A BAYESIAN APPROACH TO IDENTIFY DIFFERENTIALLY EXPRESSED GENES

Erlandson Ferreira SARAIVA[1]
Teresa Cristina Martins DIAS[2]
Luis Aparecido MILAN[2]

▪ *ABSTRACT: DNA arrays technology has become an important tool for genomic research due its capacity of measuring simultaneously the expression levels of a great number of genes or fragments of genes in different situations. An objective in gene expression data analysis is identifying genes with significant difference between the expression levels in a treatment experimental condition in relation to expression levels in a control experimental condition. We propose a Bayesian approach in order to identify differentially expressed genes based on posterior probability of difference that is calculated using the Bayes factor. The proposed approach is compared to t-test by using artificial data sets and a real data set. Results from simulation show a better performance of the proposed approach in identification of difference of means and/or variance for small samples size, usual in gene expression data analysis. The application to a real data set shows a complementarity among the methods in identification of differentially expressed genes.*

▪ *KEYWORDS: Gene expression; Bayesian approach; posterior probability; Bayes factor.*

1 Introduction

The DNA arrays technology is capable of providing gene expression levels measurements for thousands of genes simultaneously under different experimental conditions. This allows biologists study genome-wide patterns of gene expression and identify possible relationships among genes and among genes and proteins. For

[1]Universidade Federal da Grande Dourados – UFGD, Departamento de Matemática, CEP: 79825-070, Dourados, MS, Brazil. E-mail: *erlandsonsaraiva@ufgd.edu.br*
[2]Universidade Federal de São Carlos – UFSCar, Departamento de Estatística, São Carlos, SP, Brazil. E-mail: *dtmd@ufscar.br/ dlam@ufscar.br*

further discussion and additional references about DNA array technology see Arfin et al. (2000) , DeRisi et al. (1997), Hatfield et al. (2003), Lonnstedt and Speed (2001), Schena et al. (2002).

According to Baldi and Long (2001) and Hatfield et al. (2003) gene expression data can be analyzed on at least three levels of increasing complexity. In the first level, each gene is analyzed separately, where the objective is verify whether the observed expression in treatment experimental condition is significantly different from observed expression in control experimental condition. In the second level, clusters of genes are analyzed in terms of common functionalities and interactions. In the third level, the objective is infer and understand the relationship among genes and proteins.

In this text we focus in the first level of analysis. Under this focus, one of the first proposed approach was the twofold approach, in which, a gene is considered with different expression levels if the averages of treatment and control varies more than a cutoff value equals to 2 (Schena et al. (2002)). But, this approach is not adequate to yield good results since a cutoff value equal to 2 may have different significance for different observed expression levels. Another method used for gene expression data analysis is the t-test, as described in Baldi e Long (2001). The problem with the application of t-tests to this kind of data is the small size of treatment and control samples. This lead to underestimated variances and small power of tests.

We propose a Bayesian approach for gene expression data analysis. In order to identify differentially expressed genes we define for each gene the models M_0 and M_1; M_0 representing no difference, where observed expression levels in treatment and control conditions are considered as being generated from a same distribution; and M_1 representing difference, where the observed expression levels in treatment and control conditions are considered as being generated from different distributions. Then we calculate the posterior probability of difference by using the Bayes factor (BF).

We present a simulation study that show a better performance of the proposed method in relation to t-test in identification of difference of means and/or variance for situations with small sample size, usual in gene expression data analysis. We also apply both methods to a real data set, obtained from the experiment carried through with *Escherichia Coli* bacterium (Arfin et al. (2000)). In the application we also calculate the BF for a nonconjugated case via reversible jump algorithm (Green (1995)).

The paper is organized as follows. In Section 2, we define the models used with the Bayes Factor in order to calculate the posterior probability of difference. In Section 3, we apply BF and t-test to artificial data sets and a real data set. In Section 4, we discuss the results.

2 Model for gene expression data analysis

Consider a DNA array experiment with n genes and two experimental conditions which we name by control (c) and treatment (t). Suppose that control and treatment are replicated n_c and n_t times, respectively. Denote by y_{ig_h} the $i-th$ observed expression level (or its logarithm) for gene g in experimental condition h, $h \in \{c,t\}$ and $g = 1,\ldots,n$.

Consider $\mathbf{y}_{g_h} = \{y_{1g_h},\ldots,y_{n_h g_h}\}$ realizations of independent random variables $\mathbf{Y}_{g_h} = \{Y_{1g_h},...,Y_{n_h g_h}\}$ where $h \in \{c,t\}$ and $g = 1,\ldots,n$.

As is usual in gene expression data analysis, consider that the logarithm of the observed gene expression levels in control and treatment are generated from normal distributions with mean μ_{g_h} and variance $\sigma^2_{g_h}$, $Y_{ig_h} \overset{iid}{\sim} \mathcal{N}(\mu_{g_h}, \sigma^2_{g_h})$, for $i = 1,\ldots,n_h$, $h \in \{c,t\}$ and $g = 1,\ldots,n$ (Baldi and Long(2001), Fox and Dimmic(2006), Hatfield et al. (2003), Medvedovic and Sivaganesan (2002)). Denote parameters by $\theta_{g_h} = (\mu_{g_h}, \sigma^2_{g_h})$ and the parameter space by $\Theta_g = \{\boldsymbol{\theta}_g = (\theta_{g_c}, \theta_{g_t}); \theta_{g_h} \in \mathcal{R} \times \mathcal{R}^+\}$, for $g = 1,\ldots,n$ and $h \in \{c,t\}$. Other distributions, such as gamma or log-normal, could be considered with similar results for this kind of data.

The interest is verify whether gene g presents different gene expression, i.e., if $\theta_{g_t} = \theta_{g_c}$ or $\theta_{g_t} \neq \theta_{g_c}$. Under normality assumption, an usual approach to identify differentially expressed genes is the use of the t-test. This test uses the statistic

$$t_g = \frac{\bar{y}_{g_t} - \bar{y}_{g_c}}{\sqrt{\frac{s^2_{g_t}}{n_t} + \frac{s^2_{g_c}}{n_c}}}$$

which follows the Student's t distribution with

$$df = \frac{\left[\frac{s^2_{g_c}}{n_c} + \frac{s^2_{g_t}}{n_t}\right]^2}{\frac{\left(\frac{s^2_{g_c}}{n_c}\right)^2}{n_c - 1} + \frac{\left(\frac{s^2_{g_t}}{n_t}\right)^2}{n_t - 1}}$$

degrees of freedom, where $\bar{y}_{g_h}$ and $s^2_{g_h}$ are the sample mean and variance for gene g in condition $h \in \{c,t\}$. Fixed a significance level α, if $|t_g|$ is greater than a threshold $t_{1-\frac{\alpha}{2},df}$ (quantile $1-\frac{\alpha}{2}$ of Student's t distribution with df degrees of freedom) then the test conclude for different expressions.

Experimental conditions in gene expression restricts the sample size and this is a drawback for t-test. We present the Bayes factor as an approach that complement t-test and compare their performances.

2.1 Bayesian approach

In order to represent situations with and without difference between treatment and control, consider models M_0 and M_1, such that,

1. Under M_0 there is no difference between control and treatment, i.e., $\theta_{g_t} = \theta_{g_c} = (\mu_g, \sigma_g^2)$. The likelihood function is

$$L_{M_0}(\boldsymbol{\theta}_g|\mathbf{y}_g) \propto \left(\sigma_g^2\right)^{-\frac{n_g}{2}} \exp\left\{-\frac{n_g(\bar{y}_g - \mu_g)^2 + (n_g - 1)s_g^2}{2\sigma_g^2}\right\}, \qquad (1)$$

where $\mathbf{y}_g = \{\mathbf{y}_{g_c}, \mathbf{y}_{g_t}\}$, $n_g = n_c + n_t$ and $\overline{y}_g$ and s_g^2 are the sample mean and variance of the set $\mathbf{y}_g$.

2. Under M_1 there is difference, $\theta_{g_t} \neq \theta_{g_c}$. The likelihood function is

$$L_{M_1}(\boldsymbol{\theta}_g|\mathbf{y}_g) \propto \prod_{h\in\{c,t\}} \left(\sigma_{g_h}^2\right)^{-\frac{n_h}{2}} \exp\left\{-\frac{n_h(\bar{y}_{g_h} - \mu_{g_h})^2 + (n_h - 1)s_{g_h}^2}{2\sigma_{g_h}^2}\right\} \qquad (2)$$

where $\overline{y}_{g_h}$ and $s_{g_h}^2$ are the sample mean and variance of the set $\mathbf{y}_{g_h}$, $h \in \{c, t\}$.

We can identify differentially expressed genes choosing between models M_0 and M_1. We propose doing this by using Bayes Factor (BF). Bayes Factor can be defined by $B_{10} = \frac{I_1}{I_0}$, where

$$I_1 = \int_{\Theta_g} L_{M_1}(\boldsymbol{\theta}_g|\mathbf{y}_g)\pi_1(\boldsymbol{\theta}_g|M_1)d\boldsymbol{\theta}_g,$$

$$I_0 = \int_{\Theta_g^0} L_{M_0}(\boldsymbol{\theta}_g|\mathbf{y}_g)\pi_0(\boldsymbol{\theta}_g|M_0)d\boldsymbol{\theta}_g,$$

$\pi_1(\boldsymbol{\theta}_g|M_1)$ and $\pi_0(\boldsymbol{\theta}_g|M_0)$ are the prior distributions for parameters in models M_1 and M_0 respectively, $\pi(M_1)$ and $\pi(M_0)$ are the prior probabilities for models M_1 and M_0 respectively, and Θ_g^0 is the parameter space under M_0. B_{10} greater than 1 indicates M_1 as the best fitted model and B_{10} less than 1 indicates M_0. For further discussion and additional references about Bayes Factor see Kass and Raftery (1995), Aitkin(1991), Berger and Pericchi (1996), Bartolucci et al. (2006).

Using the Bayes theorem, the posterior probabilities for models M_0 and M_1 are

$$P(M_0|\mathbf{y}_g) = \frac{1}{1 + \frac{\pi(M_1)}{\pi(M_0)}B_{10}} \quad \text{and} \quad P(M_1|\mathbf{y}_g) = \frac{\frac{\pi(M_1)}{\pi(M_0)}B_{10}}{1 + \frac{\pi(M_1)}{\pi(M_0)}B_{10}}. \qquad (3)$$

Thus, if $P(M_1|\mathbf{y}_g) > P_{ref}$, where $P_{ref} \in [0.5, 1)$ is a cutoff value, we choose M_1 and the gene g presents evidence for difference between treatment and control. Otherwise, we choose M_0 and the gene g have no evidence for difference, $g = 1, \ldots, n$.

2.2 Conjugated case

In this section we calculate probabilities in (3) using conjugated prior distributions. Thus, in order to explore the fully conjugation consider the following prior distributions

$$\mu_d|\sigma_d^2, \mu_0 \sim \mathcal{N}\left(\mu_0, \sigma_d^2\right) \quad \text{and} \quad \sigma_d^2|\alpha, \beta \sim \mathcal{IG}(\alpha, \beta),$$

where μ_0, α and β are hiperparameters, for $d \in \{g, g_c, g_t\}$.

Fixing $\pi(M_0) = \pi(M_1) = \frac{1}{2}$, the Bayes factor $B_{10} = \frac{I_1}{I_0}$ can be analytically calculated, where

$$I_1 = \prod_{h=\{c,t\}} \left[\frac{1}{2\beta\pi}\right]^{\frac{n_h}{2}} \left[\frac{1}{(n_h+1)}\right]^{\frac{1}{2}} \frac{\Gamma\left(\alpha+\frac{n_h}{2}\right)}{\Gamma(\alpha)} \left[1 + \frac{(n_h-1)s_h^2}{2\beta} + \frac{n_h(\overline{y}_{g_h} - \mu_{0h})}{2\beta_h(n_h+1)}\right]^{-\alpha-\frac{+n_h}{2}}.$$

and

$$I_0 = \left[\frac{1}{2\beta\pi}\right]^{\frac{n_g}{2}} \left[\frac{1}{(n_g+1)}\right]^{\frac{1}{2}} \frac{\Gamma\left(\alpha+\frac{n_g}{2}\right)}{\Gamma(\alpha)} \left[1 + \frac{(n_g-1)s_g^2}{2\beta} + \frac{n_g(\overline{y}_g - \mu_0)}{2\beta(n_g+1)}\right]^{-\alpha-\frac{n_g}{2}}.$$

The probabilities in (3) are given by

$$P(M_0|\mathbf{y}_g) = \frac{1}{1+B_{10}} \quad \text{and} \quad P(M_1|\mathbf{y}_g) = \frac{B_{10}}{1+B_{10}}. \tag{4}$$

3 Data analysis

In this section, the proposed method is applied to artificial data sets and a real data set. The real data set was extracted from the site www.jbc.org and refers to experiment realized with *Escherichia Coli* bacterium using nylon membranes, described in details by Arfin et al. (2000).

3.1 Artficial data sets

In order to generate the artificial data sets we fix $\mu_{g_c} = -0.04$ and $\sigma_{g_c}^2 = 0.1$. These values are the average of the observed expression levels in experiment carried through with the *Escherichia Coli* bacterium. The sample sizes n_c and n_t used are $n_c = n_t = \{5, 10\}$.

To verify how the method behaves when $\theta_{g_t} = (\mu_{g_t}, \sigma_{g_t}^2)$ moves away from $\theta_{g_c} = (\mu_{g_c}, \sigma_{g_c}^2)$, we simulate its values using

$$\mu_{g_t} = \mu_{g_c} + \delta\sigma_{g_c} \quad \text{and} \quad \sigma_{g_t} = \gamma\sigma_{g_c},$$

for $\delta = \{0.0, 0.25, 0.50, 0.75, 1,\ 1.25, 1.50, 1.75, 2\}$ and $\gamma = \{1, 2, 3\}$.

The hiperparameters were fixed in the follow way: (i) observing the expressions for I_1 and I_0 above, we have the term 2β, thus we set up $\beta = 1/2$; (ii) fixed β we

choose a value for α, so that, $E[\sigma_d^2] = \frac{\beta}{\alpha-1} = \frac{R^2}{2}$, where $R = max(\mathbf{y}_g) - min(\mathbf{y}_g)$ is the length of the interval of variation of the observed data $\mathbf{y}_g$, for $d \in \{g, g_c, g_t\}$ and $g = 1, \ldots, n$. Thus, we obtain $\alpha = 1 + \frac{1}{R^2}$; (iii) The hiperparemeter μ_0 we fix as being the average of the observed in $\mathbf{y}_g$, $\mu_0 = \overline{y}_g$.

The simulation procedure is

(1) For a pair (δ, γ) and a value $n_c = n_t$ fixed, generates $S = 10,000$ samples of $\mathbf{y}_{g_c}$ and $\mathbf{y}_{g_t}$;

(2) Apply BF and consider evidence for difference if $P(M_1|\mathbf{y}_g) > 0.5$;

(3) Determine the proportion of times that difference is identified using BF and t-test,

$$P_{BF} = \frac{\sum_{s=1}^{S} \mathcal{I}(P_s(M_1|\mathbf{y}_g) > 0.5)}{S} \text{ and } P_{t-test} = \frac{\sum_{s}^{S} \mathcal{I}\{p - value_s < 0.05\}}{S},$$

where $\mathcal{I}(\cdot)$ is a indicator function.

We compare performances of BF and t-test using the proportion of cases identified with difference, denoted by "Proportion of identification" in Tables 1 to 4 below. Note that proportion of identification is an estimate of the power test of methods. For $\delta \neq 0$ or $\gamma \neq 1$, as greater is this proportion, better is the performance of the method.

Tables 1 and 2 show performances of BF and t-test in identification of difference, respectively, for $n_c = n_t = 5$. Tables 3 and 4 show for $n_c = n_t = 10$. In these tables, as we move to the right side in each line we increase the distance between control and treatment means. As we move down in columns we increase the variance of the treatment populations in relation to control.

Increasing the value of δ and γ the proportion of identification increases in both directions for BF, as showed by Tables 1 and 3. How much the treatment distribution moves away from the control distribution greater is the proportion of identification of difference. The same does not happen with the t-test, as showed by Tables 2 and 4. For this method the proportion of identification increases only as the value of δ increases, i.e, when the mean of the treatment distribution moves away from the mean of the control distribution. Increasing the variance of treatment (increasing the value of γ) the t-test present a reduction of its performance. In opposite to BF which presents an improve in its performance.

For a sample size $n_c = n_t = 10$ and $\gamma = 1$ fixed, the t-test present better performance than BF for all values of δ used. But, it does not happen for $\gamma = \{2, 3\}$. This shows a better performance of the BF when the difference refer to variance of the variable involved.

Thus, we have that cases with difference of means and similar variances are more easily identified by t-test, while cases with changes in variance with or without expressive changes in mean are more easily identified by BF. These results shows a complementary among the methods. Of the biological point of view it is interesting, because the BF may identify with difference genes which are not identified by t-test, specially, genes with difference of means and variances.

Table 1 - Proportion of identification, $n = 5$, BF

	δ								
γ	0.0	0.25	0.50	0.75	1	1.25	1.50	1.75	2
1	0.052	0,065	0,111	0,190	0,318	0,453	0,618	0,716	0,795
2	0.342	0,340	0,362	0,403	0,461	0,547	0,634	0,752	0,854
3	0.718	0,725	0,751	0,742	0,763	0,784	0,813	0,836	0,866

Table 2 - Proportion of identification, $n = 5$, t-test

	δ								
γ	0.0	0.25	0.50	0.75	1	1.25	1.50	1.75	2
1	0.041	0,056	0,091	0,162	0,259	0,377	0,512	0,646	0,764
2	0.043	0,054	0,080	0,123	0,185	0,271	0,367	0,474	0,582
3	0.046	0,055	0,075	0,105	0,152	0,212	0,284	0,374	0,461

Table 3 - Proportion of identification, $n = 10$, BF

	δ								
γ	0.0	0.25	0.50	0.75	1	1.25	1.50	1.75	2
1	0.018	0,028	0,080	0,206	0,405	0,635	0,816	0,906	0,951
2	0.490	0,497	0,536	0,603	0,688	0,760	0,843	0,927	0,978
3	0.901	0,910	0,911	0,922	0,927	0,942	0,958	0,973	0,981

Table 4 - Proportion of identification, $n = 10$, t-test

	δ								
γ	0.0	0.25	0.50	0.75	1	1.25	1.50	1.75	2
1	0.049	0,078	0,176	0,346	0,557	0,751	0,886	0,961	0,987
2	0.049	0,072	0,133	0,246	0,396	0,571	0,728	0,852	0,929
3	0.048	0,068	0,112	0,195	0,307	0,448	0,597	0,731	0,841

3.2 Real data set

Now consider the gene expression data set on *Escherichia Coli* bacterium described in Arfin et al. (2000). The data set is composed by $n = 434$ genes, each gene g has $n_c = 5$ measurements of expression levels in control condition and $n_t = 5$ in treatment condition.

Results for BF and t-test are presented in Figures 1 and 2, respectively, where "•" indicate genes without evidence for difference and "+" indicate genes with evidences for difference. The BF identifies 32 genes with evidences for difference while t-test identifies 21 genes.

Genes with means well apart are identified by BF but are not by t-test, as can be noted comparing graphics in Figures 1 and 2. An example is gene 277

(LD28084) which has $P(M_1|\mathbf{y}_{277}) = 0,67$ and $p - value(277) = 0.1211$ and are highlighted in graphics. One possible reason for this is the low performance of t-test in situations with difference of means and variances, as observed in artificial data sets. Genes with difference of means and similar variances are easily identified by t-test, it is the case of gene 383 (LP06328) which has $P(M_1|\mathbf{y}_{383}) = 0.75$ and $p - value(383) = 0.0059$ and is highlighted in Figures 1 and 2.

Table 5 shows the names of genes identified with evidences for difference by each method. Only 8 genes are identified with difference by the two methods. The BF identifies 24 genes which were not identified by t-test. As observed in artificial data sets, we have a complementarity among BF and t-test. The BF is capable of identify with difference genes which are not identified by t-test, specially, genes with difference of means and/or variances.

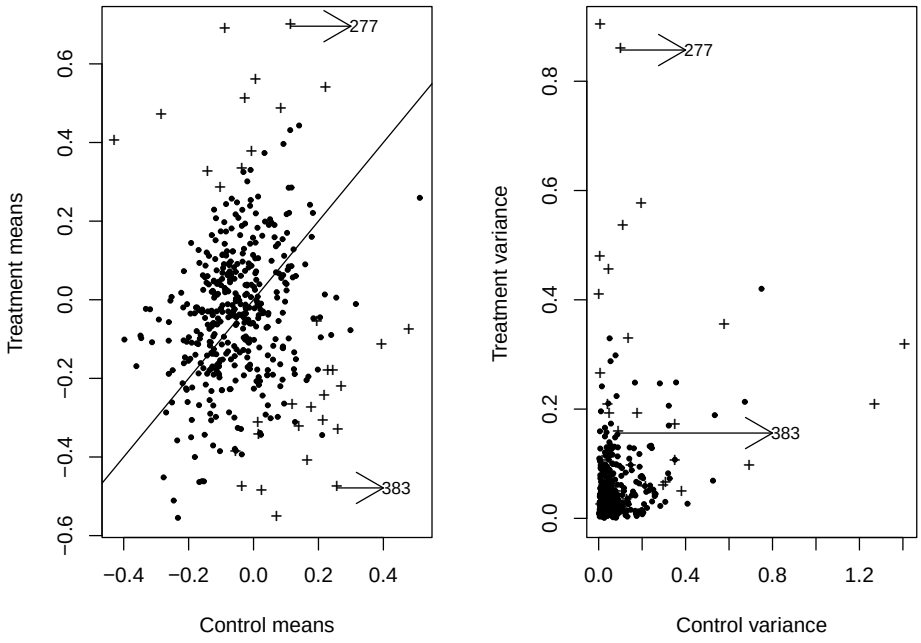


Figure 1 - Control *versus* treatment mean and variance, genes identified by BF. + indicates genes with evidence for difference.

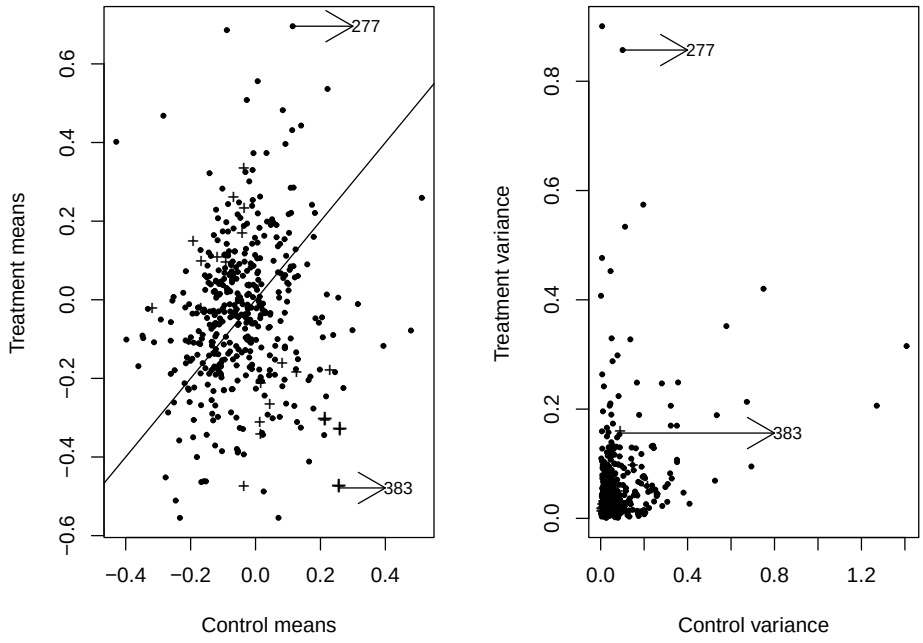


Figure 2 - Control *versus* treatment mean and variance, genes identified by t-test. + indicates genes with evidence for difference.

3.2.1 Nonconjugated case

Consider μ_d and σ_d^2 drawn independently from normal and inverse-gamma prior distributions,

$$\mu_d|\mu_0, \sigma_0^2 \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right) \quad \text{and} \quad \sigma_d^2|\alpha, \beta \sim \mathcal{IG}(\alpha, \beta),$$

where μ_0, σ_0^2, α and β are hiperparameters, for $d \in \{g, g_c, g_t\}$. Here we set up $\sigma_0^2 = s_g^2$ (sample variance) and the others hiperparameters are fixed as described earlier.

In this situation B_{10} cannot be analytically calculated, so, in order to calculate the posterior probability of difference we use an approximation of the BF obtained via reversible jump algorithm (Green (1995), Bartolucci et al. (2006), Chib Jeliazkov (2001)).

In order to simplify the notation consider $\theta_0 = (\mu_g, \sigma_g^2)$ and $\theta_1 = (\mu_{g_c}, \sigma_{g_c}^2, \mu_{g_t}, \sigma_{g_t}^2)$. In the reversible jump algorithm, to ensure reversibility of the Markov chain, we assume that there exist a diffeomorfism for models (M_0, M_1),

Table 5 - Genes identified with difference by BF and t-test

Number	Name	BF	t-test	Number	Name	BF	t-test
02	GH01059		✓	252	GM12762	✓	
12	GH01576		✓	254	GM14215		✓
83	UNK442	✓		256	GM14684	✓	
84	LD29081	✓	✓	259	HL07933		✓
86	LD29155		✓	270	LD22812	✓	
88	LD29175	✓		277	LD28084	✓	
98	LD29815		✓	309	LD37158	✓	
121	GH01770	✓		311	LD37196	✓	
133	GH05818	✓		324	LD42146	✓	
134	GH05991	✓	✓	328	LD42193	✓	
176	GH14695	✓	✓	332	LD42381	✓	
179	GH14758	✓		345	LD46223		✓
183	GH18838		✓	360	LD46862	✓	
193	GH19147	✓		375	LP05570	✓	✓
201	GH19593	✓		376	LP05614	✓	
212	GH22736	✓	✓	383	LP06328	✓	✓
213	GH22765	✓		386	SD03609		✓
227	GH25635		✓	393	SD03844		✓
233	GH25848	✓		410	SD06609		✓
244	GH26184		✓	415	SD06784	✓	
247	GH26265	✓		417	SD06811	✓	
248	GH26280	✓	✓	427	SD07170	✓	✓
251	GM12474	✓					

$\theta_1 = h(\theta_0, \mathbf{u})$ from $(\theta_0, \mathbf{u})$ to θ_1, where $\mathbf{u}$ is a suitable vector of auxiliary variables defined in a way that θ_1 and $(\theta_0, \mathbf{u})$ have the same dimension. Thus, if the current state of the Markov chain is $(\theta_0, \mathbf{u})$, model M_0, a new state θ_1, model M_1, is proposed with probability $p_{1|0}$ by generating $\mathbf{u}$ from a suitable proposal distribution $q(\mathbf{u})$ and doing $\theta_1 = h(\theta_0, \mathbf{u})$. The proposed move is accepted with probability $\Psi[M_1|M_0] = min(1, A_{1|0})$, where

$$A_{1|0} = \frac{L_{M_1}(\theta_1|\mathbf{y})}{L_{M_0}(\theta_0|\mathbf{y})} \frac{\pi(\theta_1)}{\pi(\theta_0)} \frac{\pi(M_1)}{\pi(M_0)q(\mathbf{u})} \frac{p_{1|0}}{p_{0|1}} |J(\theta_0, \mathbf{u})|, \tag{5}$$

$J(\theta_0, \mathbf{u})$ is the Jacobian of the transformation that arise from the change of variables from $(\theta_0, \mathbf{u})$ to θ_1 and $p_{a|b}$ is the probability to propose model M_a when the current model is M_b, $a, b \in \{0, 1\}$. For simplicity we fix $\pi(M_0) = \pi(M_1) = 0.5$ and $p_{1|0} = p_{0|1} = 0.5$. The reverse move, M_1 to M_0, is obtained similarly.

After a large number of iterations L, $P(M_1|\mathbf{y})$ is estimated by the number of times n_1 that chain visited the model M_1, divided by L. Thus, the standard

estimator of B_{10} based on reversible jump output is given by

$$\widetilde{B}_{10} = \frac{n_1}{n_0}$$

where n_0 he number of times that chain visited the model M_0 in L iterations.

Given $\widetilde{B}_{10}$ we estimate probabilities of models M_0 and M_1 by

$$\widetilde{P}(M_0|\mathbf{y}_g) = \frac{1}{1+\widetilde{B}_{10}} \quad \text{and} \quad \widetilde{P}(M_1|\mathbf{y}_g) = \frac{\widetilde{B}_{10}}{1+\widetilde{B}_{10}}. \tag{6}$$

Our movement of model M_0 to M_1 begins generating a bi-dimensional random vector $\mathbf{u} = (u_1, u_2)$ to specify the new parameters. We generate this values from prior distributions

$$u_1 \sim \mathcal{N}(\mu_0, \sigma_0^2) \ \text{ and } \ u_2 \sim \mathcal{IG}(\alpha, \beta)$$

and setting up

$$\mu_{g_c} = u_1, \ \ \mu_{g_t} = \mu_g, \ \ \sigma^2_{g_c} = u_2 \ \text{ and } \ \sigma^2_{g_t} = \sigma^2_g.$$

Thus, the ratio of prior densities (eq 5) cancel with candidate-generating density $q(\mathbf{u})$ and $J(\theta_0, \mathbf{u}) = 1$. The acceptance probability is given by $\Psi[M_1|M_0] = min(1, A_{1|0})$ where $A_{1|1} = \frac{L_{M_1}(\theta_1|\mathbf{y})}{L_{M_0}(\theta_0|\mathbf{y})}$.

The movement of M_1 to M_0 is obtained doing $\mu_g = \frac{\mu_{g_c}+\mu_{g_t}}{2}$ and $\sigma^2_g = \sigma^2_{g_c}$. The acceptance probability is given by $\Psi[M_0|M_1] = min(1, A_{0|1})$ where $A_{0|1} = \frac{L_{M_1}(\theta_1|\mathbf{y})}{2L_{M_0}(\theta_0|\mathbf{y})}$.

We apply this approach to *Escherichia Coli* data set, in which, for each gene g we obtain the estimate $\widetilde{B}_{10}$ using $L = 11,000$ iterations and a *burn in* $B = 1,000$. Results are showed in Figure 3. As expected, all genes identified with difference by B_{10} are also identified when we use $\widetilde{B}_{10}$. With $\widetilde{B}_{10}$ 33 genes are identified with evidence for difference. The gene 277 that is not is identified with evidence for difference by B_{10} ($P(M_1|\mathbf{y}_{227}) = 0.4193$) is considered with evidence for difference by $\widetilde{B}_{10}$ ($\widetilde{P}(M_1|\mathbf{y}_{227}) = 0,51$). This gene is also identified with difference by t-test, $p - value(227) = 0.0237$.

In order to verify convergency for some cases, we plot estimates of $\widetilde{P}(M_0|\mathbf{y})$ and $\widetilde{P}(M_1|\mathbf{y})$ across iterations, for genes 100 (without difference) and 277 (with difference), as showed by Figure 4. Note that, the number of iterations and *burn in* used is adequate to achieve stability for posterior probability of models. The acceptance ratio is of $28,52\%$ for gene 100 and $11,61\%$ for gene 277. For moves with dimension changes these proportions are satisfactory.

These results show that BF can be used for gene expression data. A disadvantage of the $\widetilde{B}_{10}$ in relation to B_{10} is the computational cost and the dependence of the specification of "good" jumping movements, need in the reversible jump algorithm.

These results show little difference between conjugated and nonconjugated prior cases in what concerns genes identified but a great difference in computational cost.

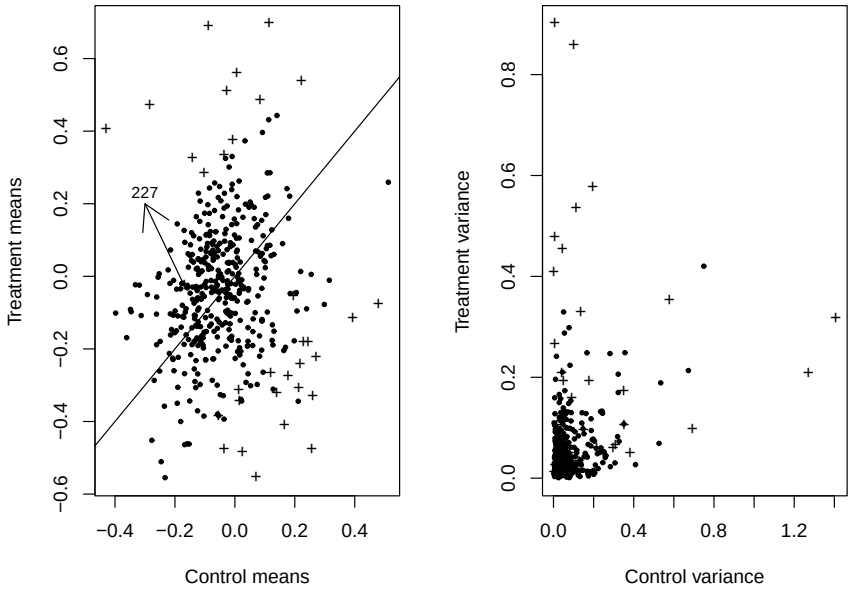
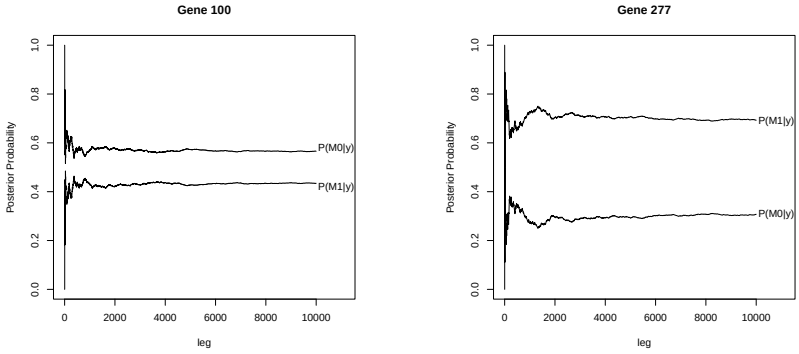


Figure 3 - Control *versus* treatment mean and variance, genes identified using $\widetilde{B}_{10}$. + indicates genes with evidence for difference.



(a) gene 100 (b) gene 277

Figure 4 - Posterior probability for models M_0 and M_1 across iterations

4 Discussion

We propose a Bayesian approach for identification of differentially expressed genes based on Bayes factor. We illustrate the application of the method using a

conjugate and nonconjugate cases, using analytical and numerical solution. The reversible jump algorithm is used to obtain the BF for the nonconjugate case.

In order to verify the performance of the BF and compare with t-test, the usual approach, we apply both methods to artificial data sets and a real data set. Results from artificial and real data sets show a better performance of BF in relation to t-test in identification of difference, mainly, in in situations with different variances.

Of the statistical point of view the method is a simple calculation of the posterior probability of difference, but results obtained show that method may identify genes which are not identified by the usual approach. The biological interest in this fact is that BF brings to light genes that are not identified when using only t-test.

Acknowledgements

The first author acknowledge support from CAPES-Brazil.

SARAIVA, E. F.; DIAS, T. C. M.; MILAN, L. A. Uma abordagem bayesiana para identificar genes diferencialmente expressos. *Rev. Bras. Biom.,* So Paulo, v.8, n.3, p.135-148, 2010.

- RESUMO: A tecnologia dos microarranjos de DNA uma ferramenta importante para a pesquisa genômica, devido a sua capacidade de medir simultaneamente os níveis de expressão de uma grande quantidade de genes ou fragmentos de genes. Um dos objetivos da análise de dados de expresso gênica a identificação de genes com níveis de expressão diferentes em condições experimentais de tratamento e controle. Neste texto, propomos uma abordagem bayesiana para identificar genes diferencialmente expressos com base na distribuição de probabilidade a posteriori, utilizando o fator de Bayes. A abordagem proposta é comparada a metodologia utilizada usualmente, que é o teste t, através de um estudo de simulação. Os resultados de simulação mostram um melhor desempenho da abordagem proposta na identificação de diferença de médias e/ou variâncias, especialmente para casos de pequenas amostras, usual em análise de dados de expressão gênica. A aplicação a um conjunto de dados reais mostra uma complementaridade entre os métodos na identificação de genes diferencialmente expressos.

- PALAVRAS-CHAVE: Expressão gênica, Abordagem Bayesiana, Fator de Bayes.

References

AITKIN, M. Posterior Bayes factor. *J. R. Stat. Soc. Ser. B*, Oxford, v.53, n.1, p.111-142, 1991.

ARFIN, S. M.; LONG, A. D.; ITO, E. T.; TOLLERI, L.; RIEHLE, M. M.; PAEGLE, E. S.; HATFIELD, G. W. Global gene expression profiling in Escherichia Coli K12. *J. Biol. Chem*, Bethesda, v.275, n.38, p.29672-29684, 2000.

BALDI, P.; LONG, D. A. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, Oxford, v.17, n.6, p.509-519, 2001.
BARTOLUCCI, F.; SCACCIA; L.; MIRA, A. Efficient Bayes factor estimation from the reversible jump output *Biometrika*, v.93, n.1, p.41-52, 2006.
BERGER, J. O.; PERICCHI, L. The intrinsic Bayes factor for model selection and prediction. *J. Am. Stat. Assoc.*, Alexandria, v.91, n.433, p.109-122, 1996.
CHIB, S.; JELIAZKOV, I. Marginal likelihood from the Metropolis-Hastings output *J. Am. Stat. Assoc.*, Alexandria, v.96, n. 453, p.270-281, 2001.
DERISI, J. L.; IYER, V. R.; BROWN, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, London, v.278, n.5338, p.680-68, 1997.
FOX, R. J.; DIMMIC, M. W. A two-sample Bayesian t-test for microarray data. *BMC Bioinform.*, London, v.7, p.126, 2006.
GREEN, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, London, v.87, n.4, p.711-732, 1995.
HATIFIELD, G. W.; HUNG, S.; BALDI, P. Differential analysis of DNA microarray gene expression data. *Mol. Microbiol.*, New York, v.47, n.4, p.871-877, 2003.
KASS, R.; RAFTERY, A. Bayes Factor. *J. Am. Stat. Assoc.*, Alexandria, v.90, n.430, p.773-795, 1995.
LONNSTEDT, I.; SPEED, T. P. Replicated microarray data. *Stat. Sinica*, Taipei, v.12, p.31-46, 2001.
MEDVEDOVIC, M.; SIVAGANESAN. Bayesian Infinite Mixture Model Based clustering of Gene Expression Profiles. *Bioinform.*, Oxford, v.18, n.9, p.1194-1206, 2002.
SCHENA, M.; SHALON, D.; DAVIS, R. W.; BROWN, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, London, v.270, n.5235, p.467-470, 1995.

Received in 03.02.2010.

Approved after revised in 02.08.2010.