

What is the response? Identifying interesting behaviour in microarray time series data

Katherine Lawler, Alvis Brazma

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD

Genome-scale expression studies using microarray platforms increasingly report results in the form of short time series. Recent examples of microarray time series in the literature include tracking the stress response of an organism after a specific stress, measuring gene expression at different times during development¹, and the well-studied early cell cycle experiments². In addition, microarray platforms are increasingly being used to measure quantities other than mRNA expression level, resulting in multiple simultaneous short time series; a current topic of wide interest is the study of mRNA kinetics by simultaneous measurement of transcription rates and expression levels as they change over time in response to a stimulus³.

Microarray timeseries typically have fewer than 20 time points – and often far fewer⁴. This is unlikely to change any time soon: experiments using microarrays are still expensive and time consuming. Timepoints are typically not uniformly sampled. It remains a challenging problem to make statistically sound inferences from the short time series generated by microarray studies, compounded by the fact that the standard time series literature overwhelmingly deals with the relatively long time series of financial data, meteorological data, and many other fields.

Here we focus on one particular problem to illustrate the importance of using adequate time series models to interrogate microarray time series data. Given a microarray timecourse, or a series of related short time series, which genes are showing a response? And what is this response? This is a direct analogy of identifying differential expression between two steady state conditions: in a time series experiment we can still ask which genes are showing differential expression between timepoints, but we can also ask which genes display a differential behaviour across the whole timecourse, and in contrast which genes are “constant” across the whole timecourse.

To allow for the many sources of error inherent in microarray data, we treat a constant response as white noise - a process with constant mean and constant variance over the whole timecourse. We investigate several methods to test whether a process is white noise, including Box-Pierce and regression tests, adapting each method where possible to deal with small numbers of timepoints (~15 time points). We discuss their applicability to short non-uniform time series and the problem of multiple testing when applying genome-scale filters. We compare these modified time series methods to the filtering and 'two-fold cutoff' methods which are used as a matter of course in differential expression studies. We show that the temporal ordering of points in a time series should not be ignored when performing initial filtering of genes, and that by combining filters which adopt different models for constant response we can generate plausible, defensible lists of unresponsive genes, and therefore generate lists of genes which are showing some response over the timecourse. We illustrate these methods using gene expression time series data for stress response in yeast, and discuss the current outstanding problems which need to be tackled for the analysis of microarray time series datasets.

¹ Sasik et al. *Bioinformatics* 18:61-66, 2002

² Spellman et al. *Mol Bio Cell* 9:3273-3297, 1998

³ Perez-Ortin et al. *Trends in Genetics* 23:250-257, 2007

⁴ Bar-Joseph *Bioinformatics* 20(16):2493-503, 2004