A Semi-parametric Bayesian Framework for Identifying up and down Regulated Genes in Subjects with Neurocysticercosis (NCC) Associated Epilepsy

Michael P. Anderson, PhD[*] Cheuk Leung,MS[*] Suzanne R. Dubnicka, PhD[†]
Douglas A. Drevets, MD[‡] Vedantam Rajshekhar, MD[§] Anna Oommen, PhD[§]
Prabhakaran Vasudevan, PhD[§] Josephin Justin Babu, RN[§]
Ramajayam Govindan, PhD[§] Hélène Carabin, DVM, PhD[*]

Abstract

Neurocysticercosis (NCC), *Taenia solium* larval infection of the brain, is the commonest cause of adult onset seizures in countries endemic for the parasitic infection. Subjects reporting to the Department of Neurological Sciences at the Christian Medical College in Vellore, India were recruited for participation in a study to determine inflammation relevant gene expression profiles specific to NCC in peripheral blood monocytes. While technology to measure and describe gene regulation has become quite sophisticated, statistical methods for analyzing such data have remained somewhat pedestrian. Traditional microarray analysis often relies on unwarranted assumptions of normality and a battery of several thousand t-tests to assist in identifying significantly up or down regulated genes between groups. In this work we propose a semi-parametric Bayesian framework as a robust, probability based alternative to the t-test approach in order to identify differentially regulated genes from a microarray analysis. We compare the top genes identified by both methods in the NCC data set.

Key Words: Micro-array Data, Gene Expression, Health Data, Neurocysticercosis, High Dimensional Data, Bayesian Classification

1. Introduction

Taenia solium larval infection of the brain (neurocysticercosis) (NCC) is the commonest parasitic infection of the central nervous system in countries endemic for the parasite. Taeniasis patients who harbor the adult worm in their intestine, the result of eating undercooked *Taenia solium* larval infected pork, shed thousands of microscopic eggs in their feces that contaminate the environment in regions of the world where open field defecation is practiced. *T. solium* eggs from the environment coupled with poor hand washing and food handling practices leads to spread and ingestion of the eggs among the population. Ingested eggs are absorbed and carried by the blood to all organs of the body especially the brain, eye, and muscle, where they develop into larva. In the brain these cysts (NCC) are the most common cause of adult onset seizures in endemic countries.

Diagnosis of NCC is typically with a MRI or CT scan in combination with serology. Radio-imaging is expensive and inaccessible to a majority of the developing world and serology lacks sensitivity for infections of a few or single cyst, the predominant infection seen in India. There is need to improve NCC diagnostics in relation to sensitivity and cost. *T. solium* infection activates the immune system and inflammatory biomarkers in blood specific to seizures due to NCC may be useful in this regard. Identification of genes

[*]Department of Biostatistics and Epidemiology, The University of Oklahoma Health Science Center
[†]Department of Statistics, Kansas State University
[‡]Department of Medicine OUHSC and the Veterans Affairs Medical Center, Oklahoma City, OK
[§]Department of Neurological Sciences, The Christian Medical College, Vellore India

in peripheral blood monocytes that are differentially expressed between NCC-associated epilepsy and non NCC-associated epilepsy may provide insight into these inflammatory biomarkers. This would be through micro-array analysis of monocyte messenger RNA (mRNA).

The process of obtaining gene expression data is sophisticated, yet the statistical methods used to analyze such data remain somewhat pedestrian. At a basic level, micro-array gene expression analysis is a multi-step process that begins with non-trivial handling of the sample. This involves isolating peripheral blood monocytes from whole blood by centrifugation followed by separation of monocytes using specific antibody labeled magnetic micro beads. RNA is isolated from the cells and stored at -70^oC until analysis.

The RNA samples are thawed and hybridized to chips containing 20,000-50,000 microscopic wells fitted with specific gene segments. The greater the gene number in the mRNA being studied the greater the hybridization and accompanying fluorophore light signals whose intensity is measured by state-of-the-art laser scanners. A host of normalization routines are employed to standardize the measured intensity levels relative to known "housekeeping" genes that are believed to have extremely specific levels of light emission in all subjects. After all this, a data set containing 20,000-50,000 rows, one for each well on the chip, and columns for each subject in the study is ready for analysis. The cost of obtaining microarray data limits most studies of this kind to investigating these genes for a few subjects in each group which results in high dimensional data sets.

Baldi and Long (2001) point out that often, the first step in the detection of differentially expressed genes is to carry out Student's t-test between groups for each of the 20,000-50,000 genes. They point out that the assumptions of normality and equal variances are often unwarranted with small sample sizes. It is also clear that outliers are problematic here in that they not only exert a strong effect on the mean but they are also difficult to detect. Combining this with multiple comparisons of tens of thousands of hypothesis tests in a single microarray analysis calls into question the validity and interpretability of the p-value. As an alternative to the p-value approach, we present a general semi-parametric Bayesian framework for detecting differentially expressed genes in microarray analyses.

2. Semi-parametric Bayesian Model

In the spirit of Anderson and Dubnicka (2014) we let $S_l^{(j)}$ define the event that a subject belongs to group l for $l = 1, \ldots, s$ and $j = 1, \ldots, n_l$, let $x^{(1)}, \ldots, x^{(p)}$ be a set of p observations that arise in sequence from that subject, and let $x^{(j)}|S_l$ define the conditional event of observing $x^{(j)}$ given the subject belongs to group l. Further suppose the initial prior probability of belonging to group l is given by $P(S_l^{(j)})$ and conditional probabilities corresponding to the conditional events, $P(x^{(j)}|S_l)$, can be computed. Then the posterior probability of belonging to group l having observed $x^{(j)}$ can be computed as

$$P(S_l^{(j)}|x^{(j)}) = \frac{P(S_l^{(j)})P(x^{(-j)}|S_l)}{\sum_{l=1}^{S} P(S_l^{(j)})P(x^{(-j)}|S_l)} \tag{1}$$

where $P(x^{(-j)}|S_l)$ is the estimated probability of observing $x^{(j)}$ obtained from a kernel density estimate of the subjects in group l without the j^{th} subject. More specifically we let $\hat{f}_h(x^{(j)}|S_l)$ be a kernel density estimate of a continuous probability density function $f(x^{(j)}|S_l)$ and we let

$$P(\mathbf{x}^{(-j)}|S_l) = \begin{cases} \int_{-\infty}^{x^{(j)}} \hat{f}_h(x^{(-j)}|S_l)dx^{(-j)} & \text{if } x^{(j)} < x_{med}^{(-j)} \\ \int_{x^{(j)}}^{\infty} \hat{f}_h(x^{(-j)}|S_l)dx^{(-j)} & \text{if } x^{(j)} \geq x_{med}^{(-j)} \end{cases}$$

where $x_{med}^{(-j)} = x$ such that $\int_{-\infty}^{x} \hat{f}_h(x^{(-j)}|S_l)dx^{(-j)} = 1/2$. In this manner observations that lie near the median of the kernel density for a group are given a higher probability of occurrence than those falling further away from the center of the kernel density. Using this leave-one-out approach, equation (1) computes a posterior probability of group affiliation for each subject. If equal priors are used and the genes under consideration are indeed differentially expressed between groups, then the above algorithm results in an increase in the posterior probabilities of the correct group. Large changes indicate a large degree of separation in the gene expression levels between the two groups whereas small changes indicate little distinction between groups. Aggregating the posterior probabilities for subjects from common groups allows us to compute the posterior probability of differential expression (PPDE) using

$$\frac{\sum_{j=1}^{n_l} P(S_l^{(j)}|x_{(j)})}{n_l} \tag{2}$$

We use the mean to aggregate the posteriors in equation (2) but one could have just as easily chosen the median instead. The optimal choice, mean or median, for the PPDE is a subject of continuing research by the authors.

3. Example

3.1 Simulated Data

As an initial comparison of this model's performance against the p-value approach, microarray data, expressed as fold values, were generated for 8 subjects (4 treatment and 4 control) for 50,000 genes under two scenarios. To keep things manageable, 10 genes were selected to be differentially expressed between the control and treatment group from $N(\mu, \sigma^2)$ with mean fold change values of μ ranging from 0.5 to 5 in increments of 0.5.

For scenario 1 we specify all gene fold change values to have normal distributions and equal variance. This setting plays well to the strengths of Student's t-test and provides an indication of how well the proposed semi-parametric model performs when the typical underlying assumptions are met. Specifically, genes 1-10 from the treatment group were generated from $N(5,1)$, $N(4.5,1)$, $N(4,1)$, $N(3.5,1)$, $N(3.1)$,$N(2.5,1)$, $N(2,1)$, $N(1.5,1)$, $N(1,1)$, and $N(0.5,1)$, respectively. The remaning genes for the treatment group, and all genes from the control group, were assigned fold change values from $N(0,1)$.

In scenario 2, all the fold change values were as specified in scenario 1 except one subject was randomly chosen to represent a mild outlier. Specifically, that subject's original fold value was replaced by $Q1 - 2.0(IQR)$ where $Q1$ is the first quartile and IQR is the inter-quartile range of the original observations. This represents an outlier in the treatment group that is more in line with the control group which should challenge both the traditional p-value approach as well as the proposed method.

3.1.1 Simulation Results

Table 1 shows the scenario 1 results of comparing the control and treatment groups using the proposed semi-parametric Bayesian (Bayes) method and the traditional p-value approach. We see the p-value approach finds the 10 differentially expressed genes and orders them correctly from largest to smallest differential expression. The p-value, which is unadjusted for multiple comparisons, indicates that Gene 1 to 7 would be identified as differentially expressed. However, any adjustment to the p-value for multiple comparisons will likely reduce this list further. The proposed method finds 8 of the 10 differentially

Table 1: Top 10 Differentially Expressed Genes: Scenario 1

	P-value		Bayes	
Rank	Gene	p-value	Gene	PPDE
1	Gene 1	0.0004	Gene 1	1.0000
2	Gene 2	0.0007	Gene 2	0.8750
3	Gene 3	0.0013	Gene 3	0.8750
4	Gene 4	0.0026	Gene 4	0.8750
5	Gene 5	0.0054	Gene 6	0.8380
6	Gene 6	0.0123	Gene 5	0.7686
7	Gene 7	0.0300	Gene 7	0.6679
8	Gene 8	0.0781	Gene 8	0.6124
9	Gene 9	0.2070	Gene 15048	0.5197
10	Gene 10	0.5060	Gene 23799	0.4848

Table 2: Top 10 Differentially Expressed Genes: Scenario 2

	P-value		Bayes	
Rank	Gene	p-value	Gene	PPDE
1	Gene 4	0.0025	Gene 4	0.8750
2	Gene 3	0.0033	Gene 1	0.8296
3	Gene 2	0.0062	Gene 2	0.8137
4	Gene 1	0.0130	Gene 3	0.6875
5	Gene 5	0.0823	Gene 7	0.6372
6	Gene 7	0.1140	Gene 29518	0.6080
7	Gene 6	0.1482	Gene 2984	0.6007
8	Gene 25323	0.2156	Gene 13197	0.5982
9	Gene 22823	0.2158	Gene 24925	0.5970
10	Gene 40988	0.2159	Gene 15816	0.5965

expressed genes in the top 10 identified using the PPDE ordered from largest to smallest. A few interesting notes here are that Gene 1 was different enough between the two groups that all subjects were correctly classified to their respective groups and an estimated probability of differential expression of 1 was obtained. Genes 2-4 all have 0.875 probability of differential expression and Genes 5 and 6 have made a curious ordering switch. Also, there is a clear distinction between the PPDE of Gene 8 and the next gene in the list (PPDE=0.6124 and 0.5197, respectively). The former represents about a 61% chance that Gene 8 is differentially expressed between the groups whereas the latter represents only a 50% chance the gene is differentially expressed between the groups. Because the fold change values of the latter gene were created so as to come from a $N(0, 1)$ for both groups, this is not surprising.

Table 2 shows the scenario 2 results of comparing the control and treatment groups in the presence of a mild outlier.

Using the p-value approach with the unadjusted p-values, Genes 1-4 would be identified as differentially expressed. Here again, we see that list would likely be reduced were we to apply an adjustment for multiple comparisons. Depending on the threshold used, the proposed method identifies Genes 1-4 with high probabilities of differential expression.

Table 3: Top 20 Differentially Expressed Genes

	Bayes Approach		P-value Approach	
	Gene	PPDE	Gene	p-value
1	XLOC003809	0.8516	XLOC003809	4.2E-05
2	C3orf58	0.8469	ZNF791	6.1E-05
3	ASAP2	0.8257	LOC100652837	6.5E-05
4	C5orf44	0.8121	C5orf44	1.2E-04
5	ZNF791	0.8091	ASAP2	1.4E-04
6	CCNL1	0.8056	C3orf58	1.6E-04
7	LOC100652837	0.8023	LOC100129960	1.8E-04
8	EIF2A	0.7874	C15orf62	2.0E-04
9	CNRIP1	0.7865	C5orf15	2.2E-04
10	USP6	0.7848	FCGR2A	2.3E-04
11	C15orf62	0.7933	EIF2A	2.7E-04
12	ATXN7	0.7826	CCNL1	2.9E-04
13	LOC100129960	0.7796	USP6	3.0E-04
14	C5orf15	0.7714	MAGI2	3.4E-04
15	KCNJ18	0.7691	CNRIP1	4.1E-04
16	FEZ2	0.7686	Q9UJ41	4.7E-04
17	SLC35A5	0.7639	LOC90784	5.2E-04
18	LOC90784	0.7597	ATXN7	5.5E-04
19	SPAG11B	0.7584	DYNC1LI2	5.8E-04
20	SLC6A19	0.7584	SPAG11B	6.0E-04

3.2 NCC Study Data

In a study funded through a US-India (NIH-DBT) joint partnership, subjects reporting to the Department of Neurological Sciences at the Christian Medical College in Vellore, India were recruited for participation in a study to determine inflammation relevant gene expression profiles specific to NCC in peripheral blood monocytes. 12 NCC-associated epilepsy subjects and 12 epilepsy controls had blood samples drawn and peripheral blood monocyte mRNA extracted. Samples were transported on dry ice to Genotypic (Pvt Ltd) where gene expression levels were measured using GXP microarrays by Agilent which contain approximately 50000 wells per chip. The proposed method of differential expression was carried out using equal prior probabilities and the top 20 genes identified were reported and compared to those obtained using the p-value approach.

3.2.1 Results

Table 3 contain the top 20 differentially expressed genes between the NCC-associated epilepsy and the epilepsy control groups. It is interesting to note the number of genes identified in both lists. In fact, there are only 4 genes (15, 16 17 and 20)on the Bayesian list not appearing on the p-value list and only 5 genes (10, 14, 16, 18, and 19) appearing on the p-value list not appearing on the Bayesian list. The PPDE ranges from about 75% to 85% for the genes in this list.

4. Discussion

In this paper we present a semi-parametric Bayesian frame work for detecting differentially expressed genes that does not rely on the t-test assumption of normality. Aside from the specification of the priors, the kernel, and the bandwidth, this approach is non parametric. As such, it is expected to be somewhat more robust to outliers and departures from normality than the p-value approach. This is an area of continued investigation. Preliminary results indicate that in settings investigated so far, the performance of the proposed method is on par with the p-value approach.

As we continue to challenge the proposed method with various scenarios to discover its strengths and weaknesses, the false discovery rate (FDR) will also be noted for comparison to the p-value approach. This may not only prove useful for establishing a sound rule-of-thumb threshold for concluding differential expression between groups but also elucidate gains achieved by avoiding and explosive type I error rate due to multiple comparisons.

One distinct advantage of the proposed method is the direct interpretability of the PPDE as the posterior probability of differential expression. The uncertainty of the decision to declare differential expression between groups or not is readily provided and intuitively understood, unlike the commonly misused and misinterpreted p-value.

This approach is admittedly computationally intensive. In the limited simulations we have explored, computing the PPDE for a set of 50000 genes ranges between 20=50 minutes. By hand type calculations to obtain the PPDE would not be advantageous due to iteratively computed kernel density estimates using the leave-one-out principle but a software program (to be developed in R) could easily handle these routines for anyone with access to a computer.

We have presented the proposed method in the context of microarray analysis comparing two groups but the methodology readily extends to two other important settings as well. First, this method need not be limited to microarray data but applies more broadly to two sample comparisons of continuous data. The computed posterior would simply be the estimated probability the two groups are detectably different. Finally, these methods easily extend to comparisons of more than two groups with some minor modifications to the specified priors. How multiple groups affect the PPDE is also a matter of continued investigation.

Acknowledgments

Research reported in this publication was supported by the National Institute of Neurological Disorders and Stroke (NINDS) of the National Institutes of Health under award number R2INS077466. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

Anderson, M.P., and Dubnicka, S.R. (2014) A Sequential Nïave Bayes Classifier for DNA barcodes. *Statistical Applications in Genetics and Molecular Biology* 13(4):423-34.

Baldi P and Long A(2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics 17(6):509-519.*