GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Combining P values to improve classification of differential gene expression in the HTself software

D.A. Cortez[1], A.P. Tonon[2], P. Colepicolo[2] and R.Z.N. Vêncio[3]

[1]Departamento de Ciências da Computação,
Instituto de Matemática e Estatística,
Universidade de São Paulo, São Paulo, SP, Brasil
[2]Departamento de Bioquímica, Instituto de Química,
Universidade de São Paulo, São Paulo, SP, Brasil
[3]LabPIB, Departamento de Computação e Matemática,
Faculdade de Filosofia Ciências e Letras de Ribeirão Preto,
Universidade de São Paulo, Ribeirão Preto, SP, Brasil

Corresponding author: R.Z.N. Vêncio
E-mail: rvencio@usp.br

Genet. Mol. Res. 10 (4): 3586-3595 (2011)
Received March 18, 2011
Accepted July 13, 2011
Published December 5, 2011
DOI http://dx.doi.org/10.4238/2011.December.5.5

ABSTRACT. HTself is a web-based bioinformatics tool designed to deal with the classification of differential gene expression in low replication microarray studies. It is based on a statistical test that uses self-self experiments to derive intensity-dependent cutoffs. We developed an extension of HTself, originally released in 2005, by calculating P values instead of using a fixed acceptance level α. As before, the statistic used to compute single-spot P values is obtained from the Gaussian kernel density estimator method applied to self-self data. Different spots corresponding to the same biological gene (replicas) give rise to a set of independent P values that can be combined by well-known statistical methods. The combined P value can be used to decide whether a gene can be considered differentially expressed or not. HTself2 is a new version of HTself that uses P values combination. It is implemented as a user-friendly desktop application to help laboratories without a bioinformatics infrastructure.

Key words: Microarray; Differential gene expression; Statistical test; Combined P value

INTRODUCTION

The study of gene differential expression in microarray studies plays a central role in bioinformatics today (Ramdayal, 2010; Sara et al., 2010). Several methods have been developed using a variety of statistical techniques (Farcomeni, 2008; Bremer et al., 2010). One of these methods is HTself (Vêncio and Koide, 2005), which was designed to deal with analysis of differentially expressed genes in low-replication contexts. This means that the ideal setup where one has as many biological and technical replicates as possible can not be fulfilled, either due to financial restrictions or due to shortage of available RNA. This is an important context for many real-life problems (Dougherty, 2001; Stevens et al, 2010).

HTself uses self-self hybridizations to derive intensity-dependent cutoffs to classify a gene as differentially expressed. Self-self experiments are commonly used in microarray analysis (Wenger et al., 2010; NCBI, 2009) and can be easily performed by labeling the same biological material with either Cy3 or Cy5 dyes and hybridizing them simultaneously on the same microarray slide. Intensity-dependent cutoffs can be represented as curves in the *A-M* diagram, where $A = \log_2 (cy3) /2 + \log_2 (cy5) /2$ and $M = \log_2 (R)$ are the usual microarray analysis variables (Yang et al., 2005). The symbols *cy*3 and *cy*5 represent the fluorescence intensities and *R* the suitably normalized intensities ratio. Therefore, *A* gives the total spot intensity and *M* the spot intensities log ratio.

Intensity-dependent cutoffs obtained from self-self data are important because they serve as a test for classifying genes of non-self-self experiments by assuming that the same random process that generated the experimental noise in the first is also acting on the last. This is the essence of what will be presented here.

Construction of cutoffs can be statistically translated to the testing of the following null hypothesis.

H_0: the spot has no differential hybridization between the two probed samples.

In HTself, cutoffs are derived empirically by using self-self experiments to derive the null probability density function (pdf) of the test. The intensity-dependent feature follows by estimating the null pdf in a sliding-window, which slides over the range of spot intensities. The estimation is obtained by applying the Gaussian kernel density estimator (Duong, 2007). Cutoff values are computed within a sliding-window by integrating the estimated pdf around the mode until a user-defined probability α is reached (α-credibility interval). The process is repeated until the window has slid over all the *A* range. Finally, non-self-self data (measured within the same technical conditions of self-self ones) can be tested against the cutoffs to decide whether they represent differentially expressed genes or not: the hypothesis test is applied to all spots and if one has a number of replicated measures for a given gene, it is evaluated if their average is above or below the intensity-dependent cutoff and classify the gene as differentially expressed. The full algorithm is described in Vêncio and Koide, 2005.

In this paper, we modify the above scheme in order to improve the classification criterion. Instead of working with a fixed α-credibility interval, we compute P values to all spots. The evaluation of single P values is based on the statistics obtained in the same manner as HTself. One estimates the null pdf using self-self data with the sliding-window process and the Gaussian Kernel Density Estimator method. For those spots corresponding to the same biological gene (replicas), we apply a combination method to generate a unique P value for that gene. At the end, one obtains a set of gene specific P values that can be sorted and used by the researcher to classify genes as differentially expressed according to one's biological intuition and the level of evidence presented.

The idea of working with P value is an old one in statistical hypothesis testing and is used in a wide range of applications (Goodman, 2001). It is therefore natural to follow such a concept to its conclusion and test it as treated here. The novelty in our context is that several independent P values can be obtained for the same gene in a single microarray experiment. The question that arises is whether the evidence contained within it can be combined to support a particular statistical hypothesis, or equivalently, can we combine P values into a single test based on a common hypothesis? Fortunately there are several methods available to produce this type of combination (Fisher, 1932) and we apply two of them to the context here. Combining P values improves the usual HTself classification method in the sense that not only consistency is taken into account, but also the strength of the evidence, giving a more reliable tool to the researcher to be used for their studies.

This paper is organized as follows: section 2 gives the details on the calculation of a single spot P value based on the same statistics used in the usual HTself. Section 3 describes two statistical methods used to combine several P value for a single gene into a unique P value. It also gives an alternate procedure to produce a fast computation for experiments with large data involved, which can result in long processing time due to the huge number of calculations required. Section 4 shows the actual implementation of the ideas described here into a complete desktop application written especially to help laboratories without a bioinformatics infrastructure. A sample analysis is presented based on real data obtained for an experiment in the context of macroalga biology. Finally, section 5 sums up with some concluding remarks.

Single-spot P value

In statistical hypothesis testing, the P value is the probability of obtaining a value of the test statistic at least as extreme as the one that was actually observed, given that a null hypothesis is true. We apply this definition in what follows.

We start with two sets of data: one is the set of total intensities and log ratios $S = \{(a^s,m^s)\}$ for the self-self experiment and the other is the corresponding $N = \{(a^n,m^n)\}$ for the non-self-self experiment. Let T_0 be the spot from a non-self-self experiment for which we want to evaluate its P value. Spot T_0 is associated with a measure (a_0,m_0) in the set N.

Since for self-self experiments, by construction, the null hypothesis H_0 (as stated in the Introduction) is true, we can construct the test statistic by collecting all self-self data in S that falls in a window of predefined length w centered at A_0. Call this set D_0. More precisely, $D_0 = \{(a^s,m^s) \in S | a_0 - w < a^s < a_0 + w\}$. We then apply the Gaussian Kernel Density Estimator

(Duong, 2007) to the set D_0 to obtain the null pdf locally:

$$\widehat{f}_h(m) = \frac{(2\pi)^{-\frac{1}{2}}}{h|D_0|} \sum_{m^s \in D_0} \exp\left[-\frac{1}{2}\left(\frac{m^s - m}{h}\right)^2\right] \qquad \text{(Equation 1)}$$

where $|D_0|$ stands for the number of elements in D_0, the "hat" over f indicates an estimator and h is the bandwidth.

Now define two one-tail P values associated with the spot T_0: $\hat{\alpha}_0^{\uparrow}$ and $\hat{\alpha}_0^{\downarrow}$. From the above estimation for the test statistic, we have

$$\widehat{\alpha}_0^{\uparrow} = \int_{m_0}^{+\infty} \widehat{f}_h(x)dx = \frac{1}{|D_0|} \sum_{m^s \in D_0} \Phi\left(\frac{m^s - m_0}{h}\right) \qquad \text{(Equation 2)}$$

and

$$\widehat{\alpha}_0^{\downarrow} = \int_{-\infty}^{m_0} \widehat{f}_h(x)dx = 1 - \widehat{\alpha}_0^{\uparrow} \qquad \text{(Equation 3)}$$

where Φ is the cumulative Standard Normal.

The interpretation of $\hat{\alpha}_0^{\uparrow}$ is that it gives the evidence level for the gene in spot T_0 to be up-regulated, while $\hat{\alpha}_0^{\downarrow}$ gives the evidence level for it to be down-regulated. Figure 1 shows a depicted version of the algorithm and serves as a summary.

The next step is to iterate the above procedure to all spots in the non-self-self experiment. Of course different spots corresponding to the same biological gene have a set $\{(\hat{\alpha}_0^{\uparrow}, \hat{\alpha}_0^{\downarrow})\}$ of independent P values which have to be analyzed to produce a single test. This will be discussed in the next section.

Combining P values

The general setup is that of combining $k \geq 2$ independent tests. The combined null hypothesis, H_0, is that each of the component null hypothesis, $H_{01}, \ldots, H_{0k}$, is true (all of them). The combined alternative, H_A, is that at least one of the alternatives, $H_{A1}, \ldots, H_{Ak}$, is true. This scenario is appropriate to our modeling needs: if $k \geq 2$ spots corresponding to the same biological gene g are tested, then combined null hypothesis is

H_0: each of the k spots corresponding to gene g has no differential hybridization between the two probed samples.

Rejection of H_0 will lead to the conclusion that g is indeed differentially expressed. Many methods have been proposed for combining P values into a single test of a common hypothesis. A good empirical comparison between some of the most common ones can be found in (Loughin, 2004).

We have chosen two well know combination strategies - the chi-square (Fisher, 1932) and the normal (Lipták, 1958) - that have a good reputation and are easy to implement. Both chi-square and normal tests are of the quantile combination type, which relies on the fact that under the null hypothesis a P value from an absolutely continuous test statistic has a uniform distribution from zero to one. Given a P value p_i, i = 1,…, k, available for each test, one selects some parametric cumulative density function (CDF), F, and transforms the P values into distributional quantiles according to $q_i = F^{-1}(p_i)$. The combining function is then $C = \Sigma\, q_i$, and the P value of the combined test is computed from the sampling distributions of C.

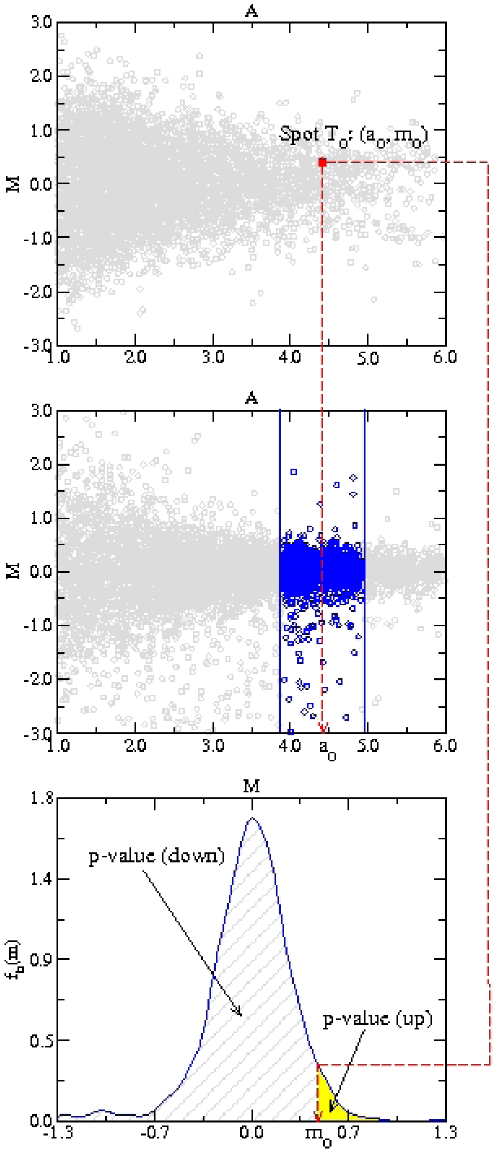


Figure 1. Algorithm to evaluate the P value of a single spot T_0. Top plot shows normalized data for the non-self-self experiment where we have selected a particular spot T_0. Middle plot shows normalized data collected from the self-self experiment to evaluate the kernel density estimator (the set D_0 of $\hat{f}_h$ in (1)). Bottom plot shows the null pdf obtained for the test and the corresponding P values $\hat{\alpha}_0^{\uparrow}$ and $\hat{\alpha}_0^{\downarrow}$.

The chi-square method uses a chi-square (2 df) to build its CDF. In this case, the sampling distribution of *C* turns out to be a chi-square (*2k* df), χ^2_{2k}. The CDF used by the normal method is based on a normal scale. The resulting sampling distribution of *C* is given by a Standard Normal, φ. Therefore, implementation of both chi-square and normal methods depends only on the evaluation of integrals of χ^2_{2k} and φ.

Application of the above ideas to the P values of a single gene is straightforward. Given $\hat{\alpha}_i$ (up or down), $1 \leq i \leq k$, obtained according to section 2, the combined P value, $\hat{\alpha}$, is given by

- Chi-square method:

$$\widehat{\alpha} = \int_{F_0}^{\infty} \chi^2_{2k}(x)\,dx\,, \text{ where } F_0 = -2\sum_{i=1}^{n} \log \widehat{\alpha}_i \qquad \text{(Equation 4)}$$

- Normal method:

$$\widehat{\alpha} = \int_{Z_0}^{\infty} \varphi(x)\,dx\,, \text{ where } Z_0 = \sum_{i=1}^{k} \frac{\Phi^{-1}(\widehat{\alpha}_i)}{\sqrt{k}} \qquad \text{(Equation 5)}$$

Biological evidence would, ideally, indicate which of these alternatives is better suited to the problem at hand. According to (Loughin, 2004), the chi-square does best when the evidence is at least moderately strong and is concentrated in a small fraction of the individual tests. This is reasonable if one, for technical or biological reasons, believes that differential gene expression actually occurs in the case where a few of the tested spots indicates differentiation with relative strength. On the other hand, also according to Loughin (Loughin, 2004), the normal combining method does well in problems where evidence against the combined null is spread among more than a small fraction of the individual tests, or when the total evidence is weak. Therefore, one would rely on this method if one believes that gene expression can be identified through consistency of observations.

Speeding up calculation time

The evaluation of all individual P values according to Equations 2 and 3 requires a great deal of computational effort due to the huge number of spots considered in a typical microarray experiment.

Although feasible in a reasonable amount of time, it is not entirely necessary to obtain all the values if we use the following heuristic: using the standard HTself method with a low credibility level, we may consider that the P values of those spots that fall under the pre-computed cutoffs are uniform from $(1 - \alpha) / 2$ to one. This is a consequence of the observation that under the null hypothesis a P value from an absolutely continuous test statistic has a uniform distribution from zero to one. The heuristic then follows from the facts that spots bellow the cutoffs constructed using such small values of α are almost sure to agree with the null hypothesis and that their P values are bounded from bellow by $\approx (1 - \alpha) / 2$.

We can, therefore, speed up calculation time if we adopt the following procedure:

1. Apply the standard HTself method with a low credibility level, say $\alpha = 0.6$.

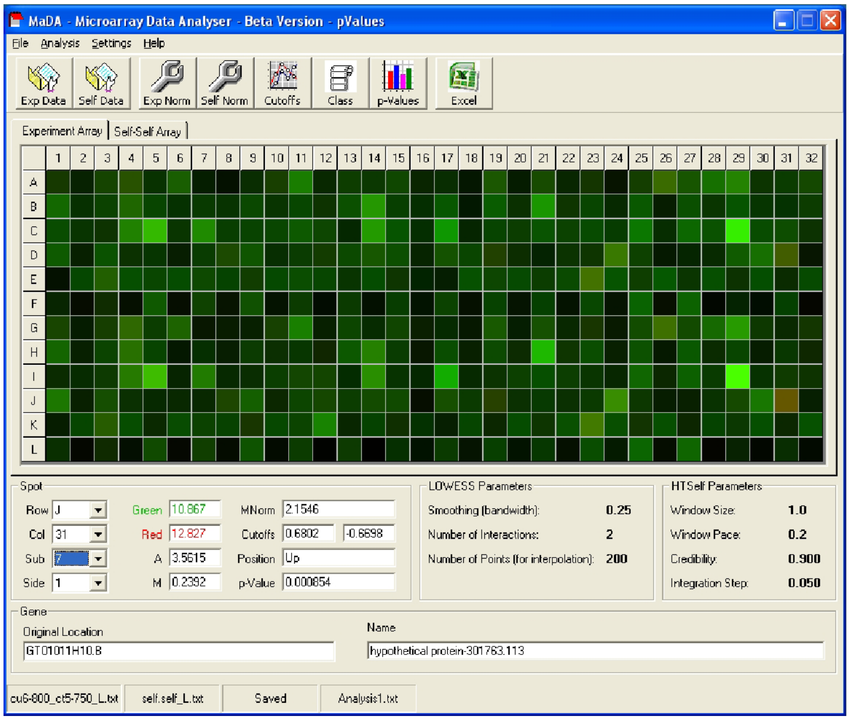
2. Select all spots that fall bellow the pre-computed cutoffs in 1.
3. Sort a random number in the interval $[a_0, 1]$, where $a_0 = (1 - \alpha) / 2$, for each spot selected in 2.
4. Assign the number sorted in 3 as the P value of the corresponding spot.

Of course the above steps give only an approximation for the actual P value. Nevertheless, since it is performed only for those spots with relatively low strength against H_0, the overall result when combining P values should not be affected by such approximation.

Application

We have implemented the ideas of HTself2 in a new microarray data analyzer software, called MaDA.

The software can be freely downloaded at http://labpib.fmrp.usp.br/MaDA. It was designed as a desktop application and written in Visual Basic 6.0, therefore running under MS Windows© operational system. It aims to give a single tool for the researcher without bioinformatics background, guiding him in the process of analyzing microarray data (in the same spirit of (Koide et al., 2006). The novelty is that it implements both the usual HTself as well as HTself2. Moreover, it is a stand alone application, so it is not necessary to have the R package installed (The R Foundation for Statistical Computing, 2006), which is well suit for laboratories with no informatics infrastructure. Figure 2 shows a picture of the main window of MaDA running a sample analysis (see next section).



Spot

Row J

Col 31

Sub 7

Side 1

Green 10.867

Red 12.827

A 3.5615

M 0.2392

MNorm 2.1546

Cutoffs 0.6802 -0.6698

Position Up

p-Value 0.000854

LOWESS Parameters

Smoothing (bandwidth): 0.25

Number of Interactions: 2

Number of Points (for interpolation): 200

HTSelf Parameters

Window Size: 1.0

Window Pace: 0.2

Credibility: 0.900

Integration Step: 0.050

Gene

Original Location GT01011H10.B

Name hypothetical protein-301763.113

cu6-800_ct5-750_L.txt | self.self_L.txt | Saved | Analysis1.txt

Figure 2. Screenshot of MaDA main window. You may see the graphical visualization of one array of the experiment being analyzed and computed data for the selected spot in the array.

The software includes, among others, the following features:

- Lets you define the number of rows and columns, as well as the number of arrays, of your experiment.
- The "configuration menu" lets you setup all parameters to be used in the analysis of your data (credibility level, LOWESS smoothing, etc) as well as to select only those spots that should be taken into account.
- Lets you open text files containing fluorescence data (background removed) arranged in any kind of format, for both self-self and non-self-self experiments.
- Lets you load a "chip map" (a formatted text file) containing the identification between spots and corresponding biological genes.
- There is a graphical visualization of your data in the arrays, with an user-friendly interface that quickly allows you to inspect computed values of the spots (A, normalized M, cutoffs, P values, etc.).
- Lets you perform data normalization using the LOWESS method (Yang, et al., 2005).
- Lets you construct HTself cutoffs and HTSelf2 combined P values.
- Gives you a complete classification list with all computed values of interest and standard errors for all genes.
- Simple integration with MS Excel©, which enables data to be worked in spreadsheets and plots to be easily constructed.

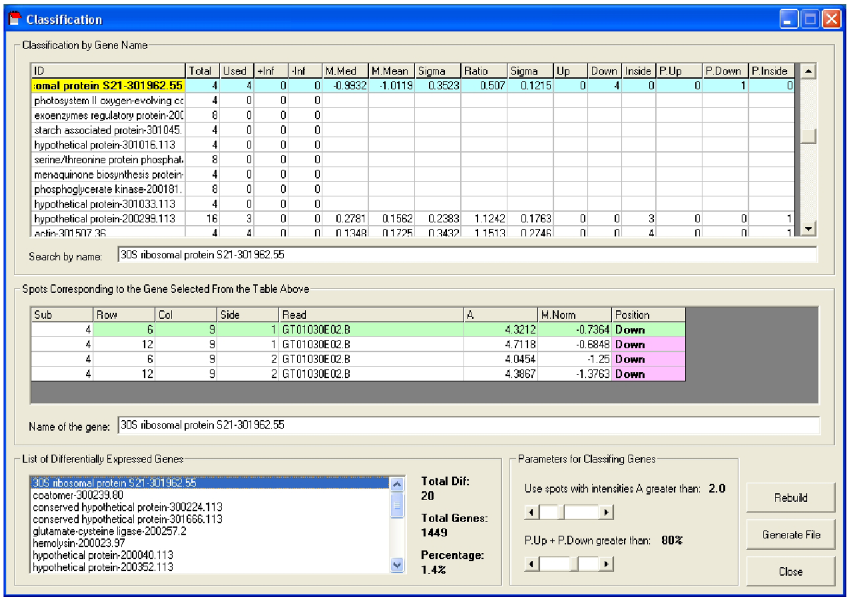
Sample analysis

MaDA has been successfully used to analyze data extracted from samples of an economically important red marine macroalgae (*Gracilaria tenuistipitata*) (Tonon and Colepicolo, 2011). The microarray was constructed using a cDNA library of the algae obtained from normal culture conditions. A stress condition was simulated by adding metal pollutants to the culture media and global gene expression was studied by hybridization of samples based on the two opposite conditions (control and metal exposed).

Figure 2 shows MaDA running with data originating from the *Gracilaria* experiment. Self-self (control condition against control condition) and non-self-self (control condition against metal exposed condition) fluorescence intensities were loaded into the program, which was previously setup with the arrays dimensions, analysis parameters and genes localization (chip map). Normalization and HTself Cutoffs were computed, generating a complete classification of the genes (see Figure 3). Finally, implementing the ideas of this paper, combined P values were evaluated using both chi-square and normal methods. As one can see from Figure 4, results are shown in a sorted order to let the user choose the most likely genes to be differentially expressed. All results can be saved exported to a MS Excel© spreadsheet, where plots can be easily constructed.

CONCLUSIONS

We have given an alternate (and improved) method to classify differential expression genes in microarray experiments using combination of P values. This extends the ideas of HTself in the sense that we do not have to work with a fixed credibility level, α. Working with P values, as usual, enables the researcher to freely decide against or in favor of the alternative



Classification by Gene Name

ID	Total	Used	+Inf	-Inf	M.Med	M.Mean	Sigma	Ratio	Sigma	Up	Down	Inside	P.Up	P.Down	P.Inside
:omal protein S21-301962.55	4	4	0	0	-0.9932	-1.0119	0.3523	0.507	0.1215	0	4	0	0	1	0
photosystem II oxygen-evolving cc	4	0	0	0											
exoenzymes regulatory protein-200	8	0	0	0											
starch associated protein-301045.	4	0	0	0											
hypothetical protein-301016.113	4	0	0	0											
serine/threonine protein phosphat.	8	0	0	0											
menaquinone biosynthesis protein-	4	0	0	0											
phosphoglycerate kinase-200181.	8	0	0	0											
hypothetical protein-301033.113	4	0	0	0											
hypothetical protein-200299.113	16	3	0	0	0.2781	0.1562	0.2383	1.1242	0.1763	0	0	3	0	0	1
actin-301507.36	4	4	0	0	0.1348	0.1725	0.3432	1.1513	0.2746	0	0	4	0	0	1

Search by name: 30S ribosomal protein S21-301962.55

Spots Corresponding to the Gene Selected From the Table Above

Sub	Row	Col	Side	Read	A	M.Norm	Position
4	6	9	1	GT01030E02.B	4.3212	-0.7364	**Down**
4	12	9	1	GT01030E02.B	4.7118	-0.6848	**Down**
4	6	9	2	GT01030E02.B	4.0454	-1.25	**Down**
4	12	9	2	GT01030E02.B	4.3867	-1.3763	**Down**

Name of the gene: 30S ribosomal protein S21-301962.55

List of Differentially Expressed Genes

- 30S ribosomal protein S21-301962.55
- coatomer-300239.80
- conserved hypothetical protein-300224.113
- conserved hypothetical protein-301666.113
- glutamate-cysteine ligase-200257.2
- hemolysin-200023.97
- hypothetical protein-200040.113
- hypothetical protein-200352.113

Total Dif: **20**

Total Genes: **1449**

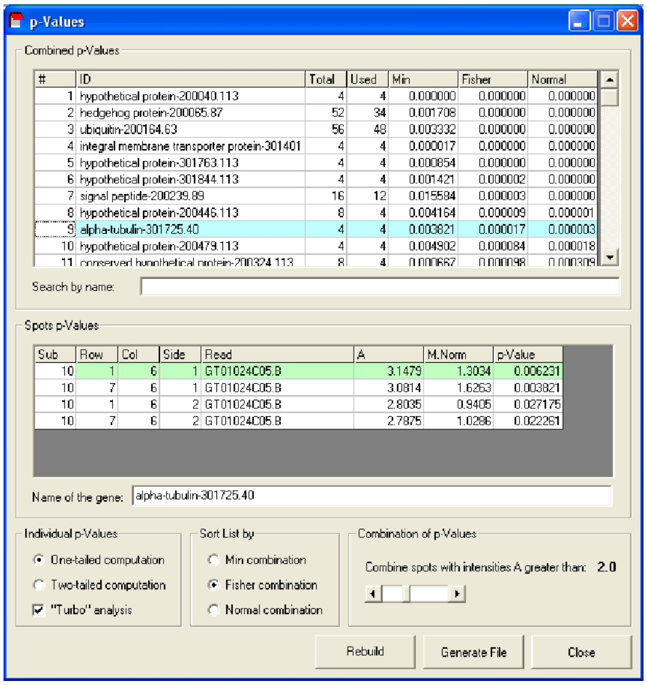
Percentage: **1.4%**

Parameters for Classifing Genes

Use spots with intensities A greater than: **2.0**

P.Up + P.Down greater than: **80%**

Figure 3. MaDA classification screen. The genes are listed showing all relevant data and standard errors. For each gene on the list you can see the corresponding spots with associated values. There is also a separated list for those genes considered differentially expressed according to the usual HTself criterion.



Combined p-Values

#	ID	Total	Used	Min	Fisher	Normal
1	hypothetical protein-200040.113	4	4	0.000000	0.000000	0.000000
2	hedgehog protein-200065.87	52	34	0.001708	0.000000	0.000000
3	ubiquitin-200164.63	56	48	0.003332	0.000000	0.000000
4	integral membrane transporter protein-301401	4	4	0.000017	0.000000	0.000000
5	hypothetical protein-301763.113	4	4	0.000854	0.000000	0.000000
6	hypothetical protein-301844.113	4	4	0.001421	0.000002	0.000000
7	signal peptide-200239.89	16	12	0.015584	0.000003	0.000000
8	hypothetical protein-200446.113	8	4	0.004164	0.000009	0.000001
9	alpha-tubulin-301725.40	4	4	0.003821	0.000017	0.000003
10	hypothetical protein-200479.113	4	4	0.004902	0.000084	0.000018
11	conserved hypothetical protein-200324.113	8	4	0.000667	0.000098	0.000309

Search by name:

Spots p-Values

Sub	Row	Col	Side	Read	A	M.Norm	p-Value
10	1	6	1	GT01024C05.B	3.1479	1.3034	0.006231
10	7	6	1	GT01024C05.B	3.0814	1.6263	0.003821
10	1	6	2	GT01024C05.B	2.8035	0.9405	0.027175
10	7	6	2	GT01024C05.B	2.7875	1.0286	0.022261

Name of the gene: alpha-tubulin-301725.40

Individual p-Values

- ● One-tailed computation
- ○ Two-tailed computation
- ☑ "Turbo" analysis

Sort List by

- ○ Min combination
- ● Fisher combination
- ○ Normal combination

Combination of p-Values

Combine spots with intensities A greater than: **2.0**

Figure 4. Screenshot of combined P values window. You may exhibit results either sorting by the chi-square values or the normal values. For each gene on the list you can see the corresponding spots with associated individual P values.

hypotheses giving his intuition for the problem and the level of evidence encountered.

MaDA was a software constructed to implement both HTself and HTself2 (with chi-square and normal methods to combine P values). It is a simple and easy to use tool.

Many researchers have been using it in our laboratories to deal with a variety of biological problems. We have shown here only one sample analysis. We hope now that the microarray community can take advantage of this useful tool.

ACKNOWLEDGMENTS

R.Z.N. Vêncio would like to thank the support from FMRP-USP. A.P. Tonon and P. Colepicolo would like to thank the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), for financial support. The costs of publication were covered by the Novos Docentes grant from Pró-Reitoria de Pesquisa USP.

REFERENCES

Bremer M, Himelblau E and Madlung A (2010). Introduction to the statistical analysis of two-color microarray data. *Methods Mol. Biol.* 620: 287-313.

Dougherty ER (2001). Small sample issues for microarray-based classification. *Comp. Funct. Genomics* 2: 28-34.

Duong T (2007). Kernel density estimation and kernel discriminant analysis for multivariate data in R. *J. Stat. Software* 21: 1-16.

Farcomeni A (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat. Methods Med. Res.* 17: 347-388.

Fisher RA (1932). Statistical Methods for Research Workers. 4th edn. London.

Goodman SN (2001). Of P-values and Bayes: a modest proposal. *Epidemiology* 12: 295-297.

Koide T, Salem-Izacc SM, Gomes SL and Vencio RZ (2006). SpotWhatR: a user-friendly microarray data analysis system. *Genet. Mol. Res.* 5: 93-107.

Lipták T (1958). On the combination of independent tests. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.* 3: 171-197.

Loughin TM (2004). A systematic comparison of methods for combining p-values from independent tests. *Comp. Stat. Data An.* 47: 467-485.

NCBI (2009). The National Center for Biotechnology Information Advances Science and Health by Providing Access to Biomedical and Genomic Information. Available at [http://www.ncbi.nlm.nih.gov/geo/]. Accessed August 23, 2009.

Ramdayal K (2010). Analytical Methods in Bioinformatics: Microarrays, Proteomics and Databases. LAP LAMBERT Academic Publishing.

Sara H, Kallioniemi O and Nees M (2010). A decade of cancer gene profiling: from molecular portraits to molecular function. *Methods Mol. Biol.* 576: 61-87.

Stevens JR, Bell JL, Aston KI and White KL (2010). A comparison of probe-level and probeset models for small-sample gene expression data. *BMC Bioinformatics* 11: 281.

The R Foundation for Statistical Computing (2006). The R project for Statistical Computing. Available at [http://www.r-project.org]. Accessed……

Tonon AP and Colepicolo P (2011). Study of acclimation of *Gracilaria tenuistipitata* under stress induced by heavy metals. To be submitted.

Vêncio RZ and Koide T (2005). HTself: self-self based statistical test for low replication microarray studies. *DNA Res.* 12: 211-214.

Wenger JW, Schwartz K and Sherlock G (2010). Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from Saccharomyces cerevisiae. *PLoS Genet.* 6: e1000942.

Yang YH, Dudoit S, Luu P, Lin DM, et al. (2005). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 30: e15.