Short Communication
doi:10.1093/dnares/dsi007

HTself: Self–Self Based Statistical Test for Low Replication Microarray Studies

Ricardo Z. N. Vêncio[1,*] and Tie Koide[2]

BIOINFO-USP—Núcleo de Pesquisas em Bioinformática, Universidade de São Paulo, Rua do Matão 1010, 05508-090 São Paulo, Brazil[1] and Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, Av. Prof. Lineu Prestes 748, 05508-090 São Paulo, Brazil[2]

(Received 7 January 2005; revised 13 April 2005)

Abstract

Different statistical methods have been used to classify a gene as differentially expressed in microarray experiments. They usually require a number of experimental observations to be adequately applied. However, many microarray experiments are constrained to low replication designs for different reasons, from financial restrictions to scarcely available RNA samples. Although performed in a high-throughput framework, there are few experimental replicas for each gene to allow the use of traditional or state-of-art statistical methods. In this work, we present a web-based bioinformatics tool that deals with real-life problems concerning low replication experiments. It uses an empirically derived criterion to classify a gene as differentially expressed by combining two widely accepted ideas in microarray analysis: self–self experiments to derive intensity-dependent cutoffs and non-parametric estimation techniques. To help laboratories without a bioinformatics infrastructure, we implemented the tool in a user-friendly website (http://blasto.iq.usp.br/∼rvencio/HTself).

Key words: microarray; self–self; homotypical; web server; statistical test; low cost; differential gene expression

DNA microarray technology has allowed the study of gene expression in a genomic scale, changing the paradigm of expression studies of a single gene to a high-throughput framework. As this technology becomes cost accessible, more laboratories can use it as a routine technique.[1] By comparing two samples labeled with different fluorescent dyes, one can classify a gene as differentially expressed (or divergent, if dealing with genomic hybridizations) using a variety of statistical methods.[2–4] The ideal design of microarray experiments consists in having as many biological and technical replicates as possible, so that the data can be analyzed using state-of-art statistical tools. Unfortunately, it is not always possible to fulfill these replication requirements.

For instance, in laboratories with financial restrictions, the microarrays are used as a high-throughput screening tool. In this case, it is preferable to perform low replicated experiments and test different biological conditions. Another example is the study of rare human diseases. This kind of research is naturally constrained to low replication, since the RNA available usually comes from only one or two patients. Although not ideally replicated, these studies are undoubtedly important. However, they will not be properly analyzed using traditional or state-of-art statistical methods that require a number of replicates and assume certain hypothesis concerning the distribution of the samples that cannot be verified.

The aim of this work is to provide an easy-to-use bioinformatics solution for the analysis of microarrays constrained to low replication. To achieve our objective, we explored simultaneously two widely accepted ideas in microarray analysis: the determination of intensity-dependent cutoffs using self–self experiments[5–7] and the use of non-parametric methods.[8–10] Our contribution is to implement a web-based tool to help the analysis of microarray datasets with low replication designs. The web-based interface is freely available at http://blasto.iq.usp.br/∼rvencio/HTself.

Communicated by Michio Oishi

[*] To whom correspondence should be addressed. Tel. +55-11-3091-6210, Fax. +55-11-3814-4135, Email: rvencio@vision.ime.usp.br
Both these authors contributed equally to this work

When analyzing a microarray data, a major question is how to classify a gene as differentially expressed. To answer this question, it is necessary to set a cutoff level for hybridization intensities ratios that permits one to decide whether a gene is differentially expressed or not. In mathematical terms, this step consists in testing the null hypothesis H_0: 'the spot has no differential hybridization between the two probed samples'.

There are many mathematical approaches to define cutoffs and reject H_0.[2–4] A simple and widely used strategy consists in arbitrarily choosing a constant ratio, commonly the 2-fold change threshold. Spotted genes with ratios above this threshold are considered as differentially hybridized. To bring some statistical rigor, it is common to perform traditional statistical tests such as the *t*-test, using log-ratios and an arbitrary threshold. It will provide a *p*-value to access the significance level of the test for a given gene. To be adequately applied, one has to verify that the log-ratios for the given gene are normally distributed and that the number of observations is not scarce. Another approach is to assume a statistical model for the whole slide behavior (commonly a *t*-student like or a normal model), define it as the null probability density function (pdf) and search for outliers.[2–4] Again, this strategy requires the data to be distributed according to some known and arbitrarily proposed model. Since this assumption does not always hold for microarray data, different non-parametric procedures have been proposed to define the null pdf of the hybridization log-ratios for a given gene.[8–10] However, since they are usually based on resampling, permutation, standard deviation estimation, order/rank statistics, etc., it might not be a good choice to derive the pdf for an individual gene with few experimental observations.

Another category of approaches to define cutoffs relies on experimental strategies such as the use of self–self hybridizations. Self–self experiments are performed by labeling the same biological material with either Cy3 and Cy5 dyes and hybridizing them simultaneously on the same microarray slide. This strategy has been used to derive intensity-dependent cutoffs to classify a gene as differentially expressed[5,6,11] or divergent in comparative genomic hybridization (CGH) studies.[7,12] The comparative analysis of constant fold change cutoffs and intensity-dependent ones has been extensively discussed, showing a superior performance of the intensity-dependent strategy.[5–7,11,12]

In our tool, we make use of self–self experiments to derive the null probability density function of the test. Since the null hypothesis 'there is no differential hybridization between the two probed samples' holds for all the genes in a self–self experiment, it is possible to escape from the gene-by-gene schema and use all the spotted genes to derive the null pdf. With an adequate amount of observations (all the spotted genes), the use of non-parametric methods is now feasible. To take into account the intensity-dependent feature of the data, the null pdf is estimated in a user-defined sliding-window, which slides over all the range of the spots' intensity measure. This procedure results in the determination of intensity-dependent cutoffs that are readily applicable to non-self–self experiments. It is implicitly assumed that the same stochastic processes that generated the experimental noise in self–self experiments are acting in non-self–self data. Therefore, log-ratios above or below the intensity-dependent cutoffs can be classified as differentially expressed. The use of these experimentally derived cutoffs relaxes the requirement of replicates, since it does not count on standard deviation estimations, resampling or permutations. Moreover, it adds an empirically derived criteria to classify a gene as differentially expressed in studies constrained to low replication.

Our web-based tool expects a normalized data set as input. Microarray data usually must be normalized due to multiplicative biases such as unequal brightness of fluorescent dyes, unequal incorporation rate of dyes, etc. Such preprocessing procedures are well discussed and web-tools to address this problem are available elsewhere.[13–15] Next, we will describe the mathematical details behind our method.

Let $\mathrm{A} = \log_2(cy3)/2 + \log_2(cy5)/2$ and $\mathrm{M} = \log_2(R)$, as usual in microarray analysis,[16] be the random variables of interest, where *cy3* and *cy5* are the fluorescence intensities and *R* is the suitably normalized intensities ratio. To represent our measurement, we prefer to use the M–A plot, where the variable A shows the dependence of the log-ratios on the average spot intensity. The procedure can be used with arbitrary reparametrizations of hybridization ratio and measurements of fluorescence intensities. An observation of a spot *s* is one realization of (A,M) and is denoted by (a_s, m_s). Therefore, self–self hybridization measurements are samples drawn from the (A,M) bidimensional null joint pdf. To find the intensity-dependent log-ratio cutoffs, we first select a sliding-window in A, which is defined by the user. The observed spots $(a_s,\ m_s)$ contained in this window will be used to define the M|A null pdf locally. This pdf is estimated by applying the gaussian Kernel Density Estimator.

The Kernel Density Estimator is a model-free method that approximates the probability density function of a random variable using observations sampled from it.[17] Let f be the pdf of a random variable X and $x_1,\ldots,x_n$, n observed samples. The estimator for f is

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x_i - x}{h}\Big)$$

where the 'hat' over f indicates an estimator, h is the bandwidth and K is the kernel function. For example, a simple histogram can be described by a particular

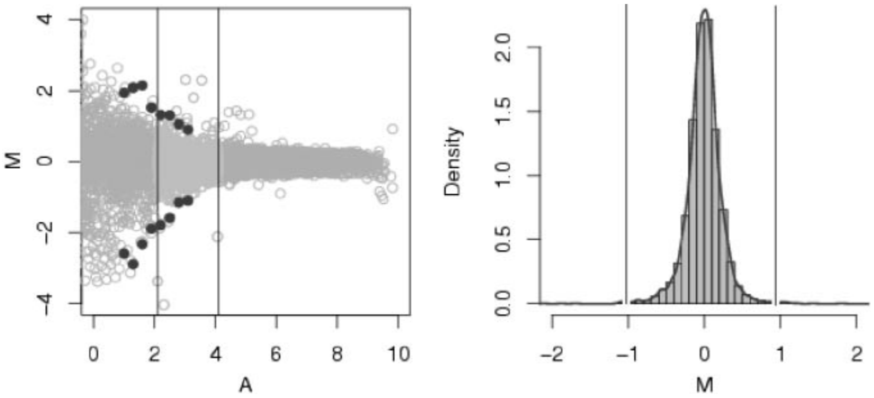

Figure 1. Snapshot of one step of the sliding window process. The left panel shows the MA-plot of the self–self data from a genomotyping CGH study.[12] The subinterval A considered in this snapshot is highlighted between the vertical lines. The histogram of M shown on the right was constructed using these highlighted observations. The Kernel Density Estimator (dark line) and the boundaries of the 99.5% credibility interval (vertical lines) are also shown. These boundaries define the intensity-dependent cutoffs, shown on the MA-plot (dark points) along with the results from previous steps. See all the steps in Supplementary Figures at the web-site (http://blasto.iq.usp.br/∼rvencio/HTself).

Kernel Density Estimator:

$$\widehat{f}_h(x) = \sum_{i=1}^{n} \frac{\mathbf{1}_{\{|x_i - x|\}} \leq h}{2nh} = \frac{\{\# x_i \text{ in } [x - h, x + h]\}}{2nh}$$

The gaussian Kernel Density Estimator is the most known and is the one used in our tool:

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \Big(\frac{x_i - x}{h} \Big)^2 \right)$$

The formulae above can be intuitively interpreted as a smoothing process for the histogram.

After estimating the null pdf of M for a given A window, the user-defined credibility interval can be determined. In short, our algorithm to define intensity-dependent cutoffs is

(i) the user defines a sliding window for A axis inputting two parameters: the window size and the walking pace. Each step of the sliding window delimits an arbitrary subinterval of A;
(ii) for each subinterval of A selected in (i), estimate the probability density function of M|A using gaussian Kernel Density Estimator;
(iii) integrate the probability density function from (ii) around the mode until the user-defined probability is reached. The intervals obtained are called the α credibility intervals;
(iv) the steps (ii) and (iii) are repeated until the window has slid over all the A range.

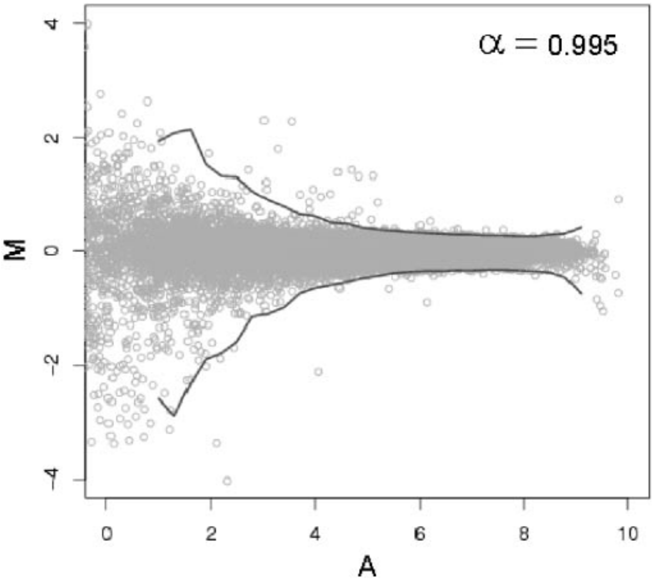

Figure 2. Web-site output for intensity-dependent cutoff determination. The self–self data used as an input example for the web-based tool is derived from a *Xylella fastidiosa* CGH study.[12] The dark lines are the upper and lower cutoffs. They were obtained by the sliding window process using a 99.5% credibility level, 0.3 pace and 1.0 window size.

Figure 1 shows a snapshot of the algorithm in an arbitrarily chosen step. It was performed using the self–self data from a genomotyping study in the bacteria *Xylella fastidiosa*. Figure 2 shows the result of the self–self derived intensity-dependent cutoffs for this data. Since we know that there should not exist true differential hybridization in self–self experiments, it is clear that the commonly used 2-fold change would be conservative for high intensity spots and permissive for low intensity ones.

After defining the intensity-dependent cutoffs, different microarray experiments made within the same technical conditions of self–self data can be evaluated. For example, suppose that a spot measurement (a, m) shows a log-ratio m outside its intensity-dependent 99% credibility cutoff. It can be classified as a differentially expressed spot since there is just 1% of chance that its measured log-ratio is due to random technical errors. This hypothesis test is applied to all spots. Since the test is applied to an individual spot, it does not depend on the number of replicates. If one has a number of replicated observations for a given gene, after applying the test to each spot, it is possible to evaluate easily if they are above or below the intensity-dependent cutoff and classify the gene as differentially expressed. Our tool has been successfully applied to a recently published gene expression study in *Xylella fastidiosa*.[18] It can also be useful for CGH studies.[12]

To use many of the available statistical tools, it is necessary to have well-replicated designs. Although many efforts have been carried out to sample as many replicates as possible, sometimes it is still difficult to achieve a well-replicated design. Financial restrictions or even biological constraints concerning rare RNA samples do not allow some researchers to analyze their microarray data according to current statistical standards. With this web-based tool, we hope to help these researchers to extract the invaluable information from their datasets constrained to low replication.

Acknowledgements: R.V. and T.K. were fellows of Fundação de Amparo a Pesquisa do Estado de Sao Paulo (FAPESP). We thank Aline M. da Silva, Paulo A. Zaini, Suely L. Gomes and Sergio Verjovski-Almeida for their involvement in the early stages of this work. We thank Diogo F.C. Patrão for his help with web-based tool implementation.

References

1. Rockett, J. C. and Hellmann, G. M. 2004, Confirming microarray data—is it really necessary? *Genomics*, **83**, 541–549.
2. Nadon, R. and Shoemaker, J. 2002, Statistical issues with microarrays: processing and analysis, *Trends Genet.*, **18**, 265–271.
3. Stolovitzky, G. 2003, Gene selection in microarray data: the elephant, the blind men and our algorithms, *Curr. Opin. Struct. Biol.*, **13**, 370–376.
4. Cui, X. and Churchill, G. A. 2003, Statistical tests for differential expression in cDNA microarray experiments, *Genome Biol.*, **4**, R210.
5. Yang, I. V., Chen, E., Hasseman, J. P., et al. 2002, Within the fold: assessing differential expression measures and reproducibility in microarray assays, *Genome Biol.*, **3**, R62.
6. Tu, Y., Stolovitzky, G., and Klein, U. 2002, Quantitative noise analysis for gene expression microarray experiments, *Proc. Natl Acad. Sci. USA*, **99**, 14031–14036.
7. Kim, C. C., Joyce, E. A., Chan, K., and Falkow, S. 2002, Improved analytical methods for microarray-based genome-composition analysis, *Genome Biol.*, **11**, R65.
8. Tusher, V. G., Tibshirani, R., and Chu, G. 2001, Significance analysis of microarrays applied to the ionizing radiation response, *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
9. Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., and Altman, R. B. 2002, Nonparametric methods for identifying differentially expressed genes in microarray data, *Bioinformatics*, **18**, 1454–1461.
10. Zhao, Y. and Pan, W. 2003, Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments, *Bioinformatics*, **19**, 1046–1054.
11. Papini-Terzi, F. S., Rocha, F. R., Vêncio, R. Z. N., et al. 2005, Transcription profiling of signal transduction-related genes in sugarcane tissues, *DNA Res.*, **12**, 27–38.
12. Koide, T., Zaini, P. A., Moreira, L. M., et al. 2004, DNA microarray-based genome comparison of a pathogenic and a non-pathogenic strain of *Xylella fastidiosa* delineates genes important for bacterial virulence, *J. Bacteriol.*, **186**, 5442–5449.
13. Park, T., Yi, S. G., Kang, S. H., Lee, S., Lee, Y. S., and Simon, R. 2003, Evaluation of normalization methods for microarray data, *BMC Bioinformatics*, **4**, 33.
14. Uchida, S., Nishida, Y., Satou, K., Muta, S., Tashiro, K., and Kuhara, S. 2005, Detection and normalization of biases present in spotted cDNA microarray data: a composite method addressing dye, intensity-dependent, spatially-dependent, and print-order biases, *DNA Res.*, **12**, 1–7.
15. Vaquerizas, J. M., Dopazo, J., and Diaz-Uriarte, R. 2004, DNMAD: web-based diagnosis and normalization for microarray data, *Bioinformatics*, **20**, 3656–3658.
16. Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. 2002, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, **12**, 111–139.
17. Silverman, B. W. 1986, Density Estimation, Chapman and Hall, London, UK.
18. Pashalidis, S., Moreira, L. M., Zaini, P. A., et al. 2004, Whole-genome expression profiling of *Xylella fastidiosa* in response to growth on glucose, *OMICS*, **9**, 77–90.