Funct Integr Genomics (2000) 1:174–185
DOI 10.1007/s101420000018

ORIGINAL PAPER

Scott A. Rifkin · Kevin Atteson · Junhyong Kim

Constraint structure analysis of gene expression

Received: 12 May 2000 / Accepted: 20 June 2000 / Published online: 25 August 2000

Abstract A microarray experiment gives a snapshot of the state of an organism in terms of the relative abundances of its mRNA transcripts, locating the organism at a point in a high dimensional state space where each axis represents the relative expression level of a single gene. Multiple experiments generate a cloud of points in this gene expression space. We present a geometric approach to analyzing the covariational properties of such a cloud and use a dataset from *Saccharomyces cerevisiae* as an illustration. In particular, we use singular value decomposition to identify significant linear sub-structures in the data and analyze the contributions of both individual genes and functional classes of genes to these major directions of variation. Analyzing the publicly available yeast expression data, we show that under all experimental conditions the variation in expression is limited to a small number of linear dimensions. Projections of individual gene axes onto the significant dimensions can order the contribution of individual genes to variation in expression within an experiment. We show that no particular groups of genes characterize particular experimental conditions. Instead, the particular structure of the coordinated expression of the entire genome characterizes a particular experiment.

Keywords Microarray · Gene expression · Structural analysis · Geometry

S.A. Rifkin · K. Atteson · J. Kim (✉)
Department of Ecology and Evolutionary Biology,
Yale University, P.O. Box 208106, New Haven, CN 06520-8106, USA
e-mail: junhyong.kim@yale.edu
Tel.: +1-203-4329917, Fax: +1-203-4323854

Present address:
K. Atteson,
Department of Electrical Engineering and Computer Science,
University of California, Berkeley, Berkeley, CA 94720, USA

Introduction

Genomes are systems of interacting components, and biological function (e.g. sporulation) may not be defined so much by the absolute activation or repression of a distinct set of genes as by a system-level coordinated pattern of gene expression (Holstege et al. 1998; Reinitz et al. 1998; Szallasi 1999). As a cartoon example, consider the case where an organism has two genes, A and B, and a trait of interest. One possibility for the relationship between the genes and the trait is that induction of gene A is associated with the trait. Alternatively, either induction of gene A coupled with repression of gene B or the converse is associated with the trait. In the latter case, an examination of marginal levels of expression of gene A or gene B will not reveal the association. When external factors or internal mutations (with respect to an ancestor) perturb a cell, it responds with a change or a series of changes in its transcriptional state. Gene interactions, or gene circuits (Hlavacek and Savageau 1996, 1997; Savageau 1999), constrain this response to fall along some substructure of a gene expression space where each dimension represents the expression level of a different gene. If the products of gene A induce gene B, we would expect the covariation of the two gene expression levels to satisfy a particular constraint structure. For example, if the gene products of A induce gene B in a linear manner, then the constraint structure would be the equation level(A)-k*level(B)=0; if the relationship were nonlinear, the constraint structure could be a host of nonlinear equations. Such constraint equations are the solution sets to the presumed dynamical system of gene interactions (Wolf and Eeckman 1998).

A microarray measurement of gene expression gives a genome-wide assay of the transcriptional state of a cell (or organism) which can be represented as a point in high dimensional gene expression space. Multiple measurements taken by microarray experiments can be used to identify the constraint structures associated with a particular developmental sequence or phenotypic state. We present an example of such an analysis using a collection

of data from the Stanford Genomics Server. We use various geometric techniques to identify the constraint structure – i.e. the structure of the scatter of measurement points – and then identify an individual gene's contribution to this constraint structure. Although in this paper we restrict our analysis to the linear constraint structures, this geometric framework is more general. Our method, therefore, differs in its approach from many clustering type of algorithms (discussed below) and other applications of linear ordination (e.g. Raychandhuri et al. 2000) because it emphasizes the global structure of gene expression and geometric decomposition of the structure. We call it constraint structure analysis (CSA) of gene circuits to emphasize its systemic view of gene function and possible generalization to non-linear structures.

Linear CSA using singular value decomposition

In the linear case, we use singular value decomposition to identify an orthonormal set of basis axes which best fit a cloud of data points by the least squares criterion (Green and Carroll 1978). In terms of matrices, singular value decomposition takes a data matrix and expresses it as the product of three other matrices: $M=VDU^T$, where M is a G×C matrix of data points (genes by conditions) with G>C and rank (or dimension) R; V is a G×R matrix; D is an R×R diagonal matrix; and U is a C×R matrix. In the case of presently available microarray data, since the number of genes is vastly more than the number of conditions, we can assume that R=C. These matrices have a geometric interpretation. The columns of V are the orthonormal basis mentioned above which best fits the data in a least squares sense; the entries of D are stretching factors; and the rows of U (the columns of U^T) are the original data expressed in this new basis. We are most interested below in the columns of V (the linear substructures in the data) and the entries of D – the singular values whose magnitudes are proportional to the standard deviation of the projected points along the respective basis axes. The entries of U allow us to gauge the fit of the orthonormal basis (our model) to the data. Principal components analysis is a special case of singular value decomposition, applicable to square, symmetric, positive definite matrices – variance-covariance matrices, for example. Given a noisy data set like microarray data, not all basis axes will be biologically meaningful. To identify significant sets of axes, we randomly permuted the input data matrices (see Materials and methods) and then used singular value decomposition on the randomized matrices to obtain a null distribution of singular values. Only the axes with singular values greater than 95% of the null distribution were kept for subsequent analysis. Adopting the terminology from principal components analysis, we call such significant axes structural component axes.

The relationship of each of the original axes – individual gene expression levels – to the structural component axes can be measured by the angle of the original axes to the structural component axes. We call the cosine of this angle the response coefficient of the gene. Response coefficients are always with respect to a particular gene, experiment, and structural component axis. A given gene's contribution to a set of structural component axes can be measured by multiplying the square of the absolute value of the response coefficient for each axis by the square of the singular value for that axis, summing over all relevant axes, and then dividing by the sum of the squares of the singular values. We call this the overall response index (ORI) of a gene *g*:

$$ORI_g = \frac{\sum_i SV_i^2 RC_{gi}^2}{\sum_i SV_i^2} \quad (1)$$

where SV_i is the *i*th singular value and RC_{gi} is the response coefficient of the gene for the *i*th axis. For a given experiment, the ORI of a gene measures the contribution of the gene to the overall covariation in the data set on a scale from 0 to 1.

Since the structural component axes are vectors in the gene expression space, the relationship between two sets of structural component axes (each for different phenotype or experimental conditions) can be measured using the multivariate technique of canonical correlation analysis (Seber 1984). Given two sets of vectors, A and B, canonical correlation analysis finds a linear combination of A and a linear combination of B such that the correlation between the two is maximized. Given two arbitrary vectors in gene expression space, for example centroid vectors, we can also find the difference vector which

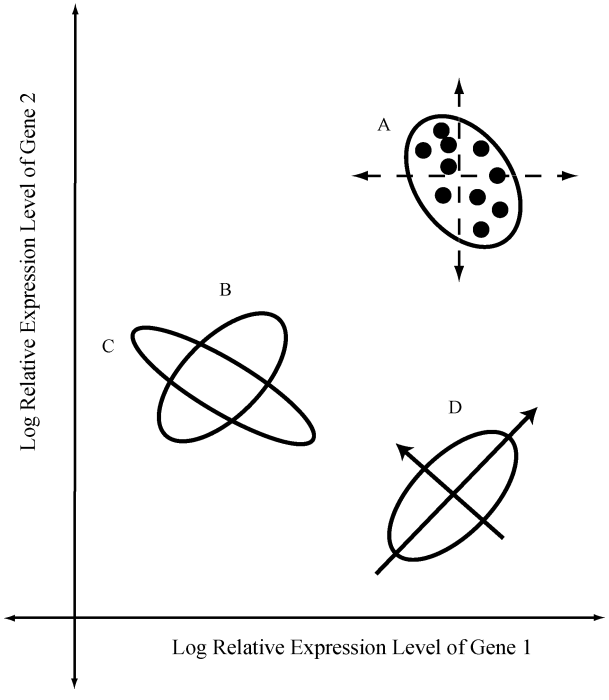

Fig. 1 Possible relationships between clouds of experiments in gene expression space (see text for details)

represents the linear direction of the differential between the conditions represented by the two original vectors.

The structural component axes extracted by linear CSA point in the directions of greatest variation of a data cloud in the gene expression space (Fig. 1, cloud D). If these data points are related – timepoints after a perturbation or measurements on a particular cell type, for example – the cloud represents the covariation of gene expression in a cell (or an organism) under a particular set of related conditions, and the structural component axes reflect the linear structure of this variation. By comparing the sets of axes from two such data clouds, we can determine whether cells under two different sets of conditions obey the same constraint structure, even if they occupy different regions of the gene expression space in an absolute sense (Fig. 1, clouds B and D). Alternatively, two clouds centered around the same point – in the same absolute region of the gene expression space – may not share the same covariational structure (Fig. 1, clouds B and C). Linear CSA enables us to distinguish clouds of measurements both by absolute location in the gene expression space and also by covariational structure, and therefore can be used to make fine distinctions between two cell types or the behavior of cells under different conditions.

Materials and methods

Data

We used three *Saccharomyces cerevisiae* datasets from the Brown and Botstein labs at Stanford University. The cell cycle experiments (alpha factor, elutriation, cdc15, cdc28) consist of measurements from 6,177 ORFs downloaded from http://genome-www.stanford.edu/cellcycle/data/rawdata/combined.txt (Spellman et al. 1998). The sporulation set consists of measurements from 6,109 ORFs downloaded from http://cmgm.stanford.edu/pbrown/-sporulation/additional/sposprea d.txt (Chu et al. 1998). The diauxic shift experiment consists of measurements from 6,121 ORFs during diauxic shift downloaded from http://cmgm.stanford.edu/-pbrown/explore/array.txt (DeRisi et al. 1997). We downloaded a list of known genes and their putative functions from ftp://genome-ftp.stanford.edu/yeast/tables/ORF_Descriptions/orf_descriptions.txt.

Each data point reflects the log-base 2 transformed mRNA abundance for each ORF at a particular condition relative to that of a reference condition. Because these are time series experiments, we interpolated the values of missing data points to make a straight line between the nearest measured timepoints, set missing values at the start and end to the first or last measured values, and discarded information on ORFs missing more than half of the timepoints by setting all of their values to zero. There were no missing datapoints in the sporulation and diauxic shift datasets, so we surmise that they were filled in as equal expression (zero) before being posted. We also averaged the expression levels of duplicated known genes. Finally, we normalized the data sets either to the centroid of the entire data matrix or to the centroid of each experimental group individually. Table 1 lists information about the datasets after these manipulations. For the covariance analysis, we discard temporal information in the data and treat each data point as an individual sample under a perturbation. For the functional contribution analyses we recalculated the axes based solely on known genes, and for the canonical correlation analysis we used only the 5,541 ORFs which all experiments shared.

Singular value decomposition and other analyses

We performed linear least squares fits to the data matrices using singular value decomposition implemented in Mathematica 4.0 (Wolfram 1999). All other computations were also implemented as functions in Mathematica 4.0.

Permutation tests

To assess the significance of the singular value and response coefficient distributions of the data, we generated 200 random matrices for each of the experiments by permuting the data matrices across conditions, holding the distribution of expression for each gene constant.

Results

Our main goal in this paper is to present a new analytical technique illustrated with publicly available data from *S. cerevisiae* (DeRisi et al. 1997; Chu et al. 1998; Spellman et al. 1998). In the discussion below, an individual microarray measurement captures the cell state under a particular *condition*. Conditions, e.g. a series of timepoints after alpha factor synchronization, comprise *experiments*. We will primarily analyze the structural components within an experiment but will also compare structural components between experiments and compare the location of experiments (clouds of points/conditions) in a common gene expression space.

A few important aspects of the dataset limit the conclusions that can be drawn. These limits also highlight experimental design issues that will enable researchers to take full advantage of the geometric framework of CSA.

Table 1 Information about the datasets

	Alpha factor	Elutriation	Cdc15	Cdc28	Sporulation	Diauxic shift	Total dataset
Number of ORFs in published data	6178	6178	6178	6178	6118	6153	6178
Number of ORFs after duplicates and poor quality ORFs removed	6074	6074	5672	6124	6109	6121	5541
Known genes	3398	3398	3179	3433	3048	3047	2790
Time points (conditions)	18	14	24	17	7	7	87
Significant axes	5	3	5	5	2	1	–
Mating protein genes	155	155	141	155	134	134	–
Cytoplasmic ribosomal protein genes	125	125	124	129	113	113	–
Cell cycle genes (Spellman et al. 1998)	791	791	763	796	790	790	–

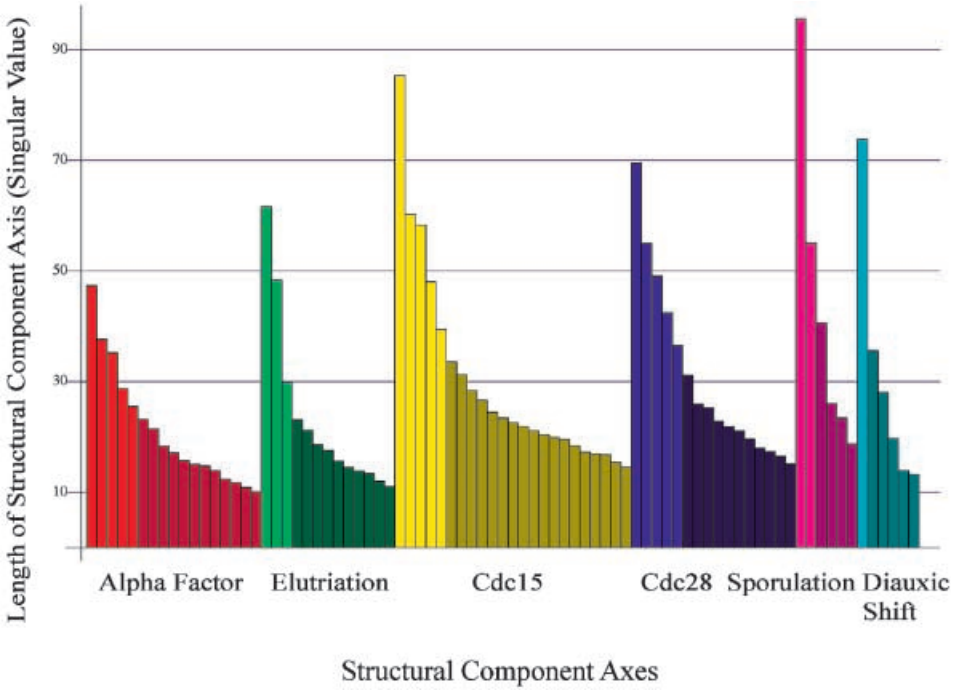

Fig. 2 Singular value distributions. The length of the structural component axis measured by the singular value (along the *y* axis) is proportional to the standard deviation of the condition points projected onto that axis. The *lighter colors* are significant axes, the *darker colors* non-significant

Each dimension of the data set represents the relative mRNA abundance of a different ORF on a logarithmic scale. If there were an overall standard reference condition or a way to measure the absolute levels of expression, we could place all data points within the same expression space, and different experiments would correspond to different clouds of points throughout the space. While the reference conditions for the cell cycle experiments were the same, those of the diauxic shift and sporulation experiments were not, and none of the experiments used the same strain, i.e. genetic background. Because each entry in our data matrix consists of a logarithm of a ratio (some measure of the fluorescence of a gene under experimental conditions relative to that under a reference condition), we are essentially subtracting a translation factor from each point. When we put different experimental groups together and analyze the entire dataset as a whole, these translation factors are no longer equivalent, making it unclear where the centers of each experimental group lie in relation to each other. Consequently, graphical ordinations such as Fig. 3 may not faithfully depict the relationships between experimental groups, and we were not able to explore the global limits on patterns of gene expression. This translation problem applies to any major factor differing between experiments, including conditions and strains. Including an accepted standard, i.e. a standard cell (organismal) state, in every future group of experiments, is essential to make disparate experiments comparable and, consequently, to constructing an integrated database.

This standardization problem only affects conclusions dependent on the centroid of each experiments. The structural component axes identified by CSA are invariant to the placement of the centroid of experiments. Therefore, we can still compare the structures of expression covariation by constructing separate but isomorphic gene expression spaces for each experiment. The center of a cloud fixes the origin of its gene expression space (Fig. 1, dashed lines in cloud A) and the original data is expressed as deviations from this center. Comparisons between CSA performed within each experiment separately (e.g. canonical correlation, response coefficients, and directions of structural component axes) are not affected by the discrepancy in reference conditions.

Structure of coordinated gene expression

The data are not distributed evenly in gene expression space (Fig. 2). The distribution of singular values shows that significant major structural component axes exist within all of the experiments. The skewed distribution of the singular values shows that a small subset of vector directions (linear subspaces) accounts for a majority of the variation. Using the pseudo-permutation test we identified five, three, five, five, two and one significant structural component axes for the six experiments, respectively. From these limited data, it is not clear whether there is a clear relationship between the number of measured data points and the number of significant axes. For example, Alpha factor, Cdc15 and Cdc28 have the same number of significant axes despite their different sample sizes. This might suggest that under a particular experimental condition, the number of significant dimensions of variation will be constant, regardless of the number of measurements. The long, flat tails of the distributions indicate that the gene expression has similar variances in the vector directions associated with these smaller singular values. This might result from yeast cell states varying randomly in gene expression space within

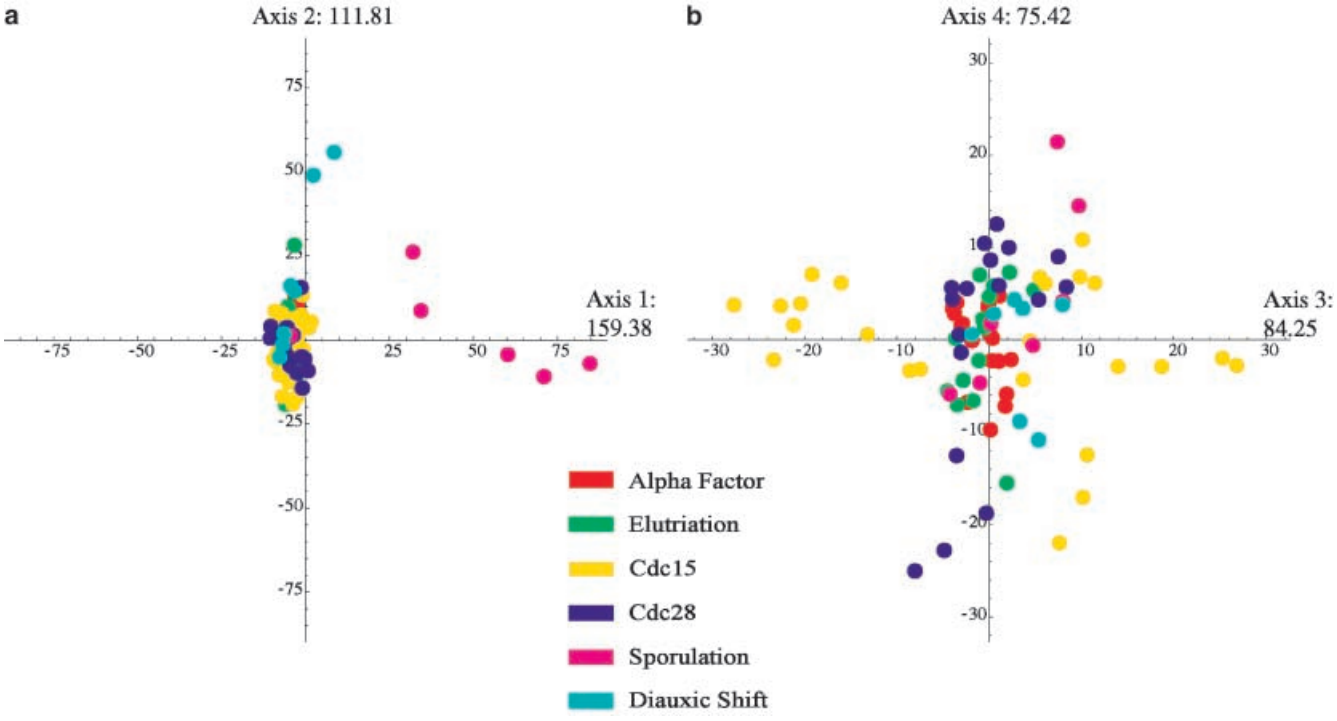

Fig. 3a, b Planes through the 5,541 dimensional subspace of all the experiments formed by **a** the first and second, and **b** the third and fourth longest structural component axes of the entire dataset analyzed together. Axes are in log-base 2 relative expression units and the lengths of the axes are marked. The *color scheme* matches Fig. 2

Table 2 Canonical correlations between the experiments

	Alpha factor	Elutriation	Cdc15	Cdc28	Sporulation	Diauxic shift
Alpha factor	–	0.35	0.60	0.49	0.18	0.32
Elutriation	–	–	0.39	0.32	0.05	0.46
Cdc15	–	–	–	0.54	0.16	0.40
Cdc28	–	–	–	–	0.19	0.29
Sporulation	–	–	–	–	–	0.20
Diauxic shift	–	–	–	–	–	–

certain small bounds, or from some normal pattern of fluctuation of gene activity for maintenance that is conserved across all experimental manipulations, or from sources of error described above. The vector directions of the larger singular values indicate linear combinations of gene expression relevant to the particular experiments.

Structure of covariation among experimental groups

For the sake of illustration, we analyzed the covariational structure of the entire dataset as a whole. The different experimental clouds occupy overlapping, but non-identical regions of the gene-expression space. Figure 3 shows the data projected onto the planes formed by the first and second (Fig. 3a), and third and fourth (Fig. 3b) longest structural component axes calculated for the entire dataset. The points corresponding to sporulation fall at a distance from the other experiments, drawing the longest structural component axis after them. The diauxic shift conditions fall along the next longest structural component axis, and the other experimental groups also begin to segregate in various directions. The spatial relations between experimental groups described by the difference vectors between the centroids of the datasets can be calculated, the components of which indicate the genes involved in positioning a cloud in one part of the space versus another. (However, we do not apply it to this data set because of the standardization problem.)

Analyzing each group individually allows us to compare their coordinated expression patterns between the experimental groups. When cells are subjected to different experimental conditions, there is a systemic response in the expression level of all the genes; this particular coordination of the genes – the interaction of the expression levels of different genes – determines the unique response of the cell to unique experimental conditions. Consequently, there is a signature pattern of *interaction* of genes for each experimental condition, rather than changes in a small cluster of individual genes. To ascertain the relationship between the gene interaction pat-

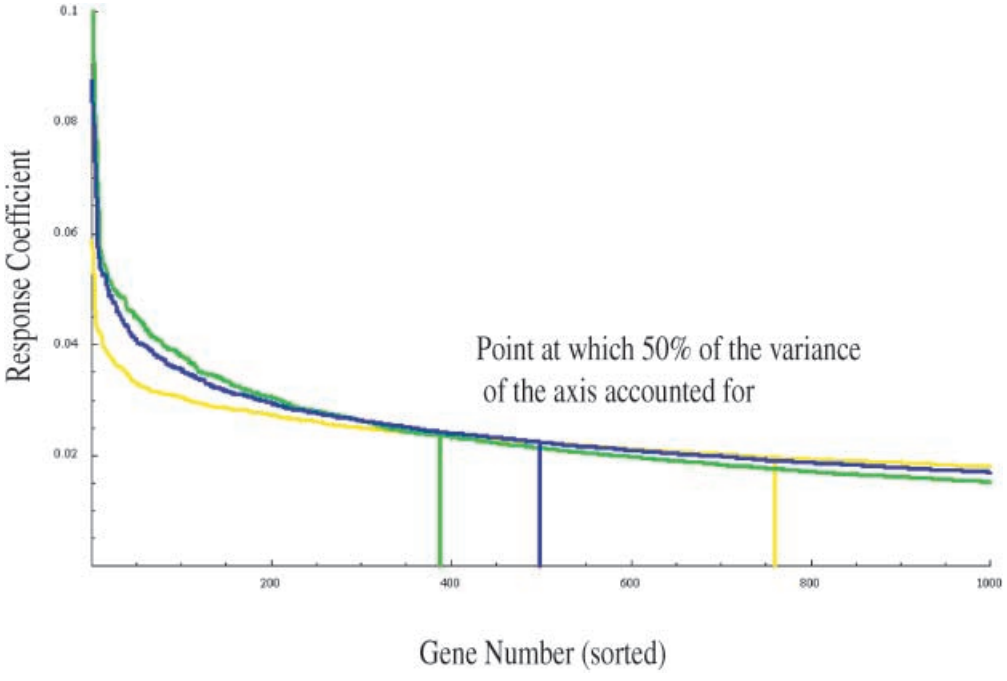

Fig. 4 Response coefficient distributions of the largest thousand response coefficients (sorted by magnitude) along the longest structural component axes of elutriation, *cdc15*, and *cdc28*. *Vertical lines* indicate the number of genes which account for the first 50% of the variance along that axis

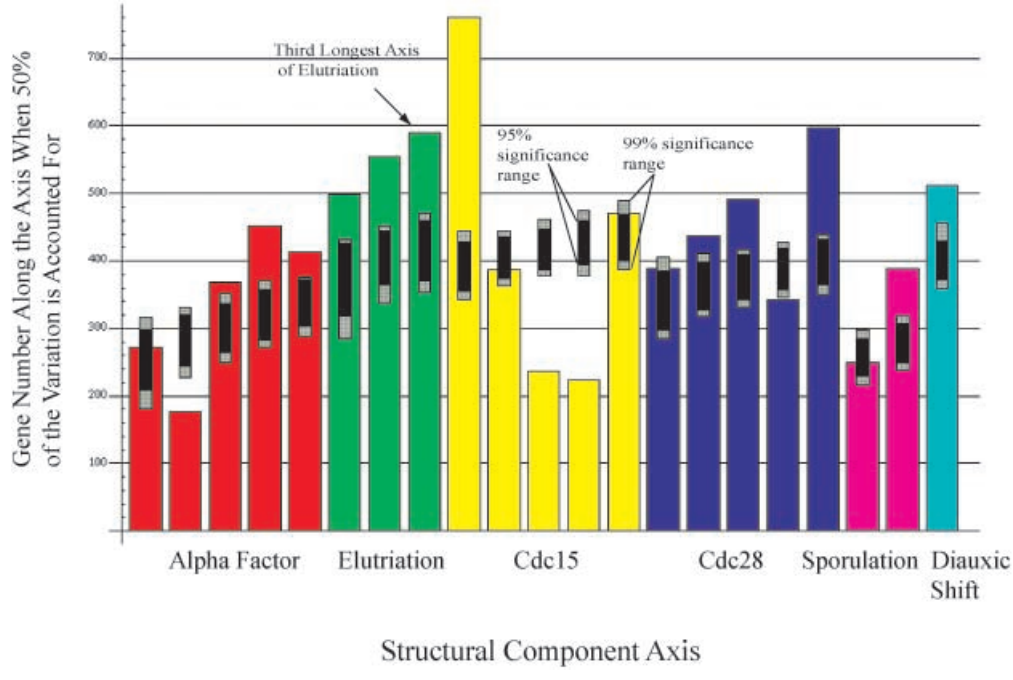

Fig. 5 Systemic contributions to the axes. The number of genes which account for the first 50% of the variance of the significant structural component axes. *Colored bars* indicate the levels for the structural component axes ordered *from left to right*. *Black bars* indicate the range of this statistic for 95% of the 200 random matrices. *Grey bars* indicate the 99% significance range

terns of the experimental groups – between their covariational patterns – we calculated canonical correlations between them (Table 2). Sporulation is least related to the other experiments and the maximal correlations between the cell cycle experiments are around 0.5.

Distributions of response coefficients along the structural component axes

Consider a system with two genes. If variation in one of the genes drives variation in the cell state as a whole during a particular experiment, the angle between the largest structural component axis and that gene's axis will be small, making the response coefficient large. Conversely, the response coefficient of the other gene will be small. This distribution of these response coefficients is more asymmetrical than one where the structural component axis lay at a 45° angle between the two. We used this asymmetry of the response coefficient distribution along a particular axis to measure the degree to which an axis lies along a particular subspace of the gene expression space. After sorting the squares of the response coefficients, we compared how many genes accounted for the

Table 3 ORFs with the 20 highest overall response indices (ORIs) in the alpha factor experiment

Rank	ORI	Gene	Functional class	Specific function	ORF (yeast final code)
1	0.0090	FIG1	Mating	Extracellular integral membrane protein	YBR040W
2	0.0065	EGT2	Cell cycle	Unknown	YNL327W
3	0.0062	AGA1	Mating	α-agglutinin anchor subunit	YNR044W
4	0.0058	MFA1	Mating	α-factor precursor	YDR461W
5	0.0056	–	–	–	YNR067C
6	0.0055	–	–	–	YCRX18C
7	0.0053	ASG7	Unknown	Unknown	YJL170C
8	0.0048	CLB2	Cell cycle	G2/M cyclin	YPR119W
9	0.0047	–	–	–	YNL279W
10	0.0045	AGA2	Mating	α-agglutinin binding subunit	YGL032C
11	0.0043	SVS1	Vanadate resistance	Unknown	YPL163C
12	0.0040	PIR1	Unknown	Unknown; Pir1p/Hsp150p/Pir3p family	YKL164C
13	0.0038	HHF1	Chromatin structure	Histone H4	YBR009C
14	0.0037	SCW11	Cell wall biogenesis	Glucanase (putative)	YGL028C
15	0.0037	MRH1	Unknown	Similar to Yro2p and Hsp30p	YDR033W
16	0.0036	CTS1	Cell wall biogenesis	Endochitinase	YLR286C
17	0.0034	CST13	Cu2+ homeostasis; chromosome stability	Unknown; required for optimal growth	YBR158W
18	0.0033	HHF2	Chromatin structure	Histone H4	YNL030W
19	0.0031	–	–	–	YPL158C
20	0.0031	HTB2	Chromatin structure	Histone H2B	YBL002W

Table 4 ORFs with the 20 highest ORIs in the elutriation experiment

Rank	ORI	Gene	Functional class	Specific function	ORF (yeast final code)
1	0.0044	HSP30	Diauxic shift	Plasma membrane heat shock protein	YCR021C
2	0.0037	CUP1–1	Cu^{2+} ion homeostasis	Metallothionein	YHR053C
3	0.0037	CUP1–2	Cu^{2+} ion homeostasis	Metallothionein	YHR055C
4	0.0036	CTS1	Cell wall biogenesis	Endochitinase	YLR286C
5	0.0034	SPI1	Unknown	Unknown; similar to Sed1p; induced in stationary phase	YER150W
6	0.0033	SRL1	Unknown	Unknown; similar to Svs1p; suppressor of Rad53 lethality	YOR247W
7	0.0026	CIS3	Unknown	Unknown; overexpression suppresses cik1 deletion	YJL158C
8	0.0026	CLB1	Cell cycle	G2/M cyclin	YGR108W
9	0.0026	CHS2	Cell wall biogenesis	Chitin synthase II	YBR038W
10	0.0024	ECM33	Cell wall biogenesis	Unknown	YBR078W
11	0.0024	–	–	–	YNR067C
12	0.0023	EGT2	Cell cycle	Unknown	YNL327W
13	0.0022	CRH1	Cell wall biogenesis (putative)	Unknown; cell wall protein	YGR189C
14	0.0022	GAS1	Unknown	Cell surface glycoprotein	YMR307W
15	0.0022	–	–	–	YOR248W
16	0.0021	GIT1	Unknown	Unknown; similar to phosphate transporter	YCR098C
17	0.0020	ECM13	Cell wall biogenesis	Unknown	YBL043W
18	0.0020	SCW10	Cell wall biogenesis	Glucanase (putative)	YMR305C
19	0.0020	–	–	–	YER124C
20	0.0019	EXG1	Cell wall biogenesis	Exo-β-1 3-glucanase	YLR300 W

first 50% of the variance along that axis compared to axes from our random matrices.

In all of the experiments, a few hundred genes account for the first 50% of the variance along any particular axis (Fig. 4). Variation in a single gene or even in a few genes does not drive the variation in cell state. Compared to the random distributions, significantly more genes account for the variation along most of the axes, indicating that these axes represent systemic variation in the cell state (Fig. 5). Moreover, the genes with high response coefficients differ between the structural component axes within an experiment. These axes only share between 10% and 25% of the genes which account for the first 50% of their respective variances.

Table 5 ORFs with the 20 highest ORIs in the cdc15 experiment

Rank	ORI	Gene	Functional class	Specific function	ORF (yeast final code)
1	0.0039	CTS1	Cell wall biogenesis	Endochitinase	YLR286C
2	0.0030	PIR1	Unknown	Unknown; Pir1p/Hsp150p/Pir3p family	YKL164C
3	0.0025	SCW11	Cell wall biogenesis	Glucanase (putative)	YGL028C
4	0.0025	SAG1	Mating	α-agglutinin	YJR004C
5	0.0024	PHO3	Thiamine uptake	Acid phosphatase; constituitive	YBR092C
6	0.0023	–	–	Similar to wheat glutenin, secalin	YBR108W
7	0.0022	MF(ALPHA)2	Mating	Alpha factor	YGL089C
8	0.0022	NCE102	Secretion	Non-classical, unknown	YPR149W
9	0.0020	ALG1	Protein glycosylation	β-1,4-mannosyltransferase	YBR110W
10	0.0020	–	–	–	YHR143W
11	0.0020	–	–	–	YER124C
12	0.0019	–	–	Similar to subtelomerically-encoded proteins	YCR007C
13	0.0019	ALD6	Ethanol utilization	Acetaldehyde dehydrogenase	YPL061W
14	0.0019	YRO2	Unknown	Putative heat shock protein	YBR054W
15	0.0019	DIE2	Glucosylation?	Glucosyltransferase	YGR227W
16	0.0019	FET3	Transport	Cell surface ferroxidase	YMR058W
17	0.0018	ECM23	Cell wall biogenesis (putative)	Unknown	YPL021W
18	0.0018	SSA1	ER and mitochondrial translocation	Cytosolic HSP70	YAL005C
19	0.0018	RME1	Meiosis	Transcription factor	YGR044C
20	0.0018	YGP1	Diauxic shift	Unknown; response to nutrient limitation	YNL160W

Table 6 ORFs with the 20 highest ORIs in the cdc28 experiment

Rank	ORI	Gene	Functional class	Specific function	ORF (yeast final code)
1	0.0083	–	–	–	YDR274C
2	0.0055	WSC4	Cell wall integrity and stress response	Unknown	YHL028W
3	0.0034	–	–	–	YOL101C
4	0.0032	PDC6	Glycolysis	Pyruvate decarboxylase 3	YGR087C
5	0.0028	–	–	Similar to human zinc finger protein PIR:JC2069	YPR031W
6	0.0024	INO1	Inositol biosynthesis	L-Myo-inositol-1-phosphate synthase	YJL153C
7	0.0024	–	–	Similar to MAL regulatory proteins	YFL052W
8	0.0024	–	–	–	YHR217C
9	0.0023	–	–	–	YKL086W
10	0.0023	–	–	–	YDR355C
11	0.0023	–	–	Similar to glycophospholipid-anchored surface glycoprotein GAS1	YOL132W
12	0.0022	–	–	–	YJR082C
13	0.0022	PDR12	Drug resistance	Transporter	YPL058C
14	0.0019	–	–	Similar to human retinoblastoma binding protein 2	YJR119C
15	0.0018	–	–	–	YNR069C
16	0.0018	CDA2	Sporulation	Chitin deacetylase	YLR308W
17	0.00178	–	–	Similar to glucan 1,4-α-glucosidase	YDL037C
18	0.0016	PRP16	mRNA splicing	RNA helicase	YKR086W
19	0.0016	–	–	–	YNR067C
20	0.0016	–	–	–	YPL222W

Classification of gene types by contributions to patterns of covariation

The components of the structural component axes vectors – the response coefficients – give us a way to measure how individual genes contribute to the overall structure of the data. Tables 3, 4, 5, 6, 7, 8 list genes ranked by their ORIs. This rank ordering indicates the importance of each gene in explaining the overall pattern of expression covariation for a given experiment. Note that no single gene dominates the significant covariation in the cell state. We could not have predicted a priori which individual genes would be highlighted by our analysis, and these genes are not necessarily the ones which are most upregulated or downregulated (data not shown). However, the categories to which these genes belong

Table 7 ORFs with the 20 highest ORIs in the sporulation experiment

Rank	ORI	Gene	Functional Class	Specific Function	ORF (Yeast Final Code)
1	0.0056	MIP6	mRNA export (putative)	RNA-binding protein	YHR015W
2	0.0046	SSP1	Meiosis	Nuclear division and spore formation	YHR184W
3	0.0046	PES4	DNA replication	Suppresses DNA polymerase epsilon mutation	YFR023W
4	0.0043	SPR28	Sporulation	Septin-related protein	YDR218C
5	0.0041	–	–	–	YOL015W
6	0.0041	–	–	–	YOR255W
7	0.0041	–	–	Similar to phosphoribulokinase precursor	YGL170C
8	0.0040	HXT10	Transport	Hexose permease	YFL011W
9	0.0040	–	–	Similar to glycophospholipid-anchored surface glycoprotein GAS1	YOL132W
10	0.0040	–	–	–	YEL023C
11	0.0039	SPS2	Meiosis	Unknown	YDR522C
12	0.0039	HXT14	Transport	Hexose permease	YNL318C
13	0.0038	SPR3	Sporulation	Septin	YGR059W
14	0.0038	–	–	–	YGR273C
15	0.0037	–	–	–	YJL038C
16	0.0035	SPO19	Sporulation	GPI-protein,meiosis-specific	YPL130W
17	0.0033	SPO21	Sporulation	Unknown	YOL091W
18	0.0033	NDT80	Meiosis	Transcription factor	YHR124W
19	0.0033	–	–	–	YAL018C
20	0.0032	–	–	–	YLR341W

Table 8 ORFs with the 20 highest ORIs in the diauxic shift experiment

Rank	ORI	Gene	Functional class	Specific function	ORF (yeast final code)
1	0.0037	JEN1	Transport	Lactate transporter	YKL217W
2	0.0032	–	–	Similar to Sur7p	YNL194C
3	0.0029	–	–	–	YDL204W
4	0.0028	–	–	Similar to Tal1p	YGR043C
5	0.0028	–	–	–	YGR236C
6	0.0028	–	–	–	YML128C
7	0.0025	GPX1	Glutathione metabolism (putative)	Glutathione peroxidase (putative)	YKL026C
8	0.0024	HSP42	Cytoskeleton assembly	Heat shock protein similar	YDR171W to HSP26
9	0.0024	SOL4	Unknown	Similar to Sol3p	YGR248W
10	0.0024	–	–	–	YGR243W
11	0.0024	–	–	–	YCR021C
12	0.0023	ACH1	Acetyl-CoA metabolism	Acetyl-CoA hydrolase	YBL015W
13	0.0022	HSP12	Glucose and lipid utilization	Heat shock protein	YFL014W
14	0.0022	CTT1	Oxidative stress response	Catalase T	YGR088W
15	0.0022	ACS1	Acetyl-CoA biosynthesis	Acetyl-CoA synthetase	YAL054C
16	0.0021	–	–	Similar to Stf2p	YLR327C
17	0.0021	–	–	–	YNL200C
18	0.0020	–	–	–	YOR215C
19	0.0020	SAM1	Methionine metabolism	S-adenosylmethionine synthetase	YLR180W
20	0.0019	OM45	Mitochondrial organization	Outer mitochondrial membrane protein	YIL136W

make sense biologically. For instance, genes which code for mating proteins, cell-cycle proteins and histones populate the alpha factor list; genes expressed in low-glucose conditions and other stress genes, cell-wall biogenesis genes, and cell cycle genes make the elutriation list; and, fittingly, sporulation and meiosis genes head the sporulation experiment list. Many of the important ORFs identified by this method are still uncharacterized.

Each structural component axis is an orthogonal direction in the gene-expression space. We examined whether each of these directions is differentiated by functional groupings of the genes by calculating the contribution of a MIPS functional class (http://www.mips.biochem.mpg.de/proj/yeast/catalogues/funcat/index.html) with respect to the variance of the structural component axes of the known genes. Figure 6a, b shows the distributions of relative contributions of two functional classes of genes across the significant structural component axes for each experiment. Figure 6c shows this analysis applied to the genes identified by Spellman et al. (1998) as cell cycle

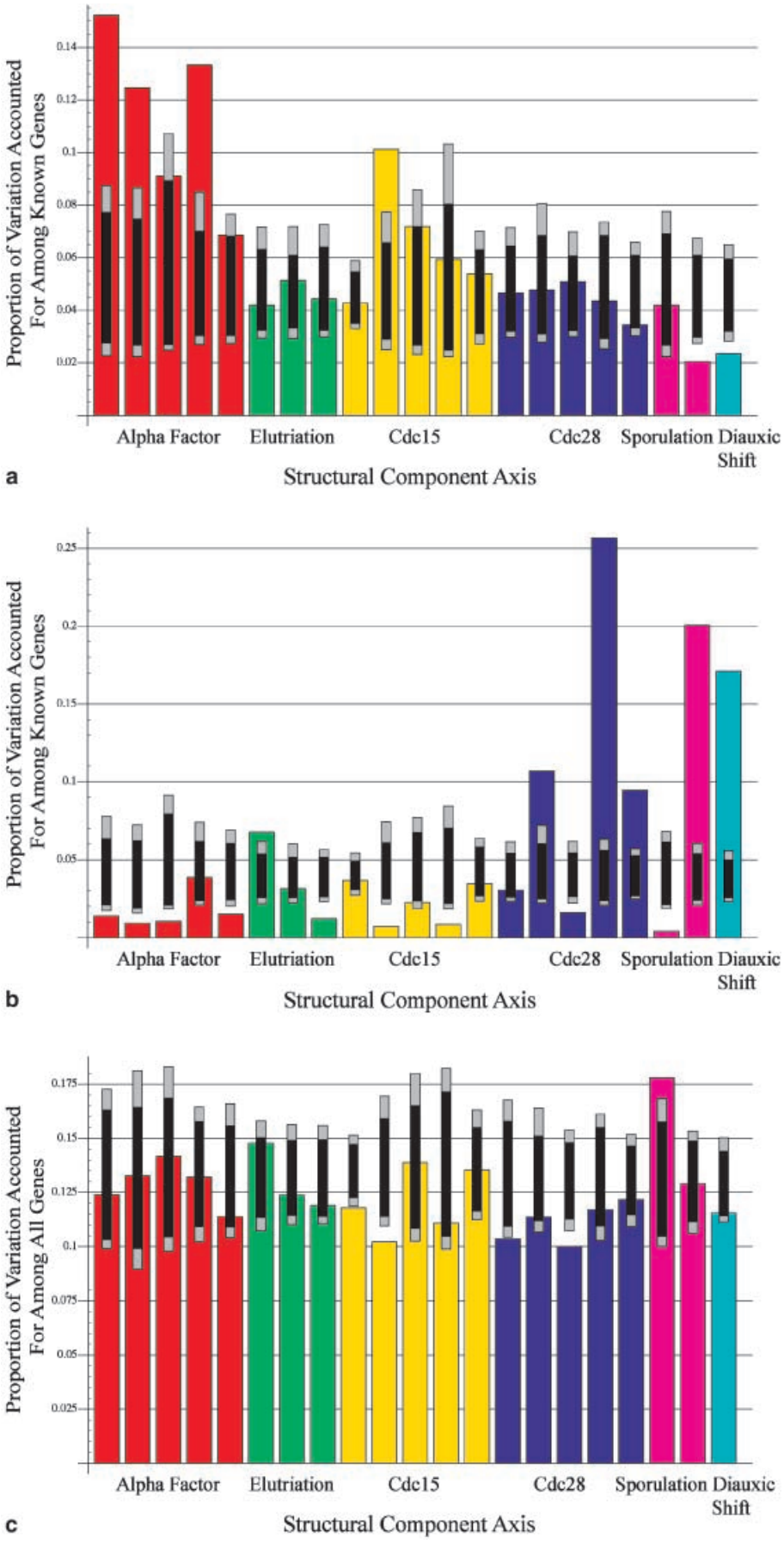

Fig. 6 Proportion of variation among known genes accounted for, **a** by mating protein genes; **b** by cytoplasmic ribosomal protein genes; and **c** by purported cell cycle regulated genes identified by Spellman et al. (1998)

regulated. The cytoplasmic ribosomal protein genes have the most extreme distribution with significantly low variation ($P<0.01$) along 8 out of 18 cell cycle structural component axes and extremely high variation ($P<<0.01$) along the second axis of sporulation and the axis of diauxic shift and markedly along one axis of the cdc28 experiment. The mating protein genes in our sample load heavily to all five alpha factor axes but only onto two of those in the other experiments. Finally, the genes identified as cell-cycle regulated by Spellman et al. (1998) did not significantly contribute as a class to any of the cell cycle axes in the cell-cycle experiments.

Discussion

Hierarchical clustering methods are currently the most prevalent methods for analyzing genomic expression data and can be interpreted within the geometric framework we present above. In gene clustering (e.g. Eisen et al. 1998; Wen et al. 1998), the data are points in a condition space where an axis is an axis of expression level in a particular hybridization and each array adds another dimension. Nearby genes share an expression pattern across conditions, and groups of genes are clustered according to the distances between their centers (which can be calculated in a variety of ways). In condition, or array, clustering (e.g. Alon et al. 1999), the data are points in a gene expression space such as the one described above, and nearby conditions share gene expression profiles. In the context of CSA, condition clustering ignores the systemic covariational pattern of a data cloud from related conditions and groups the points according to their locations. That is, conditional clustering does not yield parametric structural relationships between different experimental or phenotypic conditions; rather it produces a heuristic classification which is difficult to relate back to individual genes.

The significant contributions of functional classes to individual structural component axes, particularly the cytoplasmic ribosomal protein genes, indicate that the structural component axes identified by singular value decomposition may in fact reflect functional divisions in the yeast genome. Other classes of genes, such as genes whose proteins are involved in structuring chromatin and molecular chaperone genes, also support this correspondence since their major influences seem restricted to particular structural component axes. However, none approaches the clarity of the cytoplasmic ribosomal protein genes. This may be for several reasons. Our analyses of the response coefficient distributions indicate that the axes are picking up a systemic pattern of interaction in the genome. The linear least squares fit of orthogonal axes to the dataset by CSA may indeed isolate functionally related genes, but we may not know enough about the in vivo functions of genes and the pathways in which they operate to recognize that these axes are meaningful. In this case, the genome would be functionally more integrated than one would expect. However, it is likely that the few conditions in an experiment are insufficient to resolve these functional categories. In the alpha factor experiment, for example, the entire centered dataset forms a maximum-17 dimensional cloud in a 6,074 dimensional space. Consequently, the cell state can only vary independently in 17 dimensions, and so the structural component axes we derive may be projections of several meaningful structural component axes and not biologically meaningful in themselves. This limitation is not unique to our analysis and will disappear as the results from more microarray experiments become publicly available, and as these experiments explore widely divergent conditions – the better to amass a large set of independent data and explore the range of possible cell states.

These yeast datasets are noisy. The data points were not replicated, and errors of yet unknown magnitude can enter into every step of microarray and probe preparation, hybridization, and analysis. Although advances in the technology will certainly help to reduce error, researchers should use replicate data points in order to estimate variation due to noise, thereby making the results from their analyses more robust (e.g. White et al. 1999). The problem of noise is not unique to CSA; clustering techniques generally set an arbitrary threshold level of induction or repression to choose which genes to analyze (Eisen et al. 1998). Similar cut-offs could be applied in CSA. Since we expect the noise to be unbiased, such cutoffs would change the lengths of the structural component axes but leave the directions largely unchanged. Our permutation test identifies these more robust longest structural component axes. Noise within the measurements for genes above the cutoff presents a more serious problem which replication will ameliorate.

To summarize, building upon the notion that a single microarray experiment is a window into the gene expression state of a cell, we have developed an analytical technique based on a geometric framework to highlight structure in gene expression covariation in an experiment, to compare the gene expression states of cells or organisms under different conditions, and to find limits on how genome expression can vary under these conditions. Most importantly, we find that no particular groups of genes characterize particular experimental conditions. Instead, the particular structure of the coordinated expression of the entire genome characterizes a particular experiment.

Any developmental or physiological process depends on the coordinated expression of multiple genes. Molecular genetics has long been handicapped by its inability to probe these interactions in detail, charting instead the expression of one or a few genes at a time in a limited number of conditions. While this approach has been extremely fruitful, it has also promoted a gene-centric view where the action of each gene is viewed isolated from its context. Genome-wide expression technologies give us, for the first time, the ability to study gene action comprehensively as a dynamical system. These new data strongly show that the context-dependent systemic action of

the whole genome and not the context-independent action of individual genes governs biological function. Traditional analyses take a bottom-up approach where individual genes are identified and attempts are made to infer their interactions. The analysis we present here suggests that a more efficient solution is to take a systemic view of the genome, focusing on the state of the whole cell or the whole organism and asking how individual parts like genes participate in determining these states and their biological functions.

Acknowledgements We thank members of the Mike Snyder lab and Mike Ronemus for comments on previous versions of this paper. The paper was improved by additional comments from two anonymous reviwers. This material is based upon work supported under a National Science Foundation Graduate Fellowship to S.A.R., an NSF/DOE Postdoctoral Fellowship to K.A., and a Merck Genome Research Institute grant and an Intel Education 2000 grant to J.K.

References

Alon U, Barkai N, Notterman DA, et al (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96:6745–6750

Chu S, Derisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I (1998) The transcriptional program of sporulation in budding yeast. Science 282:699–705

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680–686

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868

Green PE, Carroll JD (1978) Mathematical tools for applied multivariate analysis. Academic Press, New York

Hlavacek WS, Savageau MA (1996) Rules for coupled expression of regulator and effector genes in inducible circuits. J Mol Biol 255:121–139

Hlavacek WS, Savageau MA (1997) Completely uncoupled and perfectly coupled gene expression in repressible systems. J Mol Biol 266:538–558

Holstege FC, Jennings EG, Wyrick JJ, et al (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95:717–728

Raychandhuri S, Stuart J, Altman R (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 5:452–463

Reinitz J, Kosman D, Vanario-Alonso CE, Sharp DH (1998) Stripe forming architecture of the gap gene system. Dev Genet 23:11–27

Savageau MA (1999) Design of gene circuitry by natural selection: analysis of the lactose catabolic system in Escherichia coli. Biochem Soc Trans 27:264–270

Seber GAF (1984) Multivariate observations. Wiley, Brisbane

Spellman PT, Sherlock G, Zhand MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol Biol Cell. 9:3273–3297

Szallasi Z (1999) Genetic network analysis in light of massively parallel biological data acquisition. Pac Symp Biocomput 4:5–16

Wen XL, Fuhrman S, Michaels GS, et al (1998) Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci USA 95:334–339

White KP, Rifkin SA, Hurban P, Hogness DS (1999) Microarray analysis of *Drosophila* development during metamorphosis. Science 286:2179–2184

Wolf DM, Eeckman FH (1998) On the relationship between genomic regulatory element organization and gene regulatory dynamics. J Theor Biol 195:167–186

Wolfram S (1999) The mathematica book. Wolfram Media/Cambridge University, Cambridge