BIOINFORMATICS

Vol. 20 no. 8 2004, pages 1272–1284
DOI: 10.1093/bioinformatics/bth082

Analysing microarray data using modular regulation analysis

R. Keira Curtis[*,†] *and Martin D. Brand*

MRC Dunn Human Nutrition Unit, Hills Road, Cambridge, CB2 2XY, UK

Received on June 24, 2003; revised on August 7, 2003; accepted on September 2, 2003
Advance Access publication February 19, 2004

ABSTRACT

Motivation: Microarray experiments measure complex changes in the abundance of many mRNAs under different conditions. Current analysis methods cannot distinguish between direct and indirect effects on expression, or calculate the relative importance of mRNAs in effecting responses.

Results: Application of modular regulation analysis to microarray data reveals and quantifies which mRNA changes are important for cellular responses. The mRNAs are clustered, and then we calculate how perturbations alter each cluster and how strongly those clusters affect an output response. The product of these values quantifies how an input changes a response through each cluster.

Two published datasets are analysed. Two mRNA clusters transmit most of the response of yeast doubling time to galactose; one contains mainly galactose metabolic genes, and the other a regulatory gene. Analysis of the response of yeast relative fitness to 2-deoxy-D-glucose reveals that control is distributed between several mRNA clusters, but experimental error limits statistical significance.

Contact: rkc24@cam.ac.uk

INTRODUCTION

Experiments using microarrays can measure the relative concentrations of all mRNA transcripts in a preparation, and illuminate many aspects of biology, including differentiation (Le Naour *et al*., 2001), cell cycle (Cho *et al*., 1998), environmental stress (Gasch *et al*., 2000) and cancer (van't Veer *et al*., 2002). Changes in gene expression are complex, sometimes thousands of mRNAs change expression between states. There are many different ways of analysing microarray data, depending upon the purpose of the experiments.

Microarray data can be used to classify samples, e.g. to discriminate between cancer types (Alizadeh *et al*., 2000). This uses supervised clustering, requiring prior knowledge about the training set used to classify the remaining samples. Principal components analysis can identify the individual genes within a disease signature that correlate best with the phenotype (Alter *et al*., 2000; Raychaudhuri *et al*., 2000), but it does not quantify the importance of those few genes for a response. Gene expression patterns can be used to predict the functions of unknown genes, because genes with similar functions may be coexpressed and, therefore, cluster together (Eisen *et al*., 1998). However, these analyses do not reveal which of the unknown genes are most important, to allow their targetting for early investigation. Finally, microarray data may be used to infer the structure of regulatory networks, predicting the regulatory interactions between genes (Rung *et al*., 2002; Schlitt and Brazma, 2002). This method can predict the direct and indirect interactions between genes, but is not quantitative. Induction kinetics can indicate which are the direct (rapid) responses and which are indirect (slow), but this method requires targeted genetic mutants and a time course (Devaux *et al*., 2001; Le Crom *et al*., 2002).

Despite abundant microarray publications, there is no means of analysing microarray data to find and quantify the important gene expression changes. If a gene of interest changes expression between two conditions, such as healthy and diseased, it is unclear whether that gene is important and involved in causing the disease response. The changed expression could be crucial in causing the symptoms, but it could also be coincidental, part of a downstream reaction to the disease, or a minor component of the response. It is also not necessarily the genes with the greatest expression changes that are most important in causing the response; it will depend on how strongly the response depends on the expression of the gene. Transcripts with weak effects may be more strongly upregulated to achieve the same response as transcripts with strong effects. Finally, while some mRNA changes may be direct, many may be indirect. None of the existing analytical techniques can discriminate between direct and indirect mRNA changes, and quantify the importance of those changes for a response. Many analysis methods require prior knowledge of the system under investigation: knockouts targeted to the pathway of interest or a classified subset of samples as a training set. We show here that control analysis (Fell, 1997; Kacser and Burns, 1973) can identify important mRNAs, quantify how much of the response they mediate and identify whether they are regulated directly or by the expression of other genes.

*To whom correspondence should be addressed.

†Present address: Department of Clinical Biochemistry, University of Cambridge, Addenbrooke's Hospital, Box 232, Hills Road, Cambridge, CB2 2QR, UK.

Metabolic control analysis is a mathematical framework for quantifying regulation of biological systems (Fell, 1997; Heinrich and Rapoport, 1974; Kacser and Burns, 1973). It has mainly been applied to complex metabolic systems such as mitochondrial bioenergetics (Brand, 1996, 1997). The modular approach to control analysis involves simplifying a complex system by grouping reactions into biologically meaningful modules, allowing the regulation of large, complex systems to be experimentally solved (Ainscow and Brand, 1999b). Regulation analysis is a subset of control analysis that allows the response to an effector to be partitioned (Brand, 1997). For example, the response to adrenaline of glucose release from hepatocytes has been partitioned, to quantify how much of the response is transmitted through each of the system's reaction blocks (Ainscow and Brand, 1999a). Modular regulation analysis also determines whether the response involves direct action of adrenaline on the glucose-producing and glucose-releasing reactions, or is indirect and mediated through other reaction blocks.

Here, we apply modular regulation analysis to microarrays using a worked example and then analyse two published datasets. We first simplify expression data by clustering mRNAs with similar expression patterns, then quantify how much of the response of some output (e.g. growth rate) to an external change is mediated by each cluster of mRNAs. Clusters that mediate a large proportion of the response contain genes that are potential targets for manipulating the response. We also show how the analysis can quantify how each cluster of mRNAs is regulated, and whether regulation by the external change is direct (e.g. via direct activation of transcription factors) or is mediated by direct changes in other mRNA clusters (e.g. via altered expression of transcription factor mRNAs). Control analysis uses relative changes in variables rather than absolute values and so is ideally suited to microarray data, which are often represented as ratios of gene expression between conditions.

METHODS

Dataset requirements

An appropriate set of experiments is required before modular control analysis can be applied. First, full microarray data are required for the reference system (such as wild-type cells), compared with the same system under the input stimulus of interest. The input might be a change in growth medium, a genetic modification, the addition of an effector (e.g. hormone or drug) or the change to a new state (e.g. disease) or physiological condition. Second, an extensive series of genetic modulations of the reference cells is required, with full expression data for these modulations compared with the reference. Third, an output of interest must be measured in each of these reference, input and modulation experiments. This can be any quantifiable response, including the rate of an enzyme or pathway; the concentration of a metabolite, signalling intermediate, protein, mRNA molecule or mRNA cluster; or a physiological marker, such as growth rate, cell volume, performance or mortality. Fourth, to allow statistical analysis of significance, repeat microarrays and output measurement are needed.

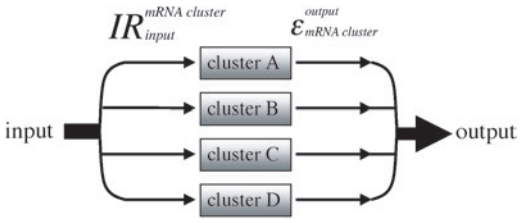


Fig. 1. Modular regulation analysis of the effect of an input on an output, showing the structure of the system. An input changes the expression of each of the four mRNA clusters as described by integrated response coefficients, *IR*. The mRNA clusters then change the output of interest, as described by elasticity coefficients, ε. The integrated response and elasticity coefficients for each cluster can be multiplied to give partial response coefficients, which quantify the relative importance of each mRNA cluster in transmitting the response of the output to the input.

Clustering of transcripts

To simplify the microarray data and make the analysis experimentally feasible, transcripts are first grouped into a relatively small number of clusters according to how they respond to experimental manipulations, as described later.

Application of control analysis

The system is conceptually organized as shown in Figure 1. The input affects the expression of the mRNA clusters. This is described using integrated response coefficients (*IR*), which represent the relative change in expression of each cluster in response to the input stimulus (Fig. 1, left). Positive integrated response coefficients indicate that expression increases; negative coefficients mean it decreases. At this stage we are not interested in whether responses are direct or indirect, but simply in quantification.

How each mRNA cluster affects the output response is described using elasticity coefficients (ε) (Fig. 1, right). Positive elasticity coefficients mean that increased expression of a cluster increases the output response; negative coefficients indicate it decreases this response.

Finally, partial response coefficients (*R*) are obtained by multiplying each integrated response coefficient (how much the cluster changes) by the relevant elasticity coefficient (how much the cluster affects the response). Partial response coefficients describe how much of the change in the output response is transmitted by each mRNA cluster. Positive partial response coefficients mean that the signal transmitted by a cluster increases the output response, e.g. the upregulation of genes that enable or downregulation of genes that oppose

the response. Negative coefficients characterize clusters that transmit decreases in the output.

In this way, the analysis partitions the response of the output to the input, and quantifies how much of that response each mRNA cluster is responsible for: how much of the response would be absent if the expression of that cluster did not change. Clusters with high partial response coefficients contain one or more mRNAs that are important for the response.

The amount of response transmitted by non-mRNA routes not requiring changes in gene expression, e.g. by allosteric alteration of enzymes, is obtained by subtracting the sum of the partial response coefficients (the response transmitted by the various mRNA routes) from the observed overall response. Any contribution of mRNAs not represented in the microarray because it is incomplete will appear as part of this non-mRNA route.

The analysis step by step

To illustrate the analysis, we simulated a dataset that fulfilled the requirements discussed above (using modified random numbers). It consists of hypothetical mRNA expression data and quantified output data for the reference and test ('input') conditions and for 11 different genetic manipulations. A worked example using this simulated dataset is shown in Figures 2 and 3 and described below.

Clustering The first step is to group mRNAs into clusters, based on their expression across all genetic manipulation experiments. Clustering reduces the number of independent experiments required in the second step for calculation of elasticity coefficients: the number of mRNA clusters cannot exceed the number of genetic modulations. The simulated dataset contains 11 genetic modulation experiments, so the mRNAs are grouped into 11 clusters using Euclidean hierarchical clustering (Fig. 2.1). The average expression of each of the 11 clusters for the reference and input conditions and for each of the 11 simulated genetic manipulations is calculated, to give a data matrix (Fig. 2.2).

Even for yeast, with its genome of about 6000 genes, 6000 independent experiments would be required if the mRNAs were not clustered. de la Fuente *et al*. (2002) and de la Fuente and Mendes (2002) used a related method to reverse-engineer gene networks from microarray data, but without clustering; this method is conceptually different as it could be used to investigate a subsystem of 50 individual genes. Our analysis looks at all the genes in a system, simplified into 50 clusters, lowering resolution but simplifying a complicated problem.

Before clustering, it may be helpful to reduce a microarray dataset by excluding mRNAs that do not change significantly in any experiments, assigning them integrated and partial response coefficients of zero.

There are various types of clustering algorithm. Hierarchical clustering is preferred because the number of clusters is dictated by cutting the dendrogram at different points (Fig. 2.1). Hierarchical clustering also simplifies merging and splitting of clusters compared with other techniques such as K-means (Tavazoie *et al*., 1999). We need to calculate the mean expression of the mRNAs in a cluster, so Euclidean hierarchical clustering (which groups using absolute values) is more appropriate than correlation-based clustering (which groups by expression trends). For example, correlational clustering groups mRNAs with the same trends of increasing or decreasing in particular experiments, but if one has much larger expression changes than the others, the mean expression level will not quantitatively represent the whole cluster. There are several clustering packages available, such as Eisen's cluster and treeview (Eisen *et al*., 1998). We use the European Bioinformatics Institute's online Expression Profiler (www.ebi.ac.uk/microarray/).

Microarray data often contain missing values. If a gene has only occasional missing values, it should be assigned the mean expression value for its cluster at those missing values. If the expression of a particular gene is missing in all experiments, any response it transmits will appear as part of the non-mRNA route.

Changing the clustering can affect the calculated partial responses, showing the importance of determining the statistical significance of the coefficients. It is not important whether the genes in a cluster share biological function (although it is likely that they will have related functions, since clustering reflects the underlying biological design). It is important, however, that the genes in a cluster have similar expression across the series of experiments performed. There is a trade-off between the number of clusters used and the accuracy of the microarray data. Decreasing the number of clusters involves forcing mRNAs together into clusters where they do not necessarily fit, making the clustering less meaningful. The minimum number of clusters is two, requiring two genetic modulation experiments, but at this very low resolution the analysis is unlikely to be informative. As the number of clusters increases, resolution improves, but experimental accuracy must also improve for statistically significant results. Conversely, microarrays are expensive, limiting the number of experiments and therefore the number of clusters.

Calculation of coefficients Integrated response coefficients (Ainscow and Brand, 1999a; Fell, 1997; Kacser and Burns, 1979; Kholodenko, 1988) describe how each mRNA cluster responds to the input stimulus. They are system responses that incorporate all the interactions of the clusters with each other and with other system components, such as metabolites. All the routes by which the change is effected are included so that knowledge of the individual routes is irrelevant. 'Integrated' shows that they apply to large step changes in the input, to distinguish them from true response coefficients that refer to infinitesimal changes. These coefficients are simply the normalized change in the average expression of each cluster between the reference sample (r) and the cells given the input

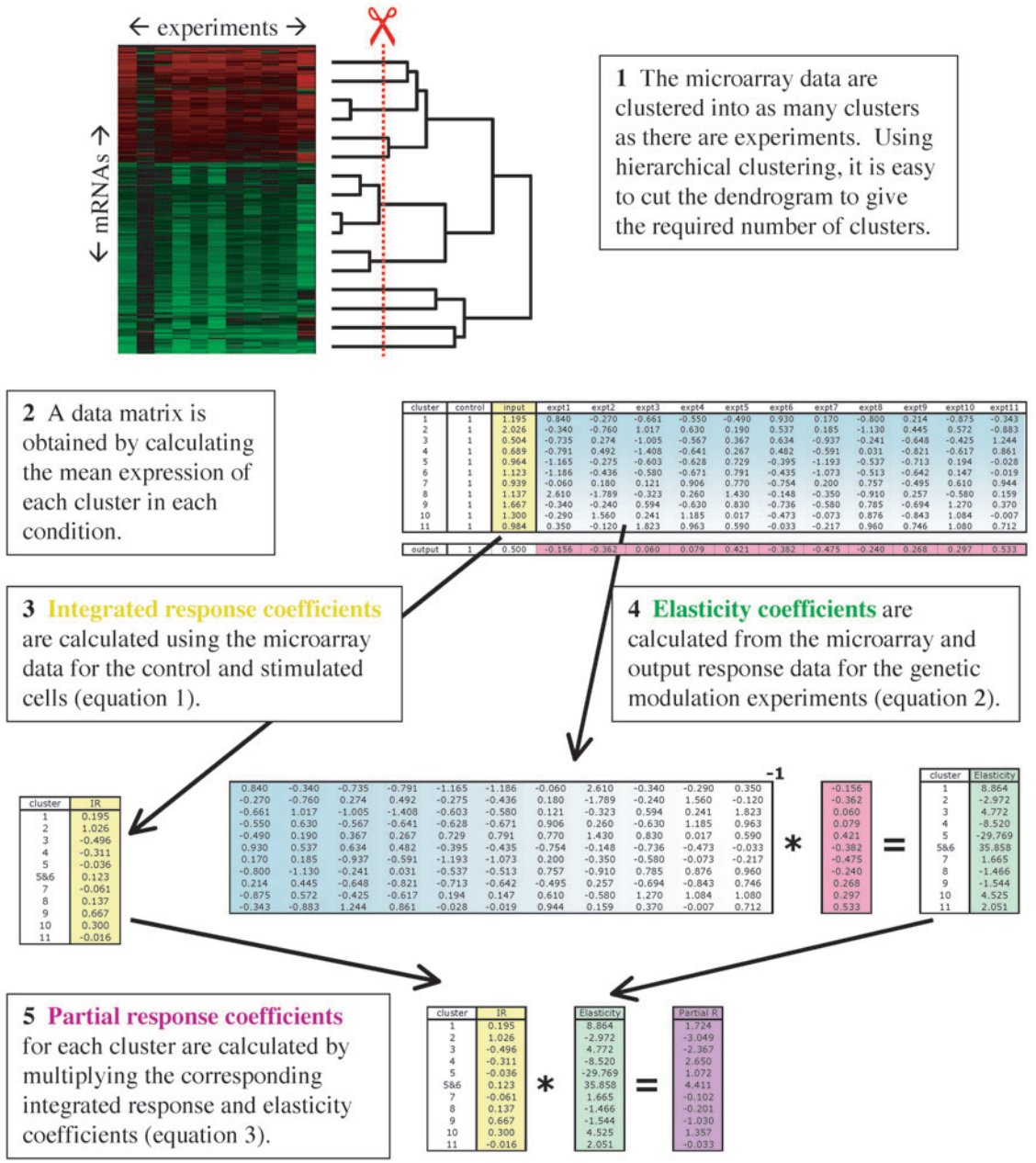
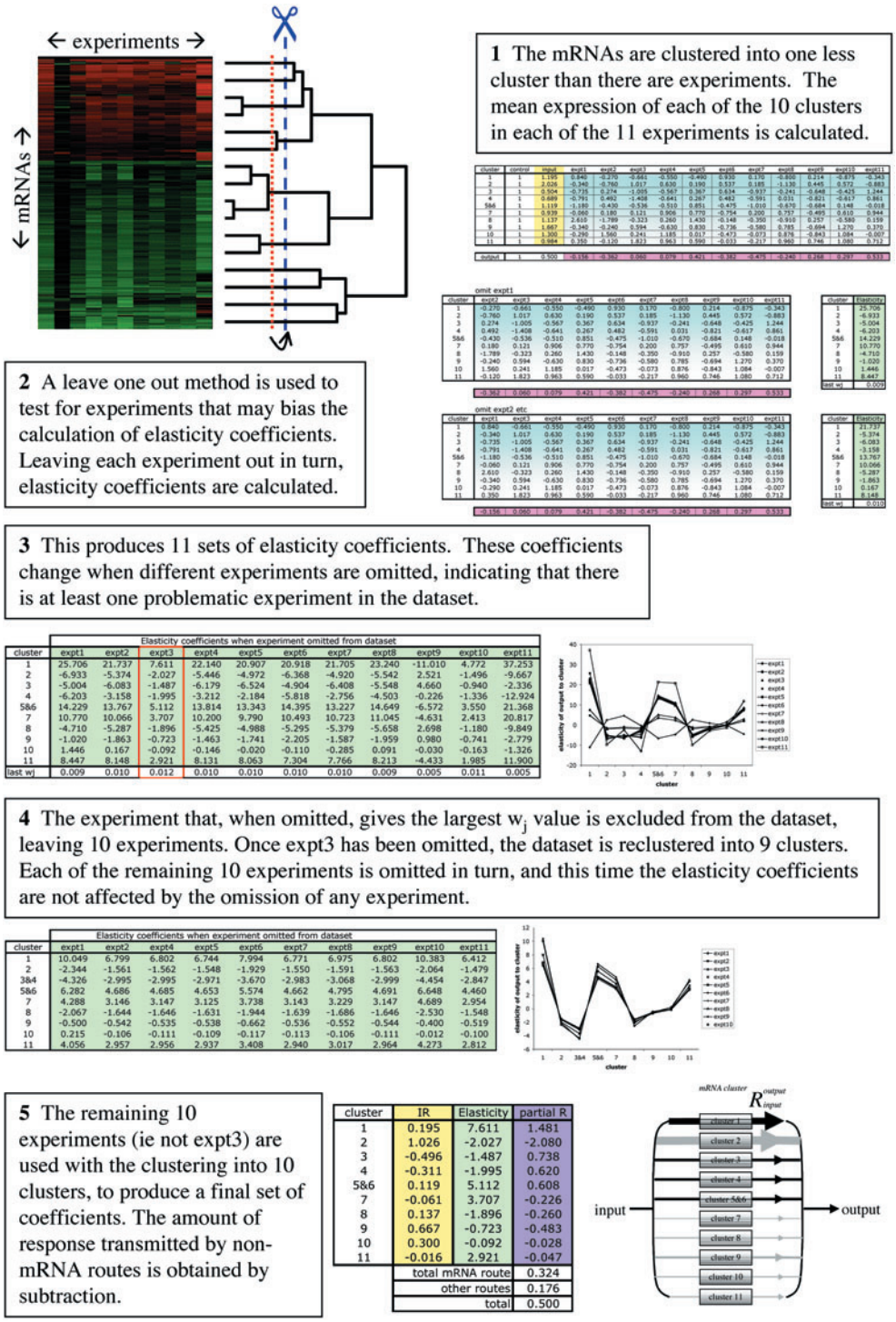


Fig. 2. Worked example showing calculation of partial response coefficients from a synthetic dataset of one input, 11 experimental modulations and a quantified output under each condition. The picture of microarray data is for illustration and is unrelated to the data used. (2.1) First, the mRNAs are clustered into 11 clusters. (2.2) The mean expression of each cluster in each experiment is calculated. (2.3) Integrated response coefficients describing how each cluster responds to the input are calculated using Equation (1). (2.4) Elasticity coefficients describing how the mRNA clusters affect the output are calculated using Equation (2). (2.5) Partial response coefficients describing how much of the response is transmitted by each mRNA cluster are calculated using Equation (3).

stimulus (i) [Equation (1)]. For each of the 11 clusters in the worked example, an integrated response is calculated using Equation (1) (Fig. 2.3).

$$IR^{\text{mRNA cluster}}_{\text{input}} = \frac{\text{ratio}(i{:}r) - \text{ratio}(r{:}r)}{\text{ratio}(r{:}r)}. \qquad (1)$$

An elasticity coefficient (Fell, 1997; Kacser and Burns, 1979) describes how a small change in the expression of one mRNA cluster affects the output response. Because the elasticities are calculated by solving a set of simultaneous equations, several independent modulations (m) of the system, different from the input experiment, are needed (Ainscow and



Elasticity coefficients when experiment omitted from dataset

cluster	expt1	expt2	expt3	expt4	expt5	expt6	expt7	expt8	expt9	expt10	expt11
1	25.706	21.737	7.611	22.140	20.907	20.918	21.705	23.240	-11.010	4.772	37.253
2	-6.933	-5.374	-2.027	-5.446	-4.972	-6.368	-4.920	-5.542	2.521	-1.496	-9.667
3	-5.004	-6.083	-1.487	-6.179	-6.524	-4.904	-6.408	-5.548	4.660	-0.940	-2.336
4	-6.203	-3.158	-1.995	-3.212	-2.184	-5.818	-2.756	-4.503	-0.226	-1.336	-12.924
5&6	14.229	13.767	5.112	13.814	13.343	14.395	13.227	14.649	-6.572	3.550	21.368
7	10.770	10.066	3.707	10.200	9.790	10.493	10.723	11.045	-4.631	2.413	20.817
8	-4.710	-5.287	-1.896	-5.425	-4.988	-5.295	-5.379	-5.658	2.698	-1.180	-9.849
9	-1.020	-1.863	-0.723	-1.463	-1.741	-2.205	-1.587	-1.959	0.980	-0.741	-2.779
10	1.446	0.167	-0.092	-0.146	-0.020	-0.110	-0.285	0.091	-0.030	-0.163	-1.326
11	8.447	8.148	2.921	8.131	8.063	7.304	7.766	8.213	-4.433	1.985	11.900
last wj	0.009	0.010	0.012	0.010	0.010	0.010	0.010	0.009	0.005	0.011	0.005

Elasticity coefficients when experiment omitted from dataset

cluster	expt1	expt2	expt4	expt5	expt6	expt7	expt8	expt9	expt10	expt11
1	10.049	6.799	6.802	6.744	7.994	6.771	6.975	6.802	10.383	6.412
2	-2.344	-1.561	-1.562	-1.548	-1.929	-1.550	-1.591	-1.563	-2.064	-1.479
3&4	-4.326	-2.995	-2.995	-2.971	-3.670	-2.983	-3.068	-2.999	-4.454	-2.847
5&6	6.282	4.686	4.685	4.653	5.574	4.662	4.795	4.691	6.648	4.460
7	4.288	3.146	3.147	3.125	3.738	3.143	3.229	3.147	4.689	2.954
8	-2.067	-1.644	-1.646	-1.631	-1.944	-1.639	-1.686	-1.646	-2.530	-1.548
9	-0.500	-0.542	-0.535	-0.538	-0.662	-0.536	-0.552	-0.544	-0.400	-0.519
10	0.215	-0.106	-0.111	-0.109	-0.117	-0.113	-0.106	-0.111	-0.012	-0.100
11	4.056	2.957	2.956	2.937	3.408	2.940	3.017	2.964	4.273	2.812

cluster	IR	Elasticity	partial R
1	0.195	7.611	1.481
2	1.026	-2.027	-2.080
3	-0.496	-1.487	0.738
4	-0.311	-1.995	0.620
5&6	0.119	5.112	0.608
7	-0.061	3.707	-0.226
8	0.137	-1.896	-0.260
9	0.667	-0.723	-0.483
10	0.300	-0.092	-0.028
11	-0.016	2.921	-0.047
		total mRNA route	0.324
		other routes	0.176
		total	0.500

Fig. 3. Worked example illustrating the testing process that finds and excludes problematic experimental modulations. The dataset described in the text and Figure 2 is used. (3.1) The number of clusters is reduced by one: from 11 to 10. (3.2) Each of the 11 experiments is omitted in turn and each time elasticity coefficients and w_j values are calculated. If omitting any experiment makes no difference to the elasticity coefficients, these 11 experiments are satisfactory. (3.3) Omitting an experiment does affect the elasticity coefficient values. (3.4) The experiment that, when omitted, gives the highest w_j value (experiment 3) is excluded permanently from the dataset. The microarray data are then reclustered into nine clusters and this process repeated. At nine clusters and 10 modulations, leaving out any experiment does not change the elasticity coefficients, indicating that the remaining 10 modulations are satisfactory. (3.5) The remaining 10 experimental modulations (i.e. not experiment 3) are used, with clustering into 10 clusters, to calculate a final set of coefficients.

Brand, 1999b; Giersch, 1994; Kacser and Burns, 1979). Use of the multiple modulation method (Giersch, 1994; Kacser and Burns, 1979) means that it is not necessary to perturb each gene individually. In each experiment, expression and the output response are measured and used to solve a set of simultaneous equations to obtain the unknowns: the elasticity coefficients [Equation (3)]. The worked example illustrates the calculation of elasticity coefficients using matrix algebra to solve the equations constructed from the 11 manipulation experiments (Fig. 2.4).

$$\frac{\text{output}_m - \text{output}_r}{\text{output}_r} = \sum_{\text{all mRNA clusters}} \left[\varepsilon^{\text{output}}_{\text{mRNA cluster}} \cdot \frac{\text{ratio}(m{:}r) - \text{ratio}(r{:}r)}{\text{ratio}(r{:}r)} \right]. \quad (2)$$

The modulations could involve gene knockout or overexpression, or other genetic manipulation of the reference cells, but not an environmental change such as altered growth conditions. This is because environmental changes could act through the non-mRNA route, adding an extra and different unknown to each of the simultaneous equations, making their solution impossible. Points from a time course are probably unsuitable because they may represent non-steady states and the non-mRNA routes may be different at different times. Genetic modulations are good, as they directly affect the mRNA clusters. Changes in genes for transcription factors may be ideal, changing many transcripts in characteristic patterns. The modulations must be independent, each inducing a different pattern of mRNA changes. Exactly what the genetic manipulations are is not important, as long as they alter gene expression in different ways. It is not even necessary to identify which gene has been overexpressed or knocked out in a modulation. However, if there is information about the system, it can be used to increase the chances of a manipulation being useful, e.g. by deliberately modifying the expression of transcription factors. If two manipulations result in very similar expression patterns, then only one of the pair should be used.

Why are elasticity coefficients used to describe how the mRNA clusters affect the output, rather than control coefficients? Typically, modular regulation analysis of a metabolic system would group enzymes into reaction modules, having concentration control coefficients over connecting metabolite intermediates and elasticities to these intermediates. Because mRNAs are also grouped into blocks or clusters, they might appear equivalent to these reaction modules. However, the mRNAs are instead equivalent to metabolite intermediates, whose concentrations are controlled by the input. All routes by which they affect the output are downstream of their concentration. The output is treated as a large module containing all these downstream mechanisms, such as translation of the mRNA into protein, and that protein's direct or indirect action on the response. This complex output module (the response) has elasticities to the concentration of each mRNA cluster.

Partial response coefficients (Ainscow and Brand, 1999a; Kholodenko, 1988) are the main result of the analysis because they describe how much of the change in the output response is transmitted through each mRNA cluster. This does not necessarily mean that every mRNA in an important cluster is involved in transmitting the response, one or more of the mRNAs in the cluster may be important. These coefficients are calculated by multiplying the integrated response and elasticity coefficients corresponding to each cluster (Ainscow and Brand, 1999a; Kholodenko, 1988) [Equation (3), Fig. 2.5]. In the worked example, a partial response coefficient is obtained for each of the 11 clusters.

$$^{\text{mRNA cluster}}R^{\text{output}}_{\text{input}} = IR^{\text{mRNA cluster}}_{\text{input}} \cdot \varepsilon^{\text{output}}_{\text{mRNA cluster}}. \quad (3)$$

Validation Some of the genetic modulation experiments may be inadequate, leading to invalid values for the coefficients. An experiment may be inadequate because it is not independent of other experiments, has unusually high experimental error, or directly changes the nature of a cluster and affects its coefficients. Figure 3 illustrates the validation of experiments.

A 'leave one out' method identifies problematic experiments. The number of clusters is reduced by one, which is easy using hierarchical clustering. In the worked example, clusters 5 and 6 are the most similar and are merged into a single larger cluster, 5&6, leaving 10 clusters and 11 genetic modulation experiments (Fig. 3.1). Each experiment is then omitted in turn, leaving 10 clusters and 10 experiments, and the elasticity coefficients are recalculated each time (Fig. 3.2). If the resulting 11 sets of elasticities are similar, then all the genetic modulation experiments are satisfactory: leaving out any experiment makes little difference to the calculation, and the 11 cluster-11 experiment solution can be used.

If, as in the worked example (Fig. 3.3), omitting an experiment does make a difference to the calculated elasticities, then the dataset contains at least one genetic modulation experiment that disagrees with the rest. This could be one with particularly large experimental error, but more probably two experiments are similar, not independent. There is insufficient information to solve the simultaneous equations if both are included, and the calculation is biased by the small differences between them. When using matrix algebra to solve simultaneous equations, a matrix containing two identical equations (identical genetic modulation experiments) is singular: its determinant is zero and the matrix cannot be inverted. A matrix containing two almost identical experiments is closer to being singular than one that does not. The last w_j value from a singular value decomposition of the matrix indicates how close to singular the matrix is: values near zero indicate a matrix almost singular. The absolute value of the determinant of the matrix could also be used for this purpose. We want to find the least independent experiment, i.e. when excluded it

makes the matrix the least singular. In the worked example of 10 clusters and 11 experiments, for each of the 11 matrices produced by leaving out each experiment in turn, the last w_j value is calculated. Experiment 3 gives the highest w_j when omitted, so it is rejected, leaving a dataset of 10 clusters and 10 experiments. The testing process is repeated, reducing the number of clusters by one, omitting each experiment in turn, until leaving out any experiment makes little difference to the calculated elasticities (Fig. 3.4). At this point, the remaining experiments are used to calculate the final elasticity coefficients. The worked example gives a final solution of 10 clusters and 10 experiments (Fig. 3.5).

Partial response coefficients measure the relative importance of the different routes through the system, and are indicated by the weights of the arrows in Figure 3.5. In the worked example, when the input rises the mRNAs in cluster 1 (and clusters 3, 4 and 5&6) are important in tending to increase the output, and those in cluster 2 are important in tending to decrease the output; other clusters are less important.

Monte Carlo analysis Microarray data are error-prone. Each step in the experiment, mRNA extraction, PCR amplification, hybridization and fluorescence scanning, introduces experimental error, perhaps from preferential labelling, or dust on the microarray slide. This results in well-documented problems with reproducibility and accuracy of the expression data (Yuen *et al*., 2002). The final step is to use Monte Carlo methods for statistical analysis of the coefficients (Ainscow and Brand, 1998). Information about experimental error obtained from the repeated microarrays is used. The repeated experiments give a mean and SD for each data point, i.e. the expression of each gene and the output response in each experiment. Each data point in the dataset is simulated, based on a normal distribution of its mean and SD. Integrated response, elasticity and partial response coefficients are then calculated for the simulated dataset. This process is repeated many times, giving a distribution for each calculated coefficient, based on the experimental error in the original dataset. If 95% of the values for a particular coefficient share a feature, such as being greater than zero, we have 95% confidence in that feature, and assign it a statistical significance of 'pseudo $P \leq 0.05$' (Ainscow and Brand, 1998).

Datasets with more error will give values for the coefficients that are less statistically significant. Elasticity coefficients are obtained by solving simultaneous equations, so for two sets of calculations from a single dataset, the set with less equations (less experiments and clusters) should produce coefficients that are more statistically significant. There is a relationship between number of clusters, amount of error in the microarray data and the statistical significance of the resulting coefficients. This relationship is data-dependent. Currently, the most accurate microarray experiments quote a coefficient of variation (SD/mean) of about 0.2 (Piper *et al*., 2002).

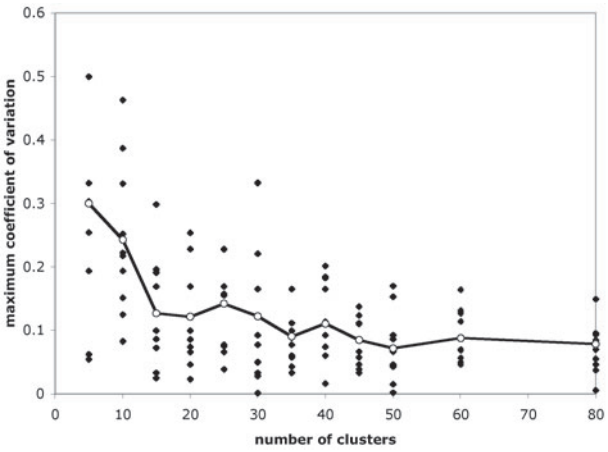


Fig. 4. Tolerable error and number of clusters. Matrices of different sizes, containing data randomly selected from a real microarray dataset (Hughes *et al*., 2000), were simulated. For each, we calculated the maximum coefficient of variation it could tolerate and still produce at least one statistically significant elasticity coefficient (pseudo-$P \leq 0.05$). Any greater experimental error would result in no significant elasticity coefficients. Closed diamonds, maximum coefficient of variation tolerated by simulated matrices; open circles, mean coefficient of variation for simulated matrices containing that number of clusters (connected by a line).

We used Monte Carlo analysis to estimate the relationship between the number of clusters and the amount of error in a dataset that the analysis can tolerate and still result in statistically significant coefficients. Simulations used matrices of different sizes, containing data randomly selected from a real microarray dataset (Hughes *et al*., 2000). We calculated the maximum coefficient of variation that the analysis could tolerate and still produce at least one statistically significant (pseudo-$P \leq 0.05$) elasticity coefficient (Fig. 4). The results indicate that, with the best reproducibility of current microarray technology, regulation analysis should use no more than 15–20 clusters. If the analysis demands many more clusters than this, the maximum coefficient of variation has to be below 0.1. This calculation underestimates the maximum tolerable error: only taking into account variation in the microarray data used to calculate elasticities, not in the measurement of the output response or calculation of integrated response coefficients.

Interpretation of results An mRNA cluster with a high absolute partial response coefficient contains one or more mRNAs important for the mRNA-mediated response. Because partial response coefficients are the product of the integrated response (how much the cluster responds to the input) and elasticity coefficients (how much that cluster affects the output response), a high partial response could be because of a large elasticity, a large integrated response or both. So it is not necessarily the mRNA cluster that changes the most that

is the most important for the response: a cluster with a small integrated response to the input may affect the output strongly and result in a large partial response coefficient.

Information about the mechanisms of the response can be obtained from the signs of the integrated response and elasticity coefficients. A cluster with positive *IR* and ε represents the upregulation of a system that acts to increase the response. If both coefficients are negative, a system that opposes the response is downregulated, and overall this will increase the response. A positive *IR* and a negative ε describes the upregulation of mechanisms that oppose the response, and so on.

Clusters containing mRNAs for regulatory pathways are unlikely to have large partial response coefficients over physiological outputs. Because they will be upstream of their target mRNAs that do have direct effects on the output, their effects will appear as part of the integrated response coefficient of the target mRNA to the input. To establish how strongly the 'regulatory' mRNAs control the target mRNAs, the data can be reanalysed with the cluster containing the target mRNAs as the output. Each cluster in turn can be used as output, to determine which mRNA clusters are involved in its regulation. In this way, the integrated response of a cluster to the input can be partitioned into the direct effects that are not transmitted by other clusters (e.g. via activation of pre-existing transcription factors) and each of the indirect effects through the other clusters (e.g. via changed expression of mRNAs coding for transcription factors).

Expanding the analysis Important clusters can be dissected using bioinformatic approaches such as searching the *Saccharomyces* Genome Database (Ball *et al*., 2000) to identify their mRNAs. They can also be subdivided to see which sub-clusters are most important, by adding additional experiments to the dataset and repeating the analysis. Once the set of elasticities is established, many different input experiments can be analysed to obtain the integrated response of each cluster to the new input. The existing elasticities can then be used with these new integrated responses to calculate partial response coefficients describing how the new input affects the mRNA-mediated response. Important transcripts (and those unexpectedly found to be much more or less important than anticipated) should be directly modulated, to confirm the result. If many transcripts that were expected to be important have low partial response coefficients, the problem will have been greatly simplified even if it has not been solved. Conversely, previously uncharacterized genes may be recognized as important, flagging them for further investigation.

The output has an elasticity to each mRNA cluster. If the individual mRNAs in the cluster have identical expression, the elasticity to the cluster is the sum of the elasticities to each of the individual mRNAs. If not, the summed elasticity is weighted by how much component mRNAs deviate from the cluster mean. As the genes in a cluster have slightly different expression, some accuracy of this summed cluster elasticity will be lost. Extra experiments could be performed to obtain higher resolution and accuracy by dividing clusters and obtaining elasticity coefficients for smaller mRNA clusters.

Control analysis is strictly valid only for infinitesimal changes, but this is not experimentally practical. Results using large changes will be acceptable if assumptions of linearity are reasonable. We propose the use of gene knockouts as modulations, but these large changes may overstep the range where responses are linearly related to inputs. The assumption of linearity is testable by carrying out the analysis following smaller experimental changes in gene expression around the condition of interest. This could involve small genetic manipulations such as 25% overexpression. If the results are similar, linearity is adequate.

RESULTS

Application of modular regulation analysis to published microarray datasets

Because microarray datasets are large and complex, we used a program written in Python (www.python.org) to automate the calculations. We applied modular regulation analysis to two published datasets that had many (but not all) of the features required for a modular regulation analysis. The published experiments had other aims, so were not designed for our analysis and therefore were not ideal. However, some useful information could be extracted.

Effect of galactose on yeast doubling time

Ideker *et al*. (2001) measured all 6200 transcripts in the yeast *Saccharomyces cerevisiae* growing in the absence and presence of galactose, to investigate regulation of the galactose utilization pathway. They investigated wild-type and nine mutant strains with different genes of the galactose utilization pathway disabled. Hybridizations were repeated four times, but experimental errors were not provided. Doubling times for each yeast strain were given. We performed a modular regulation analysis to determine which mRNA changes were important for the doubling time response to the addition of galactose.

First, the dataset was reduced from 6200 mRNAs to the 997 that changed significantly in one or more condition (Ideker *et al*., 2001). The nine different knockout strains (in the presence of galactose) were used as modulation experiments. Therefore, the maximum number of clusters was nine. To check the dataset for any problematic experiments that contained high error or were too similar to others, we followed the validation process described above and in Curtis and Brand (2002). The microarray data were grouped into one less cluster than there were experiments, i.e. eight clusters. Each experimental modulation was omitted in turn, and each time a set of elasticity coefficients was calculated. This process was repeated until a stable solution was reached, using five

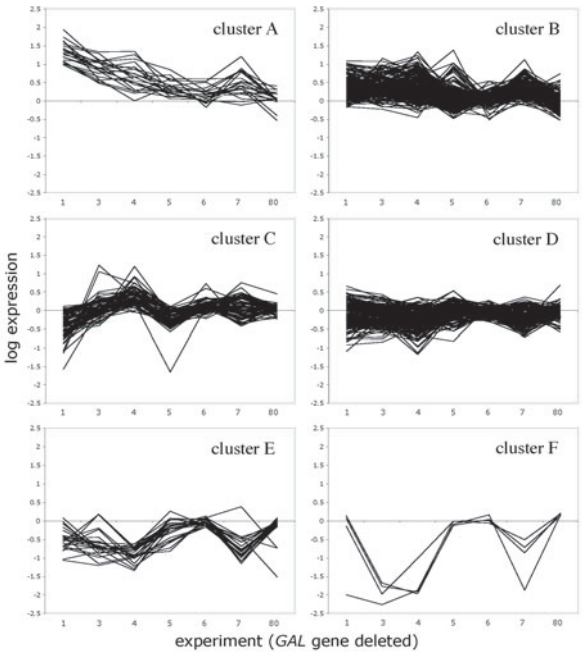


Fig. 5. Euclidean hierarchical clustering of the dataset of Ideker *et al*. (2001) into six clusters using the six validated experiments. The expression ($\log_{10}$) of each cluster across the series of experiments is shown: each line is a different mRNA. For clarity, only 200 of the 640 mRNAs in cluster D are shown.

clusters and six experiments. A final solution of six clusters and six experiments was obtained. The three experiments that were omitted (*GAL2*, *GAL6* and *GAL10* knockouts) were too similar to other experiments in the dataset. Grouping the 997 mRNAs into only six clusters gave reasonably coherent clusters (Fig. 5). Table 1 gives the composition of the clusters and the results of the modular regulation analysis. Figure 6 shows the results graphically.

Modular regulation analysis shows that two mRNA clusters (E and F) mediate the change in doubling time in response to galactose. Cluster E transmitted ~60% of the mRNA-mediated response and contained 23 genes, including the galactose metabolism regulatory gene *GAL80*. Cluster F mediated about 40% of the response and contained four genes: the metabolic genes *GAL1*, *GAL7* and *GAL10*, and the regulatory gene *GAL3*. This result was obtained without prior knowledge of the mRNA cluster contents. The two important clusters had large partial response coefficients because they had large integrated response and large elasticity coefficients. It is not necessarily the clusters that change the most that are most important, but here they were. The integrated response of cluster F to galactose was heavily biased by the presence of *GAL3*. Without *GAL3*, this cluster would have an integrated response of around 300, making cluster F the single most important cluster for the response. *GAL3* is clustered with the other three mRNAs in cluster F on the basis of its expression in the knockout experiments, not the galactose-addition experiment. This allows the clustering and calculation of elasticity coefficients to be independent of the input stimulus.

To investigate the regulation of cluster F, the system was reanalysed with cluster F as the output (data not shown). One cluster was responsible for the regulation of cluster F. A total of 9 of the 13 genes in this cluster encoded proteins involved in amino acid biosynthesis, e.g. ARG4. This suggests that increased amino acid biosynthesis is required for the large upregulation of the GAL1, 7 and 10 proteins, allowing galactose metabolism, and leading to faster growth. The important cluster did not contain transcription factors, suggesting that the proteins regulating gene expression in this system are not themselves transcriptionally regulated, instead they are allosterically activated. This analysis shows how modular regulation analysis of microarray data can discriminate between direct and indirect effects. Cluster F was found to be directly involved in the doubling time response to galactose. When the dataset was reanalysed, one cluster was found to be the most important for the regulation of cluster F. Therefore, the mRNAs in this cluster transmitted an indirect doubling time response to galactose.

The sum of the partial response coefficients predicts that the overall response transmitted by the mRNAs is a decrease in doubling time. This was observed by Ideker *et al*. (2001): the yeast grow faster with an additional carbon source. The difference between the sum of the partial response coefficients and the observed response (Table 1) is theoretically equal to the response transmitted by the non-mRNA routes. This would require the non-mRNA route to transmit an opposing large increase in doubling time, which is unrealistic. More likely, there are problems with linearity of the integrated response or elasticity coefficients: addition of galactose (or deletion of a gene) is a large step-change, rather than the small change that control analysis requires. Use of an intermediate concentration of galactose as an input could confirm the precise values of the integrated responses.

The dataset provided no experimental errors, precluding use of Monte Carlo analysis to test for statistical significance. Assuming significant results, our analysis revealed that several *GAL* genes were important for the response to galactose, without requiring any prior knowledge of the functions of genes. The analysis did not require the genetic modulations to be in the *GAL* genes; as long as the modulations were independent and resulted in different patterns of expression, the function of the knockouts was not important. All the knockouts in the dataset were of genes in the galactose utilization pathway, because this was the pathway under investigation by Ideker *et al*. (2001).

Effect of 2-deoxy-D-glucose on yeast relative fitness

Hughes *et al*. (2000) published a large dataset of 300 microarray experiments, measuring the full transcriptome of

Table 1. Coefficient values and cluster contents for the six-cluster solution of the dataset of Ideker *et al*. (2001)

Cluster	$IR^{\text{mRNA cluster}}_{\text{galactose}}$	$\varepsilon^{\text{doubling time}}_{\text{mRNA cluster}}$	$^{\text{mRNA cluster}}R^{\text{doubling time}}_{\text{galactose}}$	Percentage of partial response	*GAL* genes	Size
A	−0.90	0.21	−0.19	0.2		21
B	−0.74	−7.76	5.70	−6.9		229
C	−0.63	0.68	−0.43	0.5	5	80
D	0.16	7.81	1.23	−1.5	2, 4, 6	640
E	4.53	−11.71	−53.05	64.3	80	23
F	6.09	−5.87	−35.73	43.3	1, 3, 7, 10	4
		Total	−82.46			
		Observed	−0.20			
		Non-mRNA	82.26			

'Percentage of partial response' describes how much of the mRNA-mediated response is transmitted through each cluster. It is the partial response of a cluster divided by the sum of the partial response coefficients, displayed as a percentage. The response transmitted by the non-mRNA routes is calculated by subtracting the sum of the partial response coefficients from the observed response to galactose (Ideker *et al*., 2001).

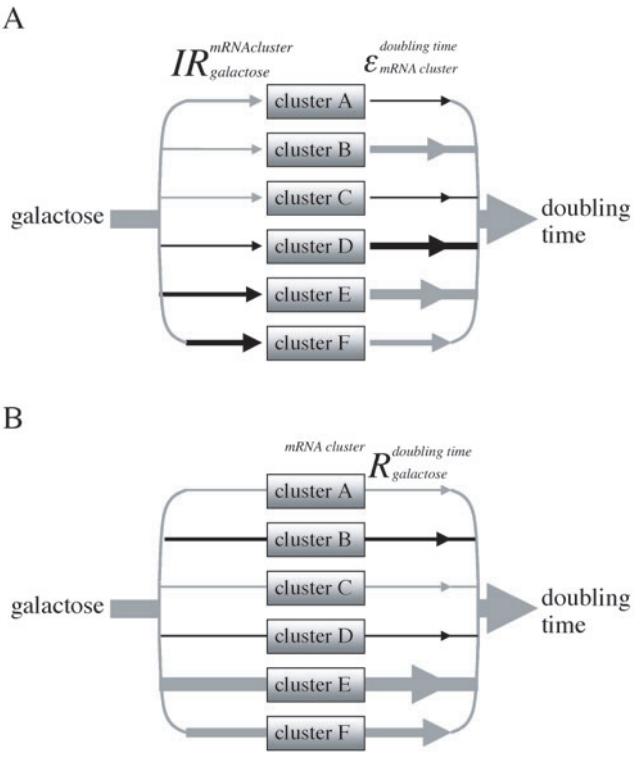


Fig. 6. The six-cluster solution of the system of Ideker *et al*. (2001). (**A**) Integrated response and elasticity coefficients. (**B**) Partial response coefficients. The weight of the line indicates the relative value of the coefficient. Grey: negative coefficient; black: positive coefficient. The values of the coefficients are shown in Table 1.

S.cerevisiae. A total of 13 experiments were chemical additions to wild-type yeast, and 276 were different knockout yeast strains. The 'relative fitness' of the knockouts was measured in a quantitative parallel growth assay. Microarray hybridizations were repeated, and using an error model, a *P*-value for each mRNA measurement in each experiment was given. The experiments were designed to predict the function of unknown genes, because genes with similar functions tend to be coexpressed and cluster together.

One of the chemical addition experiments was chosen arbitrarily to be the system input: 2-deoxy-D-glucose. This non-metabolizable glucose analogue inhibits glycolysis and causes depletion of inorganic phosphate and osmotic and cell wall problems (Kratky *et al*., 1975). Of the 276 knockouts, 120 had both replicate hybridizations and relative fitness measurements, so could be used as experimental modulations. Relative fitness was used as the output response. We performed a modular regulation analysis to determine which mRNA changes were important for the effect of 2-deoxy-D-glucose on relative fitness.

Following calculation and testing as described above, a solution was found at 78 clusters. Coefficients for the most important mRNA clusters are shown in Table 2 and Figure 7. Control of the response was distributed between several clusters, reflecting the systemic effects of 2-deoxy-D-glucose. Cluster 37 was the most strongly upregulated by 2-deoxy-D-glucose, while cluster 44 had the strongest effect on relative fitness (Fig. 7A). However, cluster 44 was only weakly upregulated by 2-deoxy-D-glucose, and the overall response to 2-deoxy-D-glucose was mediated mostly by clusters 37 and 43 and by the sum of many small positive and negative effects through other clusters. The relatively strong expression changes in the other clusters illustrated in Figure 7A had only small effects on fitness. Cluster 37 contains one gene, encoding a sodium–phosphate symporter, suggesting an attempt to correct the phosphate shortage caused by 2-deoxy-D-glucose. Cluster 43 contains five genes, one involved in cellular fusion, the other four of unknown function. The amount of response transmitted by non-mRNA routes could not be calculated as the observed relative fitness response to 2-deoxy-D-glucose was not available.

Table 2. Coefficient values for the most important clusters in the 78-cluster solution of Hughes *et al*. (2000)

Cluster	$IR^{\text{mRNA cluster}}_{\text{2-deoxy-D-glucose}}$	$\varepsilon^{\text{relative fitness}}_{\text{mRNA cluster}}$	${}^{\text{cluster}}R^{\text{relative fitness}}_{\text{2-deoxy-D-glucose}}$	Size
37	9.62	0.16	1.53	1
60	0.46	0.76	0.17	1
30	1.72	0.17	0.30	1
58	1.04	0.26	0.27	1
12	0.90	0.30	0.27	1
38	1.16	0.15	0.17	39
41	2.56	0.06	0.16	1
67	0.15	0.83	0.12	13
49	0.23	0.46	0.10	3
Other positive			0.73	60
Other negative			−0.65	46
71	−0.95	0.21	−0.20	1
4	3.66	−0.07	−0.27	2
57	0.89	−0.35	−0.32	1
44	0.07	−4.89	−0.34	6141
53	0.22	−1.62	−0.35	3
43	2.11	−0.34	−0.73	5
		Total	1.13	

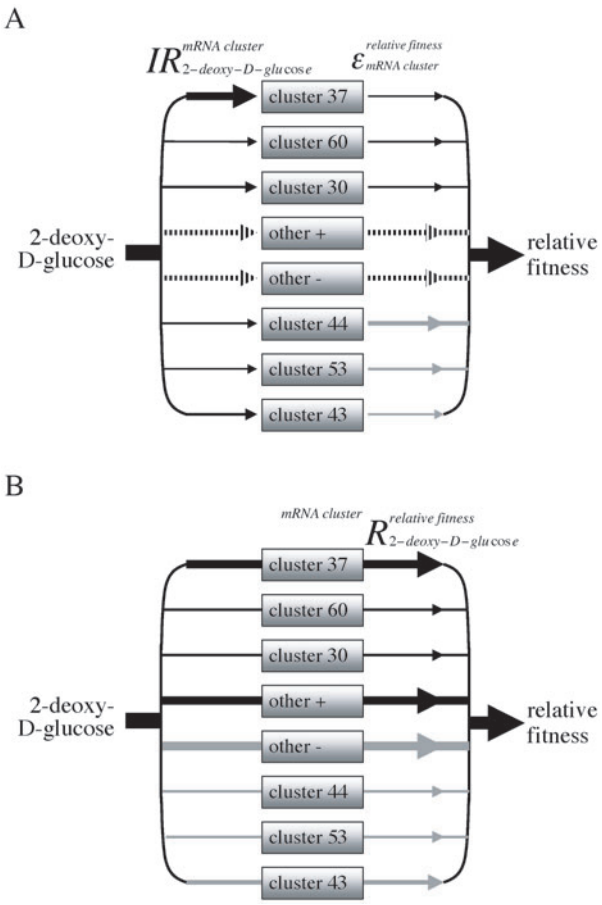


Fig. 7. Coefficients from the 78-cluster solution of the dataset of Hughes *et al*. (2000). The three clusters with the most positive and most negative partial responses are shown, as well as the sum of the remaining positive ('other +') and negative ('other −') partial response coefficients. (**A**) Integrated response and elasticity coefficients. (**B**) Partial response coefficients. The weight of the line indicates the relative value of the coefficient. Grey: negative coefficient; black: positive coefficient. Integrated response and elasticity coefficients are not summed for 'other +' and 'other −'; they are represented by unweighted dashed arrows. The values of the coefficients are shown in Table 2.

Most of the more important clusters had positive integrated responses to the input, suggesting that various systems that both facilitate and oppose the overall response were upregulated in response to 2-deoxy-D-glucose. The sum of the partial response coefficients was positive, suggesting that the relative fitness of the yeast improved in response to 2-deoxy-D-glucose. At high concentrations, 2-deoxy-D-glucose is toxic, so a negative sum of partial responses might have been be expected. The partial response coefficients in Table 2 and Figure 7 show the potential value of the regulation analysis in allowing a deeper understanding of microarray results, but were they statistically significant?

To test for significance, we performed a Monte Carlo analysis using the published experimental error, as described above. This resulted in partial response coefficients with pseudo-P-values between 0.4 and 0.6, showing that the noise in the microarray data was too great for the results to be statistically significant (pseudo-$P \leq 0.05$). We calculated that the data would need a coefficient of variation of 0.1 or better to produce at least one statistically significant elasticity coefficient. The average coefficient of variation for this microarray data was about 0.6, so the reproducibility was insufficient for a solution using so many clusters.

DISCUSSION

Modular regulation analysis is a promising method that is highly relevant to expression profiling to find mRNAs that are important in mediating responses. Our analysis of published datasets shows that it can be used to extract important information that many other methods cannot. While this type of analysis requires a series of experimental modulations, other techniques for finding mRNAs important for a response also require many experiments and are less general, quantitative and inclusive. This method is general because it can be applied to a range of problems and systems, from genetic networks to drug targets. No prior knowledge is needed about which genes are in which clusters, or what the genetic modulations are, provided that they are different from each other. Modular regulation analysis does not require the important mRNAs to have been knocked out or overexpressed; as long as their expression changes in one or more of the experimental modulations, the information is accessible. The method is

quantitative as responses are described using coefficients and their statistical significance determined. A cluster could have a large partial response if it has a large integrated response, elasticity coefficient or both. It is not necessarily the cluster that changes the most (has a high integrated response) that is the most important for the response. Our analyses show that clusters that do change the most are often important, but there are clusters with small integrated response coefficients and high elasticity coefficients that are also important. An example is cluster 53 in the 2-deoxy-D-glucose analysis; this cluster has the second most negative partial response coefficient, despite its small integrated response coefficient. These mRNAs would not be picked up by analyses that assume only large changes in expression are important. Finally, this method is inclusive as it can realistically be applied to a whole genome, rather than a subset of genes (de la Fuente and Mendes, 2002; Ideker *et al*., 2001).

This analysis describes the application of control analysis to a new field: gene expression. Control analysis of metabolic systems generally reveals that control is distributed. It is unclear whether the control of gene expression is normally localized into a few clusters, or is more widely distributed. It may be that this is context-dependent and some responses only require expression changes in a few genes, while some may require widespread alteration of expression. Application of modular regulation analysis to several example systems will reveal whether regulation of responses by gene expression is generally localized or distributed.

mRNA concentrations do not always reflect the amount of the corresponding protein in the cell. This is irrelevant to our analysis, which investigates how much of the response passes through each cluster of mRNA transcripts, regardless of the resulting expression of the corresponding proteins. Modular regulation analysis could be used to quantify the response passing through the concentrations of different proteins, but this would require accurate measurements of the concentrations of very many proteins, and proteomics has not yet attained the appropriate experimental sophistication. Any discrepancies between regulation analyses based on either mRNA or protein concentrations would illuminate relevant cases of important translational control.

A current drawback of this analysis is the high microarray accuracy required. As the reliability of microarray technology improves, modular regulation analysis of gene expression will become more applicable in practice. In principle, modular regulation analysis is also highly transferable to proteomic, metabolomic or other profiling to determine the relative importance of groups of proteins or metabolites in effecting responses.

ACKNOWLEDGEMENTS

We acknowledge helpful discussion with Christoph Giersch, Helmut Wegmann, Torsten Becker, Iain Barrass and Andrew Raine.

REFERENCES

Ainscow,E.K. and Brand,M.D. (1998) Errors associated with metabolic control analysis. Application of Monte-Carlo simulation of experimental data. *J. Theor. Biol.*, **194**, 223–233.

Ainscow,E.K. and Brand,M.D. (1999a) The responses of rat hepatocytes to glucagon and adrenaline. Application of quantified elasticity analysis. *Eur. J. Biochem.*, **265**, 1043–1055.

Ainscow,E.K. and Brand,M.D. (1999b) Top-down control analysis of ATP turnover, glycolysis and oxidative phosphorylation in rat hepatocytes. *Eur. J. Biochem.*, **263**, 671–685.

Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald,A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alter,O., Brown,P.O. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci., USA*, **97**, 10101–10106.

Ball,C.A., Dolinski,K., Dwight,S.S., Harris,M.A., Issel-Tarver,L., Kasarskis,A., Scafe,C.R., Sherlock,G., Binkley,G., Jin,H. *et al*. (2000) Integrating functional genomic information into the *Saccharomyces* genome database. *Nucleic Acids Res.*, **28**, 77–80.

Brand,M.D. (1996) Top down metabolic control analysis. *J. Theor. Biol.*, **182**, 351–360.

Brand,M.D. (1997) Regulation analysis of energy metabolism. *J. Exp. Biol.*, **200**, 193–202.

Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. *et al*. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.*, **2**, 65–73.

Curtis,R.K. and Brand,M.D. (2002) Control analysis of DNA microarray expression data. *Mol. Biol. Rep.*, **29**, 67–71.

de la Fuente,A., Brazhnik,P. and Mendes,P. (2002) Linking the genes: inferring quantitative gene networks from microarray data. *Trends Genet.*, **18**, 395–398.

de la Fuente,A. and Mendes,P. (2002) Quantifying gene networks with regulatory strengths. *Mol. Biol. Rep.*, **29**, 73–77.

Devaux,F., Marc,P., Bouchoux,C., Delaveau,T., Hikkel,I., Potier,M.C. and Jacq,C. (2001) An artificial transcription activator mimics the genome-wide properties of the yeast Pdr1 transcription factor. *EMBO Rep.*, **2**, 493–498.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci., USA*, **95**, 14863–14868.

Fell,D. (1997) *Understanding the Control of Metabolism*. Portland Press, London.

Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell.*, **11**, 4241–4257.

Giersch,C. (1994) Determining elasticities from multiple measurements of steady-state fluxes and metabolite concentrations: theory. *J. Theor. Biol.*, **169**, 89–99.

Heinrich,R. and Rapoport,S.M. (1974) A linear steady state treatment of enzymatic chains. General properties, control and effector strength. *Eur. J. Biochem.*, **42**, 89–95.

Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H.,

He,Y.D. *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Ideker,T., Thorsson,V., Ranish,J.A., Christmas,R., Buhler,J., Eng,J.K., Bumgarner,R., Goodlett,D.R., Aebersold,R. and Hood,L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Kacser,H. and Burns,J.A. (1973) The control of flux. *Symposia Soc. Exp. Biol.*, **27**, 65–104.

Kacser,H. and Burns,J.A. (1979) Molecular democracy: who shares the controls? *Biochem. Soc. Trans.*, **7**, 1149–1160.

Kholodenko,B.N. (1988) How do external parameters control fluxes and concentrations of metabolites? An additional relationship in the theory of metabolic control. *FEBS Lett.*, **232**, 383–386.

Kratky,Z., Biely,P. and Bauer,S. (1975) Mechanism of 2-deoxy-D-glucose inhibition of cell-wall polysaccharide and glycoprotein biosyntheses in *Saccharomyces cerevisiae*. *Eur. J. Biochem.*, **54**, 459–467.

Le Crom,S., Devaux,F., Marc,P., Zhang,X., Moye-Rowley,W.S. and Jacq,C. (2002) New insights into the pleiotropic drug resistance network from genome-wide characterization of the YRR1 transcription factor regulation system. *Mol. Cell Biol.*, **22**, 2642–2649.

Le Naour,F., Hohenkirk,L., Grolleau,A., Misek,D.E., Lescure,P., Geiger,J.D., Hanash,S. and Beretta,L. (2001) Profiling changes in gene expression during differentiation and maturation of monocyte-derived dendritic cells using both oligonucleotide microarrays and proteomics. *J. Biol. Chem.*, **276**, 17920–17931.

Piper,M.D., Daran-Lapujade,P., Bro,C., Regenberg,B., Knudsen,S., Nielsen,J. and Pronk,J.T. (2002) Reproducibility of oligonucleotide microarray transcriptome analyses. An interlaboratory comparison using chemostat cultures of *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **277**, 37001–37008.

Raychaudhuri,S., Stuart,J.M. and Altman,R.B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.*, 455–466.

Rung,J., Schlitt,T., Brazma,A., Freivalds,K. and Vilo,J. (2002) Building and analysing genome-wide gene disruption networks. *Bioinformatics*, **18**, S202–S210.

Schlitt,T. and Brazma,A. (2002) Learning about gene regulatory networks from gene deletion experiments. *Comp. Func. Genomics*, **3**, 499–503.

Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

van't Veer,L.J., Dai,H., van De Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van Der Kooy,K., Marton,M.J., Witteveen,A.T. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Yuen,T., Wurmbach,E., Pfeffer,R.L., Ebersole,B.J. and Sealfon,S.C. (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.*, **30**, e48.