Llama: an online microarray linear analysis tool

Routley, Ben and Muldoon, Mark R.

2007

MIMS EPrint: **2007.104**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

BIOINFORMATICS

Vol. 00 no. 0 2005, pages 1–2
doi:10.1093/bioinformatics/bti283

Application Notes

LLAMA: AN ONLINE MICROARRAY LINEAR ANALYSIS TOOL

Ben Routley[1] and Mark Muldoon[1,*]

[1]School of Mathematics of Mathematics, University of Manchester, United Kingdom.

Received on ; revised on; accepted on

Advance Access publication . . .

ABSTRACT

Motivation: We have developed a linear modelling tool for analysis of two-colour microarray data that utilises a per-spot linear model to estimate expression differences. Given the design of the experiment, the program combines all relevant data to provide the best estimate of a particular difference in expression between samples. It constructs multiple estimates based on several slides and combines them to get the most precise overall estimates of differential expression. Every effort has been made to make this tool accessible to biologists and it contains many user-friendly options.

Availability: The tool is accessible via the World Wide Web as http://exgen.ma.umist.ac.uk/llama/

Contact: ben.routley@manchester.ac.uk

1 INTRODUCTION

Identifying gene transcript (mRNA) regulation has been largely possible due to the increasing use of DNA microarrays or gene chips, first developed at Stanford University (Schena et al. 1995). A typical dual-hybridisation study compares two samples on the same microarray slide enabling the biologist to make direct, quantitative comparisons of two expression patterns. Each transcript from the two conditions is tagged with a distinct fluorescent dye and hybridised to the probes on the array. Fluorescence measurements are recorded using a scanner to yield a quantitative estimate of expression for each sample and for every probe.

2 MOTIVATION

Most microarray experiments with more than two samples use a reference sample to allow comparisons between all samples. A shortcoming of this approach is that one spends a disproportionate amount of effort on the reference, a sample that may be of little or no intrinsic biological interest. As an alternative to reference designs, loop designs involve comparing conditions to one another in a daisy-chain fashion (Churchill 2002; Yang and Speed 2002). The results from these hybridisations are direct comparisons of samples with their immediate neighbours in the loop, which can be so arranged as to focus attention on the measuring the most interesting comparisons. But a drawback of loop designs is that it can be very difficult to see how best to estimate differences in expression between conditions from opposite sides of the loop.

The linear modelling approach implemented in this tool, allows the user to choose which expression differences or *contrast-pairs* to estimate from all of the many possible alternatives. The tool works with log of ratios of expression levels, so that indirect comparisons appear as linear combinations of directly-measured ones.

3 WEB INTERFACE

This application uses the R statistical language (R Development Core Team 2005) with methods developed in (Vinciotti et al. 2004). The web interface is developed with PHP and JavaScript that communicate with R using the CGIwithR package (Firth 2003). The program accepts a tab-delimited text file containing appropriately extracted columns from scanner output files of a complete experiment; usually foreground Signal Median.

A series of linked web-pages guide the user through the various stages of an analysis. User input is required to specify the number of experimental conditions and to say whether the data has been log-transformed prior to analysis. Once the data has been verified and uploaded, the user is asked to provide names for the conditions and supply an e-mail address to receive the results. The user must then describe the experimental details by specifying, for each hybridisation, the dye and sample that was used. Based on this information, the program will choose a set of default contrast pairs on which to base the analysis. If the user wishes to estimate different contrast pairs, then these can be selected from a menu.

The results from the analysis are e-mailed to the user as a tab-delimited text file, suitable for import into Excel.

4 PROGRAM

Raw data is normalised using the smida package described in Wit and McClure 2004, and available online at http://www.stats.gla.ac.uk/~microarray/book/smida.html. Their strategy normalises for unequal dye-effects and across-array effects but does not attempt spatial-normalisation or background correction.

The tool will be illustrated for a simple loop design, figure 1. The plot represents three distinct conditions—shown as circular nodes—connected to other nodes via arrows. Each arrow represents an dual-htbridisation array: the node at the tail of the arrow is the condition to be hybridized with dye 1 and the node at the head of the arrow is the condition to be hybridised with dye 2.

*To whom correspondence should be addressed.

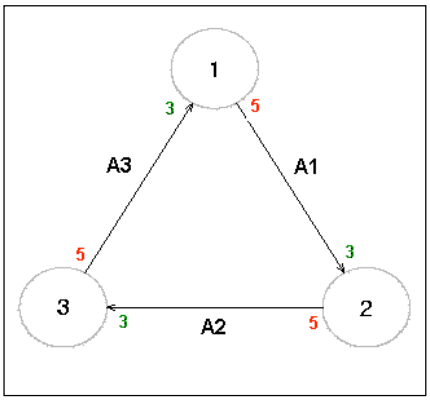

Figure 1, A simple loop design of 3 samples analysed using 3 dual-hybridisation arrays.

The log ratios that are measured directly in this experiment are thus:

$$A1 = \log (y_1^{\text{red (A1)}} / y_2^{\text{green (A1)}}) = \mu_{12}$$
$$A2 = \log (y_2^{\text{red (A2)}} / y_3^{\text{green (A2)}}) = \mu_{23}$$
$$A3 = \log (y_3^{\text{red (A3)}} / y_1^{\text{green (A3)}}) = \mu_{31}$$

where μ_{ij} is the *true* expression difference between conditions *i* and *j*. We imagine that the *observed* differences in expression for each gene across *n* observations, $y = (ya_1, \ldots .ya_n)$, can be represented as

$$y = X\mu + \epsilon,$$

where X is a *design matrix*, defining the relationship between the expression differences observed in the experiment and a set of independent parameters, μ, and ϵ is the normally-distributed error term. The independent parameters chosen for this example are μ_{12} and μ_{13}, which are the contrast-pairs that will be returned after analysis

$$\begin{matrix} \mu_{1to2} & \mu_{1to3} \end{matrix}$$
$$\begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} = X$$

Figure 2, A design matrix for the simple experiment.

The design matrix specifies implicitly all the different ways that the direct experimental measurements can be combined to obtain estimates for the chosen parameters. So, in the example above the log-ratio μ_{12} is measured directly on slide A1, but one could also obtain an indirect estimate by combining the estimates of μ_{23} and μ_{31} measured on the other two slides. A least-squares solution to the matrix equation $y = X\mu$ estimates the true expression ratios by fitting a linear equation to the observed data. It finds the closest fit (minimizing squared deviations) between the observed values (log-ratios of probe intensities) and the estimated values (linear model).From these estimates any other contrast-pair can be estimated by $\mu_{ij} = \mu_{1j} - \mu_{1i}$.

5 ACKNOWLEDGEMENTS

We thank our fellow colleagues on the ExGen:MARIE Consortium for their help and suggestions during development of this online tool. Also, we are very grateful to the biologists who have helped with usability issues and testing. This work is supported by the BBSRC. Most importantly we wish to thank the volunteer developers of the free, statistical computing language R, without which this tool would not have been possible.

REFERENCES

Churchill, G.A. 2002. Fundamentals of experimental design for cDNA microarrays. *Nat Genet* **32 Suppl:** 490-495.

Firth, D. 2003. CGIwithR: Facilities for processing web forms using R.

R Development Core Team, R.F.f.S.C. 2005. R: A language and environment for statistical computing.

Schena, M., D. Shalon, R.W. Davis, and P.O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270:** 467-470.

Vinciotti, V., R. Khanin, D. D'Alimonte, X. Liu, N. Cattini, G. Hotchkiss, G. Bucca, O. De Jesus, J. Rasaiyaah, C.P. Smith, P. Kellam, and E. Wit. 2004. An experimental evaluation of a loop versus a reference design for two-channel microarrays. *Bioinformatics*.

Yang, Y.H. and T. Speed. 2002. Design issues for cDNA microarray experiments. *Nat Rev Genet* **3:** 579-588.