $\mathbf{Y}$ from new $\mathbf{X}$.

2.2.1 Scaling

Since variables are not originally uniform, they must be scaled before applying the PLS model. One of the most used methods is Auto-scaling. Auto-scaling mean-centers and variance scales the variables to mean value zero and relative variance one. This is done by $\mathbf{x}_k^{scaled} = (\mathbf{x}_k - \bar{\mathbf{x}}_k)/s_k$, where $\mathbf{x}_k$ is one column of $\mathbf{X}$, $\bar{\mathbf{x}}_k$ is the column mean, and s_k is the column standard deviation.

As well as using the auto-scaling, a method called Moving Median Normalisation was also used on the datasets before analysis. The main idea behind this method is that samples should be linearly correlated [4]. This normalisation was already done when the datasets were received and thus the basics of this method will not be discussed here.

2.2.2 Variable Influence on Projection and Variable Selection

Variable Influence on Projection (VIP) is a way of finding the variables that contribute most to the PLS model. By calculating a VIP-score (see Equation 1), the variables that are most relevant for explaining $\mathbf{Y}$ and $\mathbf{X}$ can be obtained (*i.e.* variable selection). This method enables reduction of the dataset since irrelevant variables can be removed without reducing the predictive ability of the model.

There are two main reasons for using variable selection; to improve the interpretation of the model [8] and to remove noise (and thereby increase the predictive power of the model). When variables with low VIP-score (and thus not contributing to the prediction) are removed, it is likely that the predictive power of the model increases. (The variable space needed

to be modelled is reduced, which makes it easier to model the "complete" [reduced] system.) This is a balancing act though, since when too many variables are removed, the model becomes overfit and loses its predictive ability (see Section 2.2.6 for more on overfitting).

$$VIP_k = \sqrt{\sum_{a=1}^{A} \left(W_{ak}^2 * SSY_a * \frac{K}{SSY_{tot}} \right)} \tag{1}$$

Where k is the variable number, K is the total number of variables, a is the PLS component number, W is the PLS weights, SSY is the explained variance (in %) and SSY_{tot} is the cumulative explained variance (in %) [6].

2.2.3 Cross Validation

Cross Validation (CV) is a way to validate the performance of a prediction model (see Section 2.2.6 for further validation analysis). CV is basically a method where the dataset is divided into two sets, one training and one test set. The prediction model is built on the training set and validated on the test set. In this way prediction performance estimates can be obtained for how the model would perform on unseen data.

The CV method is usually done by $1/N$ splits, *i.e.* the dataset is divided into N subsets where $N-1$ are used as a training set and the last one used as a test set. This process is repeated until all N sets have been used as test sets. Special cases such as Leave-One-Out (when N equals the number of samples in the dataset) can be used when the number of objects are few [6].

2.2.4 Prediction Performance

Prediction performance (PP) is a measure of how many samples the model classifies correctly. A sample can be classified in four different ways; True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

The model predicts a sample to be either positive or negative. By knowing the actual sample class, this prediction can then be evaluated to either True or False, depending on if the model prediction was correct (True) or if it was wrong (False). Prediction performance is calculated by Equation 2. Table 1 shows the relationship between the model predictions and the real classes.

$$\text{PP} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

Table 1: Relationship between model prediction and real classes.

		Model	
		T	F
Reality	T	TP	FN
	F	FP	TN

Closely connected to Prediction Performance is Sensitivity (Equation 3) and Specificity (Equation 4). These are measures of the portion of all Positive samples that are classified as Positive and the portion of all Negative samples are classified as Negative respectively.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

2.2.5 Cut-off

As **Y**-values are often discrete and the predicted values of a PLS model are always continuous, some distinction between whether a predicted value should be classified as Positive or Negative must be used. This is done using a cut-off that discriminates the continuous values of the PLS model.

The cut-off is placed somewhere between the real values of the samples. A prediction value of the PLS model is then treated as Negative if it is lower than the cut-off and Positive if it is higher.

2.2.6 Over-fitting

Overfitting is when the model explains **X** but has little or no predictive power of **Y** [17]. Not overfitting a model on training-data is one of the most important aspects when building a general prediction model. An overfitted model is basically a model with too many parameters (such as PLS components *etc*). A simple example of overfitting is when a polynomial function is fitted on linear data.

A way of measuring overfit is Q^2 (see Equation 5) and R^2, which are both illustrated in Figure 4. R^2 is the goodness of fit (how well the model

explains the training-data) and Q^2 is the goodness of prediction (how well the model explains test-data) [6].

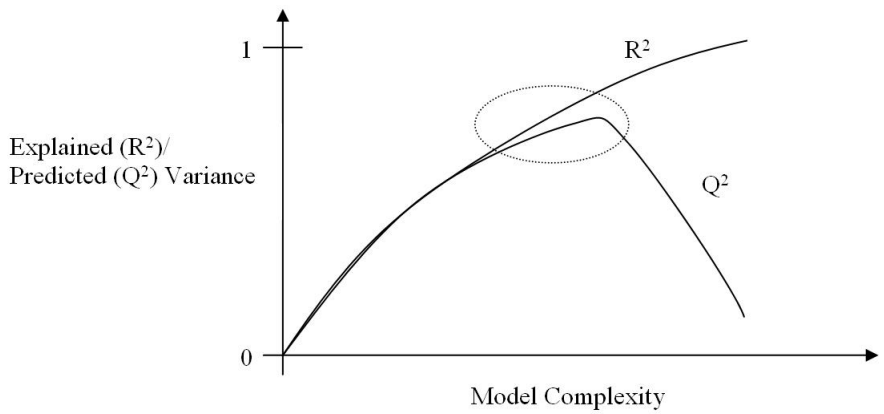

Figure 4: A plot showing R^2 and Q^2 as a function of model complexity (number of components, parameters et cetera). The model with optimal complexity is obtained within the dotted oval, where Q^2 starts to decrease. This illustration has been modified from Eriksson *et al.* [6].

Q^2 can be seen as a measure related to Prediction Performance but instead of generating actual values of how correct the PLS model is, it is a measure of how close to the real values the model predictions are. Q^2 can vary from negative values (no model prediction at all) to 1 (perfect model prediction), but values regarded as good are usually somewhere around $0.5-0.7$ [6].

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_{i,CV})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

where y_i is the real value, $y_{i,CV}$ is the predicted value and $\bar{y}$ is the mean over all y-values.

2.3 Protocol

In this section, a description of the steps taken to obtain the results presented further on, are presented. These include some steps that have already been described in former sections as well as steps that are described in latter sections of this project.

2.3.1 Scaling the dataset

The initial step involved in most data analyses is scaling the data so that objects can be compared. This was done, for both the purely mathematical and the literature reference approach, by using the moving median normalisation and consecutively the auto-scaling method (see Section 2.2.1).

2.3.2 Mathematical Approach

The PLS calculations applied to the scaled dataset were done in Matlab using the PLS toolbox (http://www.eigenvector.com/). These calculations were incorporated into a script that looped over each cross validation subset, a set of PLS components (1-4), and in each step applying the VIP score to reduce the dataset by 10% until it only included 250 variables.

The cross validation subsets were obtained by applying the "Leave-two-out" version described in Section 3.1.1. By applying this cross validation method, 10 samples were used to train the PLS model in each loop and two samples were used to test the model. This approach to the mathematical method resulted in $6 \times 4 \times 42 \approx 1000$ rounds of iterations (six from the number of cross validation subsets, 4 from the number of PLS components, and 42 from the number of times the dataset was reduced by 10%).

2.3.3 Literature Approach

The corresponding literature method to the VIP scores, that were used to mathematically reduce the number of variables, was performing searches and applying functions provided by IPA. The criteria for these searches and functions are described in Section 3.4.

By applying these methods a small set of protein names (100 to 800 names) could be obtained from IPA. Mapping these names to the Affymetrix probe names that were present in the dataset allowed for massive reduction in variables.

The PLS modelling applied was done using the same script as described above, for the mathematical approach, but with 1-6 PLS components and instead of reducing the dataset by 10% in each round, the number of variables were kept constant. This resulted in $6 \times 6 = 36$ iterations (six from the number of cross validation subsets, and 6 from the number of PLS components).

2.3.4 Obtaining Results

During the iterations, the PLS predictions were saved for forthcoming calculations. These calculations included finding the optimal Q^2 values (Section 2.2.6) and evaluating the Prediction Performance, Sensitivity and Specificity

(Section 2.2.4) at each cut-off (cut-off varied from 0.0 to 1.0 with an increment of 0.05 for the literature approach predictions and an increment of 0.01 for the mathematical approach predictions).

The numerical results were extraced from Matlab and visualised in Spotfire (http://spotfire.tibco.com/). Several of these graphs are presented further on in this project.

2.3.5 Program Versions used

Program	Version	Usage
Matlab	7.1	PLS modelling
IPA	5.1	Literature references and Protein Pathways
Spotfire	8.1	Visualising PLS results

2.4 Datasets

2.4.1 Affymetrix

A DNA microarray provides a simple way of analysing expressions from several thousand genes at once. The microarrays used to obtain the data for this thesis were based on Affymetrix GeneChip DNA microarrays (Human Genome U133A). These chips contain around 23000, features where each feature consists of a number (6-11) of probe cells and each probe cell contains an oligonucleotide probe of approximate length of 25bp.

In short, mRNA is extracted from a biological sample and converted to labeled complementary DNA (cDNA). This cDNA is applied to the microarray and allowed to hybridize with complementary probes. Signal intensities for each probe are thereafter obtained by confocal scanning, and determined to be "present", "absent" or "minimal". All probes classified as "absent" or "minimal" are removed from the dataset in order to only have reliable signal intensities (http://www.affymetrix.com).

2.4.2 Original dataset

The dataset to be studied was based on fibroblast samples from six individuals. These samples were divided into two groups, "treated" and "control" which were treated with substance + vehicle, and only vehicle, respectively (resulting in two samples from each patient - "treated" and "control"). This resulted in a dataset with 12 objects and 22215 reliable variables ("genes").

It might seem a little odd to use fibroblast cells when AD is brain related but the reason for doing this is that the GSK3β protein is also present in other tissues. Therefore, measuring the effects of a GSK3β-inhibitor can just as well be done in many other tissues than brain. Also, it would provide a much simpler way of measuring response if the samples could be taken from skin tissue instead of brain tissue.

3 Analysis & Results

3.1 Method Modifications

In order to build reliable prediction models on the dataset, some modifications to the methods described in Section 2 had to be applied.

3.1.1 "Leave-two-out"-CV

For two reasons, the small sample size of the dataset, and the fact that the variation between patients was greater than the variation between "treated" and "control" of the same patient, a special type of CV had to be applied to this dataset. This method was called "Leave-two-out" (LTO).

The LTO method builds the PLS model on N-2 samples, and predicts the remaining two. The essential part of this method is that both samples that are to be predicted are from the same patient. By doing this, the PLS-model is allowed to concentrate on explaining only *treatment variations* instead of also having to explain patient variations.

3.1.2 Mathematical Variable Selection

The approach used in this paper was to reduce the number of variables by 10% in each round (*i.e.* 10% of the variables from the previous dataset were removed and the process of creating a PLS prediction model on the now reduced dataset, evaluating it on the same unseen objects, and calculating new VIP-scores, were repeated). In this way, fewer variables are removed when the number of variables decrease, until some number of variables is reached. The final set of variables are the ones that are mathematically most important for the description of $\mathbf{X}$ and the prediction of $\mathbf{Y}$, and thereby the ones that can be further studied as possible biomarkers.

3.2 Differences in the Mathematical Approach Compared to the Literature Reference Approach

As the mathematical method was based on VIP-scores for variable selection, the literature reference approach was based purely on literature findings in IPA. Apart from this, the prediction modelling was done in the same way for both approaches. In order to get an overview of this, a flowchart is presented in Figure 5 showing the basic steps in the two approaches.

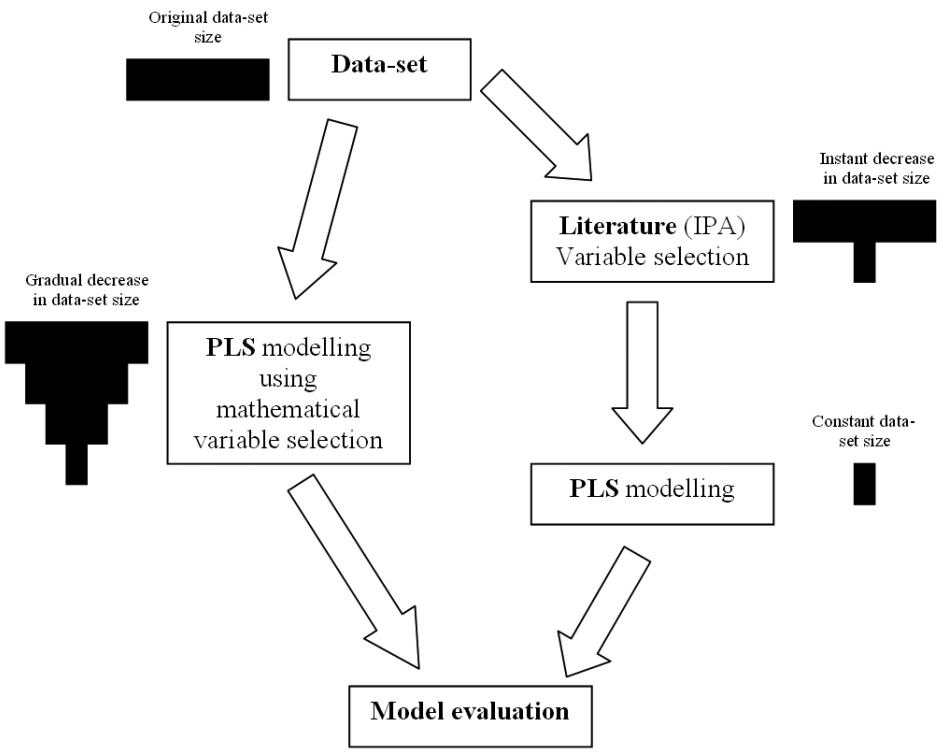

Figure 5: Flow chart showing the steps involved in creating and evaluating the basic PLS model using mathematical variable selection, compared to the literature references approach, as well as some basic information on dataset size changes is each step.

3.3 VIP Variable Selection applied to the Original dataset

Variable selection was applied according to Section 2.2.2 for 1-4 PLS components on the original datas-set. The resulting Q^2-values can be seen in Figure 6.

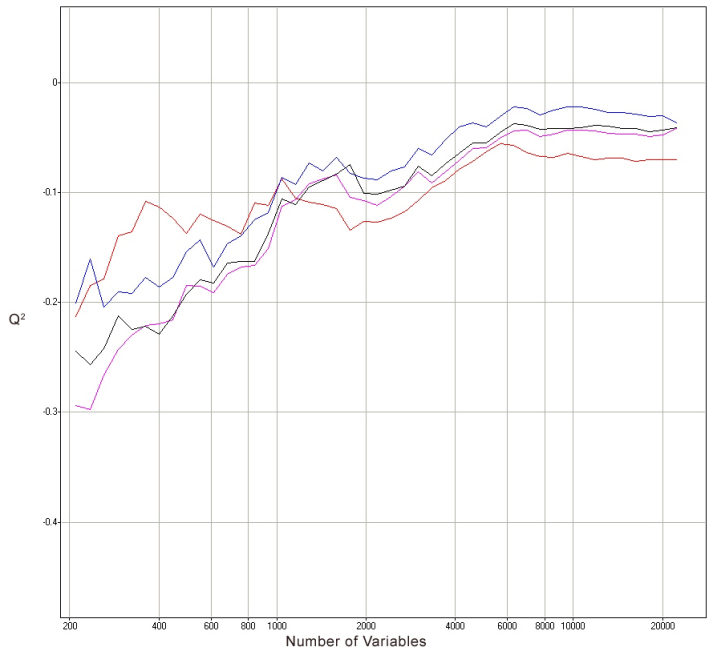

Figure 6: Q^2 vs Variables on original dataset; Coloring by PLS components: Red = 1, Blue = 2, Purple = 3, Black = 4. The Q^2-values obtained for this dataset are negative over all variables and thus do not have any predictive power at all.

3.4 IPA Variable Selection applied to the Original dataset

As described in Section 2.1, IPA provides ways of performing literature variable selection (as compared to the mathematical VIP-score) through searches and biomarker filters. By doing this, variables of proven linkage to the treatment are included in the model, and variables with no linkage are removed. Two ways of performing variable selection through IPA were conducted - biomarker filter and searches.

3.4.1 IPA Biomarker Filter dataset

The criteria for the applied biomarker filter:

Tissue	:	Epidermis (fibroblast)
Species	:	Human
Related Disease	:	Neurological

These criteria resulted in 797 selected variables. A plot showing Sensitivity, Specificity and Prediction Performance *vs* Cut-off on this data can be seen in figure 7.

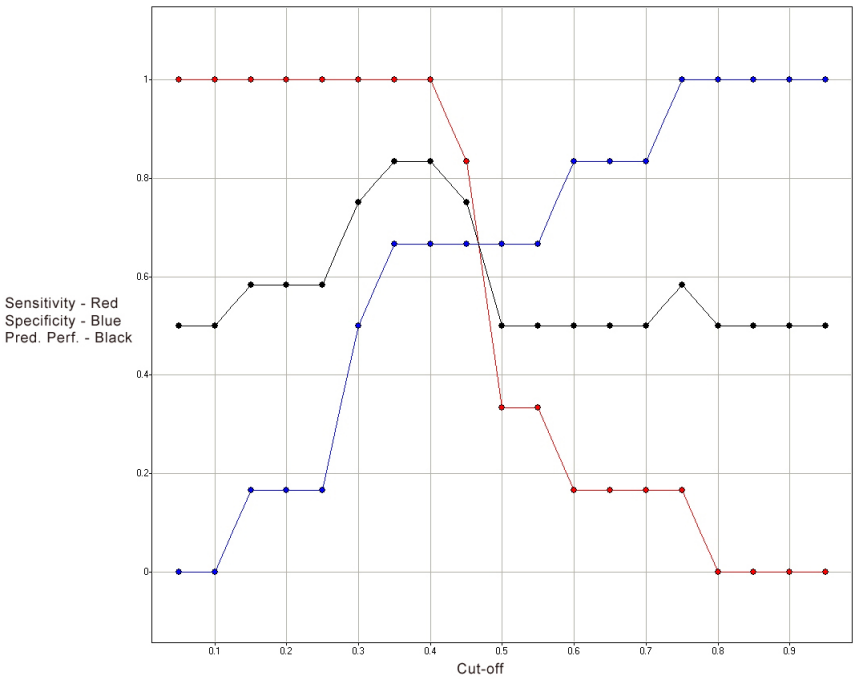

Figure 7: Sensitivity, Specificity and Prediction Performance vs. Cut-off for the biomarker dataset at 1 LV and 797 variables. The highest Prediction Performance (0.83) is obtained at cut-off 0.4.

3.4.2 Searches dataset

Two searches were performed - for genes related to AD, and genes downstream of GSK3β. These resulted in 113 and 138 variables, respectively, with a grand total of 251 selected variables (no overlap).

LTO-CV runs using the resulting four datasets (one from the biomarker filter and three from the searches) were performed and Q^2 values were calculated (see Figure 8).

Numerical values for each dataset can be seen in Table 2.

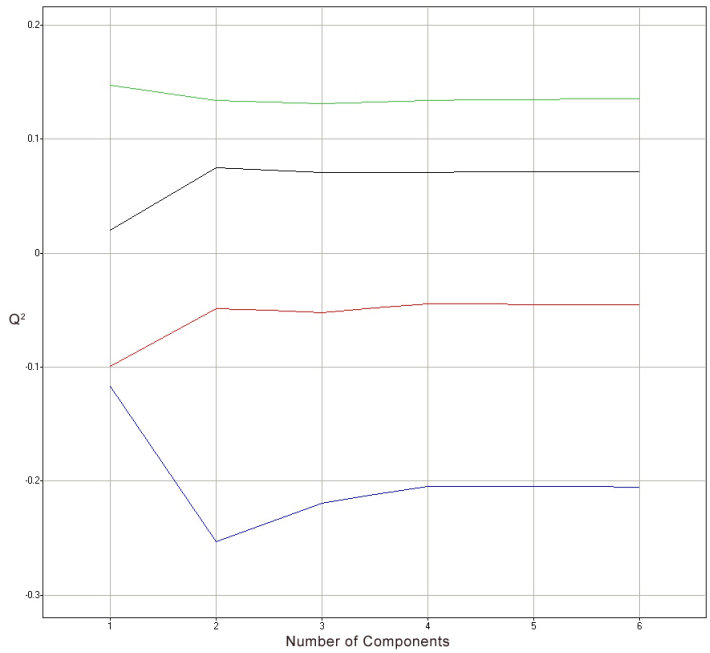

Figure 8: Q^2 *vs* PLS Components on IPA variable selection datasets; Coloring by datasets: Blue = Alzheimer's Disease (113 variables), Black = GSK3β + Alzheimer's Disease (251 variables), Red = GSK3β (138 variables), Green = Biomarker filter (797 variables). The highest Q^2-value (0.15) is obtained for the biomarker dataset at one PLS component.

Table 2: Related values for the highest Q^2-values for each dataset.
DS = Down Stream, i.e. all proteins GSK3β affects, located in both brain and skin cells.
AD = Alzheimer's Disease, located only in brain.

Dataset	PLS Components	Variables	Q^2	Pred. Perf. @ Cut-off
Original	2	9559	-0.02	0.75 @ 0.37
Biomarker Filter	1	797	0.15	0.83 @ 0.4
GSK3β DS	3	138	-0.05	0.75 @ 0.5
AD	1	113	-0.12	0.58 @ 0.5
GSK3β DS + AD	2	251	0.07	0.75 @ 0.55

None of these Q^2-values are close to what is regarded as a good model (Q^2 around 0.5-0.7), but the biomarker filter approach can be regarded as having at least some predictive power.
It is no coincidence that the AD dataset received negative Q^2-values (no predictive power at all), since these genes are not even present in skin cells.

3.5 Validation

3.5.1 Validation by Randomisation

From the analysis (Section 3), it can be seen that the prediction values varied among the datasets. In order to validate that these values were not mere coincidence, randomisation runs were conducted.

These runs included choosing a number of randomised variables from the original dataset, as well as randomising the response values ("treated" [0] and "control" [1]). Since the best prediction values were obtained from the biomarker dataset, this was the only set that was validated.

3.5.1.1 Randomising Response Since the sample size was small, the number of possible permutations of response values was limited. This gave the possibility to actually use all different permutations of the response space.

Since the two samples from every patient ("treated" or "control") need to be different, only two combinations exist per patient ($[1 \quad 0]$ or $[0 \quad 1]$). Thus the total number of response permutations is $2^6 = 64$. PLS runs using the biomarker dataset (797 variables), VIP-selection with 10% removed in each round and 1-6 PLS components were conducted (resulted in roughly 10000 rounds). The results can be seen in Figure 9.

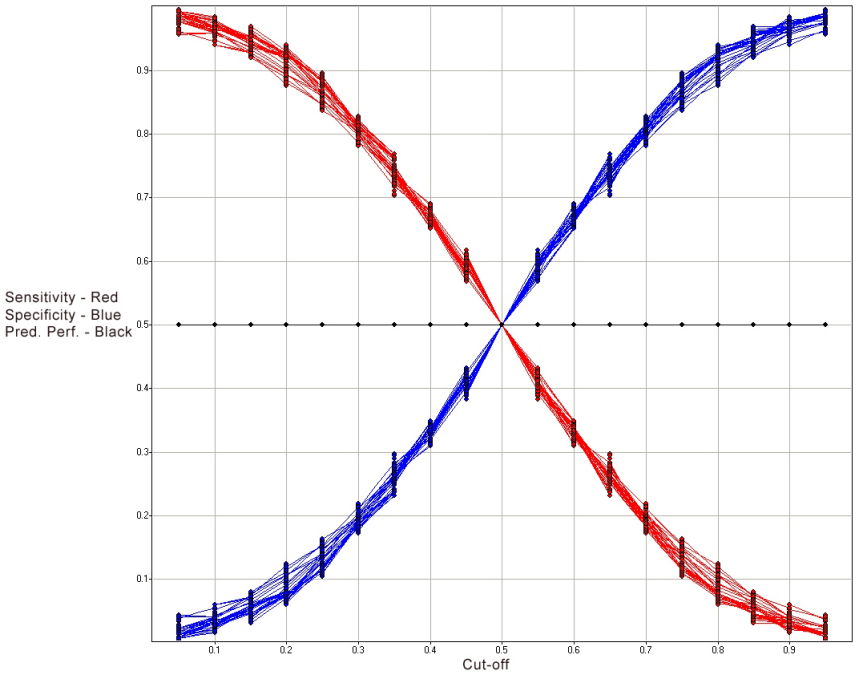

Figure 9: Randomised response, all 64 permutations. 1-6 components, 47-797 variables. This figure shows all combinations of the stated parameters and, as can be seen, the prediction performance is constantly 50% showing that randomising the response gives the same prediction performance as pure guessing would.

3.5.1.2 Randomising Variables To exclude the possibility that any set of 797 variables would give as good prediction values as the biomarker dataset, runs using random variables needed to be done.

From the original dataset, 500 rounds of selecting 797 random variables for one PLS component (the number of components that gave the highest Q^2-value in Section 3) were conducted. The results are shown in Figure 10.

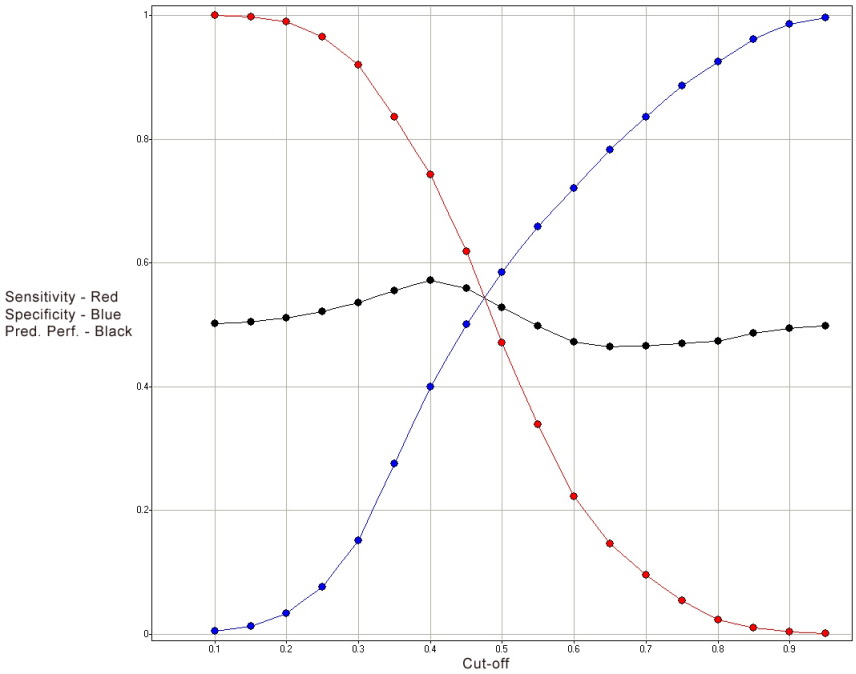

Figure 10: Prediction performance, Sensitivity, Specificity for 797 randomized variables from the original dataset (mean over 500 loops). The rise in prediction performance at 0.4 shows that "treated" samples still have resemblance to "control" samples making it difficult for the model to classify correctly. It should be noted that this cut-off is the same as was optimal for the biomarker filter approach, showing that the literature approach was able to identify underlying data that the mathematical approach could not.

3.5.1.3 Randomisation Conclusion These validations show that the PLS model built on the biomarker filter dataset was able to explain the underlying data as well as indicating that the variables constituting this dataset can not be extracted randomly. Assuming that these interpretations are correct, the next step is to make a biological interpretation of the variables most important for the prediction model.

3.6 Biological Interpretation

3.6.1 Genes with high VIP

From the model built on the biomarker filter from IPA, all genes were ranked according to the VIP-score. Out of the 797 genes, the 100 highest were uploaded to IPA for further analysis of their connection to Alzheimer's Disease genes and GSK3β.

A search was conducted in IPA for all genes related to Alzheimer's. This resulted in 79 genes. Along with these 79 genes and GSK3β the 100 genes from the VIP-scoring table was added to a pathway and connected by IPA. Any genes that were not connected were removed. The final pathway can be seen in Figure 11.

The most interesting genes are the ones connected to GSK3β, namely **VDAC1, CSNK1ϵ, CDKN1A** and **NF-κBIA** (NF-κBIA was not found by IPA, but some investigation (Susanne Fabre, AstraZeneca, personal communication) shows that it is in fact connected to GSK3β [9].

According to IPA, these proteins were not related to AD, thus searches were conducted in PubMed (http://www.ncbi.nlm.nih.gov/sites/entrez) for articles regarding these proteins to investigate if there was any connection to AD. These findings are shown below.

- VDAC1
  The voltage-dependent anion-selective channel proteins (VDACs), are found in the mitochondrial membranes of all eukaryotes. According to a study by Yoo *et al.* there is a decrease of VDAC1 in AD brain. This may lead to decreased synaptic loss and also be linked to apoptosis in cortex regions, two issues both involved in AD [20].

- CSNK1ϵ
  Expression increase of Casein kinase 1 ϵ (an isoform of CK1) has been described in human AD brain. The reason for this is probably that it leads to an increase in Aβ production [7].

- CDKN1A
  Expression increase of Cyclin-dependent kinase inhibitor 1A (a.k.a. p21/WAF1) has been described in AD fibroblast samples [14].

- NF-κB
  Several studies [9, 16, 13] show that NF-κB is a critical component of neuronal function. A study by Paris *et al.* shows that NF-κB inhibitors decrease both Aβ_{1-40} and Aβ_{1-42} production [16].

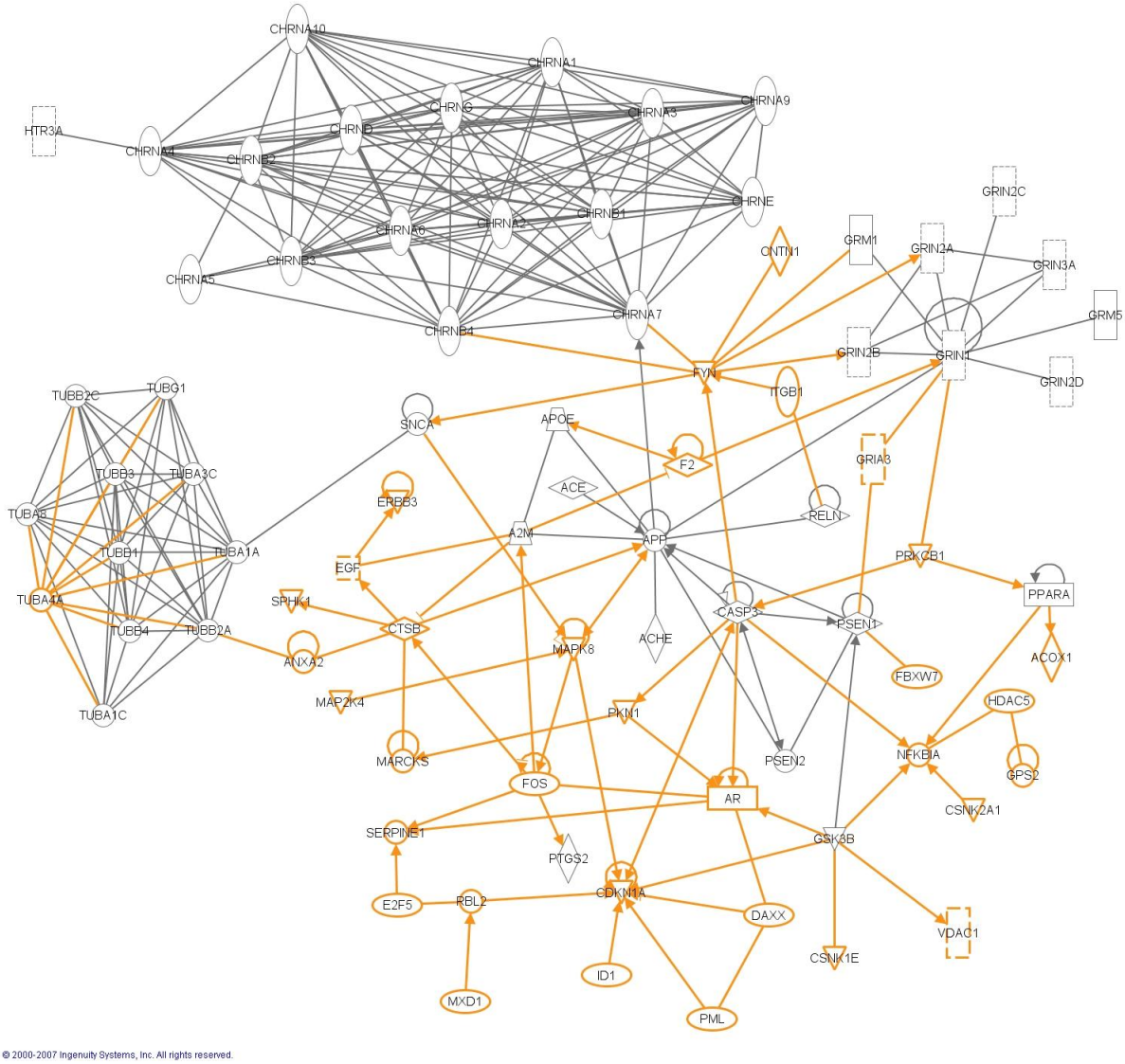

Figure 11: Alzheimer's related genes according to IPA (in black), high VIP-scoring genes (in orange).

Even though the exact functions of these proteins and their relationship to AD are not covered here, the fact that links between them and AD could easily be found in PubMed shows that there is at least some predictability in the mathematical method with literature references, and these proteins may indeed be potential biomarkers for GSK3β inhibitor drugs.

4 Discussion

4.1 Substitute Prediction Performance Measurements

Even though Prediction Performance is widely used for evaluating mathematical prediction models, other measures such as Positive Predicted Value (PPV, Equation 6) and Negative Predicted Value (NPV, Equation 7) also exist. These measures are more related to individual patients than the sample space as a whole [1].

$$PPV = \frac{Sens. \times Prev.}{Sens. \times Prev. + (1 - Spec.) \times (1 - Prev.)} \tag{6}$$

$$NPV = \frac{Spec. \times (1 - Prev.)}{(1 - Sens.) \times Prev. + Spec. \times (1 - Prev.)} \tag{7}$$

For instance if the Sensitivity (0.83) and Specificity (0.67) at cut-off 0.4 from Figure 7 were to be used, then $PPV = 0.72$ and $NPV = 0.80$ (Prevalence is 0.5 since the whole sample space is made up of 6 "treated" and 6 "control"). These values can be read as if one sample was classified as positive, the chance of it being "treated" would be 71% and if one sample was classified as negative, the chance of it being "control" would be 80%. If these measurements are more informative or suited for this specific thesis project I cannot say, but at least they convey another aspect of interpretation.

4.2 Validating Ingenuity Pathways Analysis

Even though the mathematical methods may be fairly easy to validate, what is most important is the system from which the information of the variable selection originated. To validate all relationships extracted from IPA is not within the scope of this thesis, but still, some kind of discussion around this is required.

According to IPA, all edges in a pathway diagram are supported by at least one literature reference (http://www.ingenuity.com). Even though these references have all been published in more or less well-respected journals, can you actually trust relationships that are only based on one single article?

In Figure 11 several of the relationships presented are based on a single reference. And, even though all genes in this figure are present in humans, some relationships have only been studied in mice. In this figure there are also two clusters of genes all pairwise connected: TUBxx & CHRNxx. These clusters are, according to sources at AstraZeneca, partially artefacts (Hugh Salter, personal communication).

In Figure 12, mammalian proteins phosphorylated by GSK3β are shown [9]. Of these relationships, 16 out of 30 are not found in IPA, a sign that significantly more information can be added to the knowledgebase.

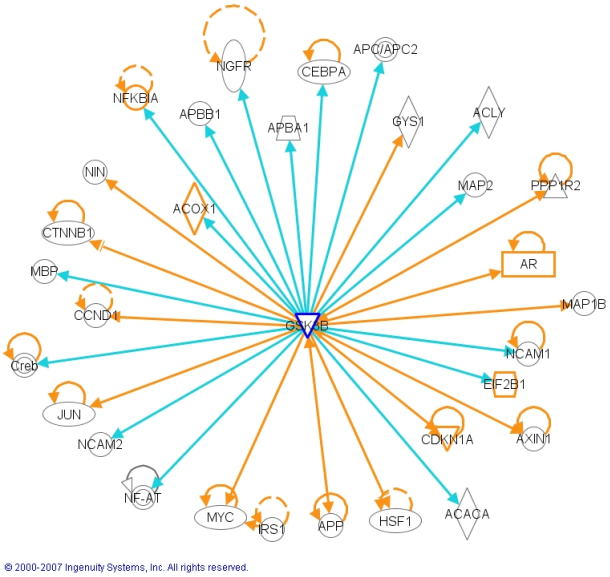

Figure 12: Genes phosphorylated by GSK3β. Orange connections are found through IPA, Turquoise connections from literature (http://www.ingenuity.com), [9] .

Another aspect of literature is that it is always interpreted by the reader. In this way relationships that do not exist may be found, and relationships that do exist may be neglected.

The entire prediction model relies on the literature reference being well investigated and trustworthy, and even though there are flaws in IPA the results presented in this thesis show that it is still a reliable reference, and as research continues, it will probably only become more reliable.

5 Conclusion

The use of literature as a substitute or complement to mathematical methods may increase prediction performance of PLS models. Taking literature in to account means that genes without association to the treatment may be omitted from the PLS models, thus the amount of variables decrease and the results may be easier to interpret. Evidently, this method relies on that literature being used has good coverage so that no variables are missed. If it not so, variables that are in fact related to the treatment and thus important for the prediction, may be left out. Another problem that arises when omitting genes with no previously found relation to the treatment is that no *new* biomarkers can be found, since they are already discarded.

These issues are clearly shown when looking at the resulting Q^2-values (Table 2). Here the AD dataset obtained the most negative Q^2-value - due to the simple fact that these genes cannot be affected by a substance applied to skin-cells since they are only present in the brain, whereas the biomarker dataset that only included genes from skin-cells obtained the highest Q^2 and PP value (0.15 and 0.83, respectively). Comparing these values to the values of the mathematical approach (Q^2 at -0.02) show that the predictive power of a mathematical model rooted on literature references can be both beneficial and destructive; the vital part in using a literature reference is that it should hold enough information about the subject being investigated so that nothing is neglected.

6 Acknowledgements

I would like to thank my supervisors at AstraZeneca - Dr. Hugh Salter, *Associate Director*, and Kerstin Nilsson, *Senior Research Scientist* - for helping me throughout this Master thesis, as well as my co-worker Sara Grey, *Research Scientist*, with whom I discussed many aspects of the project.

References

[1] D. G. Altman and J. M. Bland. Statistics notes: Diagnostic tests 2: predictive values. *Brittish Medical Journal*, 309:102, 1994.

[2] R. V. Bhat and S. L. Budd. Gsk3β signalling: Casting a wide net in alzheimer's disease. *Neurosignals*, 11:251–261, 2002.

[3] AL. Boulesteix and K. Strimmer. Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8:32–44, 2007.

[4] W. S. Cleveland. Robust locally weighted regression and smoothing scatter plots. *Journal of The American Statistical Association*, 74:829–836, 1979.

[5] L. Eriksson, E. Johansson, N. Kettaneh-Wold, C. Wikström, and S. Wold. *Design of Experiments - Principles and Applications*. Umetrics Academy, 2000.

[6] L. Eriksson, E. Johansson, N. Kettaneh-Wold, and S. Wold. *Multi- and Megavariate Data Analysis*. Umetrics Academy, 2001.

[7] M. Flajolet, G. He, M. Heiman, A. Li, A. C. Nairn, and P. Greengard. Regulation of alzheimer's disease amyloid-β formation by casein kinase i. *Neuroscience*, 104:4159–4164, 2007.

[8] E. Freyhult, P. Prusis, M. Lapinsh, J. ES. Wikberg, V. Moulton, and M. G. Gustafsson. Unbiased descriptor and parameter selection confirms the potential of proteochemometric modelling. *BMC Bioinformatics*, 6:50, 2005.

[9] C. A. Grimes and R. S. Jope. The multifaceted roles of glycogen synthase kinase 3β in cellular signaling. *Progress in Neurobiology*, 65:391–426, 2001.

[10] R. S. Jope and G. V. W. Johnson. The glamour and gloom of glycogen synthase kinase-3. *Trends in Biochemical Sciences*, 29:95–102, 2004.

[11] J. LaBaer. So, you want to look for biomarkers (introduction to the special biomarkers issue). *Journal of Proteome Research*, 4:1053–1059, 2005.

[12] R. B. Maccioni, J. P. Muñoz, and L. Barbeito. The molecular bases of alzheimer's disease and other neurodegenerative disorders. *Archives of Medical Research*, 32:367–381, 2001.

[13] M. P. Mattson and S. Camandola. Nf-κb in neuronal plasticity and neurodegenerative disorders. *The Journal of Clinical Investigation*, 107:247–254, 2001.

[14] J. Naderi, C. Lopez, and S. Pandey. Chronically increased oxidative stress in fibroblasts from alzheimer's disease patients causes early senescence and renders resistance to apoptosis by oxidative stress. *Mechanisms of Ageing and Development*, 127:25–35, 2006.

[15] H. Pang, A. Lin, M. Holford, BE. Enerson, B. Lu, MP. Lawton, E. Floyd, and H. Zhao. Pathway analysis using random forests classification and regression. *Bioinformatics*, 22:2028–2036, 2006.

[16] D. Paris, N. Patel, A. Quadros, M. Linan, P. Bakshi, G. Ait-Ghezala, and M. Mullan. Inhibition of aβ production by nf-κb inhibitors. *Neuroscience Letters*, 415:11–16, 2007.

[17] S. Wold, M. Sjöström, and L. Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58:109–130, 2001.

[18] G. Wolf-Klein, R. Pekmezaris, L. Chin, and J. Weiner. Conceptualizing alzheimer's disease as a terminal medical illness. *American Journal of Hospice & Palliative Medicine*, 24:77–82, 2007.

[19] ZR. Yang and R. Hamer. Bio-basis function neural networks in protein data mining. *Current Pharmaceutical Design*, 13:1403–1413, 2007.

[20] B. C. Yoo, M. Fountoulakis, N. Cairns, and G. Lubec. Changes of voltage-dependent anion-selective channels proteins vdac1 and vdac2 brain levels in patients with alzheimer's disease and down syndrome. *Electrophoresis*, 22:172–179, 2001.