BIOINFORMATICS

Vol. 19 no. 18 2003, pages 2461–2464
DOI: 10.1093/bioinformatics/btg357

Using credibility intervals instead of hypothesis tests in SAGE analysis

Ricardo Z.N. Vêncio[1,*], Helena Brentani[2,3] and Carlos A.B. Pereira[1]

[1]Departamento de Estatística, Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo 04601-003, Brazil, [2]Fundação Antônio Prudente, São Paulo 01509-900, Brazil and [3]Ludwig Institute for Cancer Research, São Paulo 01509-010, Brazil

Received on March 26, 2003; revised on June 17, 2003; accepted on July 17, 2003

ABSTRACT

Motivation: Statistical methods usually used to perform Serial Analysis of Gene Expression (SAGE) analysis are based on hypothesis testing. They answer the biologist's question: 'what are the genes with differential expression greater than r with P-value smaller than P?'. Another useful and not yet explored question is: 'what is the uncertainty in differential expression ratio of a gene?'.

Results: We have used Bayesian model for SAGE differential gene expression ratios as a more informative alternative to hypothesis tests since it provides credibility intervals.

Availability: The model is implemented in *R* statistical language script and is available under GNU/GLP copyleft at supplemental web site.

Supplementary information: http://www.ime.usp.br/~rvencio/SAGEci/

Contact: rvencio@ime.usp.br

INTRODUCTION

In Statistics it is known that there is a close relationship between Hypothesis Test and Confidence/Credibility Intervals. However, the dominance of the first approach in Serial Analysis of Gene Expression (SAGE) (Velculescu *et al.*, 1995) research field is quite obvious. See Ruijter *et al.* (2002) for a good review.

In spite of using different strategies, the goal of hypothesis testing methods is to reject/accept the rule for H_0: 'the gene G is not differentially expressed between SAGE libraries' with an associated significance level. Users are able to consider only genes with expression ratios greater than some arbitrary ratio r, with reliability expressed by P-values smaller than an arbitrarily chosen P.

Unfortunately, these hypothesis testing approaches lose one potentially relevant quantitative aspect of SAGE technology since they deal only with punctual estimation for 'expression ratio' random variable. See Stollberg *et al.* (2000) for a good survey into several SAGE quantitative issues.

Inherent technical limitations from SAGE methodology, as GC content bias (Margulies *et al.*, 2001), sequencing errors (Stern *et al.*, 2003) as well as the possibility of non-unique tags, could be detected/corrected using some bioinformatics procedures. As these systematic errors need to be handled, we believe that random sampling errors should have the same attention. Posterior analyses could be highly improved with knowledge of the error-bars. We believe that it could be intuitive for biologists to get their answers about a gene's relative expression as a number with error-bar, i.e. estimation by interval. This motivated us to use a model that gives credibility intervals for SAGE analysis.

SYSTEMS AND METHODS

Let $\pi \in [0; 1]$, the unknown parameter of interest, be the abundance of some gene G in the whole SAGE library studied, and let $N/T = p$ be its usual estimate, where N are the counts for G tag and T the total number of sequenced tags. The union of several related SAGE libraries (same cancerous tissue, for example) could be regarded in this model as one single library with N and T as the sum of component libraries.

For a given gene G, counting tags is a Bernoulli Process that generates our observation (N, T). Thus the likelihood function is:

$$L(\pi|N,T) \propto \pi^N (1-\pi)^{(T-N)} \tag{1}$$

The natural choice for an a priori distribution of π is the Beta(a, b) distribution because it belongs to conjugate class of distributions:

$$\pi \sim \text{Beta}(a,b) \tag{2}$$

This leads to the well-known Bayesian result:

$$\pi|N,T \sim \text{Beta}(N+a, T-N+b) \tag{3}$$

When biologists lack prior knowledge about gene abundance, π can assume equally any value in [0; 1] and $a = b = 1$, meaning the non-informative uniform prior. With these particular parameters equation (3) is the likelihood and reaches

*To whom correspondence should be addressed.

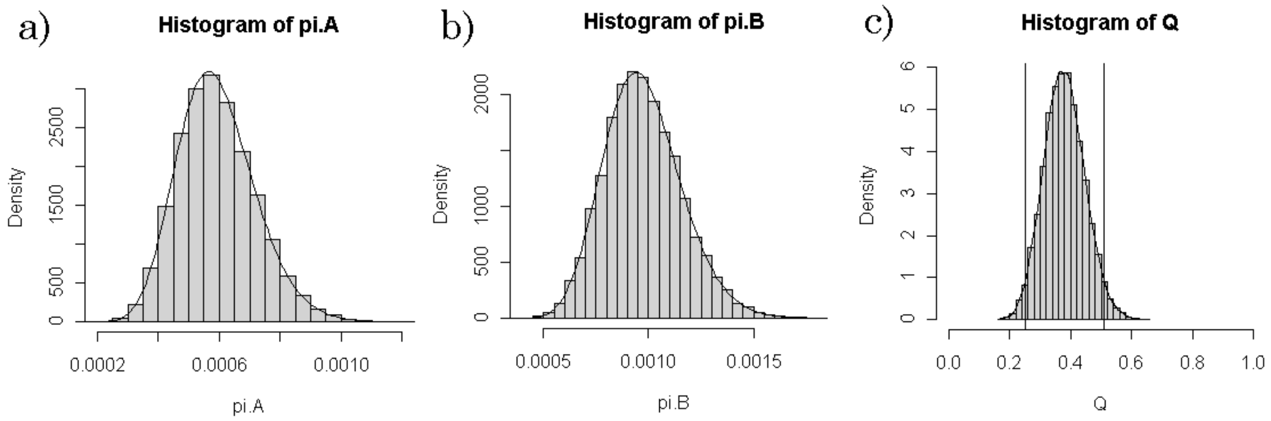

Fig. 1. Example of analysis flow. Gene example: HLA-C, `GTGCGCTGAG` tag, $N_A = 21$, $N_B = 27$, $T_A = 37\,121$, $T_B = 28\,719$, non-informative a priori, $k = 20\,000$, and $q = 0.95$. Frame **(a)** and **(b)** show results for Beta random variates (stage i of the algorithm described in the text). Solid lines are the theoretically known distribution. Frame **(c)** shows result for Q distribution. Solid line is the Kernel Density estimation (algorithm's stage ii) and the vertical lines delimit the 95% credibility interval [0.23; 0.50] (algorithm's stage iv). Since biologists familiar ratio is $R = Q/(1 - Q)$, the interval is presented as [0.3; 1.0]. This example gene is considered not differentially expressed since the interval contains ratio one. Rising credibility stringency expand the interval, but at lower credibility this example could be considered differentially expressed.

its maximum when $\pi = N/T = p$. It is possible to incorporate biologists' prior knowledge (Kuznetsov *et al.*, 2002) into the model by means of a different a priori distribution. The non-informative priori is well suited because SAGE problems typically have sufficient observations to overcome typical priori assessment issues.

As usual, we define relative gene expression between two libraries A and B as the ratio $R = \pi_{\mathrm{A}}/\pi_{\mathrm{B}}$. π_{A} and π_{B} are completely independent.

It is not easy to derive an analytical form for the posteriori probability distribution of R random variable, so we propose a computational approach:

(i) draw from a pseudo-random number generator k samples following a Beta distribution described in equation (3) for A and B studied SAGE libraries;

(ii) for all $j \in [1; k]$ samples from step (i) calculate the ratios $R_j, \log_2(R_j)$, or $1/(1 + 1/R_j)$;

(iii) estimate the probability distribution of (ii) with Kernel Density Estimator;

(iv) integrate around the kernel estimated distribution's mode until the desired probability (credibility level) q is reached.

Figure 1 shows the results of all algorithm stages for a tag in the two library examples considered in Results section.

Although the ratio R is our final aim, several researchers, (Nadon and Shoemaker, 2002), prefer to deal with $\log_2(R)$ because it makes expression ratios symmetrical around zero. We prefer to deal with $Q = 1/(1+1/R)$ because it is limited as $Q \in [0; 1]$ and naturally deals with extreme cases when transcripts are present in one library and absent in the other ($Q = 0$ and $Q = 1$). Nevertheless, any of these forms can be rewritten as usual ratios. Here we use Q as our random variable, find credibility intervals and then return to biologists familiar ratio R notation. Figure 2 shows some important features of this formulation.

IMPLEMENTATION

The method described above was implemented as a script in **R** statistical language (www.r-project.org). **R** has efficient and easy-to-use built-in functions to draw pseudo-random numbers and Kernel Density Estimator, and can be easily integrated into analysis pipe-line or database.

RESULTS AND DISCUSSION

To show this model in action we applied it to arbitrarily chosen SAGE data taken from The Cancer Genome Anatomy Project (CGAP) public database (http://cgap.nci.nih.gov/SAGE): normal mammary gland library SAGE_Breast_normal_AP_Br_N, with $T_A = 37\,121$ tags, versus breast cancer library SAGE_Breast_carcinoma_MD_DCIS-2, with $T_B = 28\,719$.

To obtain posteriori as the likelihood, we used non-informative uniform a priori distributions for every gene, i.e. $a_A = b_A = a_B = b_B = 1$. We have chosen arbitrarily $q = 95\%$ for credibility level, but this is one of the free parameters of the script. The 95% interval is traditionally taken, so as 0.05 significance levels, by unknown reasons. A better choice could be guided by external confirmatory results or biologists' experience.

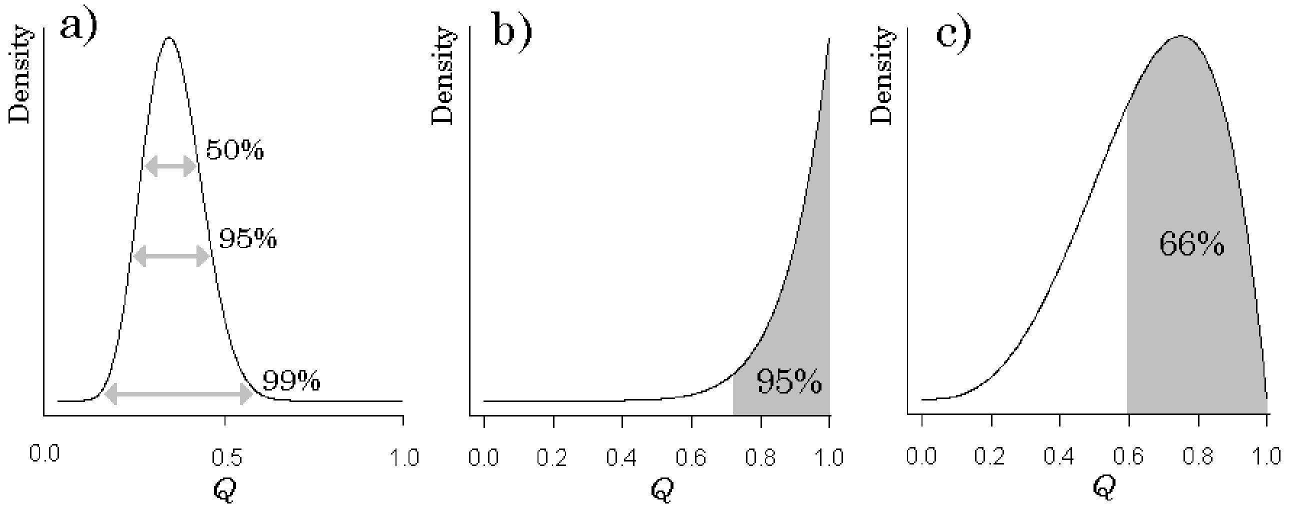

Fig. 2. Schematic view of model's features. Frame **(a)** shows that some gene could be regarded as differentially expressed depending on desired credibility stringency. In this example, choosing $q = 50\%$ or $q = 95\%$ leads to 'differentially' conclusion and $q = 99\%$ leads to 'not differentially', since the last contains $Q = 0.5$ (equivalent to ratio $R = 1$) and the others do not. Frame **(b)** shows method's behavior at extreme case, when transcript is present in one pool and absent in other with $Q = 1$ or $R = \infty$. The 95% credibility is [0.71; 1.0] meaning that ratio is at least 2.4. Frame **(c)** shows strategy similar to gene selection in hypothesis testing methods. In this example, the probability of expression ratio being greater than 1.5 (or $Q > 0.6$) is 66%.

Table 1. Some results of a hypothesis testing and credibility intervals approaches

Tag	Gene name	N_A	N_B	r	P-value	95% c.i.	*Diff.* hyp. test	*Diff.* 95% c.i.
GTGCGCTGAG	HLA-C	21	27	0.6	0.05	[0.3; 1.0]	Yes	No
ACCATCCTGC	IER3	12	16	0.6	0.10	[0.3; 1.2]	No	No
GGGGACTGAA	QP-C	5	11	0.4	0.05	[0.1; 0.9]	Yes	Yes
AGTGTCTGTG	CYR61	8	1	6.2	0.07	[1.1; 99.0]	No	Yes
AATATGTGGG	COX6C	2	230	0.01	0.00	[0.00; 0.02]	Yes	Yes
ACTCAGCCCG	TNFAIP2	14	1	10.8	0.01	[2.0; ∞]	Yes	Yes
ATAATAAAAG	CXCL3	82	0	∞	0.00	[15.7; ∞]	Yes	Yes
CTTCGAAACT	NDUFV2	7	0	∞	0.05	[1.2; ∞]	Yes	Yes
GCGGTGTCCG	BG752037	2	0	∞	0.28	[0.3; ∞]	No	No
GTGGTGCCGC	FBXO26	0	3	0.0	0.14	[0.0; 1.3]	No	No

Tags examples were chosen among 23 907 analyzed tags (see supplementary material for all results). A pool, with $T_A = 37\,121$, is the normal library and B pool, with $T_B = 28\,719$, is the cancerous library. P-values were estimated by SAGE Genie's DGED tool. c.i. is the 95% credibility interval. *Diff* status shows if the gene was considered differentially expressed by each method. Other symbols are defined in the text.

A gene not differentially expressed is intuitively identified since its credibility interval contains ratio $R = 1$, i.e. it is impossible to know if $\pi_A < \pi_B$ or if $\pi_A > \pi_B$ with the desired credibility, as for example in ACCATCCTGC tag, IER3 gene, with $r = 0.6$ and [0.3; 1.2] credibility interval (see results for all tags in the supplemental material). To avoid very sharp decision boundaries, one can determine other credibility levels, (90 or 99%), and separate their differentially expressed genes in reliability classes. However, once intervals are defined, there is no way to avoid such arbitrariness, even in traditional P-value hypothesis test approaches.

Table 1 shows some of our results along side results provided by the hypothesis testing method implemented in SAGE Genie Digital Genetic Expression Display (DGED) tool (Boon *et al.*, 2002; Lal *et al.*, 1999) but we avoid naive comparisons since methods are conceptually distinct and return numbers with different interpretations. See Man *et al.* (2000) for a detailed comparison between hypothesis testing approaches.

Our approach works entirely and solely at the parametric space instead of sample space, as in frequentist procedures. All inference is made only upon observed data, instead of upon 'data that could be observed but was not', thus do not violate Likelihood Principle (Basu and Ghosh, 1988). This means that P-value itself, comparisons with it, adjustments for it (such as Bonferroni), and so on, are meaningless in our approach.

In this study we explore only credibility intervals, but the same statistical principles can be used to construct a fully Bayesian significance test, with more convenient significance measure, avoiding P-value adjustments and paradoxes (Pereira and Stern, 1999).

We use tag counting as unique input data assuming that errors detectable with some previous bioinformatics processing were corrected.

Further development of our method will consider uncertainty due to eventual counting errors or tag misclassification (Colinge and Feger, 2001) using count correction principles proposed recently in Stern *et al*. (2003), mainly for rare transcripts; and biological variability between different SAGE libraries and thus consider $\boldsymbol{\pi}_A = (\pi_1, \ldots, \pi_n)$ and $\boldsymbol{\pi}_B = (\pi_1, \ldots, \pi_m)$ instead of scalars.

ACKNOWLEDGEMENTS

We thank Tie Koide for critical reading of manuscript and Junior Barrera, from BIOINFO-USP, for authors collaboration catalysis. R.V. is supported by FAPESP Msc. 02/04698-8 fellowship.

REFERENCES

Boon,K., Osório,E.C., Greenhut,S.F., Schaefer,C.F., Shoemaker,J., Polyak,K., Morin,P.J., Buetow,K.H., Strausberg,R.L., Souza,S.J. and Riggins,G.J. (2002) An anatomy of normal and malignant gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 11287–11292.

Colinge,J. and Feger,G. (2001) Detecting the impact of sequencing errors on SAGE data. *Bioinformatics*, **17**, 840–842.

Basu,D. and Ghosh,I. (1988) Statistical information and likelihood: a collection of critical essays by Dr. Basu. *Lecture Notes in Statistics* 45. Springer-Verlag, New York.

Kuznetsov,V.A., Knott,G.D. and Bonner,R.F. (2002) General statistics of stochastic process of gene expression in eukaryotic cells. *Genetics*, **161**, 1321–1332.

Lal,A., Lash,A.E., Altschul,S.F., Velculescu,V.E., Zhang,L., McLendon,R.E., Marra,M.A., Prange,C., Morin,P.J., Polyak,K. *et al*. (1999) A public database for gene expression in human cancers. *Cancer Res.*, **59**, 5403–5407.

Man,M.Z., Wang,X. and Wang,Y. (2000) POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinfomatics*, **16**, 953–959.

Margulies,E.H., Kardia,S.L. and Innis,J.W. (2001) Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Res.*, **29**, e60.

Nadon,R. and Shoemaker,J. (2002) Statistical issues with microarrays: processing and analysis. *Trends Genet.*, **18**, 265–271.

Pereira,C.A. and Stern,J. (1999) Evidence and credibility: a full Bayesian test of precise hypothesis. *Entropy*, **1**, 104–115.

Ruijter,J.M., Kampen,A.H.C. and Baas,F. (2002) Statistical evaluation of SAGE libraries: consequences for experimental design. *Physiol. Genomics*, **11**, 37–44.

Stollberg,J., Urschitz,J., Urban,Z. and Boyd,C.D. (2000) A quantitative evaluation of SAGE. *Genome Res.*, **10**, 1241–1248.

Stern,M.D., Anisimov,S.V. and Boheler,K.R. (2003) Can transcriptome size be estimated from SAGE catalogs?. *Bioinformatics*, **19**, 443–448.

Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.