Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

BayBoots: a model-free Bayesian tool to identify class markers from gene expression data

Ricardo Z.N. Vêncio[1,2], Diogo F.C. Patrão[3], Cassio S. Baptista[4], Carlos A.B. Pereira[1] and Bianca Zingales[4]

[1]BIOINFO-USP Núcleo de Pesquisas em Bioinformática and Departamento de Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo, Rua do Matão, 1010, 05508-090 São Paulo, SP, Brasil
[2]Instituto Israelita de Ensino e Pesquisa Albert Einstein, Hospital Israelita Albert Einstein, Av. Albert Einstein, 627, 05651-901 São Paulo, SP, Brasil
[3]Hospital do Câncer A.C. Camargo, R. Prof. Antonio Prudente, 109, 01509-010 São Paulo, SP, Brasil
[4]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, Av. Prof. Lineu Prestes, 748, 05508-000 São Paulo, SP, Brasil
Corresponding author: R.Z.N. Vêncio
E-mail: rvencio@vision.ime.usp.br

Genet. Mol. Res. 5 (1): 138-142 (2006)
Received January 10, 2006
Accepted February 17, 2006
Published March 31, 2006

ABSTRACT. One of the goals of gene expression experiments is the identification of differentially expressed genes among populations that could be used as markers. For this purpose, we implemented a model-free Bayesian approach in a user-friendly and freely available web-based tool called BayBoots. In spite of a common misunderstanding that Bayesian and model-free approaches are incompatible, we merged them in the BayBoots implementation using the Kernel density estimator and Rubin's Bayesian Bootstrap. We used the Bayes error rate (BER) in-

stead of the usual P values as an alternative statistical index to rank a class marker's discriminative potential, since it can be visualized by a simple graphical representation and has an intuitive interpretation. Subsequently, Bayesian Bootstrap was used to assess BER's credibility. We tested BayBoots on microarray data to look for markers for *Trypanosoma cruzi* strains isolated from cardiac and asymptomatic patients. We found that the three most frequently used methods in microarray analysis: *t*-test, non-parametric Wilcoxon test and correlation methods, yielded several markers that were discarded by a time-consuming visual check. On the other hand, the BayBoots graphical output and ranking was able to automatically identify markers for which classification performance was consistent. BayBoots is available at: http://www.vision.ime.usp.br/~rvencio/BayBoots.

Key words: Microarray, Bioinformatics, Statistics, Web tool, Gene expression, Bayesian bootstrap

INTRODUCTION

An important challenge in gene expression analysis is the decision of whether a particular gene is differentially expressed between two populations and could therefore be used as a marker for one of the two classes. Multi-class comparisons can also be partitioned and reduced to this paradigm. To our knowledge, there are no model-free solutions in a Bayesian statistical framework available for high-throughput detection of biomarkers.

We focused on gene expression data obtained in a high-throughput fashion from microarrays; however, the rationale is readily applicable to expression data obtained from other technologies with a good level of measurement replication. Microarray statistical analysis is known to be particularly challenging, due to the many sources of random and systematic errors that affect hybridization measurements (Zhang and Shmulevich, 2002). The advantages of the Bayesian over the Frequentist statistical framework for microarray analysis have been thoroughly discussed elsewhere (Yang et al., 2004).

Bayesian statistics often relies heavily on modeling because we need to know the likelihood function associated with a probabilistic model that describes the data, in order to subsequently use the Bayes rule, and finally ask inferential questions about the parameters, given the data. Conversely, model-free approaches rely only on observations, bypassing the proposition of a probabilistic model for the data (Troyanskaya et al., 2002). These features create some intuitive understanding that Bayesian analysis and model-free analysis are incompatible approaches (Ferguson et al., 1992; Müller and Quintana, 2004). In spite of this misunderstanding, a non-parametric determination of the *a posteriori* probability density function of interesting quantities can be achieved by the utilization of Rubin's version of the non-parametric Bootstrap technique (Efron, 1979), called Bayesian Bootstrap (Rubin, 1981). Our contribution was to couple

these two well-known paradigms, the model-free and the Bayesian approaches, into a freely available, easy-to-use, web-based statistical analysis tool called BayBoots.

MATERIAL AND METHODS

BayBoots is available at http://www.vision.ime.usp.br/~rvencio/BayBoots and allows multi-user "BLAST-like" job requests, e-mail warnings and data privacy. To identify probes that could be used as class markers, we used the Bayes error rate (BER), instead of the P values from formal statistical tests, since the former has a simple and graphical interpretation. BER's properties have been discussed previously (Duda et al., 2000; Vêncio et al., 2004). Briefly, BER measures how "far apart" the probability density functions of both classes are. For a given probe, the probability density function of each class is estimated, in the model-free paradigm, by the Kernel density estimator (KDE), using Silverman's rule for optimal bandwidth selection (Silverman, 1986), implemented in the R package (R Development Core Team, 2004). The robustness of the BER that is obtained is assessed according to the Bayesian paradigm, determining credibility intervals ("error-bars") on BER's predictive probability density function yielded by the Rubin's Bayesian Bootstrap, using methods described in detail elsewhere (Vêncio et al., 2003). It is important to keep in mind that model-free analysis, including Bootstrap-like techniques, is meaningless if applied to a dataset having a very low replication level (Chernick, 1999; Polansky, 2000).

For an illustration of BER logic, we simulated the behavior of two genetic marker candidates in a simple example that does not lead to controversy if compared to the usual methods (Figure 1A and B). In microarray applications, the input data is the normalized hybridization $\log_2$-ratio (M) of a particular probe. BER values closer to zero mean distributions are "far apart" and suggest the best candidates for class markers. Otherwise, BER values are closer to mean superimposed distributions and suggest the worst candidates. An adequate cut-off for BER error values could be defined in each experimental set, ranking the results of the probes from zero to one and performing some independent calibration experiments, such as Northern blot or RT-PCR to disclose the biological meaning (and not only the statistical significance) of each BER error level. This approach avoids the definition of cut-offs based solely on statistical assumptions that cannot match the precision of subsequent validation techniques (Rockett and Hellmann, 2004), such as the usual approaches: type I/II error analysis, multiple testing corrections, false discovery ratio, and so on.

To illustrate the utility of our web-based tool for dealing with real data in a microarray context, we used the data from a *Trypanosoma cruzi* experiment in which we hybridized the cDNAs obtained from two populations of parasite strains isolated from patients with cardiac manifestations of Chagas' disease and from asymptomatic patients. Complete information about the microarray slide and the hybridization is available at the GEO database (http://www.ncbi.nlm.nih.gov/geo) under the accession number GSE1828. The first class was composed of three strains from asymptomatic individuals and the second class, of three strains from cardiac patients. All hybridizations were performed using one of the asymptomatic strains as the common reference. Hybridizations were performed in duplicate, and at least six replicates of each probe were spotted on the slide, yielding at least $3 \times 2 \times 6 = 36$ measurements for each probe. The scanned images were submitted to quality control, intensity extraction and normalization processes, as previously described (Baptista et al., 2004).

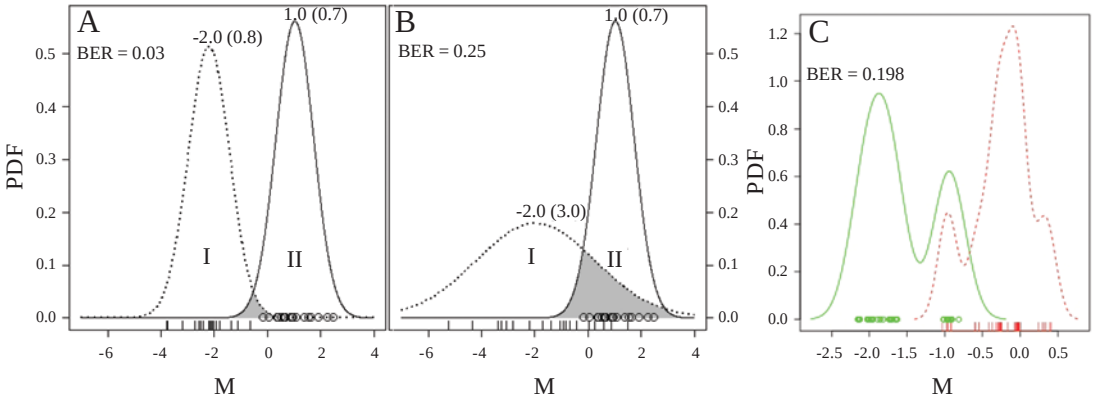

Figure 1. BayBoots approach to rank class markers. **A** and **B** show illustrative simulated data of two marker candidates. Twenty expression $\log_2$-ratios (M) were simulated according to a normal and arbitrarily definition for class I (ticks) and class II (circles). Probability density functions (PDF) of the simulated data are represented as a dashed line (class I) and a solid line (class II). The mean and standard deviation of the M values are indicated above the peaks. The shaded area in the overlapping probability density functions is the visual representation of the Bayes error rate (BER). **C.** The experimental data for probe 3465. Actual observations and the estimated probability density function of cardiac (circles and solid line) and asymptomatic (ticks and dashed line) classes are shown.

RESULTS AND DISCUSSION

Traditional methods, such as the *t*-test, the non-parametrical Wilcoxon test and correlation techniques, widely used by the microarray community, failed to automatically and selectively detect convincing marker candidates. The well-known *t*- and Wilcoxon tests ask if the means of two datasets are equal. The correlation method measures the Pearson correlation between the expression ratio results and the class labels (e.g., defining "cancer" as 1 and "normal" as 0, or 1 for "cardiac" and 2 for "asymptomatic", and so on). These methods yielded several candidate markers with very small P values or significant non-zero correlation, suggesting differential expression. However, most of the candidates could be readily discarded by time-consuming visual inspection of the M scattering, which revealed a high degree of superposition among the observations of each class.

The results for all probes are available at a supplemental website (http://www.vision.ime.usp.br/~rvencio/BayBoots). In particular, we highlighted results from illustrative examples of good markers and several examples of suspicious markers that were considered significant by the traditional methods.

Figure 1C shows a handpicked illustrative example of a probe that was detected as a promising class marker candidate using traditional statistical methods, but was rejected since it showed a clear superposition between the two classes. Based on the usual statistical methods, this probe could be regarded as a relatively good class marker since it showed a *t*-test P value $\leq 6 \times 10^{-13}$, non-parametric Wilcoxon test P value $\leq 5 \times 10^{-11}$ (both highly significant) and correlation with class labels $\rho = -0.841$. However, a clear superposition of observations was detected by BayBoots, suggesting that this probe is not a good marker. Several other similar examples are available at the supplemental website.

The kind of graphical output generated by BayBoots could be very useful for avoiding that candidates with poor performance, identified by microarray alone, be sent to time- and/or

resource-consuming subsequent validation steps, since they are probably statistical artifacts (false-positives). Although visual inspection is always an objective way to check consistency of results obtained by any statistical method, it is not a desirable solution in a high-throughput context.

For our particular data set, independent Northern blot experiments were able to validate markers, giving error levels of BER < 0.05. Markers with BER values greater than this empirically derived cut-off were not validated by independent Northern blot, yielding hybridization bands with the same pattern among the strains probed. However, a similar cut-off "rule" could not be defined using the output of the traditional methods, since there is no simple relation with the Northern blot experiments (Baptista CS, Vencio RZ, Abdala S, Silva MN, Pereira CAB and Zingales B, unpublished results).

We conclude that BayBoots interpretability and availability make it a valuable tool for biomarker discovery.

ACKNOWLEDGMENTS

We thank Sarah Abdala and Marcelo Nunes for technical assistance. R.Z.N. Vêncio and C.S. Baptista were supported by FAPESP, and B. Zingales and C.A.B. Pereira were partially supported by CNPq.

REFERENCES

Baptista CS, Vêncio RZ, Abdala S, Valadares MP, et al. (2004). DNA microarrays for comparative genomics and analysis of gene expression in *Trypanosoma cruzi*. *Mol. Biochem. Parasitol*. 138: 183-194.

Chernick MR (1999). Bootstrap methods: A practioner guide. Wiley-Interscience Press, New York, NY, USA.

Duda RO, Hart PE and Stork DG (2000). Pattern classification. 2nd edn. Wiley-Interscience Press, New York, NY, USA.

Efron B (1979). Bootstrap methods: another look at the jackknife. *Ann. Stat*. 7: 1-26.

Fergunson TS, Phadia EG and Tiwari RC (1992). Bayesian nonparametric inference. In: Current issues in statistical inference: Essays in honor of D. Basu (Ghosh M and Pathak PK, eds.). Institute of Mathematical Statistics, Beachwood, NJ, USA, pp. 127-150.

Müller P and Quintana FA (2004). Nonparametric Bayesian data analysis. *Stat. Sci*. 19: 95-110.

Polansky AM (2000). Stabilizing bootstrap-t confidence intervals for small samples. *Can. J. Stat*. 28: 501-516.

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rockett JC and Hellmann GM (2004). Confirming microarray data - is it really necessary? *Genomics* 83: 541-549.

Rubin DB (1981). The Bayesian bootstrap. *Ann. Stat*. 9: 130-134.

Silverman BW (1986). Density estimation. Chapman and Hall Ltd., London, England.

Troyanskaya OG, Garber ME, Brown PO, Botstein D, et al. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* 18: 1454-1461.

Vêncio RZ, Brentani H and Pereira CAB (2003). Using credibility intervals instead of hypothesis tests in SAGE analysis. *Bioinformatics* 19: 2461-2464.

Vêncio RZ, Brentani H, Patrao DF and Pereira CAB (2004). Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). *BMC. Bioinformatics* 5: 119.

Yang D, Zakharkin SO, Page GP, Brand JP, et al. (2004). Applications of Bayesian statistical methods in microarray data analysis. *Am. J. Pharmacogenomics* 4: 53-62.

Zhang W and Smulevich I (2002). Computational and statistical approaches to genomics. Kluwer Academic Publishers, Boston, MA, USA.