Research Article

An Empirical Bayesian Method for Detecting Differentially Expressed Genes Using EST Data

Na You, Junmei Liu, and Chang Xuan Mao

Department of Statistics, University of California, Riverside, CA 92521, USA

Correspondence should be addressed to Chang Xuan Mao, cmao@stat.ucr.edu

Received 18 September 2007; Accepted 22 December 2007

Recommended by Xinping Cui

Detection of differentially expressed genes from expressed sequence tags (ESTs) data has received much attention. An empirical Bayesian method is introduced in which gene expression patterns are estimated and used to define detection statistics. Significantly differentially expressed genes can be declared given detection statistics. Simulation is done to evaluate the performance of proposed method. Two real applications are studied.

Copyright © 2008 Na You et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

It is important to detect differentially expressed genes, for example, exploring the key genes related to certain diseases. As the EST sequencing technology develops, a large number of EST databases from a variety of tissues are available. Enormous EST collections provide opportunities to quantify gene expression levels [1]. Efficient statistical methods are in great demand.

Several methods have been proposed to detect significantly differentially expressed (SDE) genes from EST data [2]. Fisher's exact test was used by the Cancer Genome Anatomy Project [3]. Audic and Claverie [4] developed a Bayesian method. GT statistic [5] and R statistic [6] were proposed for multilibrary comparison. In each method, gene-specific detection statistics quantify differences of gene expression levels and SDE genes are declared by their rankings.

An empirical Bayesian method is proposed to detect SDE genes. The relative gene expression abundances are estimated in each library, and a new detection statistic is derived for each gene. In Section 2, simulation experiments suggest that the proposed method outperforms those existing methods. Real applications are also studied in Section 2. Statistical methods are described in Section 3. The possibility of extending the method for multiple libraries is indicated in Section 4.

2. RESULTS

Let $(\pi_{11}, \pi_{12}, \dots, \pi_{1c})$ and $(\pi_{21}, \pi_{22}, \dots, \pi_{2c})$ be the gene expression patterns in two libraries, where π_{ji} is the relative abundance of gene i in library j . The absolute difference between relative abundances is $D_i = |\pi_{1i} - \pi_{2i}|$. Given a sample of ESTs from library j , an empirical Bayes estimator $\hat{\pi}_{ji}$ for π_{ji} is defined in Section 3. Given gene i seen in both samples, define $\hat{D}_i = |\hat{\pi}_{1i} - \hat{\pi}_{2i}|$. Given gene i seen in only one sample, for example, sample 2, define $\hat{D}_i = |\hat{\pi}_{1i} - \hat{\pi}_{2i}|$ if $\hat{\pi}_{1i} < \hat{\pi}_{2i}$ and $\hat{D}_i = 0$ otherwise, which is conservative in the sense that $\hat{D}_i$ possibly underestimates D_i . Gene i is declared to be SDE if $\hat{D}_i$ is relatively large.

2.1. Simulation

In a simulation experiment, EST frequencies are generated from a multinomial distribution with sample size s_j and probability vector $(\pi_{j1}, \pi_{j2}, \dots, \pi_{jc})$, where $c = 1000$, $\pi_{ji} = \lambda_{ji} / \sum_{k=1}^c \lambda_{jk}$, $(\lambda_{11}, \lambda_{12}, \dots, \lambda_{1c})$ from G_1 , $(\lambda_{21}, \lambda_{22}, \dots, \lambda_{2c})$ from G_2 , and G_1 and G_2 are two distributions over $(0, \infty)$. The proposed methods, Fisher's exact test, χ^2 test, AC statistic, and R statistic, are studied. Given a cutoff point τ , the efficiency of a statistical method is measured by p_τ , the expected percentage of the true first τ SDE genes being correctly declared as the first τ SDE genes. The average of estimated p_τ is calculated from 500 replications.

FIGURE 1: Simulation results of Fisher's exact test ($\circ$), χ^2 test (Δ), AC statistic (+), R statistic ($\times$), and the proposed statistic ($\bullet$) in detecting SDE genes using two EST samples of the same size.

In the first four experiments, $s_1 = s_2 = 2000$ and the results are presented in Figure 1. Note that $G_1 = U(0, 10)$, $\text{Beta}(2, 5)$, $0.2\delta(2) + 0.4\delta(5) + 0.2\delta(10)$, $\text{Gamma}(3, 0.1)$ and $G_2 = \text{Beta}(2, 1)$, $\text{Beta}(2, 5)$, $\text{Beta}(2, 2)$, $\text{Beta}(2, 2)$, respectively, where $U(a, b)$ is the uniform distribution on (a, b) , $\delta(a)$ is degenerate at a , $\text{Beta}(a, b)$ is transformed from the beta distribution with shape parameters a and b by $\lambda = p/(1 - p)$ for $p \in (0, 1)$, and $\text{Gamma}(a, b)$ is the gamma distribution with shape a and scale b . For each cutoff point $\tau = 10, 20, \dots, 100$, p_τ are calculated. Clearly the proposed method has better performance than others.

In the second four experiments, $(s_1, s_2) = (2000, 4000)$, $(4000, 2000)$, $(2000, 4000)$, and $(4000, 2000)$, respectively, and the results are presented in Figure 2. Note that $G_1 = \text{Gamma}(3, 0.1)$ and $G_2 = \text{Beta}(2, 2)$ in Figures 2(a) and 2(b) and $G_1 = U(0, 10)$ and $G_2 = \text{Beta}(2, 1)$ in Figures 2(c) and 2(d). The proposed method is usually the best one among all methods studied.

2.2. Real applications

One example concerns Chinese spring wheat drought stressed leaf cDNA library (7235) and root cDNA library (#ASP), available at TIGR gene indexes database (down-

loaded at <http://www.tigr.org/tdb/tgi>, 01/06/2006). In each EST sample, there are totally 790 and 1306 sequenced ESTs, respectively. After removing the unannotated 103 and 194 ESTs, the annotated ESTs are clustered into 465 and 804 groups with each group associated with a unique gene. Only those well-annotated ESTs are used. The first 20 SDE genes by the proposed method are listed in Table 1, among which 7, 7, 7, and 7 genes are in the set of first 20 SDE genes by Fisher's exact test, χ^2 test, AC statistic, and R statistic, respectively.

Another example concerns pinus gene expression level comparison in root gravitropism April 2003 test library (#FH3) and root control 2 (late) library (#FH4), also from TIGR, in which 2513 and 1132 ESTs associated with 1211 and 605 genes are well annotated and clustered. Table 2 lists the first 20 SDE genes by the proposed method, among which 4, 4, 5, and 3 genes are in the set of the first 20 SDE genes by Fisher's exact test, χ^2 test, AC statistic, and R statistic, respectively.

3. METHODS

Suppose that there are c genes in a library. Let x_i be the number of ESTs from gene i , a Poisson variable with mean λ_i .

FIGURE 2: Simulation results of Fisher's exact test ($\circ$), χ^2 test (Δ), AC statistic (+), R statistic ($\times$), and the proposed statistic ($\bullet$) in detecting SDE genes using two EST samples of different sizes.

Given a prior distribution G on the λ_i , the posterior mean of λ_i is $E(\lambda_i | x_i) = (x_i + 1)h_G(x_i + 1)/h_G(x_i)$, where $h_G(x) = \int \lambda^x/x! e^{-\lambda} dG(\lambda)$ is a Poisson mixture. A gene is observed if and only if $x_i \geq 1$. Conditioning on $x_i \geq 1$, x_i follows a zero-truncated Poisson mixture $h_G(x)/(1 - h_G(0))$ or a mixture $f_Q(x)$ of truncated Poisson, where

$$f_Q(x) = \frac{h_G(x)}{1 - h_G(0)} = \int \frac{\lambda^x}{x!(e^\lambda - 1)} dQ(\lambda), \quad (1)$$

$$dQ(\lambda) = \frac{(1 - e^{-\lambda})dG(\lambda)}{\int (1 - e^{-\eta})dG(\eta)}.$$

Let $\theta(Q) = h_G(0)/(1 - h_G(0))$ be the odds that a gene is unseen. Write $E(\lambda_i | x_i) = f_Q(1)/\theta(Q)$ if $x_i = 0$ and $E(\lambda_i | x_i) = (x_i + 1)f_Q(x_i + 1)/f_Q(x_i)$ otherwise.

Let n_x denote the number of genes with exactly x ESTs in the sample. The nonparametric maximum likelihood estimator $\hat{Q}$ for Q is

$$\hat{Q} = \operatorname{argmax}_{x \geq 1} \sum n_x \log f_Q(x), \quad (2)$$

whose calculation is discussed in [7]. It is difficult to estimate $\theta(Q)$ well [8]. There are lower bound estimators, for example, $\widehat{\theta(Q)} = n_1(n_1 - 1)/\{2n(n_2 + 1)\}$ [9], where $n = \sum_{x \geq 1} n_x$ is the number of observed expressed genes. An empirical Bayes estimator for λ_i is

$$\hat{\lambda}_i = E(\lambda_i | x_i) = \begin{cases} \frac{f_{\hat{Q}}(1)}{\widehat{\theta(Q)}}, & x_i = 0, \\ \frac{(x_i + 1)f_{\hat{Q}}(x_i + 1)}{f_{\hat{Q}}(x_i)}, & x_i \geq 1. \end{cases} \quad (3)$$

As the relative abundance π_i satisfies $\pi_i = \lambda_i / \sum_{k=1}^c \lambda_k$, let $\hat{\pi}_i = \hat{\lambda}_i / \hat{s}$, where

$$\hat{s} = \sum_{k=1}^{\hat{c}} \hat{\lambda}_k = n f_{\hat{Q}}(1) + \frac{\sum_{x \geq 1} n_x (x + 1) f_{\hat{Q}}(x + 1)}{f_{\hat{Q}}(x)}, \quad (4)$$

$$\hat{c} = n \{1 + \widehat{\theta(Q)}\}.$$

TABLE 1: The first 20 SDE genes in wheat leaf and root libraries by the proposed method (x_{1i} -the EST number of gene i from leaf library, x_{2i} -that from root library, 0/1-absence/presence in the set of the first 20 SDE genes).

Gene	x_{1i}	x_{2i}	$1000D_i$	Fisher	χ^2	AC	R
TC24953	19	0	27.10	1	1	1	1
TC23443	8	2	7.88	1	1	1	1
TC23215	1	8	4.87	0	0	0	0
TC26419	1	8	4.87	0	0	0	0
TC26431	1	8	4.87	0	0	0	0
TC24980	5	0	3.40	1	1	1	1
TC23786	0	6	2.62	0	0	0	0
TC26436	0	6	2.62	0	0	0	0
TC24819	0	6	2.62	0	0	0	0
TC26455	7	12	1.85	0	0	0	0
TC23314	1	5	1.59	0	0	0	0
TC24981	1	5	1.59	0	0	0	0
TC24795	0	5	1.57	0	0	0	0
TC24804	0	5	1.57	0	0	0	0
TC26553	0	5	1.57	0	0	0	0
TC26356	4	1	1.37	0	0	0	0
TC23560	4	0	1.37	1	1	1	1
TC24669	4	0	1.37	1	1	1	1
TC24679	4	0	1.37	1	1	1	1
TC26379	4	0	1.37	1	1	1	1

TABLE 2: The first 20 SDE genes in #FH3 and #FH4 by the proposed method (x_{1i} -the EST number of gene i from #FH3, x_{2i} -that from #FH4, 0/1-absence/presence in the set of the first 20 SDE genes).

Gene	x_{1i}	x_{2i}	$1000D_i$	Fisher	χ^2	AC	R
TC40351	4	9	7.62	1	1	1	1
TC40355	6	10	6.25	1	1	1	0
TC51779	19	2	5.12	0	0	1	0
TC40566	7	7	4.03	0	0	0	0
TC51682	7	7	4.03	0	0	0	0
TC46290	14	2	3.79	0	0	0	0
TC46372	15	5	3.70	0	0	0	0
TC40768	13	3	3.40	0	0	0	0
TC40912	13	4	3.36	0	0	0	0
TC51995	12	3	2.94	0	0	0	0
TC40420	12	5	2.56	0	0	0	0
TC51708	18	12	2.44	0	0	0	0
TC40405	11	4	2.43	0	0	0	0
TC40361	0	6	2.34	1	1	1	1
TC46426	0	6	2.34	1	1	1	1
TC46276	19	12	2.12	0	0	0	0
TC40388	9	1	1.82	0	0	0	0
TC40647	9	2	1.82	0	0	0	0
TC40350	9	3	1.81	0	0	0	0
TC40731	8	2	1.63	0	0	0	0

4. DISCUSSION

A new statistical method is proposed to compare the gene expression patterns in two cDNA libraries. It can be extended to multilibrary comparison, for example, considering all pairwise comparisons among multiple libraries [3].

REFERENCES

- [1] S. Mekhedov, O. M. de Ilárduya, and J. Ohlrogge, "Towards a functional catalog of the plant genome. A survey of genes for lipid biosynthesis," *Plant Physiology*, vol. 122, no. 2, pp. 389–402, 2000.
- [2] C. Romualdi, S. Bortoluzzi, and G. A. Danieli, "Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests," *Human Molecular Genetics*, vol. 10, no. 19, pp. 2133–2141, 2001.
- [3] C. O'Brien, "Cancer genome anatomy project launched," *Molecular Medicine Today*, vol. 3, no. 3, p. 94, 1997.
- [4] S. Audic and J.-M. Claverie, "The significance of digital gene expression profiles," *Genome Research*, vol. 7, no. 10, pp. 986–995, 1997.
- [5] L. D. Greller and F. L. Tobin, "Detecting selective expression of genes and proteins," *Genome Research*, vol. 9, no. 3, pp. 282–296, 1999.
- [6] D. J. Stekel, Y. Git, and F. Falciani, "The comparison of gene expression from multiple cDNA libraries," *Genome Research*, vol. 10, no. 12, pp. 2055–2061, 2000.
- [7] C. X. Mao, "Inference of the number of species geometric lower bounds," *Journal of American Statistical Association*, vol. 101, no. 476, pp. 1663–1670, 2006.
- [8] C. X. Mao and B. G. Lindsay, "Estimating the number of classes," *Annals of Statistics*, vol. 35, no. 2, pp. 917–930, 2007.
- [9] A. Chao, "Nonparametric estimation of the number of classes in a population," *Scandinavian Journal of Statistics*, vol. 11, no. 4, pp. 265–270, 1984.