Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

A Statistical and Biological Approach for identifying misdiagnosis of incipient Alzheimer patients Using Gene expression Data

Sandeep Joseph* ,Kelly R. Robbins* and Romdhane Rekaya

Abstract—**A latent-threshold model and misclassification algorithm were implemented to examine potential misdiagnosis among 16 Alzheimer's disease (AD) subjects using gene expression data. Results obtained without invoking the misclassification algorithm showed limited predictive power of the model. When the misclassification algorithm was invoked, four subjects were identified as being potentially misdiagnosed. Results obtained after adjustment of the AD status of these four samples showed a significant increase in the model's predictive ability. Mixed model analysis detected no AD related genes as differentially expressed when using original classifications; conversely, multiple AD genes were identified using the new classifications. These results suggest that this algorithm can identify misclassified subjects which, in turn, can increase power to predict disease status and identify disease related genes.**

I. Introduction

Alzheimer's disease (AD) is an incurable and debilitating condition which, along with other neurodegenerative diseases, represents one of the largest areas of unmet need in modern medicine. Over four million Americans are currently stricken with AD and by the middle of this century, the baby boomers could take that number to 14 million. AD is the most common form of dementia and leads to irreversible neurodegenerative damage of the brain. The disease progression of AD is slow, and it may take several years from onset of cognitive decline to diagnosis. Current diagnostic tools like MMSE (Mini-Mental State Examination) and NFT (Neurofibrillary Tangles) scores have poor sensitivity, especially for the early stages of the disease and do not allow for diagnosis until the disease has lead to irreversible brain damage [1], [2]. Diagnosing AD in its early stages can be difficult due to the similar clinical symptoms of AD and more rare degenerative dementias. As a result of such problems, diagnosis of patients with incipient AD has proven to be somewhat difficult.

Recent advances in molecular genetics and functional genomics have provided a powerful tool to study the genetic bases of several complex diseases. The use of microarray expression profiling for the classification and subtype discovery of diseases has been proposed and investigated [3], however, such algorithms work under the assumption that all classifications in the training set are correct. In cases where the diagnostic tools for classification of training samples are not highly reliable, this assumption may not hold true. Reference [4] found that when misclassification occurred in the training set, the predictive ability of the algorithm used was greatly reduced. In cases where the certainty of training sample classification is in question, a method to identify and correct potential misclassifications may be needed to effectively predict disease status and identify disease related genes This, coupled with data that indicate the brain changes associated with AD begin well before clinical symptoms appear [5], suggest that gene expression data may be used to identify potentially misclassified subjects. For this study a threshold model and misclassification algorithm were implemented to identify subjects that were misclassified using clinical diagnostic tools, to predict whether they are healthy or incipient AD patients. Further more, a statistical and biological approach was implemented in order to validate or approve the effects of misclassification on healthy and incipient AD patients and AD related gene discovery.

II. Materials and Methods

A. Gene expression data

Hippocampal specimens used in this study were obtained from the autopsy of 31 subjects through the Brain Bank of the Alzheimer's disease Research Center at the University of Kentucky. The affymetrix gene expression data is publicly available in the Gene expression database at NCBI. Methods for mRNA extraction, and disease classification, based on MMSE and NFT scores, are fully described by [6]. The Affymetrix human GeneChip, HG-U133A, containing 22283 targets was used. Table 1 provides a summary description of the original data. In this study, only control and incipient samples were used. For this study, the average difference data as well as the probe level data was used.

B. Pre-processing of the probe level data

The affy package of the Bioconductor was used for pre-processing the raw data, which has many built-in pre-processing methods for background adjustments as well as

Manuscript received April 24, 2006.

Kelly R. Robbins is with the Rhodes Centre for Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA. (e-mail: krobbin1@uga.edu).

Sandeep Joseph is with the Rhodes Centre for Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA. (e-mail: sandeep@uga.edu).

Romdhane Rekaya is with the Rhodes Centre for Animal and Dairy Science, Department of Statistics, Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA. (correspondence: phone: 1-706-542-0949; fax: 1-706-583-0274; e-mail: rrekaya@uga.edu).

* First and second authors contributed equally.

data normalization specifically for Affymetrix microarray data. The log base 2 transformation was performed on the background adjusted and normalized PM intensities as suggested by previous studies [7].

TABLE 1: Mean and standard deviation of age, MMSE and NFT scores for healthy and Alzheimer's patients[a]

Status	Mean MMSE	MMSE SE	Mean NFT	NFT SE
Healthy	27.7	0.5	2.7	3.1
Incipient AD	24.3	3.0	17.5	21.7

[a]AD= Alzheimer's Disease; MMSE=Mini-Mental Status Examination; NFT=Nerofibrilliary tangles; For this study, only control and incipient AD subjects were used for analysis.

C. Statistical Analysis

The regression on AD status was done using a latent variable model such that:

$$y_i = \begin{cases} 1 & \text{if } l_i \geq 0 \\ 0 & \text{if } l_i < 0 \end{cases}$$

where $yi = 1$ indicates an incipient AD status for subject i.

The liability *l*i was modeled using a linear regression. After dimension reduction of **X**, the regression equation was:

$$\boldsymbol{l = QD\gamma + e}$$

Where **X** is the *n* (number of samples) by *m* (number of genes) matrix of the logs of the average difference of gene expression (LGE), **Q** is an *n* by *n* orthogonal matrix, **D** is an ordered *n* by *n* diagonal matrix**,** and **γ** is an *nx1* vector of "super" gene effects. Using the probit link function the following relationship is obtained:

$$y_i = \begin{cases} 1 & \text{if } \boldsymbol{\Phi}(\mathbf{Q}_i\mathbf{D}\boldsymbol{\gamma}) \geq 0.5 \\ 0 & \text{if } \boldsymbol{\Phi}(\mathbf{Q}_i\mathbf{D}\boldsymbol{\gamma}) < 0.5 \end{cases}$$

The predictive ability of this model was tested using a "leave one out" validation procedure. To examine the possibility of misdiagnosis, the probability of miscoding (PM) was calculated for each sample in the validation data set using a Bayesian approach derived by [8] as:

$$\text{PM} = p(m_i = 1 \mid \boldsymbol{\gamma, \pi, m}_{-i}\mathbf{, Q, D, r})$$

Where **m** is a vector of binary variables with $m_i = 1$ indicating a miscoding event for subject i; **r** is a vector of the unobserved, true binary disease status; and π is the probability of observing a miscoding event.

Using the binomial approximation the normal distribution, z-scores were calculated from the PM for each sample in the training set. The z-score with the highest absolute value (z*) was selected and then compared to the threshold z_α. If $z^* \geq z_\alpha$, using a one-sided test (α=0.05), the subject was identified as being potentially misclassified.

To examine the accuracy of the reclassification process, an analysis to identify differentially expressed genes was performed on probe level data using the original and recoded classifications. The analysis was nested within gene using the following mixed linear model:

$$y_{ijk} = \mu + P_i + T_i + A_k + e_{ijk}$$

where y_{ijk} is the $\log_2$ transformed intensity for probe P_i (i=1,2….20) generated under treatment or disease class T_j (j=1,2) in array A_k (k=1,2,….16) and e_{ijk} is the residual term. It is important to note that the number of probes per gene ranged from 11 to 20.

The contrast statement was used to estimate and test linear combinations of controls and incipient AD subjects in the model. Least square estimates of both treatments for each single gene were calculated using the LSMEANS statement of SAS. The fold-change for each gene was calculated by taking the ratio of the least square estimates of incipients and controls. Probes having an absolute fold change greater than two were outputed.

Due to the large number of genes tested in microarray studies a method to asses the number of false positives is of great importance. In this study, the Benjamini Hochberg method for false discovery rate (FDR) calculation was used in order to account for the multiple comparisons.

Initial analysis was performed using gene expression data classified according to [6] called as D1 classification. To examine potential misclassification in the initial diagnosis of several AD subjects, statistical analysis was conducted after switching the health status of four patients identified as potentially misclassified (D2) according to the misclassification algorithm. Finally, an analysis was conducted after removal of the four potential misclassified patients (D3).

III. RESULTS AND DISCUSSION

The ranges of MMSE and NFT scores, given in Table 2, show large overlapping regions between the scores for control and incipient AD groups.

Clearly there are no well defined borders separating control and incipient AD subjects, which could make diagnosis difficult when using these clinical tests. This seems to be reflected by the initial validation results found in Table 3.

TABLE 2: Ranges of MMSE and NFT scores

AD status [a]	MMSE		NFT	
	LB	UB	LB	UB
Control	26	30	0	8
Incipient AD	20	29	5.5	65.8

[a]AD=Alzheimer's Disease. MMSE=Mini-Mental Status Examination. NFT=Nerofibrilliary tangles. LB = Lower Bound and LU = Upper Bound

The model had little power to correctly predict disease status given the original disease classification based on MMSE and NFT scores. In fact, 8 out of 16 samples had their disease status correctly predicted for a classification accuracy of 50%.

When potential misdiagnosis in the training set was postulated in the statistical model, four samples, subjects 4, 10, 12, and 15, were identified as being misclassified

TABLE 3: Validation results

Subject	Original status[a]	P(Status =1)[b]	Recoded status[c]	P(Status =1)[d]
1	1	0.12	1	0.72
2	1	0.90	1	0.97
3	0	0.14	0	0.11
4	1	0.05	0	0.04
5	0	0.12	0	0.11
6	0	0.50	0	0.50
7	0	0.41	0	0.87
8	0	0.50	0	0.56
9	0	0.44	0	0.07
10	1	0.55	0	0.27
11	1	0.31	1	0.37
12	0	0.83	1	0.92
13	1	0.73	1	0.86
14	0	0.30	0	0.30
15	0	0.93	1	0.91
16	1	0.32	1	0.83

[a] Alzheimer's disease status based on clinical tests (0 = healthy; 1= incipient AD); [b] predicted probability of an individual being incipient AD using original status; [c] Alzheimer's disease status after reclassification (0=healthy; 1=incipient AD); [d] Predicted probability of an individual being incipient AD using the recoded status.

Table 3 shows that, after iteratively reclassifying (switching of their binary status) of these four subjects, there were large increases in the prediction accuracy of the model. It can be seen that in addition to the four reclassified subjects, the predictions for subjects 1 and 16 went from being incorrect to correct after reclassification. Clearly, the consideration of potential misdiagnosis in the statistical model yielded a large increase in prediction accuracies, from 50% to 75%.

An absolute fold change of greater than 2 as well as an FDR of 7%, to account for multiple testing, were used to identify differentially expressed genes after the analysis of the probe level data using the mixed linear model. Using fold change, 4, 58 and 43 genes were found to be differentially expressed using D1, D2 and D3 data sets, respectively. There were 31 genes in common between the D2 and D3 classification, while none of the genes in the D1 classification were present either in the D2 or D3 classification. A detailed examination of the genes in common between D2 and D3 classifications reveled that 22 out of the 31 common genes were directly related to AD and the central nervous system as indicated in table 4. In fact, several genes related to synaptic loss and neurofibrillary pathology of the central nervous system, membrane proteins, and vesicle fusion of the synaptic membrane were found to be differentially expressed using D2 and D3, while none were found using D1.

It is well documented that cognitive alterations in patients with AD are closely associated with synaptic loss and neurofibrillary pathology of the central nervous system which might accelerate in later stages of AD [9], [10] and could be used as a predisposing cause for the progression of early AD. Molecular biomarkers of active synapses like synaptotagmin[11],synaptophysin, Chromogranin B (secretogranin 1), alpha Synuclein (ASN) and Synaptosomal-associated protein 25(SNAP-25) could be good indicators of early synaptic damage. The results of this study (Table 4) show that genes responsible for the expression of these synaptic proteins (biomarkers) like SNAP25 (Synaptosomal-associated protein, 25K), SYT1 (Synaptotagmin 1), and CHGB (Chromogranin B) have been found to be differentially expressed in both the D2 and D3 classification, while ASN (alpha Synuclein) was differentially expressed in the D2 classification only. None of these synaptic proteins were found to be differentially expressed in D1 classification.

The high correspondence between D2 and D3 classifications in identifying relevant genes (>80%) associated with AD or to the central nervous system coupled with the fact that none of these genes were identified using the original classification in D1 clearly indicates the seriousness of misdiagnosis regarding the status of healthy and incipient AD subjects based on MMSE and NFT scores as provided by [6]. It is apparent that such misdiagnosis can greatly affect the power to detect differentially expressed genes. The extremity of the results between D2, D3, and D1 is in part due to the fact the number of potential misclassified samples in D1 (4 samples) represents almost 25% of the input data set. Furthermore, the results of this study provide additional evidence for the poor sensitivity of the method used for diagnosis in the D1 classification.

It is worth mentioning that several AD related genes were identified as differentially expressed only using D2 classification (Table 5). The list included genes such as SCG-10 which is a neuron-specific signal transduction molecule that plays a role in cellular differentiation and proliferation [12], Alpha Synuclein (ASN) which is a heat-stable protein present more abundantly in telencephalon [13] and was recently considered as a major factor in the pathogenesis of AD [14], a mutant of the ubiquitin gene (UCH-L1) which might lead to dysfunction of the neuronal ubiquitination/de-ubiquitination machinery, causing synaptic deterioration and neuronal degeneration in AD brains [15], and the mitogen-activated protein kinase 1 gene (MEK1) that seems to be involved in initial symptoms of AD through an apparent nuclear accumulation suggesting that abnormal nuclear trafficking may contribute to the pathogenesis of AD [16].

The fact that these genes were not identified in the D3 classification could be due in part to the reduced sample size after the removal of the 4 potentially misclassified samples. It was found that the parameter estimates of genes present in both D2 and D3 had high similarity between the two data classes mentioned above.

TABLE 4: Important differentially expressed genes with known relationship to AD or nervous system diseases in common between D2 and D3 classifications

Gene symbol	Gene description	D2		D3	
		FC[$]	P[$]	FC	p
PPP3CA	Calcineurin A alpha	-2.35	<0.01	-2.26	0.01*
SNAP25	Synaptosomal-associated protein	-2.64	<0.01	-2.33	0.06*
SYT1	Synaptotagmin 1	-2.50	<0.01	-2.20	0.04*
CHGB	Chromogranin B	-2.20	<0.01	-2.09	<0.01
GAP43	Growth associated protein 43	-2.14	<0.01	-2.05	0.02*
NELL2	Neural Epidermal Factor	-2.48	<0.01	-2.45	0.01*

*indicates genes that are not differentially expressed at a FDR level of 7%.
$ FC=Fold change, P=P-values

This suggests that the decrease in power to identifying differentially expressed genes was mainly due to the reduction of 25% of the expression data in D3. Based on these results it is clear that, when the classifications of a large proportion of samples are questionable, it is better to reclassify questionable subjects, using the available information (D2), rather than excluding them from analysis (D3).

TABLE 5: Important differentially expressed genes with known relationship to AD or nervous system diseases found only in D2 classification

Gene Symbol	Gene Description	D2 classification	
		Fold change	p-value
UCHL1	Ubiquitin carboxyl-terminal esterase L1	-2.080	<0.01
MEK1	Mitogen-activated protein kinase 1	-2.086	<0.01
SCG10	Stthmin-like 2	-2.05	<0.01
ASN	Synuclein Alpha	-2.001	<0.01
MYTIL	Myelin transcription factor 1	-2.022	<0.01

IV. CONCLUSIONS

The use of LGE, coupled with the reclassification algorithm can greatly increased the accuracy of disease diagnosis in the presence of potentially misclassified subjects, as seen with the increase from 50% to 75% accuracy in AD diagnosis. While the use of gene expression data showed improved performance over traditional AD makers, utilization of such data for disease diagnosis may not always be practical due to prohibitive cost or the invasiveness of tissue collection. However, to fully understand and effectively treat genetically influenced diseases, a better understanding of the genes involved in disease development will be required. As such, the use of this algorithm as a preprocessing step for analysis of microarray data could yield a substantial increase in the power to detect important disease related genes when misclassified test subjects are present in the data.

REFERENCES

[1] Galasko D, Klauber MR, Hofstetter CR, Salmon DP andLasker B Thal LJ, The Mini-Mental State Examination in the early diagnosis of Alzheimer's disease. Arch. Neurol. Vol. 47, 1990, pp. 49 52.

[2] Haroutunian V, Purohit DP, and Perl DP, Neurofibrillary tangles in nondemented elderly subjects and mild Alzheimer disease. Arch. Neurol. Vol. 56, 1999, pp. 713 718.

[3] Golub TR, Slonim DK, Tamayo P, Huard C, Gassenbeek M, Mesirov P et al Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science vol. 286,1999, pp. 531-537.

[4] Zhang W, Rekaya R and Bertrand JK A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer. Bioinformatics vol . 22(3), 2006, pp .317 325.

[5] Price, D. L. and Sisodia, S. S, Mutant genes in familial Alzheimer's disease and transgenic Models, Annual Review of Neuroscience,vol. .21, 1998,pp.479 505.

[6] Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR andLandfield PW, Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc. Natl. Acad. Sci. vol. 101, 2004. pp. 2173-2178.

[7] *Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B and Speed TP Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res.vol. 31(4),2003, pp. e15.*

[8] Rekaya R., Weigel K. A. and Gianola D., Threshold model for misclassified binary responses with applications to animal breeding, Biometrics, vol . 57, 2001, pp 1123 1129.

[9] Masliah E, Mallory M, Alford M, DeTeresa R, Hansen LA, McKeel DW Jr, Morris JC Altered expression of synaptic proteins occurs early during progression of Alzheimer's disease. Neurology vol. 56(1), 2001, pp. 127 129.

[10] Heinonen O, Soininen H, Sorvari H, Kosunen O, Paljarvi L, Koivisto E, Riekkinen PJ Sr, Loss of synaptophysin-like immunoreactivity in the hippocampal formation is an early phenomenon in Alzheimer's disease. Neurosci. Vol . 64(2), 1995, pp. 375-384.

[11] Davidsson P, Jahn R, Bergquist J, Ekman R and Blennow K, Synaptotagmin, a synaptic vesicle protein, is present in human cerebrospinal fluid: a new biochemical marker for synaptic pathology in Alzheimer disease? Mol. Chem. Neuropathol. Vol. 27(2), 1996, pp. 195 210.

[12] Mori N, Stein R, Sigmund O and Anderson DJ, A cell type-preferred silencer element that controls the neural-specific expression of the SCG10 gene. Neuron. Vol . 4(4), 1990, pp. 583 594.

[13] Iwai A, Yoshimoto M, Masliah E and Saitoh T, Non-A beta component of Alzheimer's disease amyloid (NAC) is amyloidogenic**.** Biochemistry vol . 34(32), 1995, pp.10139 10145.

[14] Masliah E, Mechanisms of synaptic pathology in Alzheimer's disease. J. Neural. Transm. Vol. 53 (Suppl), 1998, pp. 147 158.

[15] Choi J, Levey AI, Weintraub ST, Rees HD, Gearing M, Chin LS, Li L Oxidative modifications and down-regulation of ubiquitin carboxyl-terminal hydrolase L1 associated with idiopathic Parkinson's and Alzheimer's diseases. J. Biol. Chem. Vol. 279(13), 2004, pp. 13256 13264

[16] Zhu X, Sun Z, Lee HG, Siedlak SL, Perry G, Smith M A Distribution, levels, and activation of MEK1 in Alzheimer's disease. J. Neurochem. Vol . 86(1), 2003, pp. 136 142.

.