A Friendly Statistics Package for Microarray Analysis

P. Sykacek [a,b]*, R. A. Furlong [b], G. Micklem [a,c]

[a] Dept. of Genetics, [b] Dept. of Pathology, [c] Cambridge Computational Biology Institute at the Dept. of Applied Mathematics and Theoretical Physics, University of Cambridge

ABSTRACT

Summary: The friendly statistics package for microarray analysis (FSPMA) is a tool that aims to fill the gap between simple to use and powerful analysis. FSPMA is a platform-independent R-package that allows efficient exploration of microarray data without the need for computer programming. Analysis is based on a mixed model ANOVA library (YASMA) that was extended to allow more flexible comparisons and other useful operations like k nearest neighbour imputing and spike-based normalisation. Processing is controlled by a definition file that specifies all the steps necessary to derive analysis results from quantified microarray data. In addition to providing analysis without programming, the definition file also serves as exact documentation of all the analysis steps.

Availability: The library is available under GPL 2 license and, together with additional information, provided at http://www.ccbi.cam.ac.uk/software/psyk/fspma.html.

Contact: peter@sykacek.net

INTRODUCTION

The number of analysis packages for microarray data is vast and yet one is still faced with the problem of how to best analyse any particular dataset. Easy to use tools are appealing but many are only available commercially. More elaborate packages like LIMMA (Smyth *et al.*, 2003) or YASMA (Wernisch *et al.*, 2003) require programming skills and are thus out of reach for non-specialists. The friendly statistics package for microarray analysis (FSPMA) aims to fill the gap between simple to use, yet powerful analysis. It is a set of R-scripts based on YASMA that makes it possible to explore data efficiently without computer programming. The entire process is controlled by a definition file that specifies all steps to generate analysis results from microarray data. The analysis is centred around an existing tool for mixed model ANOVA (Analysis of Variance) for balanced experiments. Mixed model ANOVA was chosen as this allows for correct treatment of nested effects that would otherwise be regarded as independent *identically* distributed samples. We thus obtain more realistic p-values in the ANOVA table and in

---

*to whom correspondence should be addressed

subsequent tests. The library introduced here provides some useful extensions of the original YASMA package; to allow for more general comparisons, gene ranking is based on contrasts. We provide a k nearest neighbour (knn) based method to impute missing values and also spike-based normalisation which can equally well be used with "housekeeping genes". The tool operates on quantified single and two channel microarray data whether normalised or not, as long as the experiment is a balanced reference design. In addition to providing analysis without programming, the definition file serves as an exact documentation of all analysis steps, which is important in its own right.

DATA LOADING

Analysis requirements in a microarray laboratory can be rather diverse. Experiments are typically done with single or two colour arrays and sometimes the data have been pre-processed; e.g. conversion to log ratios or application of a favourite normalisation method. To obtain a generic solution these various data sources have to be standardised. This is done by having default values for unavailable channels, a boolean dye swap indicator for each file and a flag indicating whether the data is log transformed or not. Headers are ignored and data columns are identified via their column names, so that column order is unimportant and use of heterogeneous file structures in one analysis run does not matter.

IMPUTING AND NORMALISATION

To accommodate poor quality flagged spots or missing information, the library provides an implementation of knn-imputation, (Troyanskaya *et al.*, 2001). Alternatively all such spots can be taken care of by removing their corresponding genes. In terms of normalisation, the library uses YASMA's functionality to provide removal of within-slide location and scale, or removal of the amplitude dependent mean by subtracting a loess fit. In addition, FSPMA allows normalisation based on RNAs of known concentration, spiked into the RNA samples, where the spike residual log ratio (i.e. the difference between actual and theoretical log ratio of spike concentration) is used to normalise the data. Options for spike based

normalisation include removing a global mean or a loess fit, and/or adjusting the variance of each slide to the global variance across all slides. The loess fit can be based on spot position (spatial effects), subgrid number (pin effect) or spot intensity as well as interactions of the above.

ANOVA AND CONTRAST BASED RANKING

In the context of microarray data a mixed model ANOVA was introduced by YASMA (Wernisch *et al.*, 2003). The approach taken in FSPMA requires all effects of an experiment to be specified. Each effect is either *random* or *fixed*. Random effects are variables where the experiment does not contain instances of all possible levels (e.g. biological replicate). Fixed effects are those variables where all possible levels are part of the experiment, or other levels are not of interest (e.g. pathological classification or time point in a longitudinal study). The description of an experiment is automatically converted to an ANOVA model equation, where each effect is considered hierarchically and modelled as an interaction term with the previous grouping. As an example, gene, G, within slide replicate, r, technical replicate, s, and time point, t, where gene and time point are fixed effects and technical replicate and within slide replication are random effects, will result in the ANOVA model equation

$$y_{G,t,s,r} = \mu_G + \alpha_{G,t} + B_{G,t,s} + \epsilon_{G,t,s,r}.$$

We use $y_{G,t,s,r}$ as the expression value, μ_G as the gene specific global mean and $\mu_G + \alpha_{G,t}$ as the mean of each gene-time interaction. Variable $B_{G,t,s}$ is a Gaussian random variable that represents interactions of gene and time with the random effect "technical replicate". Finally $\epsilon_{G,t,s,r}$ is the residual. Such equations are used to calculate the ANOVA table and variance components using the functions provided by YASMA. If the ANOVA table allows rejection of the null hypothesis for the fixed effect of interest (e.g. time), the user may further assess the differences between groups. In order to do that, the library allows for general contrasts, such that evaluations beyond pairwise comparisons are possible. We

Table 1. Time course of mammary gland development, gene expression at day of lactation and hours (hrs.) into involution.

level	pair wise	apoptosis II
lactation day 1	0	-1
lactation day 5	0	-1
lactation day 10	1	-1
involution hrs. 12	-1	-1
involution hrs. 24	0	1
involution hrs. 48	0	1
involution hrs. 72	0	1
involution hrs. 96	0	1

illustrate this in table 1 using a longitudinal study of mammary gland development (Clarkson *et al.*, 2004): the first column shows the time points of the experiment; the second column illustrates a contrast for pairwise comparisons between the last lactation day and involution onset; the third column is a more general contrast that tests for significant differences between groups of time points and is here indicative for causes of type II apoptosis.

In addition a gene-based ANOVA rank list can be produced. This ranks genes by the p-values of an F-statistic that is obtained from the null hypothesis that all levels of the corresponding effect have identical mean. The total number of comparisons within a definition file is used to adjust p-values for multiple testing. For each comparison an ordered gene list is written into a separate tab-delimited file.

DISCUSSION

The R-package FSPMA should serve two purposes. Since it does not require programming, FSPMA based analysis of microarray experiments is accessible to non-experts with a basic understanding of ANOVA, random and fixed effects and contrasts. For the expert, FSPMA allows efficient analysis of balanced reference designs by providing pre-defined definition files. In non-standard situations that go beyond what is possible with mixed effects ANOVA, the library can still serve as a front end for data loading and normalisation. The normalised experiment is then returned to the R command line or the calling script for further analysis.

Acknowledgements

REFERENCES

Clarkson, R. W. E., Wayland, M. T., Lee, J., Freeman, T. & Watson, C. J. (2004) Gene expression profiling of mammary gland development reveals putative roles for death receptors and immune mediators in post-lactational regression. *Breast Cancer Res.,* **6** (2), R92–R109.

Smyth, G. K., Yang, Y.-H. & Speed, T. P. (2003) Statistical issues in microarray data analysis. In *Functional Genomics: Methods and Protocols*, (Brownstein, M. J. & Khodursky, A. B., eds),. Humana Press Totowa, NJ. pp. 111–136.

Troyanskaya, G. O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. & Altman, R. B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics,* **17** (6), 520–525.

Wernisch, L., Kendall, S. L., Soneji, S., Wietzorrek, A., Parish, T., Hinds, J., Butcher, P. G. & Stoker, N. G. (2003) Analysis of whole-genome microarray replicates using mixed models. *Bioinformatics,* **19** (1), 53–61.