Data mining of gene expression changes in Alzheimer brain

Walker, P. Roy; Smith, Brandon; Liu, Qing Yang; Famili, A. Fazel; Valdés, Julio; Liu, Ziying; Lach, Boleslaw

Publisher's version / Version de l'éditeur:
https://doi.org/10.1016/j.artmed.2004.01.008
Artificial Intelligence in Medicine, 31, 2, pp. 137-154, 2004-06-01

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

*Data Mining of Gene Expression Changes in Alzheimer Brain**

Walker, P.R., Smith, B., Liu, Q.Y., Famili, F., Valdes, J., Liu, Z., and Lach, B.
June 2003

* published in Publication: Artificial Intelligence in Medicine, Elsevier Science. NRC 45838.

Canada

Data Mining of Gene Expression Changes in Alzheimer Brain

P. Roy Walker [a], Brandon Smith [a], Qing Yan Liu [a],

[a] *NeuroGenomics Group, Institute for Biological Sciences*
National Research Council of Canada
1200 Montreal Rd. Ottawa, ON. K1A 0R6 Canada

A. Fazel Famili [b], Julio J. Valdés [b], Ziying Liu [b] and

[b] *Integrated Reasoning Group, Institute for Information Technology*
National Research Council of Canada
1200 Montreal Rd. Ottawa, ON. K1A 0R6 Canada

Boleslaw Lach [c]

[c] *Department of Laboratory Medicine, The Ottawa Hospital Civic Campus*
1053 Carling Avenue. Ottawa, ON. K1Y 4E9 Canada

Abstract

Genome wide transcription profiling is a powerful technique for studying the enormous complexity of cellular states. Moreover, when applied to disease tissue it may reveal quantitative and qualitative alterations in gene expression that give information on the context or underlying basis for the disease and may provide a new diagnostic approach. However, the data obtained from high-density microarrays is highly complex and poses considerable challenges in data mining. The data requires care in both pre-processing and the application of data mining techniques.

This paper addresses the problem of dealing with microarray data that come from two known classes (Alzheimer and normal). We have applied three separate techniques to discover genes associated with Alzheimer disease (AD). The 67 genes identified in this study included a total of 17 genes that are already known to be associated with Alzheimers or other neurological diseases. This is higher than any of the previously published Alzheimer's studies. Twenty known genes, not previously associated with the disease, have been identified as well as 30 uncharacterized

Expressed Sequence Tags (ESTs). Given the success in identifying genes already associated with AD, we can have some confidence in the involvement of the latter genes and ESTs.

From these studies we can attempt to define therapeutic strategies that would prevent the loss of specific components of neuronal function in susceptible patients or be in a position to stimulate the replacement of lost cellular function in damaged neurons.

Although our study is based on a relatively small number of patients (4 AD and 5 normal), we think our approach sets the stage for a major step in using gene expression data for disease modelling (i.e. classification and diagnosis). It can also contribute to the future of gene function identification, pathology, toxicogenomics, and pharmacogenomics.

Key words: Data Mining, Genomics, Gene identifications, Gene expression, Alzheimer's disease and Microarray
PACS:

1 Introduction

Alzheimer's disease (AD) is an incurable, chronic, progressive, debilitating condition which, along with other neurodegenerative diseases, represents the largest area of unmet need in modern medicine. Progress in understanding these diseases is hampered by their complexity, but there is now renewed hope that genomics technologies, particularly gene expression profiling, can have an impact. Genome-wide expression profiling of thousands of genes provides rich datasets that can be mined to extract information on the genes that best characterize the disease state.

Gene expression profiling using microarrays is a complex task subject to many variables that can obliterate the subtle differences that exist between the normal and diseased states [13] and [33]. The best results are usually obtained when numerous samples are available and are all analyzed, in replicate, at the

⋆ The authors would like to acknowledge the contributions of all members of the BioMine project, Alan Barton, Youlian Pan, Junjun Ouyang from IIT, Melanie Lehman, Marianna Sikorska and Marilena Ribecco from IBS and a number of former students who worked in this project.

Email addresses: roy.walker@nrc.gc.ca (P. Roy Walker), brandon.smith@nrc.gc.ca (Brandon Smith), qing_yan.liu@nrc.gc.ca (Qing Yan Liu), fazel.famili@nrc.gc.ca (A. Fazel Famili), julio.valdes@nrc.gc.ca (Julio J. Valdés), ziying.liu@nrc.gc.ca (Ziying Liu), boleklach@hotmail.com (Boleslaw Lach).

same time using the same lots of RNA (riboucleic acid) extraction and hybridization reagents and the same lot of microarrays. However, clinical samples from AD and normal, aged matched individuals are usually acquired in small numbers over a prolonged period of time and analyzed at different times. To date, only a few microarray studies relevant to Alzheimer's disease have been published [5], [7], [12], [21], [25] and [24]. All of these studies used small numbers of samples ranging from 1 to 6 AD patients. Significantly, there is little, and in some cases no, overlap in the genes identified between these studies and genes already known to be associated with or differentially expressed in Alzheimer's disease are seldom picked up. It is critical, therefore, to develop data processing and data mining strategies that can account for discrepancies in the data that are due to experimental variability from the true differences between the normal and disease states in small sample sets (see also [18] and [32]).

In this paper we use a data mining strategy that can derive useful information from a small number of samples acquired and analysed at different times. In section 2 we explain the problem of gene expression profiling and in section 3 we provide details on our data preparation process. Section 4 discusses why understanding gene expression and discovering useful genes is important. Section 5 explains our data mining process and in section 6 we provide the results of this research. We conclude the paper in section 7.

2 The problem of gene expression profiling and data reduction

Gene expression analysis using high-density cDNA (complementary deoxyribonucleic acid) arrays presents a number of problems, both in terms of execution of the experiments and analysis of the data [10] and [40]. The primary concern is the quality of the microarrays compromised by variability in both the quality of the spotted features and imperfections in the substrate that lead to variability in the level of background. This variability can exist even within a batch of microarrays printed at the same time and there can be even more substantial variability from batch to batch. The second source of variability is in the hybridization reaction. Most microarray experiments are conducted as a competitive hybridization between a control sample (i.e. normal tissue) labelled with one fluorochrome (usually Cy3) and a test sample (i.e. disease tissue) labelled with a second fluorochrome (usually Cy5). Fluorochrome labelling is carried out using enzymes that have bias both in terms of incorporation efficiency of the different fluorochromes and in the efficiency with which they can transcribe any given sequence. The hybridization reaction, itself, is also biased in the degree to which any given probe sequence can hybridize to its cognate target and in the overall stringency of the reaction (i.e. the degree to which non-specific hybridization occurs). All of these

sources of variability contribute to the generation of artifacts, some of which are remarkably reproducible.

To derive useful information from such experiments, samples must be analyzed in replicate and the data appropriately normalized and filtered to reject poor quality spots. Although improving the overall quality of the data, such pre-processing will not remove any "global" systemic biases such as those introduced by performing experiments at different institutions or at considerably different times. Methods must also be developed to assess and remove such biased data.

3 Experimental Procedures and Data preparation

This section includes details of our experiments and data preparation process. We briefly explain where our samples are obtained and how we performed our hybridization process. We also elaborate on the microarray data acquisition and data preprocessing steps that we have taken.

3.1 Patient samples and RNA extraction

A total of 4 clinically diagnosed AD patients and 5 normal patients of similar age were used in this study. Post mortem intervals ranged from $4-8$ hours. Previous studies [37] demonstrated that there is little loss of RNA integrity during this timeframe and the RNA quality is suitable for use in microarray analysis experiments. However, we cannot rule out the loss of some short half-life transcripts. All of the work was performed under a protocol approved by the National Research Council of Canada Human Ethics committee. Detailed patient information is listed in Figure 3. Total RNA was extracted from the frontal cortex of samples of post mortem brain using the Tri Reagent (MRC Inc. Cincinnati, OH). There was no obvious RNA degradation in these samples as judged by agarose gel electrophoresis. The purity of total RNA was assessed by optical density ratios, A260/A280, which ranged from 1.8-2.0. The extractions were performed according to the manufacturer's instructions for tissue samples. Poly $(A)^+$ messenger RNA (mRNA) was then obtained from the total RNA samples using the Oligotex kit (Qiagen, Mississauga, ON Canada).

3.2 cDNA microarray hybridization

Fluorescently-labelled AD and normal patient cDNAs were hybridized to the Human 19K microarray slides obtained from The Microarray Center of the

University Health Network, Toronto, Canada (http://www.microarrays.ca/). The two slide set contains 19,200 characterized and unknown human ESTs together with a number of control features. The slides are designated as Slide A and Slide B and each slide has 9,600 ESTs spotted in duplicate, organized into 32 sub-arrays of 600 spots each. The hybridizations were carried out using a common control sample created by pooling equal amounts of RNA obtained from the 4 normal patients. This sample was labeled with Cy3. The samples obtained from the 4 AD patients and the 5 normal patients (labelled with Cy5) were arrayed individually against this pooled normal control in a competitive hybridization reaction. At least three replicate hybridizations were performed for each sample. Fluorescence-labelled first strand cDNA probes were generated from $1\mu g$ of mRNA in a $40\mu l$ reaction mix, containing 1 X first strand buffer, 150 pmole AncT (5' T(20) VN 3') primer, 20 mM each of dATP, dGTP and dTTP, 2mM dCTP, 1 mM Cyanine 3-dCTP (Cy3, control samples) or Cyanine 5-dCTP (Cy5, individual normal and AD samples) and 0.4M DTT. The reaction mixture was first heated to $65°C$ for 5 min and then cooled to $42°C$ for another 5 min to denature the RNA and anneal the AncT primer. Reverse transcription was accomplished by adding $2\mu l$ of Superscript II reverse transcriptase (Invitrogen Life Technologies, Burlington, ON) and $1\mu l$ of RNase inhibitor (Promega, Madison, WI) and incubation at $42°C$ for 2-3 h. The reaction was stopped by adding 5 ml of 50 mM EDTA and the RNA templates were hydrolyzed by adding $2\mu l$ of 10 N NaOH to the cDNA reaction, followed by an incubation at $65°C$ for 20 min. The reaction was then neutralized by adding $4\mu l$ of 5 M acetic acid. Before hybridization, the Cy3 and Cy5 probes were combined and precipitated with an equal volume of isopropanol. The pellet was washed with 70% ethanol and air-dried in the dark. The labelled cDNA probe mix was then resuspened in $5\mu l$ water and combined with $80\mu l$ of DIG Easy Hyb buffer (Boehringer Manheim, Germany) containing $0.5\mu g/ml$ yeast tRNA and salmon sperm DNA. This hybridization solution was heated to $65°C$ for 2 min, cooled to room temperature and injected between a paired set of the human 19K microarray slides. Hybridization was carried out in the dark at $37°C$ for 18 h. After hybridization, the slides were washed three times in 1x SSC containing 10% SDS for 10 min, plus a final wash with 1x SSC alone. The slides were dried by centrifugation at 40xg in a Sigma 4K 15 centrifuge for 5 min.

3.3 Microarray data acquisition

The slides were scanned using a ScanArray 5000 confocal scanner (Packard BioScience, Meriden CT, USA) with excitation/emission wavelengths of 543 nm / 570 nm for Cy3 and 633 nm / 670 nm for Cy5, at 10 μm resolution. The resulting 16-bit grayscale image files, one for each channel, were quantitated together with QuantArray v3.0 (Packard BioScience) using an adaptive spot

finding method to generate spot intensities from mean pixel values and local area background measurements were derived from a background mask (doughnut) surrounding the spot. Poor quality spots were flagged manually by the user and recorded in the output file to be used as an "ignore spot" filter. The tab delimited text data files produced were subsequently pre-processed using macros in Microsoft Excel 2000 (Microsoft Corporation, Redmond, WA).

3.4 Microarray data pre-processing

For each sample replicate there was a data file for each of the slides, A and B. Each data file contained intensity data for 19200 features (i.e. 9,600 ESTs in duplicate) measured in two channels, Ch1 and Ch2. Ch1 data represents intensity measurements from the pooled control sample and Ch2 data represents intensity measurements from the sample prepared from each individual patient who was either clinically diagnosed with AD or an age matched normal. Median subarray background values were calculated for each channel and subtracted from the respective intensity values. Spots flagged by the user during quantization (the "ignore" filter) and spots failing to meet the following criteria; intensity $>$ 2.5-fold background and intensity $> 5^{th}$ and $< 98^{th}$-percentile of all intensities for each channel, were filtered out and not used in the computation of normalization correction factors. The corrected intensity data were logged (base 2) and corrected for dye bias (normalized) using a linear-regression correction applied to the Ch2 intensities for all the spots in each subarray. This correction yielded a Ch2 versus Ch1 scatter plot with a linear regression best-fit line having slope 1 and intercept 0. Log2 ratios representing expression values for sample versus pooled control were then calculated by subtracting the Log2 Ch1 intensity from the corrected Log2 Ch2 Intensity values for each spot. Data for control spots were removed and finally the spot duplicates were averaged. The resulting data set contained 9600 Slide A and 9381 Slide B relative expression values in triplicate (or more) for the 4 AD patients and 5 age-matched controls. For this study only the data from slide A were used for gene discovery. Figure 1 shows intensity scatter (a) and M vs. A (b) plots of data from a typical microarray from a hybridization performed in this study. The raw data is shown in yellow and the processed data in blue. Normalization and filtering using the criteria described removes anomalous data and improves the fit. This is demonstrated more dramatically in the pseudo-array images of raw and processed data (Figure 2). The bias caused by preferential dye incorporation is removed and the data improved considerably.

The next task was to identify the total number of missing values in the data. The distribution was not even across the samples, ranging from 7.8-37.7%. Overall, 15.68% of all the data was missing. Using BioMiner (our data mining

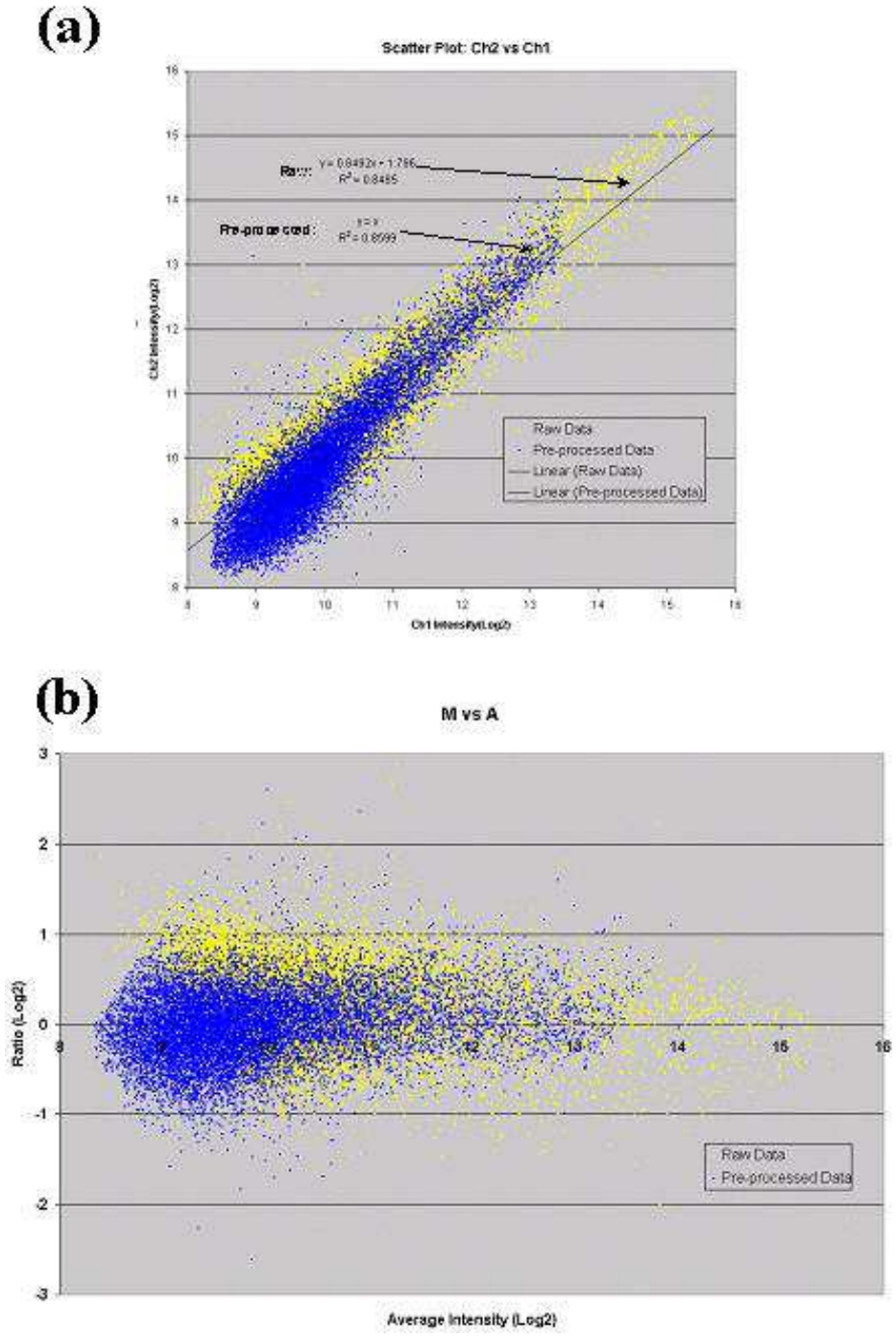

Fig. 1. The Intensity Scatter (a) and M vs. A (b) plots above show the effect of pre-processing the raw intensity data. Raw data is shown in yellow and pre-processed data in blue. (a) shows that the background correction, linear regression, normalization and filtering results in a best-fit line with $y = x$, and a slight improvement in the correlation (R^2). The M vs. A plot (b) is used to more clearly show differential expression (Log_2 Ratio) and can reveal intensity dependent bias in the data.

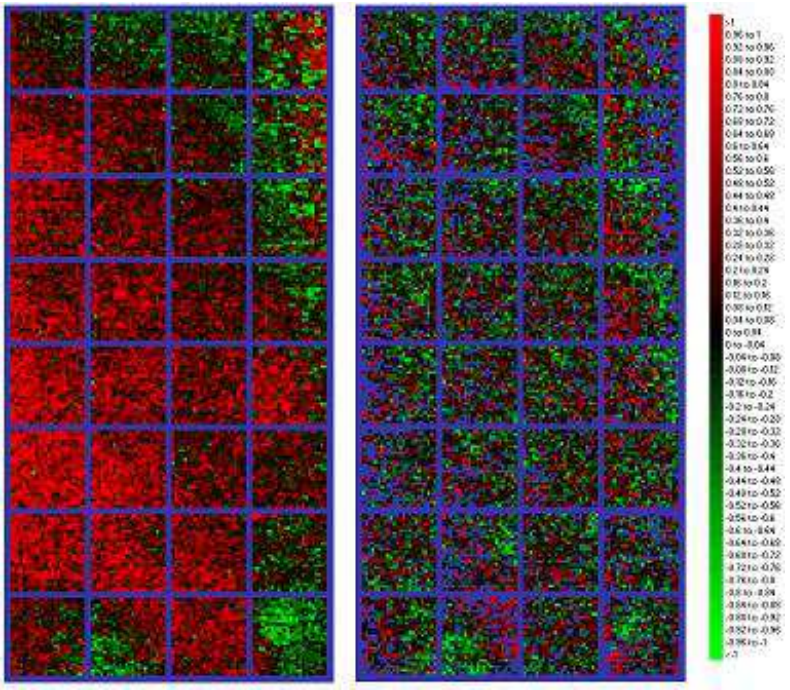

Fig. 2. The PseudoArray is used to plot raw or pre-processed data back onto the original microarray layout. Spatial biases, such as edge effects, background problems, poorly hybridized areasand localised dye bias can be observed using this representation. The plots above show raw (left) and pre-processed (right) Log_2 ratio data. The scale shows the colours used to represent the ratio values, green for down- and red for up-regulation. Blue spots represent filtered data.

software introduced in Section 5) to compute statistics and generate virtual reality representations of the ratio data from each hybridization, a total of 6 samples that were poor representatives of the whole data set were identified. Based on these criteria, the 6 microarrays were rejected and the composition of the final dataset is shown in Figure 3. This data set, called F0, was used for subsequent gene discovery experiments.

4 Understanding gene expression data and discovering useful genes

Gene expression data mining involves studies that combine the use of domain knowledge with data obtained from two or more classes (e.g. disease and normal) to discover genes that are associated with a particular problem. Since our data consists of two classes, those with Alzheimer's Disease and those without, the investigation is focused on using inductive and statistical techniques to identify the most informative genes amongst all the genes in the data sets. This is done through searching for patterns and relations that exist in the data. We used the following approaches to perform this search:

Patient	Sex	Postmortem	Pathology	Sample_ID	Rejected	Replicates
AD 109-96	male	4-8 hours	Dementia, senile changes of AD type moderate	AD1.1		4
				AD1.2		
				AD1.3		
				AD1.4		
AD 112-96	male	6 hours	Subdural hematomas, senile changes of AD type	AD2.1		3
				AD2.2		
				AD2.3		
				AD2.4	X	
AD 227-94	male	4-8 hours	Dementia, senile changes of AD type moderate	AD3.1		3
				AD3.2		
				AD3.3		
AD 90-96	female	4-8 hours	AD, dementia	AD4.1		2
				AD4.2		
				AD4.3	X	
Normal 102-97	male	6 hours	Normal	N1.1		2
				N1.2		
				N1.3	X	
Normal 154-94	female	4-8 hours	Normal	N2.1		2
				N2.2		
				N2.3	X	
Normal 211-95	male	4-8 hours	Normal	N3.1		2
				N3.2		
				N3.3	X	
Normal 67-97	male	4.5 hours	Normal	N4.1		3
				N4.2		
				N4.3		
Normal 88-96	female	4-8 hours	Normal	N5.1		2
				N5.2		
				N5.3	X	

Fig. 3. Table containing the number of patients and replicates used to construct the dataset.

- **Pattern recognition**: (see Section 5.1 for details) We arbitrarily repeated each data mining experiment 20 times to identify the most informative genes. The gene(s) identified in each experiment were then removed from the data set and the experiment was repeated until 20 experiments were completed. By doing this, we forced the algorithm to only focus on the available genes in each run. Knowing that in each run, the gene(s) with the highest information value are identified and reported, we therefore forced the algorithm to discover all important genes that could be identified in 20 runs. This approach has been used in a previous study related to gene identifications using leukemia data and has generated interesting results [11].
- **Individual dichotomization**: This is a search technique performed within the scope of the information associated individually with each gene (i.e. the intensity or the ratio values). The goal is to find the best level which partitions the expression values into two sets which are maximally related (in a probabilistic sense), with two previously defined groups (in the present case,

the Alzheimer and the Normal classes). Section 5.2 describes the method in detail.

- **P-value and ratio thresholding**: In section 5.3, a combination of two-tailed one-sample and 2-sample t-tests assuming unequal variance and ratio thresholding was applied to the gene expression data to identify gene expression changes that are both statistically significant and biologically relevant.
- **Virtual Reality for visualizing Databases**: Relational database tables usually contain information about large collections of objects described in terms of many properties (attributes, fields). In general these attributes are composed of numeric and non-numeric information (real-valued, qualitative information, etc.), and often many of them are missing. This technique constructs a visual representation of the heterogeneous and multidimensional space of the original database objects, in the form of a virtual reality space trying to preserve as much structure of the database as possible. The result is a virtual reality environment where one can navigate and visually inspect the main features of the data. Section 5.4 describes the technique in more detail.

5 Gene discovery

This section includes our data mining process where we introduce our research in discovering patterns in our data and identifying genes associated with Alzheimer. We also introduce our data mining software used in these experiments.

5.1 Pattern recognition

Figure 4 shows the structure of the data sets each consisting of a matrix containing p genes for n samples and an attribute vector containing labels for all samples. In each data set p=9600 and n=23. The overall goal of the research reported here was to identify from all the genes: (i) the most informative genes that are correlated with classification of AD vs normal, and (ii) models (set of rule(s)) consisting of one or more genes that contain a particular threshold to be used for accurate discrimination of AD samples from others.

5.1.1 The BioMiner software

The BioMiner data mining software was used for the data mining experiments reported in this paper. This software has been designed and built in house to provide support for biologists and bioinformaticians performing data mining

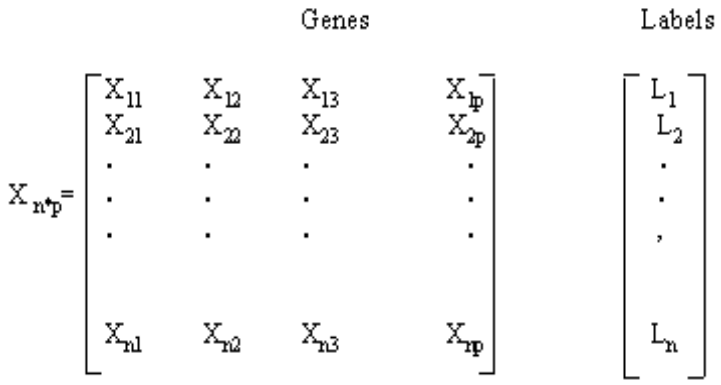

Fig. 4. Structure of gene expression data (Labels = Patient classes (i.e. AD=1 or Control=2)).

research in functional genomics. One of the key advantages of the software is that all available forms of data pre-processing and analysis functionality are integrated into one environment. The pre-processing and data analysis modules consist of a collection of algorithms and tools to support data mining research activities in an interactive and iterative manner. Figure 5 shows the interface for the BioMiner software.

5.1.2 Overview of data mining experiments performed using BioMiner

The following experiments were performed on the data:

- A number of data preprocessing experiments to examine the data. These helped us to obtain a general view of the data. It was also possible to identify missing values, abnormal or interesting characteristics, such as anomalies in the data.
- Hierarchical clustering to identify the inherent class characteristics of the data.
- A set of machine learning experiments, using inductive algorithms to identify the most informative genes with an associated threshold for classification along with a measure of strength.
- Use of data visualization techniques to better investigate the overall patterns of identified genes.

5.1.3 Data selection for pattern recognition using BioMiner

In addition to the pre-filtered dataset, designated F0, which was generated as described in Section 3.4, a total of 3 additional data subsets were generated and used as input to the BioMiner Pattern Recognition module. These data subsets, called F1, F2 and F3, employ statistical tests designed to re-

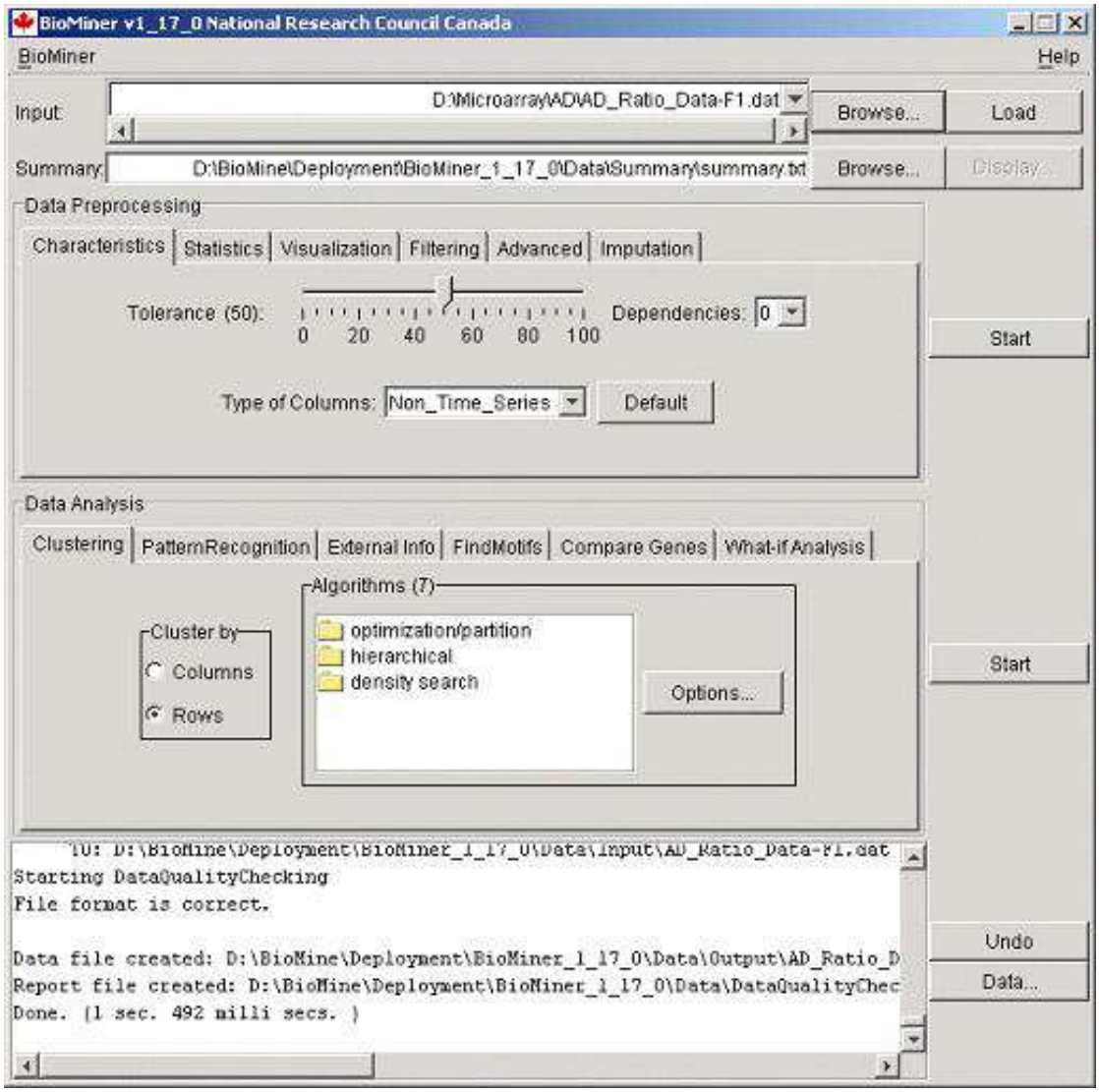

Fig. 5. The BioMiner interface.

move from consideration, genes with expression profiles that are biologically meaningless (See Figure 6).

Datasets F1 and F2 were filtered using single-sample t-test p-value thresholds combined with expression ratio cut-offs on the normal patient samples only. Two different p-value/ratio threshold combinations were used: Filter F1 employed thresholds of $\alpha < 0.05$ for the single sample t-test and an absolute value of the average Log2 ratio > 0.6. The settings for Filter F2 were $\alpha < 0.01$ for the single sample t-test and an absolute value of the average Log2 ratio > 0.5. These filters reject data that show apparent differential expression where none is expected, i.e., genes with mean expression across normal samples significantly different from zero. Lowering the α value for the t-test decreases the number of genes filtered out, while lowering the threshold for the absolute value of the Log2 ratio increases the number of genes filtered out. Datasets F1 and F2 filtered 222 and 220 genes, respectively, from the F0 dataset. Of

Data set	Criteria for selection	No. of genes filtered	No. of genes
F0	Pre-filtered dataset	57	9543
F1	1 sample t-test of N vs. 0, $\alpha < .05$ AND $\lvert \bar{N} \rvert > 0.6$	279	9321
F2	1 sample t-test of N vs. 0, $\alpha < .01$ AND $\lvert \bar{N} \rvert > 0.5$	277	9323
F3	$\lvert \bar{AD} \rvert < 0.5$	9519	81

Fig. 6. Table containing the data for data mining experiments.

these, 106 were filtered by F1, 104 by F2 and 116 by both. Dataset F3 was a much more aggressive filter applied to the AD samples only. All genes with an absolute value of the average Log2 ratio < 0.5 were filtered out, removing 9519 genes from the 9600 gene dataset. This filter selects for genes with an average 1.4-fold change in gene expression or greater in AD samples.

5.2 Individual Dichotomization

The transformation of numeric attributes into discrete values (discretization) is a very useful technique, especially in relation to supervised classification and the use of induction-based methods of machine learning. Moreover, the characterization of the different attributes in terms of discrete categories simplifies the interpretation of the data and especially the relationships between single or combined attributes and the class structure. Here a simple screening algorithm was used with the purpose of finding individual relevant genes from the point of view of their ability to differentiate the class of samples having Alzheimer's disease from the normal ones. The inputs for the algorithm are: a) the values of a given attribute A (gene) for all the studied objects (in this case, the ratio between channel-2 and channel-1 for all samples, b) the classes C_1, C_2 associated with each sample (Alzheimer vs Normal), and c) a probability threshold p_T . The algorithm then proceeds as follows:

(1) construct the set of distinct values of A (call it $\triangle$). That is, if O is the set of objects and $A(o)$ is the value of the attribute for any object $o \in O$, $\triangle = \{\delta_1, \ldots, \delta_p\}$ with the following properties: $(\forall \delta_i, \delta_j \in \triangle, \delta_i \neq \delta_j)$, $(\forall o \in O, \exists \delta \in \triangle \text{ s.t. } A(o) = \delta)$ and $(\forall \delta \in \triangle, \exists o \in O \text{ s.t. } A(o) = \delta)$.
(2) sort $\triangle$ in increasing order.

(3) construct the set $\hat{\triangle}$ composed by the mean of all consecutive values of $\triangle$. That is, for every pair δ_i, δ_{i+1} , compute $\hat{\delta} = (\delta_i + \delta_{i+1})/2$. Clearly, $\hat{\triangle}$ has one element less than $\triangle$.
(4) use each $\hat{\delta} \in \hat{\triangle}$ as the binary threshold for the values of attribute A . This divides the set of objects into two disjointed classes A_1, A_2.
(5) compute the contingency table of A_1, A_2 vs C_1, C_2.
(6) on the table, compute the conditional probabilities $p_1 = P(C_1/A_1)$, $p_2 = P(C_1/A_2)$ and retain $p_{max} = max(p_1, p_2)$.
(7) if $p_{max} \geq p_T$ select the attribute as relevant, and discard it otherwise.

The process is repeated for all attributes describing the objects, and the resulting set of selected attributes gives an indication on how many of them (genes in this case) contain a differentiation power equal or better than the pre-set probability threshold p_T. Specifically, if $p_T = 1$ the algorithm will give a set of genes such that each of them will perfectly differentiate the corresponding classes (Alzheimer/Normal).

5.3 P-value and ratio thresholding

A standard method for determining whether two means are significantly different is the t-test. While this test, when conducted over a number of samples, can provide confidence in a mean, its application to gene expression data it may yield expression changes which are too small to be biologically meaningful. It has been suggested that simply observing a difference in the mean can serve as a proxy for more rigorous statistical approaches if the difference is large [20], however using a large difference in means as a threshold for differential gene expression excludes many small, but significant changes. The F0 dataset was filtered using a combination of p-value and ratio thresholding according to the following criteria:

(1) Data points were rejected if the p-value from a two-tailed single-sample t-test of the Log_2 Ratio of the AD samples versus 0 was greater than 0.01.
(2) Data points were rejected if the p-value from a two-tailed two-sample t-test with unequal variance of the Log_2 Ratios of the AD samples versus the normal samples was greater than 0.01.
(3) Data points were rejected if the absolute value of the mean of the Log_2 Ratios of the AD samples was less than 0.41 (1.33-fold change).

5.4 Use of Virtual Reality

This is a technique for visual data mining of heterogeneous databases based on virtual reality (http://www.hybridstrategies.com), [38] and [39]. The purpose is to facilitate the process of understanding the underlying structure of single or compound databases of a general kind. The method is based on parameterized mappings between the heterogeneous space $\hat{H}$ representing the original data and the virtual reality space. They can also be constructed for unions of information systems (e.g. heterogeneous and incomplete data sets together with knowledge bases composed by decision rules), simplifying the process of discovery of interesting patterns as well as relationships between the original data and the symbolic expressions representing the structured knowledge.

A virtual reality space Ω is composed of different sets and functions in the following way: $\Omega =< \underline{O}, G, B, \Re^m\, g_0, l, g_r, b, r >$, where $\underline{O}$ is a relational structure (a set of objects and attributes, endowed with a set relations Γ^ν defined over the objects), G is a non-empty set of geometries representing the different objects and their relationship in the visual space (an empty or invisible geometry is a possibility), B is a non-empty set of behaviors (i.e. ways in which the objects from the virtual world will express themselves: movement, response to stimulus, etc.), $\Re^m$ is a metric space of dimension m ($\Re \subseteq \mathbb{R}$, the reals), which will be the actual virtual reality geometric space (usually $m = 3$). The rest of the elements are mappings: $g_0 : O \to G$, $l : O \to \Re^m$, $g_r : \Gamma^\nu \to G$, and r is a collection of characteristic functions for Γ^ν.

The representation of an extended information system (i.e. database) $\hat{S}$ implies the construction of another one $\hat{S}^\nu$ in the virtual world. It requires the specification of several sets and a collection of extra mappings. There are many ways in which it can be done. A desideratum for the virtual reality heterogeneous space $\hat{H}^\nu$ is to keep as many properties from $\hat{S}$ as possible, in particular, the similarity structure of the original data. In this sense, the idea is to maximize some metric/non-metric structure preservation criteria as in multidimensional scaling [4] and [16], or minimize some error measure of information loss. If δ_{ij} is a dissimilarity measure between any two objects i, j , and ξ_{ij} is another dissimilarity measure defined on objects i^ν, j^ν in the virtual reality space (the images of the original objects) , an error measure frequently used is the *Sammon error* $= \frac{1}{\sum_{i<j} \delta_{ij}} \frac{\sum_{i<j} (\delta_{ij} - \xi_{ij})^2}{\delta_{ij}}$ [30]. The transformation l is implicit, as no functional representations are found.

The possibilities derived from this approach are practically unlimited, since the number of different similarity, dissimilarity and distance functions definable for the different kinds of source sets is immense. Moreover, similarities and distances can be transformed into dissimilarities according to a wide variety of schemes. This provides a rich framework where appropriate measures

capable of detecting interrelationships hidden in the data can be found, more suited to both its internal structure and to external criteria. The virtual reality representation of heterogeneous data sets is a technique available within the BioMiner software.

6 Results

This section includes results from all of our experiments. We start with our results with hierarchical clustering in which we investigated the separation between the two classes. We then continue with other results from the use of virtual reality and pattern recognition techniques to search for interesting genes among mean genes available in these experiments.

6.1 Hierarchical clustering

Since class attributes were known for each sample, we used the agglomerative hierarchical clustering algorithm with Euclidean as a distance measure and single linkage (Ward method) as options on the entire ratio data (F0) to identify the hierarchical partitioning properties among all cases. Figure 7 shows the detailed dendogram representation of the results, which illustrates how perfectly the neighboring samples are grouped together. From these results, the ratio data has only one sample (N5.1, indicated by a blue arrow) that has been incorrectly included in the first group. The red line in this figure indicates the natural separation between the two classes.

6.2 Virtual reality representation

Virtual reality representations of different data sets were constructed in order to obtain an idea of the structure of the data. For obvious reasons, it is impossible to illustrate appropriately the look, feel and immersion of a virtual reality 3D environment within the limits imposed by printed paper. Screen snapshots from different application examples are presented only to give a rough idea. The design of the virtual reality spaces was kept simple in terms of the geometries used (spheres for representing the objects and colors for representing the classes), and in particular, behaviors were excluded (objects in the virtual world are inanimate). In all cases the snapshots were simplified with regard to the information included in the corresponding virtual world to avoid information overload.

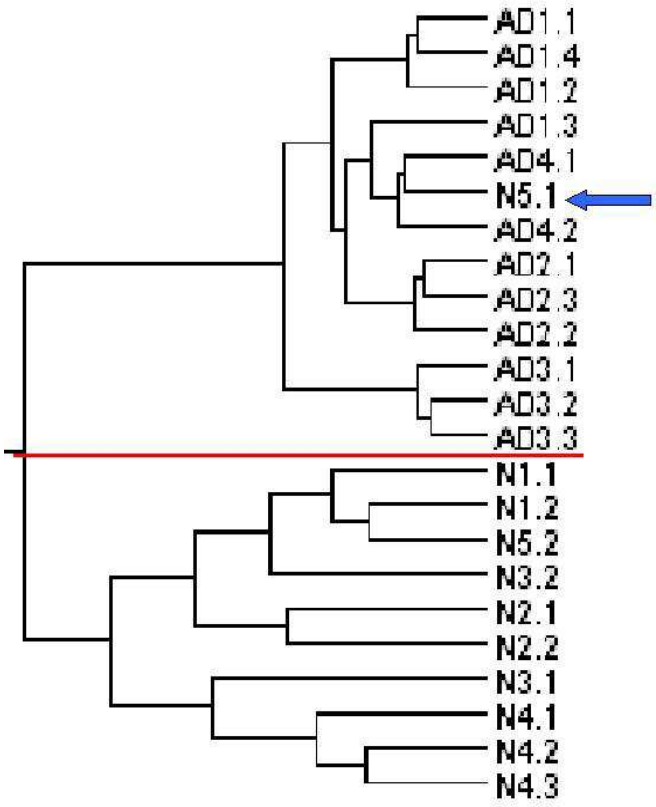

Fig. 7. Results of hierarchical clustering of Ratio data. The red line indicates the separation of two classes. The blue arrow indicates a misclassified sample.

In the case of F0 ratio data, each object (sample) is described by a collection of 9600 attributes (genes), therefore, each one is a vector from a 9600 dimensional space, also containing missing values. The virtual reality space (VR) is a 3D Euclidean space with the images of these objects and with non-missing values. In the first experiment, the entire F0 set was used in computing the virtual reality space. After 335 iterations, the absolute error obtained was 0.1034 (with an absolute difference = 9.9e-07). This error level indicates that the considerable non-linear dimensionality reduction which took place when going from 9600 to 3 attributes, satisfactorily retained most of the similarity structure present in the original data. The presence of a meaningful structure in the VR space characterized by a small information loss with respect to the original data (see below), clearly suggests that there is a subset of informative or relevant genes within the sample. A snapshot of the virtual worlds is shown in Figure 8. The samples corresponding to the Alzheimer's class are colored red, and the normal ones green. In this case it is clearly seen that the samples corresponding to the Alzheimer's class appears as more homogeneous and compact (i.e. more similar to each other) than those from the normal class. Moreover, the Alzheimer class appears "wrapped" by the normal class, which is more irregular.

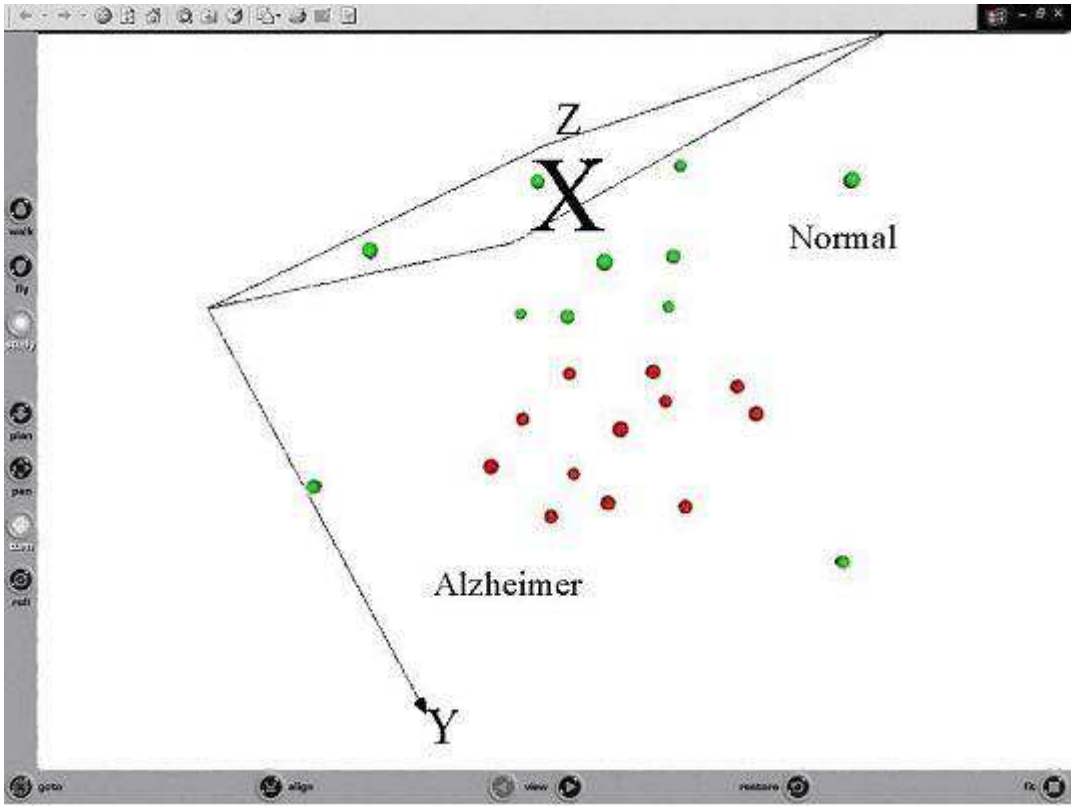

Fig. 8. VR space of the ratio data. Green: Alzheimer; Red: Normal.

6.3 Searching for genes and patterns using BioMiner software

With this understanding of the data, acquired from preprocessing, and knowing that the data was labelled, we then initiated our search for patterns through the Pattern Recognition module in BioMiner. This was done primarily to identify the most informative genes, from amongst all 9600 gene expressions. The pattern recognition module of BioMiner software provides support for various forms of supervised techniques which include discrimination and prediction, in which one can develop models from historical data to predict future cases. There are several algorithms for discrimination and prediction that are mainly from WEKA [41] and the J48 Decision Tree induction algorithm was used for these experiments. Decision trees and rule learners generate tree structures or rules directly for class assignments. A decision tree can be used to classify a case by starting at the root of the tree and moving through it until a leaf is encountered [26]. Rules are in $L \rightarrow R$ formats, in which L represents attributes-based tests and R is a class [26]. Rules could be derived from decision trees.

A number of machine learning experiments were performed to search for genes in the data. In each series of experiments, the J48 [41] rule learner was used 20 times to analyze the Ratio data and to identify the target genes from all the

genes available in the data set. Figures 9-10 shows the genes discovered from all data mining experiments. The % values given for each gene indicate how accurately the gene discriminates between the two classes for that experiment. There are four groups of genes in this table that are interesting to pay attention to. Group 1 (in green) are three genes that were discovered, in all four experiments, regardless of our filtering criteria. These three genes were able to correctly discriminate 22/23 (96%) of cases (samples). Group 2 (in red) are three genes that were identified in the first three experiments. All these genes were able to discriminate between the two classes by themselves with 100% accuracy. Group 3 (in blue) are all the genes that were discovered in the first three experiments, all with 96% accuracy (22/23 samples). And finally, group 4 consists of the remainder of the genes identified. Some of these genes are also important in this gene discovery process. For example, genes 16 and 17 were identified with 96-100% accuracy in the first two runs. In fact, gene 16 (in purple) has been identified as the first (perhaps the most informative of all genes) with 100% accuracy for discrimination in the first two runs.

Following the process of identification of informative genes from the ratio data, we plotted different groups of genes for both classes and from all experiments. These plots illustrate interesting patterns between all the genes for both class 1 (AD) and 2 (normal) samples. Similarly, we used these genes to create intensity spectrum plots which also show interesting patterns. Figures 11 and 12 show plots of genes from all ratio data for both classes. Figure 13 is a spectrum plot of the top 20 genes discovered from all ratio data and Figure 14 shows the spectrum intensity pattern of the three groups of genes in Figures 9-10 (green, red and blue). In these figures, the color code is a small square for each element in the vector representing the gene. The red square represents a value greater than the center point value, and the value less than the center point is colored green. By default zero is the center point. The intensity of the color is a measure of the relative difference between values at the same side of center point. For example, the lowest values will have lightest green color and the highest ones the lightest red.

6.4 Individual Dichotomization

The Individual Dichotomization algorithm (Section 5.2) was applied to the F0 data set described in Section 3.4, with a series of probability thresholds ranging from 0.1 to 1. The dependency between the number of perfectly differentiating genes and the probability threshold is shown in Figure 15.

In the case of a probability threshold equal to 1 (perfect classification between the Alzheimer and the Normal classes), from the 9600 genes only 4 were found having the perfect dichotomic property: (Gene #s 4,5,6 and 16 in

Figures 9-10). When only these genes are considered, the overall sample set drastically collapses to a data set of 23 objects and only 4 attributes. The virtual reality representation of this data set is shown in Figure 16 (representation error = 0.002). The Alzheimer class (red) and the Normal class (green) are displayed with a wrapping semi-transparent membrane covering all objects

Gene #	Gene Name	Symbol	Run_F0	Run_F1	Run_F2	Run_F3	p-Val	DiCh	Fold Change	Association*
1	dystrobrevin, alpha	DTNA	F0(96%)	F1(96%)	F2(96%)	F3(96%)	PV	DC1	1.9	
2	EST		F0(96%)	F1(96%)	F2(96%)	F3(96%)	PV	DC1	2.0	
3	**beta-2-microglobulin**	**B2M**	F0(96%)	F1(96%)	F2(96%)	F3(96%)	PV	DC1	1.5	AD (8)
4	**amyloid beta precursor-like protein 1**	**APLP1**	F0(100%)	F1(100%)	F2(100%)		PV	DC0	1.8	AD (2)
5	EST		F0(100%)	F1(100%)	F2(100%)			DC0	-1.4	
6	**complement component 4B**	**C4B**	F0(100%)	F1(100%)	F2(100%)			DC0	1.5	SC (26)
7	LIM and senescent cell antigen-like 2	LIMS2	F0(96%)	F1(96%)	F2(96%)			DC1	1.5	
8	EST		F0(96%)	F1(96%)	F2(96%)			DC1	-1.3	
9	EST		F0(96%)	F1(96%)	F2(96%)			DC1	-1.4	
10	EST		F0(96%)	F1(96%)	F2(96%)			DC1	1.8	
11	EST		F0(96%)	F1(96%)	F2(96%)			DC1	1.3	
12	immunoglobulin heavy constant mu	IGHM	F0(96%)	F1(96%)	F2(96%)			DC1	1.4	
13	EST		F0(96%)	F1(96%)	F2(96%)			DC1	1.5	
14	EST		F0(96%)	F1(96%)	F2(96%)			DC1	-1.6	
15	glucose regulated protein, 58kDa	GRP58	F0(96%)	F1(96%)	F2(96%)			DC1	-1.3	
16	**keratin 8**	**KRT8**	F0(100%)	F1(100%)				DC0	1.7	AD (32)
17	EST		F0(96%)	F1(96%)				DC1	-1.8	
18	activating transcription factor 4	ATF4		F1(96%)				DC1	1.6	
19	EST		F0(96%)					DC1	-1.8	
20	EST		F0(96%)					DC1	-1.7	
21	**Ran GTPase activating protein 1**	**RANGAP**	F0(96%)					DC1	1.8	AD (12)
22	EST			F1(96%)	F2(96%)			DC1	-1.5	
23	EST			F1(96%)	F2(96%)			DC1	1.5	
24	EST				F2(96%)			DC1	1.4	
25	**glutathione S-transferase M2**	**GSTM2**			F2(100%)				1.6	AD (21)
26	EST								1.2	
27	EST				F2(100%)				-1.4	
28	EST								1.1	
29	TU3A protein	TU3A				F3(100%)	PV		1.6	
30	**adducin 3 (gamma)**	**ADD3**					PV		1.3	AD (28)
31	**ferritin, light polypeptide**	**FTL**				F3(91%)	PV		2.6	AD (1)
32	hemoglobin, beta	HBB				F3(100%)	PV		1.9	
33	**clusterin**	**CLU**					PV		2.4	AD (6)
34	EST					F3(100%)	PV		2.0	
35	hemoglobin, gamma G	HBG2					PV		2.0	

Fig. 9. Table containing the results from all experiments (part 1).

Gene #	Gene Name	Symbol	Run_F0	Run_F1	Run_F2	Run_F3	p-Val	DiCh	Fold Change	Association
36	EST					F3(100%)	PV		3.0	
37	angiotensinogen	AGT					PV		1.7	
38	**proteolipid protein 1**	**PLP1**				F3(87%)	PV		2.7	AD (25)
39	EST					F3(100%)	PV		2.6	
40	**proteolipid protein 1**	**PLP1**					PV		1.9	AD (25)
41	hemoglobin, beta	HBB				F3(96%)	PV		2.5	
42	**adducin 3 (gamma)**	**ADD3**					PV		1.4	AD (28)
43	**peroxiredoxin 1**	**PRDX1**				F3(100%)	PV		1.4	AD (14)
44	**6-pyruvoyltetrahydropterin synthase**	**PTS**							1.4	AD (31)
45	hemoglobin, beta	HBB				F3(96%)	PV		1.6	
46	**6-pyruvoyltetrahydropterin synthase**	**PTS**							2.5	AD (31)
47	hemoglobin, beta	HBB				F3(96%)	PV		1.7	
48	EST								1.4	
49	hemoglobin, gamma A	HBG1				F3(96%)	PV		1.6	
50	proprotein convertase *s/k* 1 inhibitor	PCSK1N							1.5	
51	dystrophia myotonica-protein kinase	DMPK				F3(96%)	PV		-1.5	
52	EST								-2.6	
53	prenylcysteine lyase	PCL1				F3(91%)			-1.5	
54	ribosomal protein L31	RPL31							1.3	
55	**metallothionein 1G**	**MT1G**							1.5	AD (24)
56	EST					F3(96%)			1.4	
57	**chimerin (chimaerin) 2**	**CHN2**							2.0	WS (17)
58	EST					F3(96%)	PV		1.3	
59	**ferritin, light polypeptide**	**FTL**							1.5	AD (1)
60	oxysterol binding protein-like 3	OSBPL3				F3(96%)	PV		1.2	
61	pleckstrin homology domain B1	PLEKHB1					PV		-1.2	
62	eukaryotic translation elongation factor 1 a1	EEF1A1							1.2	
63	**proteolipid protein 1**	**PLP1**					PV		2.2	AD (25)
64	**proteolipid protein 1**	**PLP1**					PV		2.0	AD (25)
65	EST						PV		1.5	
66	**glutathione S-transferase M2**	**GSTM2**					PV		1.3	AD (21)
67	EST						PV		1.6	
68	major histocompatibility complex, class I, B	HLA-B					PV		1.9	
69	EST						PV		1.7	
70	RAP1, GTP-GDP dissociation stimulator 1	RAP1GDS1					PV		1.4	
71	**CD81 antigen**	**CD81**					PV		1.6	SCI (9)
72	**cell division cycle 10 homolog**	**CDC10**					PV		1.3	AD (18)
73	tetraspan 3	TSPAN-3					PV		1.3	
74	**neural expressed, devel. down-reg 5**	**NEDD5**					PV		1.3	AD (15)
75	EST						PV		-1.5	
76	EST						PV		1.2	
77	EST							DC1	-1.6	

Fig. 10. Table containing the results from all experiments (part 2).

belonging to the corresponding class. In this virtual reality space both classes are very clearly differentiated.

When the probability threshold is lowered by allowing one misclassified sample out of the original 23 (threshold = 0.9565), then 25 genes are found relevant

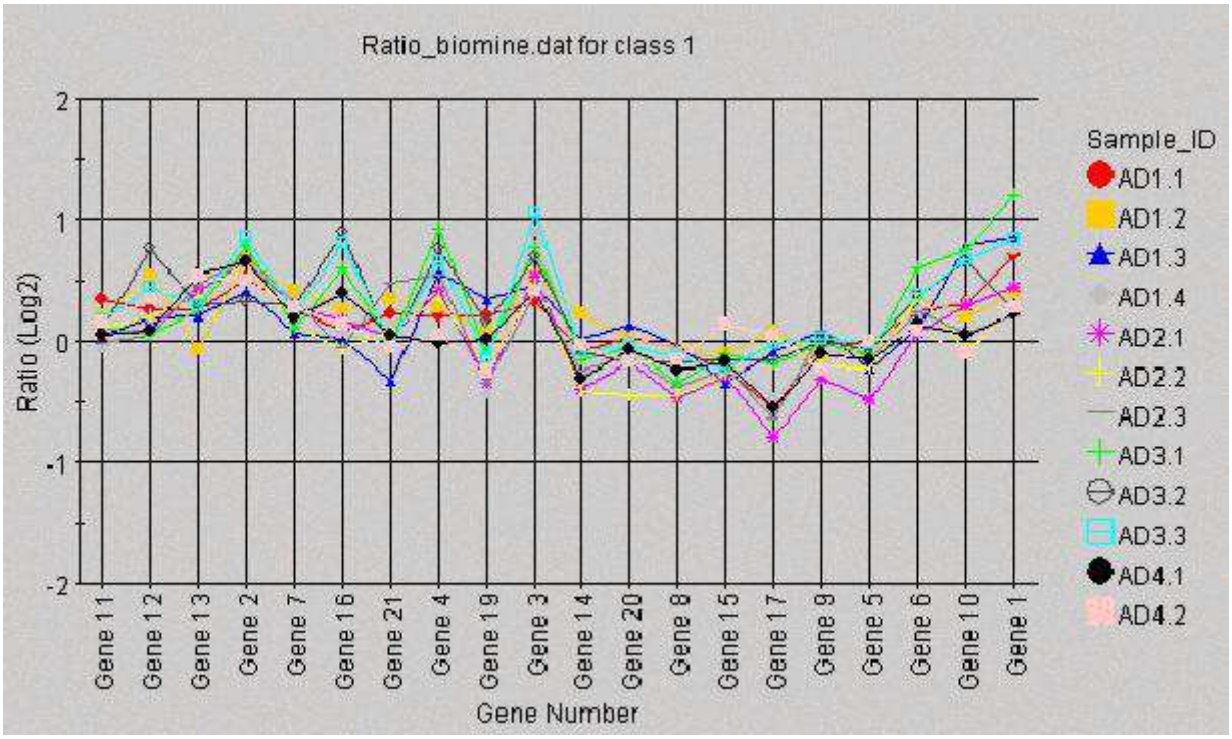

Fig. 11. Genes pattern plot for top 20 genes - Ratio data (F0), Class 1.

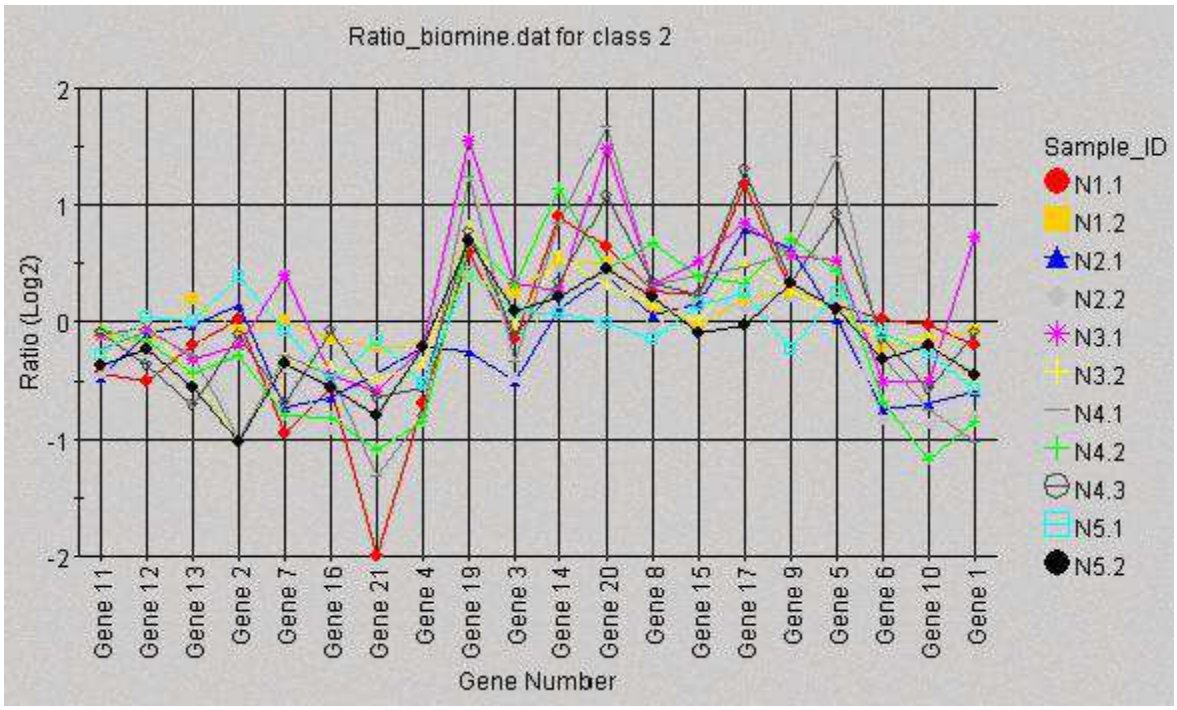

Fig. 12. Genes pattern plot for top 20 genes - Ratio data (F0) for Class 2.

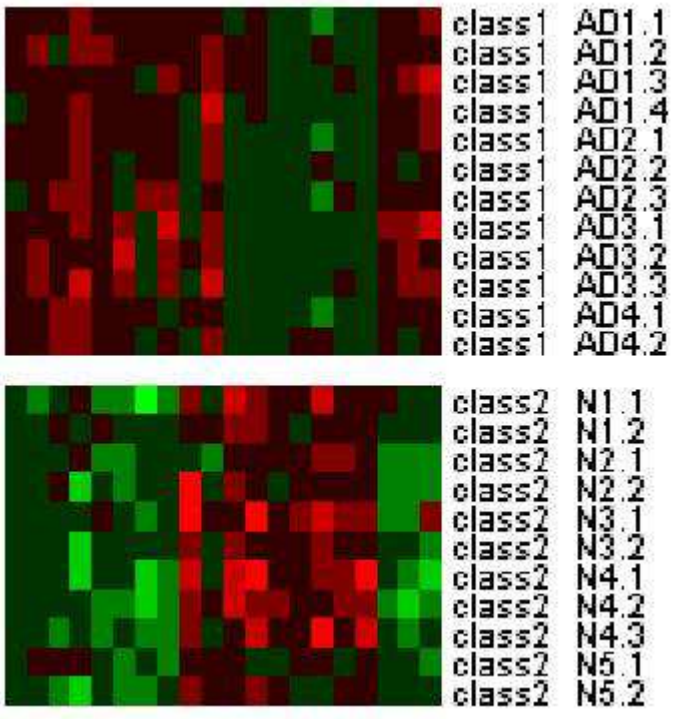

Fig. 13. Intensity spectrum plot for top 20 .

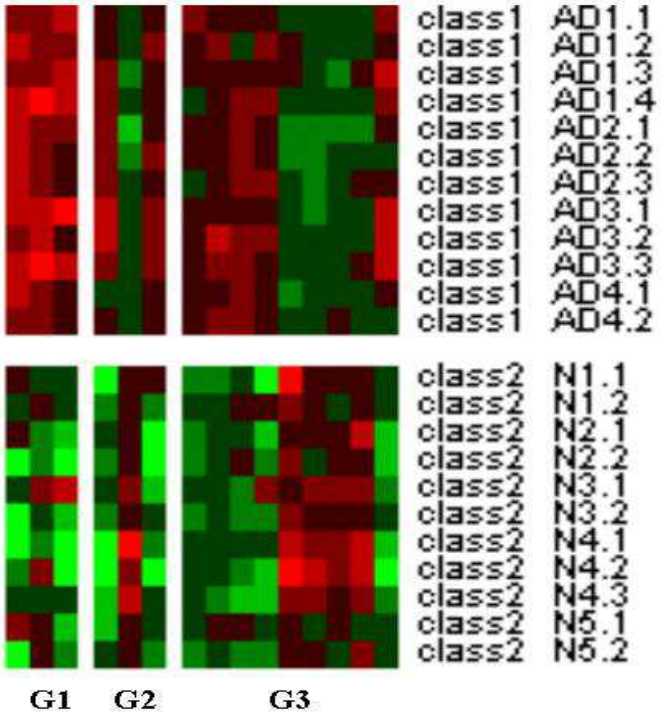

Fig. 14. Intensity spectrum plot for three groups of genes - (G1: group1, G2: group 2, G3: group 3)

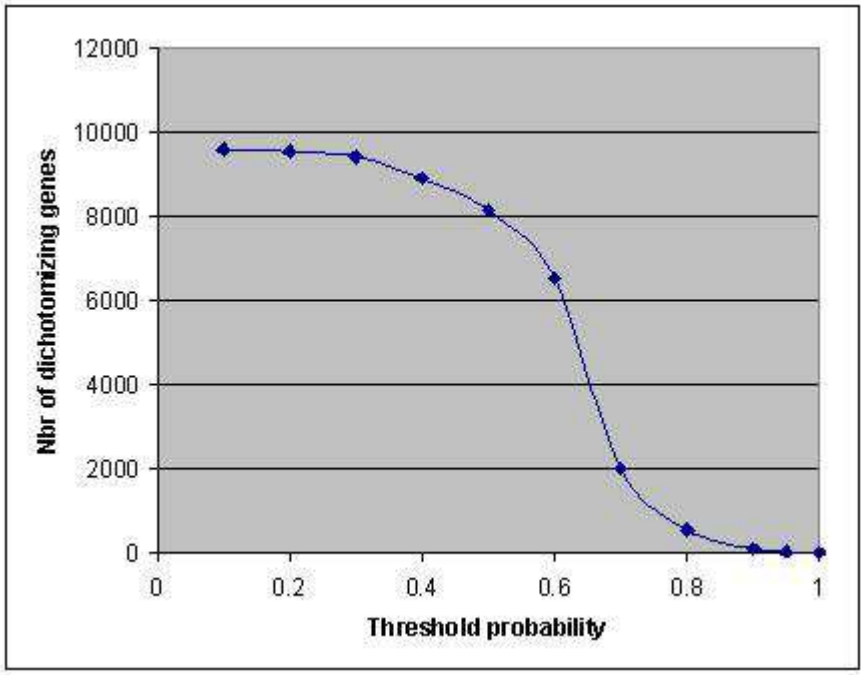

Fig. 15. Gene dichotomization vs probability threshold.

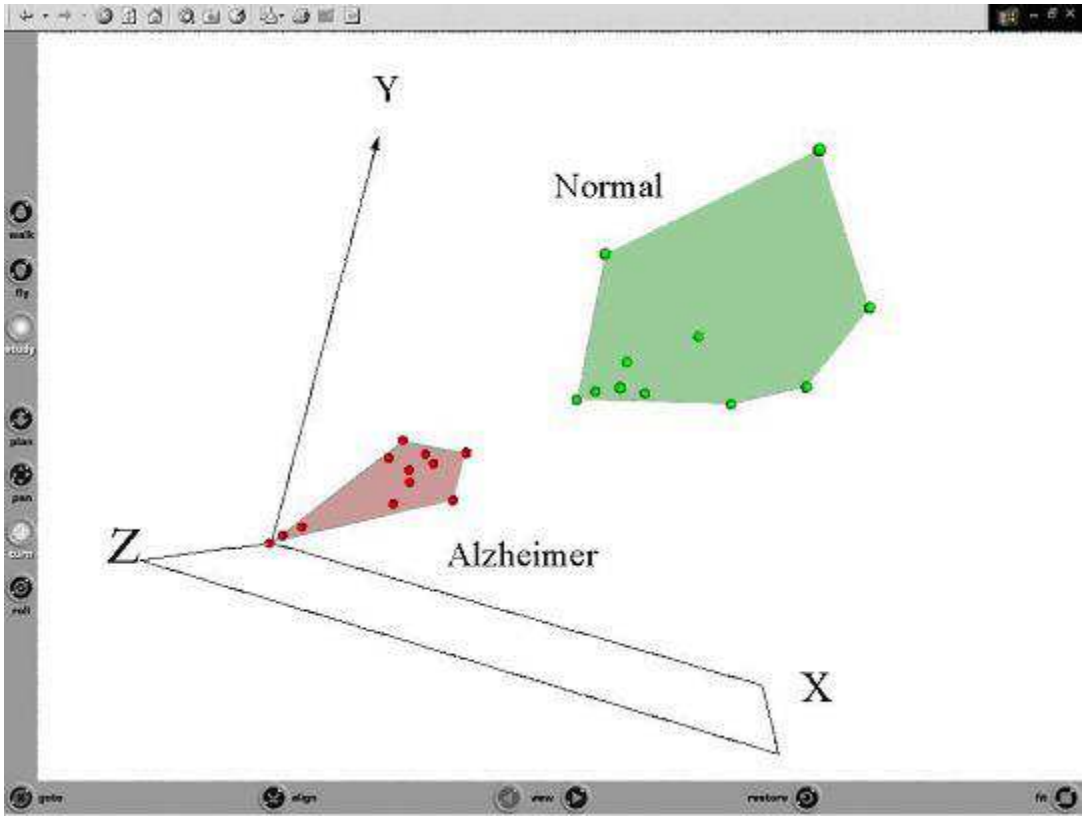

Fig. 16. The VR representation for Ratio data set, 23 samples and the 4 genes with probability 1.

(Figures 9-10). Lowering the threshold to allow two misclassifications (threshold = 0.9130) revealed 94 genes (data not shown).

6.5 P-value and ratio thresholding

The thresholding strategy described in section 5.3 was applied to ratio data set F0, yielding 40 genes. Of these, 10 were different clones representing just 3 genes. The cDNA microarrays used in this study contain a number of such redundancies, i.e., genes represented by cDNA from multiple different clones. Clone redundancy on the microarray can prove useful in terms of result validation. We can be more confident that a gene discovery is real when it is corroborated by independent clones that report the same expression behaviour. Genes PLP1 and HBB, were detected by 4 different clones and ADD3 was detected by 2 clones. Significantly, this represents all known clones on the microarray that target these 3 genes adding further confidence to the result. After removing clone redundancies from the list, there were 11 different genes known to be associated with AD, 12 not previously associated with the disease and 10 ESTs. Six of the AD associated genes were not discovered by either the pattern recognition or dichotomization methods used in this study. A complete list of genes discovered by this method is shown in Figures 9-10 and Figure 17 presents a summary of the gene discoveries for all the methods. Both of these tables include the redundant gene clones. The p-value and ratio cut-offs applied to the data in this strategy are somewhat arbitrary and obviously the list of genes produced by this method would be altered somewhat by choosing different thresholds. The thresholds used were selected for their potential to discover genes with significant differential expression in AD combined with a significant difference between AD and normal sample means. Retrospective analysis of the results indicated that all 15 AD-related gene discoveries would have been detected with thresholds of AD vs. 0 $\alpha < 0.0005$, AD vs. normal $\alpha < 0.006$ and $|AD| > 0.411$.

7 Conclusion

This paper reports the results of our data mining research in which we have searched for patterns in microarray data that come from two known classes (AD and normal). Applying data mining techniques (classical machine learning) is probably the most promising way to identify genes and their behaviour, as the samples come from classes that are known in advance. The results would therefore be more meaningful and easier to interpret, validate and apply. From the data mining point of view, a measure of success in identifying disease-associated, differentially expressed genes is the extent to which the algorithms

	Run_F0	Run_F1	Run_F2	Run_F3	p-Val	DiCh	Any Method	1 method	All methods
ESTs	11 (0)	11 (0)	14 (3)	8 (3)	10 (5)	15 (1)	30	12	1
Associated with AD	5 (0)	4 (0)	4 (1)	13 (5)	15 (6)	4 (0)	24	12	1
No previous association	4 (0)	5 (0)	4 (0)	16 (4)	15 (3)	6 (0)	23	7	1
Total	20 (0)	20 (0)	22 (4)	37 (12)	40 (14)	25 (1)	77	31	3

Note: Numbers in brackets () are results unique to that method.

Fig. 17. Table containing results summary.

identify genes that have previously been associated with disease using independent methodologies. However, from the medical point of view, success is ultimately measured in terms of a more accurate prediction and diagnosis of a disease.

In this research we applied 3 separate techniques to discover genes associated with Alzheimer's disease. Figure 17 shows the summary of the results for all of our experiments. In the first four runs (F0-F3), we used a machine learning algorithm 20 times, to identify the most informative genes through a discover-and-mask approach, in which a gene identified in a run was removed from the data set, before the next run, until all experiments were completed. This identified 20 or more of the most informative genes, as some experiments from runs F2 and F3 identified 2-3 genes. In the second technique (pVal), a statistical thresholding technique, aimed at improving the biological relevance of the results, was applied to the F0 data set. Forty genes were identified, 15 of which have been previously associated with Alzheimer's or other neurological diseases. The high yield of disease-relevant genes discovered clearly demonstrates the potential of this relatively simple method. In the third approach (DiCh), the individual dichotomizer algorithm was applied. It identified individual genes with high classification power associated with Alzheimer and normal and also the optimal ratio values differentiating these classes.

Of the 77 clones identified in this study, 24 represented 17 different genes that are already known to be associated with Alzheimer's or other neurological diseases (see Table 3 and references therein). This is higher than any of the previously published Alzheimer studies. Five of the 17 genes were represented by multiple clones. Twenty-three clones representing 20 different known genes (1 gene was represented by 4 clones), not previously associated with the disease, have been identified as well as 30 uncharacterised ESTs. The number of AD associated genes discovered using these data mining strategies is significantly higher than one would expect by chance. A literature study of 3 sets of 50 genes selected randomly from the microarray data revealed that 10 percent of the known genes had an association with AD, compared to 53 percent in the data mining study. Given the success in identifying genes already associated with AD, we can have some confidence in the involvement of the latter genes and ESTs. For example, one of the genes, identified by all of the methods, but not previously associated with AD, is dystrobrevin. Dystrobrevin is

a dystrophin-associated protein found in dystrophin-associated protein complexes in the brain [1]. Given that one third of Duschenne muscular dystrophy patients, a disease caused by mutations in the dystrophin gene, have a mild dementia, it is possible that altered expression of dystrobrevin could be related to dementia of the Alzheimer type. Similarly, there is a high probability that proprotein convertase inhibitor (PSK1N) could be AD-associated given the role of proprotein convertase in β-amyloid processing [34]. The biological significance of the AD-associated genes found in this paper will be discussed further in a subsequent publication.

We think our approach sets the stage for a major step in using gene expression data for disease classification and diagnosis. It can also influence the future of gene function identification, pathology, toxicogenomics, and pharmacogenomics.

References

[1] Bartzokis G., Tishler T.A.: MRI evaluation of basal ganglia ferritin iron and neurotoxicity in Alzheimer's and Huntingon's disease. Cell Mol Biol (Noisy-le-grand) 2000 June; 46(4):821-33.

[2] Bayer T.A., Paliga K., Weggen S., Wiestler O.D., Beyreuther K., Multhaup G.: Amyloid precursor-like protein 1 accumulates in neuritic plaques in Alzheimer's disease. Acta Neuropathol (Berl) 1997 Dec;94(6):519-24.

[3] Blake D.J., Hawkes R., Benson M.A., and Beesley P.W.: Different dystrophin-like complexes are expressed in neurons and glia. J Cell Biol,1999, 147: 645-58.

[4] Borg I., Lingoes, J.: Multidimensional Similarity Structure Analysis, Springer-Verlag 1987.

[5] Brown V. M, Ossadtchi A., Khan A.H., Cherry S.R., Leahy R.M. and Smith D.J., High-throughput imaging of brain gene expression. Genome Res, 2002, 12: 244-54.

[6] Calero M., Rostagno A., Matsubara E., Zlokovic B., Frangione B., Ghiso J. Apolipoprotein J: (clusterin) and Alzheimer's disease. Microsc Res Tech, 2000 Aug 15;50(4):305-15.

[7] Colangelo V., Schurr J., Ball M.J., Pelaez RP, Bazan NG, Lukiw WJ.: Gene expression profiling of 12633 genes in Alzheimer hippocampal CA1: transcription and neurotrophic factor down-regulation and up-regulation of apoptotic and pro-inflammatory signaling. J Neurosci Res, 2002, 70: 462-73.

[8] Davidsson P., Westman-Brinkmalm A., Nilsson C.L., Lindbjer M., Paulson L., Andreasen N., Sjogren M., Blennow K.: Proteome analysis of cerebrospinal fluid proteins in Alzheimer patients. Neuroreport, 2002 Apr 16;13(5):611-5.

[9] Dijkstra S., Geisert E.E. JR, Gispen W.H., Bar P.R., Joosten E.A.: Up-regulation of CD81 (target of the antiproliferative antibody; TAPA) by reactive microglia and astrocytes after spinal cord injury in the rat. J Comp Neurol, 2000 Dec 11;428(2):266-77.

[10] Drobyshev A.L., Machka C., Horsch M., Seltmann M., Liebscher V., Hrabe De Angelis M., and Beckers J.: Specificity assessment from fractionation experiments (SAFE): a novel method to evaluate microarray probe specificity based on hybridisation stringencies. Nucleic Acids Res, 2003 Jan 15;31(2):E1-11.

[11] Famili A. and Ouyang. J.: Data mining: understanding data and disease modeling, Proceedings of IASTED-AI-03 Conference, Innsbruck, Austria, Feb. 10-13, 2003.

[12] Hata R., Masumura M., Akatsu H., Li F, Fujita H, Nagai Y, Yamamoto T, Okada H, Kosaka K., Sakanaka M. and Sawada T.: Up-regulation of calcineurin Abeta mRNA in the Alzheimer's disease brain: assessment by cDNA microarray. Biochem Biophys Res Commun, 2001, 284: 310-6.

[13] Holloway A.J., van Laar R.K., Tothill R.W. and Bowtell D.D.: Options available–from start to finish–for obtaining data from DNA microarrays II. Nature Genetics, 2002 Dec;32 Suppl:481-9.

[14] Kim S.H., Fountoulakis M., Cairns N., Lubec G.: Protein levels of human peroxiredoxin subtypes in brains of patients with Alzheimer's disease and Down syndrome. J Neural Transm Suppl 2001;(61):223-35.

[15] Kinoshita A., Kinoshita M., Akiyama H., Tomimoto H., Akiguchi I., Kumar S., Noda M., Kimura J.: Identification of septins in neurofibrillary tangles in Alzheimer's disease. Am J Pathol 1998 Nov;153(5):1551-60.

[16] Kruskal, J. B.: Multidimensional scaling by optimizing goodness of fit to non-metric hypothesis. Psychometrika, 29, 1-27. 1964.

[17] Lecka-Czernik B., Moerman E.J., Jones R.A., Goldstein S. Identification of gene sequences overexpressed in senescent and Werner syndrome human fibroblasts. Exp Gerontol 1996 Jan-Apr;31(1-2):159-74.

[18] Lee, M.-L., T., Kuo, F. C., Whitmore, G. A. and Sklar, J.: Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations PNAS, 2000, 97: 9834-39, .

[19] Liauw J., Nguyen V., Huang J., St George-Hyslop P., Rozmahel R.: Differential display analysis of presenilin 1-deficient mouse brains. Brain Res Mol Brain Res, 2002, Dec 30;109(1-2):56-62.

[20] Long A.D., Mangalam H.J, Chan B.Y., Tolleri L., Hatfield G.W. and Baldi P.: Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. J Biol Chem 2001 June 8;276 (23):19937-44.

[21] Loring J.F., Wen X., Lee J.M., Seilhamer J. and Somogyi R.: A gene expression profile of Alzheimer's disease. DNA Cell Biol, 2001 20: 683-95.

[22] Lovell M.A., Xie C., Markesbery W.R.: Decreased glutathione transferase activity in brain and ventricular fluid in Alzheimer's disease. Neurology, 1998 Dec;51(6):1562-6.

[23] Lukiw W.J, Carver L.A., LeBlanc H.J., Bazan N.G.: Analysis of 1184 gene transcript levels in AD CA1 hippocampus: synaptic signaling and transcription factor deficits and upregulation of pro-inflammatory pathways. Alzheimer Reports, 2000 3:161-167.

[24] Mufson E.J., Counts S.E., Ginsberg S.D.: Gene expression profiles of cholinergic nucleus basalis neurons in Alzheimer's disease. Neurochem Research, 2002, 27:1035-1048, .

[25] Pasinetti G.M.: Use of cDNA microarray in the search for molecular markers involved in the onset of Alzheimer's disease dementia. J. Neurosci Res, 2001 65: 471-6.

[26] Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann, San Mateo, CA, 1993.

[27] Richarz A.N., Bratter P.: Speciation analysis of trace elements in the brains of individuals with Alzheimer's disease with special emphasis on metallothioneins. Anal Bioanal Chem 2002. Feb;372(3):412-7.

[28] Roher A.E., Weiss N., Kokjohn T.A., Kuo Y.M., Kalback W., Anthony J., Watson D., Luehrs D.C., Sue L., Walker D., Emmerling M., Goux W., Beach T.: Increased A beta peptides and reduced cholesterol and myelin proteins characterize white matter degeneration in Alzheimer's disease.Biochemistry 2002 Sep 17;41(37):11080-90.

[29] Rudduck C., Beckman L., Franzen G., Jacobsson L., Lindstrom L.: Complement factor C4 in schizophrenia. Hum Hered, 1985 35, 223-6.

[30] Sammon, J. W.: A non-linear mapping for data structure analysis. IEEE Trans. Computers, 1969, C-18, 401-408.

[31] Sangerman J., Kakhniashvili D., Brown A., Shartava A., Goodman SR.: Spectrin ubiquitination and oxidative stress: potential roles in blood and neurological disorders Cell Mol Biol Lett 2001;6(3):607-36.

[32] Sendera T.J., Dorris D., Ramakrishnan R., Nguyen A., Trakas D., Mazumder A.: Expression profiling with oligonucleotide arrays: technologies and applications for neurobiology. Neurochem Research, 2002 27:1005-1026.

[33] Spruill S.E., Lu J, Hardy S. and Weir B.: Assessing sources of variability in microarray gene expression data. Biotechniques Oct; 2002, 33(4):916-20, 922-3.

[34] Thomas G.: Furin at the cutting edge: from protein traffic to embryogenesis and disease. Nat Rev Mol Cell Biol, 2002 3, 753-66

[35] Thony B, Auerbach G, Blau N.: Tetrahydrobiopterin biosynthesis, regeneration and functions. Biochem J. 2000 Apr 1;347 Pt 1:1-16.

[36] Trojanowski J.Q., Newman P.D., Hill W.D., Lee V.M.: Human olfactory epithelium in normal aging, Alzheimer's disease, and other neurodegenerative disorders. J Comp Neurol, 1991 Aug 15;310(3):365-76.

[37] Trotter S.A., Brill L.B. 2nd, Bennett J.P. Jr.: Stability of gene expression in postmortem brain revealed by cDNA gene array analysis. Brain Res, 2002 942:120-3.

[38] Valdés J.J.: Virtual reality representation of relational systems and decision rules: an exploratory tool for understanding data structure. Theory and Applications of relational Structures as Knowledge Instruments (TARSKI). COST Action 274, Prague Nov 2002 14-16.

[39] Valdés, J.J.: Virtual Reality Representation of Information Systems and Decision Rules: An exploratory technique for understanding data and knowledge structure. In Proc. 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing RSFDGrC'2003, Chongqing, China, May 2003, 26-29, (to appear).

[40] Yang I.V., Chen E., Hasseman J.P., Liang W., Frank B.C., Wang S., Sharov V., Saeed A.I., White J., Li J., Lee N.H., Yeatman T.J. and Quackenbush J.: Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol 2002 Oct 24;3(11):research0062.

[41] Witten I. and Eibe F.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann. San Mateo, CA, 1999.