Differential proportionality – an alternative to differential gene expression not requiring sample normalization

I. Erb1, T. Quinn2, D. Lovell3, and C. Notredame1

1Centre for Genomic Regulation (CRG), C/Dr Aiguader 88, 08003 Barcelona, Spain; ionas.erb@crg.eu
2Deakin University, Geelong, Victoria, Australia
3Queensland University of Technology, Brisbane, Queensland, Australia

Abstract

In gene expression, the emergence of large aggregated data sets along with new single-cell technologies have led to a heterogeneity of samples that makes normalization extremely difficult. The few existing log-ratio applications to gene expression analysis (Fernandes and others, 2013; Lovell and others, 2015) do not fully overcome the problem of sample heterogeneity as their results depend crucially on the choice of a reference in the form of a gene or gene set (Erb and Notredame, 2016).

Here we propose a differential analysis of all possible gene ratios. More precisely, considering $n$ samples coming from two different conditions, we propose a statistic to detect proportionality (i.e. log-ratio variance close to zero) between genes $\mathbf{x}$ and $\mathbf{y}$ in one condition that differs in the proportionality factor in the other condition:

$$\vartheta(\mathbf{x}, \mathbf{y}) = \frac{k \cdot \text{var } L_{\mathbf{x}, \mathbf{y}}^{\{1, \ldots, k\}} + (n - k) \cdot \text{var } L_{\mathbf{x}, \mathbf{y}}^{\{k + 1, \ldots, n\}}}{n \cdot \text{var } L_{\mathbf{x}, \mathbf{y}}^{\{1, \ldots, n\}}},$$

where by $L_{\mathbf{x}, \mathbf{y}}^{\{1, \ldots, k\}}$ we denote the log ratio of $\mathbf{x}$ and $\mathbf{y}$ over the indices $\{1, \ldots, k\}$. $\vartheta$ can be obtained from a decomposition of log-ratio variance into between and within group variance. (The numerator corresponds to the latter, and $\vartheta$ values fall between 0 and 1, with smaller values indicating better separation.) Note that $\vartheta$ is related to the statistic $F$ underlying one-way ANOVA by $F = (1 - \vartheta)/\vartheta$.

In fact, a standard differential expression framework can now be applied (applied, however, on ratios) using false discovery rates from permutation tests to detect significant values of $\vartheta$. As an example, we apply this framework to a data set of 98 post-mortem brain samples (Lonsdale, J. and others, 2013) from cortex and cerebellum. Unlike in classical differential expression studies, where the main result is a list of genes whose read counts differ between conditions, here we obtain a list of gene pairs whose ratio of co-expression differs between conditions. This allows for a subsequent network analysis, cf. (Tesson and others, 2010) for the classical equivalent called differential correlation.

We also derive an alternative to $\vartheta$ that can handle zeroes and compares with the use of pseudo counts. For this statistic, the three terms of the form $k \cdot \text{var } L_{\mathbf{x}, \mathbf{y}}^{\{1, \ldots, k\}}$ in (1) are replaced respectively by

$$\sum_{i=1}^{k} \left( \frac{x_i^\alpha}{\frac{1}{k} \sum_{j=1}^{k} x_j^\alpha} - \frac{y_i^\alpha}{\frac{1}{k} \sum_{j=1}^{k} y_j^\alpha} \right)^2.$$

This is inspired by the observation that chi-square distances converge to log-ratio variances when applying a Box-Cox transformation with the parameter $\alpha \to 0$ (Greenacre, 2009). We supplement this work with an R package that provides a fast and efficient implementation of these analyses.

References


