Extracting meaning from microarray data

R.K. Curtis1 and M.D. Brand
MRC Dunn Human Nutrition Unit, Hills Road, Cambridge CB2 2XY, U.K.

Abstract
Gene expression is complex: many mRNAs change in abundance in response to a new condition. But while some of these expression changes may be direct, many may be downstream, indirect effects. One of the major problems of microarray data analysis is distinguishing between these changes. Some of the most common methods of analysis are discussed, in the context of their ability to distinguish between direct and indirect expression changes. The application of modular control analysis to microarray data in order to partition and quantify the importance of mRNA clusters in mediating responses is described.

Introduction
Microarrays can be used to profile the expression of every gene in a sample. Their use is increasing and microarrays have been used to investigate many aspects of biology, such as the cell cycle [1], differentiation [2], and how environmental stress [3] and diseases such as cancer [4] affect gene expression. Microarray experiments reveal that hundreds and sometimes thousands of mRNAs change in abundance between experimental conditions. The datasets produced are very large, and there are many different ways of analysing the data produced, depending upon the aim of the experiments. For example, Alizadeh et al. [4] were interested in diagnosing cancer subgroups by classifying samples, Hughes et al. [3] predicted the function of novel genes, and Ideker et al. [5] used expression data to investigate the regulation of a metabolic pathway.

Visual inspection
The simplest way to analyse microarray data is to look at the mRNAs that change. But if the expression of a gene of interest changes, you still do not know if this is a direct change or a downstream indirect effect, or how important it is in causing the response. However, this is the commonest method in practice, with large changes in the expression of a set of mRNAs thought to be relevant greeted with satisfaction, and the lack of such changes greeted with disappointment and frustration. Nonetheless, this method does throw up unexpected genes for further study.

Clustering
Clustering is often used to predict the function of unknown genes, on the basis that genes with similar functions tend to be co-expressed and, in that sense, cluster together [6]. There are several types of cluster analysis. Hierarchical clustering involves repeatedly merging mRNAs or mRNA clusters, based on a distance measure, to form a new cluster. The distance measures reflect the absolute expression values (Euclidean), or expression trends (correlation-based). Genes of similar function tend to cluster together, and clustering is widely used to predict the function of novel genes [3], and to classify samples such as tumours [4]. Two-dimensional clustering (or biclustering) is used by Kluger et al. [7]. This involves clustering by both gene expression and by sample in order to improve the overall result.

Microarray data analysis
When cells are given a stimulus (such as the addition of a hormone) and a response (such as growth rate) is measured, there may be more than one pathway involved in effecting that response. Some of these pathways may require transcription of genes. If gene expression during a response has been profiled using microarrays, often many mRNAs change in abundance, revealing that gene expression is complex. The stimulus may have direct effects on the mRNAs, for example a hormone binds a transcription factor, which directly increases the expression of three mRNAs. There may also be indirect effects: the transcription factor regulates those three mRNAs, which in turn indirectly regulate hundreds of other genes. Both the direct and indirect effects may be involved in causing the eventual response to the new experimental condition. One of the major problems of microarray data analysis is finding a method of analysis that is able to distinguish between these changes, and can quantify the importance of each of these direct and indirect effects on the end response, and on the cascade of mRNA responses.

Key words: clustering, metabolic control analysis, microarray, modular control analysis.

©2003 Biochemical Society

1To whom correspondence should be addressed (e-mail keira.curtis@mrc-dunn.cam.ac.uk).
(variables), but does not provide information about how those changes propagate to a system response from a stimulus.

**Genetic networks**

Gene expression data from knockout experiments can be used to infer genetic networks [5,10]. Using information about the expression of which genes change when another is knocked out, a graph of nodes (genes) connected by edges (interactions between the genes) is constructed. This predicts relationships between genes, such as transcriptional regulation, and some methods are able to distinguish between direct and indirect interactions. However, it is semi-quantitative at best, and as with cluster analysis, does not describe the transmission of signals/responses through a system.

Despite all the methods described above, there is still a need for a means of analysing this type of microarray data that is able to distinguish direct and indirect effects, and can find and quantify pathways that affect a response. Once important mRNAs have been identified, this may help to indicate the mechanisms of a response, such as the change to a disease state, and those genes that are targets for modifying that response.

**Modular control analysis**

We have been applying modular control analysis [11,12] to microarray data. This is a subset of metabolic control analysis [13], which has been used to investigate the control and regulation of biological systems such as metabolic pathways. The modular approach involves simplifying a system by dividing it into a small number of biologically meaningful modules. Ainscow and Brand [14] investigated the response of hepatocyte metabolism to the effectors glucagon and adrenaline. Using control analysis of a nine-module system, the response was partitioned and quantified, revealing which metabolic modules responded directly to the effector. Indirect routes of the response were also quantified, revealing, for example, that glucagon acted directly on the glucose-release module, while adrenaline acted by increasing the glucose 6-phosphate concentration, which in turn stimulated glucose release [14].

**Application to microarray data**

Modular regulation analysis can be applied to microarray data in order to partition and quantify the importance of expression changes that are involved in a response [15]. First, the microarray data are simplified by grouping mRNAs together into clusters, based on similarities in their expression patterns. Next, how the input to the system affects the mRNA clusters is calculated. The input might be a change in growth medium, a genetic modification, the addition of an effector (e.g., hormone, drug or protein), or the change to a new state (e.g., disease). This acts to change the expression of each mRNA cluster and is described using integrated response coefficients, which are easy to calculate from the microarray data.

How the mRNA clusters affect the output response is described using elasticity coefficients. The output can be any quantifiable response, such as the rate of an enzyme; the concentration of a metabolite; or a physiological marker, such as growth rate, cell volume or mortality. Expression data from a series of genetic modulation experiments (knockouts) are used to calculate elasticity coefficients, which describe how much an mRNA cluster matters for the output response.

Once integrated responses (how the input affects each cluster) and elasticity coefficients (how much a cluster matters for the response) have been calculated, these are multiplied together to give partial response coefficients. These are the main result of the analysis and describe how much of the response to the input is transmitted by each mRNA cluster. Clusters with large partial response coefficients are important for the response. Once important clusters have been identified, they can be opened up to see which mRNAs they contain; these mRNAs are then possible targets for modifying the response.

**Conclusions**

Unlike other methods, modular control analysis of microarray data does not require any prior information about the function of genes or the position of knockouts. It can be used to quantify the importance of mRNAs in physiological responses such as growth. It is highly applicable to microarray data and is transferable to proteome and metabolome data.

**References**


Received 30 June 2003