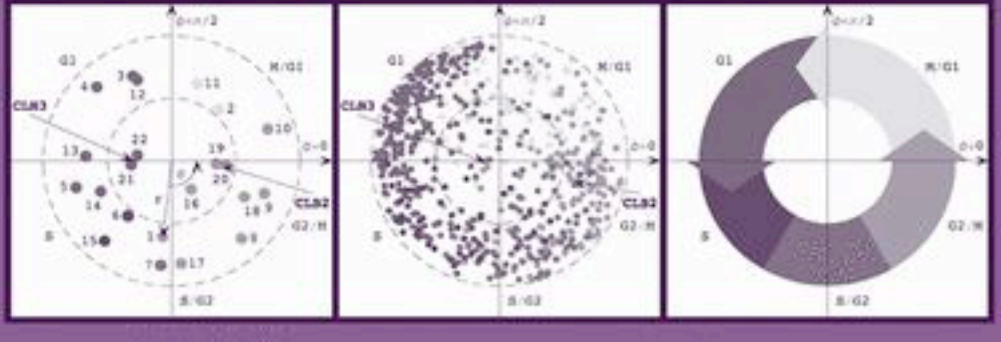
METHODS IN MOLECULAR BIOLOGY™ 377

Microarray Data Analysis

Methods and Applications

Edited by

Michael J. Korenberg



HUMANA PRESS

2

Genomic Signal Processing: From Matrix Algebra to Genetic Networks

Orly Alter

Summary

DNA microarrays make it possible, for the first time, to record the complete genomic signals that guide the progression of cellular processes. Future discovery in biology and medicine will come from the mathematical modeling of these data, which hold the key to fundamental understanding of life on the molecular level, as well as answers to questions regarding diagnosis, treatment, and drug development. This chapter reviews the first data-driven models that were created from these genome-scale data, through adaptations and generalizations of mathematical frameworks from matrix algebra that have proven successful in describing the physical world, in such diverse areas as mechanics and perception: the singular value decomposition model, the generalized singular value decomposition model comparative model, and the pseudoinverse projection integrative model. These models provide mathematical descriptions of the genetic networks that generate and sense the measured data, where the mathematical variables and operations represent biological reality. The variables, patterns uncovered in the data, correlate with activities of cellular elements such as regulators or transcription factors that drive the measured signals and cellular states where these elements are active. The operations, such as data reconstruction, rotation, and classification in subspaces of selected patterns, simulate experimental observation of only the cellular programs that these patterns represent. These models are illustrated in the analyses of RNA expression data from yeast and human during their cell cycle programs and DNA-binding data from yeast cell cycle transcription factors and replication initiation proteins. Two alternative pictures of RNA expression oscillations during the cell cycle that emerge from these analyses, which parallel well-known designs of physical oscillators, convey the capacity of the models to elucidate the design principles of cellular systems, as well as guide the design of synthetic ones. In these analyses, the power of the models to predict previously unknown biological principles is demonstrated with a prediction of a novel mechanism of regulation that correlates DNA replication initiation with cell cycle-regulated RNA transcription in yeast. These models may become the foundation of a future in which biological systems are modeled as physical systems are today.

From: *Methods in Molecular Biology, vol. 377, Microarray Data Analysis: Methods and Applications*
Edited by: M. J. Korenberg © Humana Press Inc., Totowa, NJ

Key Words: Singular value decomposition (SVD); generalized SVD (GSVD); pseudoinverse projection; blind source separation (BSS) algorithms; genome-scale RNA expression and proteins' DNA-binding data; cell cycle; yeast *Saccharomyces cerevisiae;* human HeLa cell line; analog harmonic and digital ring oscillators.

1. Introduction

1.1. DNA Microarray Technology and Genome-Scale Molecular Biological Data

The Human Genome Project, and the resulting sequencing of complete genomes, fueled the emergence of the DNA microarray hybridization technology in the past decade. This novel experimental high-throughput technology makes it possible to assay the hybridization of fluorescently tagged DNA or RNA molecules, which were extracted from a single sample, with several thousand synthetic oligonucleotides ***(1)*** or DNA targets ***(2)*** simultaneously. Different types of molecular biological signals, such as DNA copy number, RNA expression levels, and DNA-bound proteins' occupancy levels, that correspond to activities of cellular systems, such as DNA replication, RNA transcription, and binding of transcription factors to DNA, can now be measured on genomic scales (e.g., **refs. *3*** and ***4***). For the first time in human history it is possible to monitor the flow of molecular biological information, as DNA is transcribed to RNA, RNA is translated to proteins, and proteins bind to DNA, and thus to observe experimentally the global signals that are generated and sensed by cellular systems. Already laboratories all over the world are producing vast quantities of genome-scale data in studies of cellular processes and tissue samples (e.g., **refs. *5–9***).

Analysis of these new data promises to enhance the fundamental understanding of life on the molecular level and might prove useful in medical diagnosis, treatment, and drug design. Comparative analysis of these data among two or more organisms promises to give new insights into the universality as well as the specialization of evolutionary, biochemical, and genetic pathways. Integrative analysis of different types of these global signals from the same organism promises to reveal cellular mechanisms of regulation, i.e., global causal coordination of cellular activities.

1.2. From Technology and Large-Scale Data to Discovery and Control of Basic Phenomena Using Mathematical Models: Analogy From Astronomy

Biology and medicine today, with these recent advances in DNA microarray technology, may very well be at a point similar to where physics was after the advent of the telescope in the 17th century. In those days, astronomers were

compiling tables detailing observed positions of planets at different times for navigation. Popularized by Galileo Galilei, telescopes were being used in these sky surveys, enabling more accurate and more frequent observations of a growing number of celestial bodies. One astronomer, Tycho Brahe, compiled some of the more extensive and accurate tables of such astronomical observations. Another astronomer, Johannes Kepler, used mathematical equations from analytical geometry to describe trends in Brahe's data, and to determine three laws of planetary motion, all relating observed time intervals with observed distances. These laws enabled the most accurate predictions of future positions of planets to date. Kepler's achievement posed the question: why are the planetary motions such that they follow these laws? A few decades later, Isaac Newton considered this question in light of the experiments of Galileo, the data of Brahe, and the models of Kepler. Using mathematical equations from calculus, he introduced the physical observables mass, momentum, and force, and defined them in terms of the observables time and distance. With these postulates, the three laws of Kepler could be derived within a single mathematical framework, known as the universal law of gravitation, and Newton concluded that the physical phenomenon of gravitation is the reason for the trends observed in the motion of the planets ***(10)***. Today, Newton's discovery and mathematical formulation of the basic phenomenon that is gravitation enables control of the dynamics of moving bodies, e.g., in exploration of outer space.

The rapidly growing number of genome-scale molecular biological datasets hold the key to the discovery of previously unknown molecular biological principles, just as the vast number of astronomical tables compiled by Galileo and Brahe enabled accurate prediction of planetary motions and later also the discovery of universal gravitation. Just as Kepler and Newton made their discoveries by using mathematical frameworks to describe trends in these large-scale astronomical data, also future predictive power, discovery, and control in biology and medicine will come from the mathematical modeling of genome-scale molecular biological data.

1.3. From Complex Signals to Simple Principles Using Mathematical Models: Analogy From Neuroscience

Genome-scale molecular biological signals appear to be complex, yet they are readily generated and sensed by the cellular systems. For example, the division cycle of human cells spans an order of one day only of cellular activity. The period of the cell division cycle in yeast is of the order of an hour.

DNA microarray data or genomic-scale molecular biological signals, in general, may very well be similar to the input and output signals of the

central nervous system, such as images of the natural world that are viewed by the retina and the electric spike trains that are produced by the neurons in the visual cortex. In a series of classic experiments, the neuroscientists Hubel and Wiesel ***(11)*** recorded the activities of individual neurons in the visual cortex in response to different patterns of light falling on the retina. They showed that the visual cortex represents a spatial map of the visual field. They also discovered that there exists a class of neurons, which they called "simple cells," each of which responds selectively to a stimulus of an edge of a given scale at a given orientation in the neuron's region of the visual field. These discoveries posed the question: what might be the brain's advantage in processing natural images with a series of spatially localized scale-selective edge detectors? Barlow ***(12)*** suggested that the underlying principle of such image processing is that of sparse coding, which allows only a few neurons out of a large population to be simultaneously active when representing any image from the natural world. Naturally, such images are made out of objects and surfaces, i.e., edges. Two decades later, Olshausen and Field ***(13***; *see also* Bell and Sejnowski, **ref.** ***14***) developed a novel algorithm, which separates or decomposes natural images into their optimal components, where they defined optimality mathematically as the preservation of a characteristic ensemble of images as well as the sparse representation of this ensemble. They showed that the optimal sparse linear components of a natural image are spatially localized and scaled edges, thus validating Barlow's postulate.

The sensing of the complex genomic-scale molecular biological signals by the cellular systems might be governed by simple principles, just as the processing of the complex natural images by the visual cortex appear to be governed by the simple principle of sparse coding. Just as the natural images could be represented mathematically as superpositions, i.e., weighted sums of images, which correlate with the measured sensory activities of neurons, also the complex genomic-scale molecular biological signals might be represented mathematically as superpositions of signals, which might correspond to the measured activities of cellular elements.

1.4. Matrix Algebra Models for DNA Microarray Data

This chapter reviews the first data-driven predictive models for DNA microarray data or genomic-scale molecular biological signals in general. These models use adaptations and generalizations of matrix algebra frameworks ***(15)*** in order to provide mathematical descriptions of the genetic networks that generate and sense the measured data. The singular value decomposition (SVD) model formulates a dataset as the result of a simple linear network **(Fig. 1A)**: the measured gene patterns are expressed mathematically as superpositions of the effects of a few independent sources, biological or experimental, and the

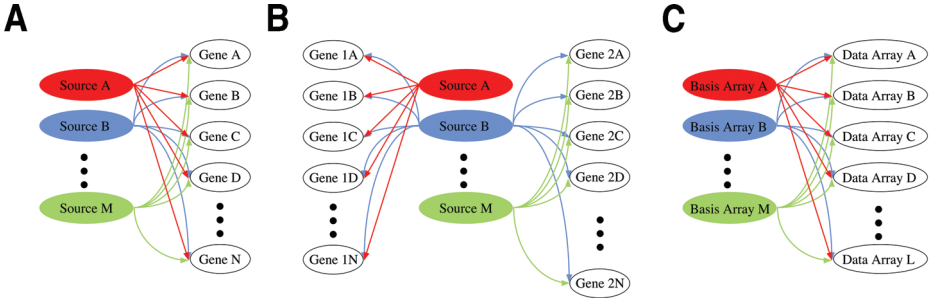


Fig. 1. The first data-driven predictive models for DNA microarray data. **(A)** The singular value decomposition (SVD) model describes the overall observed genome-scale molecular biological data as the outcome of a simple linear network, where a few independent sources, experimental or biological, and the corresponding cellular states, affect all the genes and arrays, i.e., samples, in the dataset. **(B)** The generalized SVD (GSVD) comparative model describes the two genome-scale molecular biological datasets as the outcome of a simple linear comparative network, where a few independent sources, some common to both datasets whereas some are exclusive to one dataset or the other, affect all the genes in both datasets. **(C)** The pseudoinverse projection integrative model approximates any number of datasets as the outcome of a simple linear integrative network, where the cellular states, which correspond to one chosen "basis" set of observed samples, affect all the samples, or arrays, in each dataset.

measured sample patterns, as superpositions of the corresponding cellular states ***(16–18)***. The comparative generalized SVD (GSVD) model formulates two datasets, e.g., from two different organisms such as yeast and human, as the result of a simple linear comparative network **(Fig. 1B)**: the measured gene patterns in each dataset are expressed mathematically simultaneously as superpositions of a few independent sources that are common to both datasets, as well as sources that are exclusive to one of the datasets or the other ***(19)***. The integrative pseudoinverse projection model approximates any number of datasets from the same organism, e.g., of different types of data such as RNA expression levels and proteins' DNA-binding occupancy levels, as the result of a simple linear integrative network **(Fig. 1C)**: the measured sample patterns in each dataset are formulated simultaneously as superpositions of one chosen set of measured samples, or of profiles extracted mathematically from these samples, designated the "basis" set ***(20,21)***.

The mathematical variables of these models, i.e., the patterns that these models uncover in the data, represent biological or experimental reality. The "eigengenes" uncovered by SVD, the "genelets" uncovered by GSVD, and the pseudoinverse correlations uncovered by pseudoinverse projection, correlate with independent processes, biological or experimental, such as observed

genome-wide effects of known regulators or transcription factors, the cellular elements that generate the genome-wide RNA expression signals most commonly measured by DNA microarrays. The corresponding "eigenarrays" uncovered by SVD and "arraylets" uncovered by GSVD, correlate with the corresponding cellular states, such as measured samples in which these regulators or transcription factors are overactive or underactive.

The mathematical operations of these models, e.g., data reconstruction, rotation, and classification in subspaces spanned by these patterns also represent biological or experimental reality. Data reconstruction in subspaces of selected eigengenes, genelets, or pseudoinverse correlations, and corresponding eigenarrays or arraylets, simulates experimental observation of only the processes and cellular states that these patterns represent, respectively. Data rotation in these subspaces simulates the experimental decoupling of the biological programs that these subspaces span. Data classification in these subspaces maps the measured gene and sample patterns onto the processes and cellular states that these subspaces represent, respectively.

Because these models provide mathematical descriptions of the genetic networks that generate and sense the measured data, where the mathematical variables and operations represent biological or experimental reality, these models have the capacity to elucidate the design principles of cellular systems as well as guide the design of synthetic ones (e.g., **ref. *22***). These models also have the power to make experimental predictions that might lead to experiments in which the models can be refuted or validated, and to discover previously unknown molecular biological principles ***(21,23)***. Ultimately, these models might enable the control of biological cellular processes in real time and in vivo ***(24)***.

Although no mathematical theorem promises that SVD, GSVD, and pseudoinverse projection could be used to model DNA microarray data or genome-scale molecular biological signals in general, these results are not counterintuitive. Similar and related mathematical frameworks have already proven successful in describing the physical world, in such diverse areas as mechanics and perception ***(25)***.

First, SVD, GSVD, and pseudoinverse projection, interpreted as they are here as simple approximations of the networks or systems that generate and sense the processed signals, belong to a class of algorithms called blind source separation (BSS) algorithms. BSS algorithms, such as the linear sparse coding algorithm by Olshausen and Field ***(13)***, the independent component analysis by Bell and Sejnowski ***(14)*** and the neural network algorithms by Hopfield ***(26)***, separate or decompose measured signals into their mathematically defined optimal components. These algorithms have already proven successful in modeling natural signals and computationally mimicking the activity of the brain as it expertly perceives these signals, for example, in face recognition ***(27,28)***.

Second, SVD, GSVD, and pseudoinverse projection can be also thought of as generalizations of the eigenvalue decomposition (EVD) and generalized EVD (GEVD) of Hermitian matrices, and inverse projection onto an orthogonal matrix, respectively. In mechanics, EVD of the Hermitian matrix, which tabulates the energy of a system of coupled oscillators, uncovers the eigenmodes and eigenfrequencies of this system, i.e., the normal coordinates, which oscillate independently of one another, and their frequencies of oscillations. One of these eigenmodes represents the center of mass of the system. GEVD of the Hermitian matrices, which tabulate the kinetic and potential energies of the oscillators, compares the distribution of kinetic energy among the eigenmodes with that of the potential energy. The inverse projection onto the orthogonal matrix, which tabulates the eigenmodes of this system, is equivalent to transformation of coordinates to the frame of reference, which is oscillating with the system (e.g., **ref. *29***). SVD, GSVD, and pseudoinverse projection are, therefore, generalizations of the frameworks that underlie the mathematical theoretical description of the physical world.

In this chapter, the mathematical frameworks of SVD, GSVD, and pseudoinverse projection are reviewed with an emphasis on the mathematical definition of the optimality of the components, or patterns, that each algorithm uncovers in the data. These models are illustrated in the analyses of RNA expression data from yeast and human during their cell cycle programs and DNA-binding data from yeast cell cycle transcription factors and replication initiation proteins. The correspondence between the mathematical frameworks and the genetic networks that generate and sense the measured data is outlined in each case, focusing on the correlations between the mathematical patterns and the observed cellular programs, as well as between the mathematical operations in subspaces spanned by selected patterns and the experimental observation of the cellular programs. Two alternative pictures of RNA expression oscillations during the cell cycle that emerge from these analyses are considered, and parallels between these pictures and well-known designs of physical oscillators, namely the analog harmonic oscillator and the digital ring oscillator, are drawn to convey the capacity of the models to elucidate the design principles of cellular systems, as well as guide the design of synthetic ones. Finally, the power of these models to predict previously unknown biological principles is demonstrated with a prediction of a novel mechanism of regulation that correlates DNA replication initiation with cell cycle-regulated RNA transcription in yeast.

2. SVD for Modeling DNA Microarray Data

This section reviews the SVD model for DNA microarray data ***(16–18, 22–24)***. SVD is a BSS algorithm that decomposes the measured signal, i.e., the measured gene and array patterns of, e.g. RNA expression, into mathematically decorrelated

and decoupled patterns, the "eigengenes" and "eigenarrays." The correspondence between these mathematical patterns uncovered in the measured signal and the independent biological and experimental processes and cellular states that compose the signal is illustrated with an analysis of genome-scale RNA expression data from the yeast *Saccharomyces cerevisiae* during its cell cycle program ***(6)***. The picture of RNA expression oscillations during the yeast cell cycle that emerges from this analysis suggests an underlying genetic network or circuit that parallels the analog harmonic oscillator.

2.1. Mathematical Framework of SVD

Let the matrix $\hat{e}$ of size *N*-genes × *M*-arrays tabulate the genome-scale signal, e.g., RNA expression levels, measured in a set of *M* samples using *M* DNA microarrays. The vector in the *m*th column of the matrix $\hat{e}$, $|a_m\rangle \equiv \hat{e}|m\rangle$, lists the expression signal measured in the *m*th sample by the *m*th array across all *N* genes simultaneously. The vector in the *n*th row of the matrix $\hat{e}$, $\langle g_n| \equiv \langle n|\hat{e}$, lists the signal measured for the *n*th gene across the different arrays, which correspond to the different samples.*

SVD is a linear transformation of this DNA microarray dataset from the *N*-genes × *M*-arrays space to the reduced *L*-eigenarrays × *L*-eigengenes space **(Fig. 2)**, where $L = \min\{M,N\}$,

$$\hat{e} = \hat{u}\hat{\varepsilon}\hat{v}^T. \tag{1}$$

In this space, the dataset or matrix $\hat{e}$ is represented by the diagonal nonnegative matrix $\hat{\varepsilon}$ of size *L*-eigenarrays × *L*-eigengenes. The diagonality of $\hat{\varepsilon}$ means that each eigengene is decoupled of all other eigengenes, and each eigenarray is decoupled of all other eigenarrays, such that each eigengene is expressed only in the corresponding eigenarray.

The "fractions of eigenexpression" $\{p_l\}$ are calculated from the "eigenexpression levels" $\{\varepsilon_l\}$, which are listed in the diagonal of $\hat{\varepsilon}$,

$$p_l = \frac{\varepsilon_l^2}{\sum_{k=1}^{L} \varepsilon_k^2}. \tag{2}$$

These fractions of eigenexpression indicate for each eigengene and eigenarray their significance in the dataset relative to all other eigengenes and eigenarrays in terms of the overall expression information that they capture in the data. Note that each fraction of eigenexpression can be thought of as the probability for any given gene among all genes in the dataset to express the corresponding

*In this chapter, $\hat{m}$ denotes a matrix, $|v\rangle$ denotes a column vector, and $\langle u|$ denotes a row vector, such that, $\hat{m}|v\rangle$, $\langle u|\hat{m}$, and $\langle u|v\rangle$ all denote inner products and $|v\rangle\langle u|$ denotes an outer product.

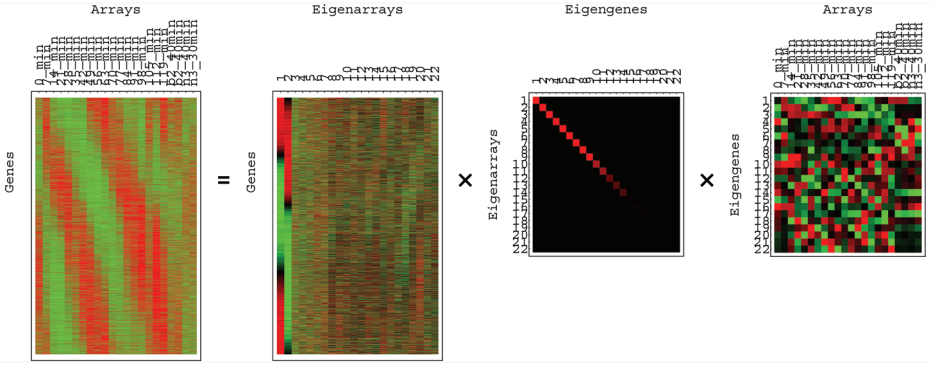


Fig. 2. Raster display of the SVD of the yeast cell cycle RNA expression dataset, with overexpression (red), no change in expression (black), and underexpression (green) around the steady state of expression of the 4579 yeast genes. SVD is a linear transformation of the data from the 4579-genes × 22-arrays space to the reduced diagonalized 22-eigenarrays × 22-eigengenes space, which is spanned by the 4579-genes × 22-eigenarrays and 22-eigengenes × 22-arrays bases.

eigengene, and at the same time, the probability for any given array among all arrays to express the corresponding eigenarray.

The "normalized Shannon entropy" of the dataset,

$$0 \le d = -\frac{1}{L}\sum_{k=1}^{L} p_k \log(p_k) \le 1, \tag{3}$$

measures the complexity of the data from the distribution of the overall expression information between the different eigengenes and corresponding eigenarrays, where $d = 0$ corresponds to an ordered and redundant dataset in which all expression is captured by one eigengene and the corresponding eigenarray, and $d = 1$ corresponds to a disordered and random dataset where all eigengenes and eigenarrays are equally expressed.

The transformation matrices $\hat{u}$ and $\hat{v}^T$ define the N-genes × L-eigenarrays and the L-eigengenes × M-arrays basis sets, respectively. The vector in the lth column of the matrix $\hat{u}$, $|\alpha_l\rangle \equiv \hat{u}|l\rangle$, lists the genome-scale expression signal of the lth eigenarray. The vector in the lth row of the matrix $\hat{v}^T$, $\langle\gamma_l| \equiv \langle l|\hat{v}^T$, lists the signal of the lth eigengene across the different arrays. The eigengenes and eigenarrays are orthonormal superpositions of the genes and arrays, such that the transformation matrices $\hat{u}$ and $\hat{v}^T$ are both orthogonal,

$$\hat{u}^T\hat{u} = \hat{v}^T\hat{v} = \hat{I}, \tag{4}$$

where $\hat{I}$ is the identity matrix. The signal of each eigengene and eigenarray is, therefore, not only decoupled but also decorrelated from that of all other

eigengenes and eigenarrays, respectively. The eigengenes and eigenarrays are unique up to phase factors of ±1 for a real data matrix $\hat{e}$, such that each eigengene and eigenarray captures both parallel and antiparallel gene and array expression patterns, except in degenerate subspaces, defined by subsets of equal eigenexpression levels. SVD is, therefore, data driven, except in degenerate subspaces.

2.2. SVD Analysis of Cell Cycle RNA Expression Data From Yeast

In this example, SVD is applied to a dataset that tabulates RNA expression levels of 4579 genes in 22 yeast samples, 18 samples of a time course monitoring the cell cycle in an α factor-synchronized culture, and two samples each of yeast strains where the genes *CLN3* and *CLB2*, which encode G_1 and G_2/M cyclins, respectively, are overexpressed or overactivated ***(6)***.

2.2.1. Significant Eigengenes and Corresponding Eigenarrays Correlate With Genome-Scale Effects of Independent Sources of Expression and Their Corresponding Cellular States

Consider the 22 eigengenes of the α factor, *CLB2*, and *CLN3* dataset **(Fig. 3A)**. The first eigengene, which captures about 80% of the overall expression signal **(Fig. 3B)**, and describes sample-invariant expression, is inferred to represent steady-state expression **(Fig. 3C)**. The second and third eigengenes, which capture about 9.5% and 2% of the overall expression signal, respectively, describe initial transient increase and decrease in expression, respectively, superimposed on time-invariant expression during the cell cycle. These eigengenes are inferred to represent the responses to synchronization by the pheromone α factor. The fourth through ninth and 11th eigengenes, which capture together about 5% of the overall expression information, show expression oscillations of two periods during the α factor-synchronized cell cycle, and are inferred to represent cell cycle expression oscillations **(Fig. 3D–F)**.

The corresponding eigenarrays are associated with the corresponding cellular states. An eigenarray is parallel and antiparallel associated with the most likely parallel and antiparallel cellular states, or none thereof, according to the annotations of the two groups of n genes each, with largest and smallest levels of signal, e.g., expression, in this eigenarray among all N genes, respectively. A coherent biological theme might be reflected in the annotations of either one of these two groups of genes. The p-value of a given association by annotation is calculated using combinatorics and assuming hypergeometric probability distribution of the K annotations among the N genes, and of the subset of $k \subseteq K$ annotations among the subset of $n \subseteq N$ genes,

$$P(k;n,N,K) = \binom{N}{n}^{-1} \sum_{l=k}^{n} \binom{K}{l} \binom{N-K}{n-l},$$

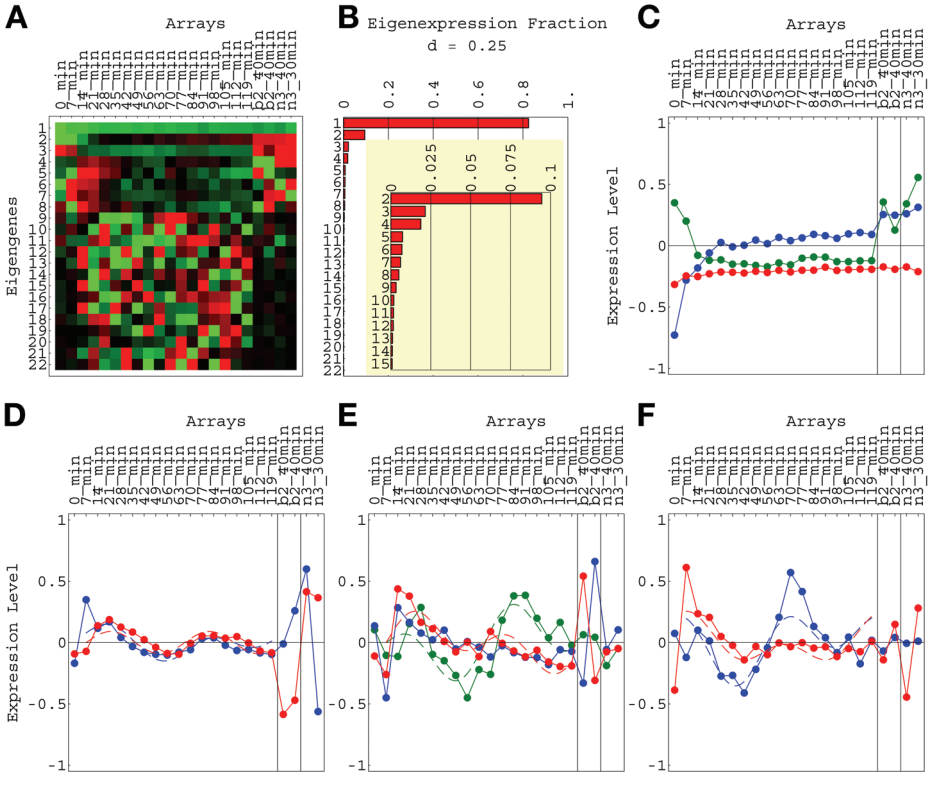


Fig. 3. The eigenegenes of the yeast cell cycle RNA expression dataset. **(A)** Raster display of the expression of 22 eigengenes in 22 arrays. **(B)** Bar chart of the fractions of eigenexpression, showing that the first eigengene captures about 80% of the overall relative expression. **(C)** Line-joined graphs of the expression levels of the first eigenegene (red), which represents the steady expression state, and the second (blue) and third (green) eigengenes, which represent responses to synchronization of the yeast culture by α factor. **(D)** Expression levels of the fourth (red) and seventh (blue) eigengenes, **(E)** the fifth (red), eighth (blue), and 11th (green) eigengenes, and **(F)** the sixth (red) and ninth (blue) eigengenes, all fit dashed graphs of sinusoidal functions of two periods superimposed on sinusoidal functions of one period during the time course.

where $$\binom{N}{n} = N!n!^{-1}(N-n)!^{-1}$$

is the Newton binomial coefficient ***(30)***. The most likely association of an eigenarray with a cellular state is defined as the association that corresponds to the smallest *p*-value.

Following the *p*-values for the distribution of the 364 genes, which were microarray-classified as α factor regulated ***(31)*** and that of the 646 genes, which were traditionally or microarray-classified as cell cycle-regulated ***(6)*** among all 4579 genes and among each of the subsets of 200 genes with the largest and smallest levels of expression, respectively, the second and third eigenarrays are associated with the cellular states of the α factor response program, whereas the fourth through ninth and 11th eigenarrays are associated with the cellular states of the cell cycle program.

2.2.2. Filtering Out of Eigengenes and Eigenarrays Simulates the Experimental Suppression of the Cellular Processes and States That These Eigengenes and Eigenarrays Represent

Any eigengene $\langle\gamma_l|$ and corresponding eigenarray $|\alpha_l\rangle$ can be filtered out, without eliminating genes or arrays from the dataset, by setting their corresponding eigenexpression level in $\hat{e}$ to zero, $\varepsilon_l = 0$, and reconstructing the dataset according to **Eq. 1**, such that $\hat{e} \to \hat{e} - \varepsilon_l|\alpha_l\rangle\langle\gamma_l|$. The α factor, *CLB2*, and *CLN3* dataset is normalized by filtering out the first eigengene, which represents the additive steady-state expression level, the second and third eigengenes, which represent the α factor synchronization response, as well as the 10th and 12th through 22nd eigengenes. After filtering out the first eigengene, the expression pattern of each gene is approximately centered at its time-invariant level. Similarly, the expression of each gene is then approximately normalized by its steady scale of variance ***(16,17)***. The normalized dataset tabulates for each gene an expression pattern that is of an approximately zero arithmetic mean, with a variance which is of an approximately unit geometric mean.

Consider the eigengenes of the normalized α factor, *CLB2*, and *CLN3* dataset **(Fig. 4A)**. The first, second, and third normalized eigengenes, which are of similar significance, capture together about 60% of the overall normalized expression **(Fig. 4B)**. Their time variations fit normalized sine and cosine functions of two periods superimposed on a normalized sine function of one period during the cell cycle **(Fig. 4C)**. Although the first and third normalized eigengenes describe underexpression in both *CLB2*-overactive arrays, and overexpression in both *CLN3*-overactive arrays, the second normalized eigengene describes the antiparallel expression pattern of overexpression in both *CLB2*-overactive arrays and underexpression in both *CLN3*-overactive arrays. These normalized eigengenes are inferred to represent expression oscillations during the cell cycle superimposed on differential expression because of *CLB2* and *CLN3* overactivations. The corresponding eigenarrays are associated by annotation with the corresponding cellular states.

None of the significant eigengenes and eigenarrays of the normalized dataset represents either the steady-state expression or the response to the α factor

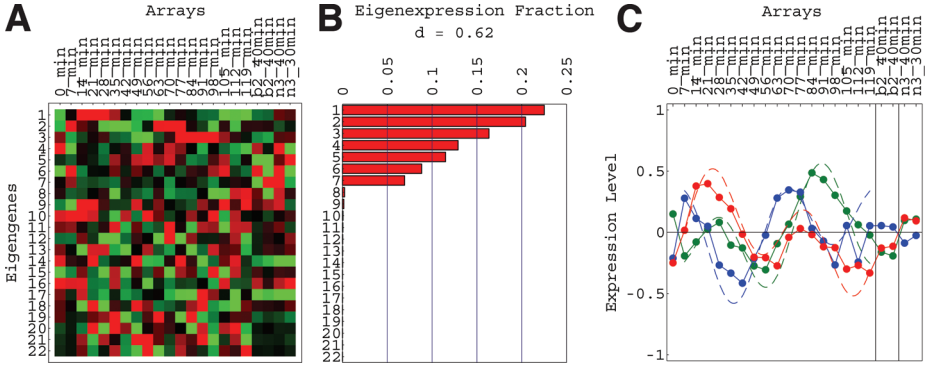


Fig. 4. The eigengenes of the normalized yeast cell cycle RNA expression dataset. **(A)** Raster display. **(B)** Bar chart of the fractions of eigenexpression, showing that the first, second, and third normalized eigengenes capture approximately 20% of the overall normalized expression information each, and span an approximately degenerate subspace. **(C)** Line-joined graphs of the expression levels of the first (red), second (blue), and third (green) normalized eigengenes, fit dashed graphs of two-period sinusoidal functions superimposed on one-period sinusoidal functions during the time course.

synchronization. The normalized dataset simulates an experimental measurement of only the cell cycle program and the differential expression in response to overactivation of *CLB2* and *CLN3*.

2.2.3. Rotation in an Almost Degenerate Subspace Simulates Experimental Decoupling of the Biological Programs the Subspace Spans

The almost degenerate subspaces spanned by the first, second, and third eigengenes and corresponding eigenarrays are approximated with degenerate subspaces, by setting each of the corresponding eigenexpression levels equal, $\varepsilon_1, \varepsilon_2, \varepsilon_3 \to \sqrt{(\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2)/3}$, and reconstructing the dataset according to **Eq. 1**. With this approximation, the three eigengenes and corresponding eigenarrays can be rotated, such that the same expression subspaces that are spanned by these eigenegenes, and eigenarrays will be spanned by three orthogonal superpositions of these eigengenes and eigenarrays, i.e., by three rotated eigengenes and eigenarrays. Requiring two of these three rotated eigengenes to describe equal expression in the *CLB2*-overactive samples as in the *CLN3*-overactive samples, so that only the one remaining rotated eigengene captures the differential expression between these two sets of arrays, gives unique angles of rotations in the three-dimensional subspaces of eigengenes and eigenarrays, and therefore also unique rotated eigengenes and eigenarrays.

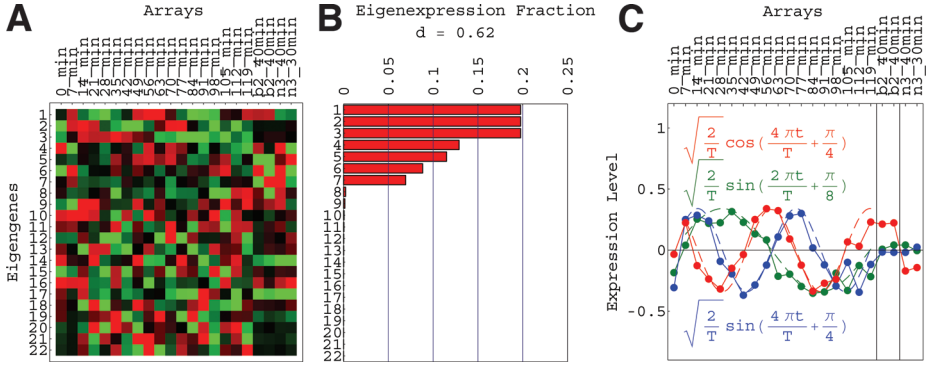


Fig. 5. The rotated eigengenes of the normalized yeast cell cycle RNA expression dataset. **(A)** Raster display. **(B)** Bar chart of the fractions of eigenexpression, showing that the first, second, and third rotated eigengenes span an exactly degenerate subspace. **(C)** Line-joined graphs of the expression levels of the first (red) and second (blue) rotated eigengenes fit normalized sine and cosine functions of two periods, and the third rotated eigengene (green) fits a normalized sine of one period during the time course.

Consider the eigengenes of the normalized and rotated α factor, *CLB2*, and *CLN3* dataset **(Fig. 5A)**, where the first, second, and third fractions of eigenexpression are approximated to be equal **(Fig. 5B)**. The time variations of the first and second rotated eigengenes fit normalized sine and cosine functions of two periods during the cell cycle **(Fig. 5C)**. The time variation of the third rotated eigengene fits a normalized sine function of one period during the cell cycle, suggesting differences in expression between the two successive cell cycle periods, which may be due to dephasing of the initially synchronized yeast culture. Although the second and third rotated eigengenes describe steady-state expression in the *CLB2*- and *CLN3*-overactive arrays, the first rotated eigengene describes underexpression in the *CLB2*-overactive arrays and overexpression in the *CLN3*-overactive arrays. The first rotated eigengene, therefore, is inferred to represent cell cycle expression oscillations that are *CLB2*- and *CLN3*-dependent, whereas the second rotated eigengene is inferred to represent cell cycle expression oscillations that are *CLB2*- and *CLN3*-independent. The third rotated eigengene is inferred to represent variations in the cell cycle expression from the first period to the second, which also appear to be *CLB2*- and *CLN3*-independent. The first, second, and third rotated eigenarrays are associated by annotation with the corresponding cellular states.

The rotation of the data, therefore, simulates decoupling of the differential expression owing to *CLB2* and *CLN3* overactivation from at least one of the cell

cycle stages. It also simulates decoupling of the variation between the first and the second cell cycle periods from the cell cycle stages and from the *CLB2* and *CLN3* overactivation.

2.2.4. Classification of the Normalized Yeast Data According to the Rotated Eigengenes and Eigenarrays Gives a Global Picture of the Dynamics of Cell Cycle Expression

Consider the normalized expression of the 22 α factor, *CLB2*, and *CLN3* arrays in the subspace spanned by the first and second rotated eigenarrays, which represents approximately all cell cycle cellular states **(Fig. 6A)**. Sorting the arrays according to their correlations with the second rotated eigenarray along the *y*-axis, $\langle \alpha_2 | a_m \rangle / \sqrt{\langle a_m | a_m \rangle}$, vs that with the first rotated eigenarray along the *x*-axis, $\langle \alpha_1 | a_m \rangle / \sqrt{\langle a_m | a_m \rangle}$, reveals that all except for five arrays have at least 25% of their normalized expression in this subspace. This sorting gives an array order that is similar to that of the cell cycle time-points measured by the arrays, an order that describes the progression of the cell cycle from the M/G1 stage through G_1, S, S/G_2, and G_2/M and back to M/G_1 twice. The first rotated eigenarray is correlated with samples that probe the cellular state of cell cycle transition from G_2/M to M/G_1, which is simulated experimentally by *CLB2* overactivation. This eigenarray is also anticorrelated with the cellular state of transition from G_1 to S, which is simulated by *CLN3* overactivation. Similarly, the second rotated eigenarray is correlated with the transition from M/G_1 to G_1, and anticorrelated with S/G_2, both of which appear to be *CLB2* and *CLN3* independent.

Consider also the normalized expression of the 646 yeast genes in this dataset that were traditionally or microarray-classified as cell cycle regulated **(Fig. 6B)**. Sorting the genes according to their correlations with the first and second rotated eigengenes reveals that 551 of these genes have at least 25% of their normalized expression in this subspace. This sorting gives a classification of these genes into the five cell cycle stages, which is in good agreement with both the traditional and microarray classifications. The first rotated eigengene is correlated with the observed expression pattern of *CLB2* and its targets, genes for which expression peaks at the transition from G_2/M to M/G_1. This eigengene is also anticorrelated with the observed expression of *CLN3* and its targets, genes for which expression peaks at the transition from G_1 to S. The second rotated eigengene is correlated with the cell cycle oscillations, which peak at the transition from M/G_1 to G_1 and anticorrelated with these which peak at S/G_2, both of which appear to be independent of the genome-scale effects of *CLB2* and *CLN3*.

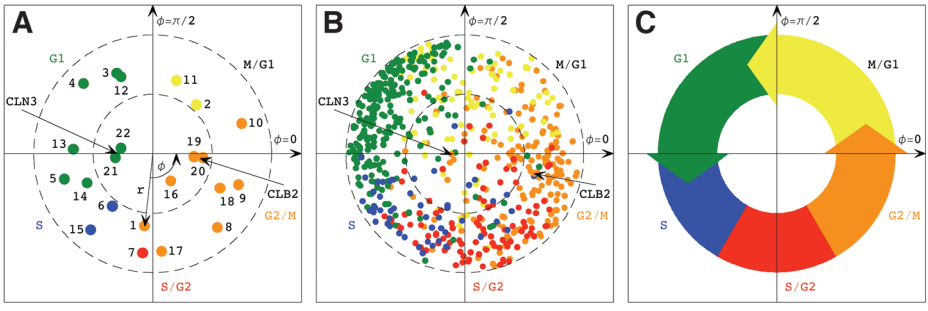


Fig. 6. The normalized yeast RNA expression in the SVD cell cycle subspace. **(A)** Correlations of the normalized expression of each of the 22 arrays with the first and second rotated eigenarrays along the *x*- and *y*-axes, color-coded according to the classification of the arrays into the five cell cycle stages: M/G_1 (yellow), G_1 (green), S (blue), S/G_2 (red), and G_2/M (orange). The dashed unit and half-unit circles out-line 100% and 25% of overall normalized array expression in this subspace. **(B)** Correlations of the normalized expression of each of the 646 cell cycle-regulated genes with the first and second rotated eigengenes along the *x*- and *y*-axes, color-coded according to either the traditional or microarray classifications. **(C)** The SVD picture of the yeast cell cycle.

Classification of the yeast arrays and genes in the subspaces spanned by these two rotated eigenarrays and corresponding eigengenes gives a picture that resembles the traditional understanding of yeast cell cycle regulation ***(32)***: G_1 cyclins, such as *CLN3*, and G_2/M cyclins, such as *CLB2*, drive the cell cycle past either one of two antipodal checkpoints, from G_1 to S and from G_2/M to M/G_1, respectively **(Fig. 6C)**.

2.3. SVD Model for Genome-Wide RNA Expression During the Cell Cycle Parallels the Analog Harmonic Oscillator

With all 4579 genes sorted, the normalized cell cycle expression approximately fits a traveling wave, varying sinusoidally across both genes and arrays **(Fig. 7A)**. The normalized expression in the *CLB2*- and *CLN3*-overactive arrays approximately fits standing waves, constant across the arrays and varying sinusoidally across the genes only, which appear anticorrelated and correlated with the first eigenarray, respectively. The gene variations of the first and second rotated eigenarrays fit normalized cosine and sine functions of one period across all genes, respectively **(Fig. 7B,C)**. In this picture, all 4579 genes, about three-quarters of the yeast genome, appear to exhibit periodic expression during the cell cycle. This picture is in agreement with the recent observation by Klevecz et al. ***(33***; *see also* Li and Klevecz, **ref. *34*)** that DNA replication is gated by genome-wide RNA expression oscillations, which suggests that the whole yeast genome might exhibit expression oscillations during the cell cycle.

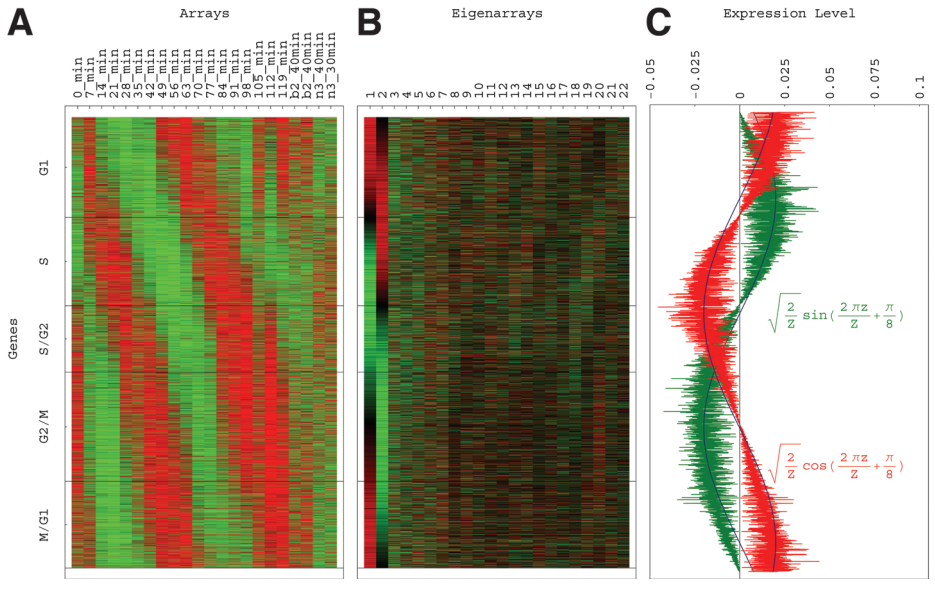


Fig. 7. The sorted and normalized yeast cell cycle RNA expression dataset and its sorted and rotated eigenarrays. **(A)** Raster display of the normalized expression of the 4579 genes across the 22 arrays. The genes are sorted by relative correlation of their normalized expression patterns with the first and second rotated eigengenes. This raster display shows a traveling wave of expression during the cell cycle and standing waves of expression in the *CLB2*- and *CLN3*-overactive arrays. **(B)** Raster display of the rotated eigenarrays, where the expression patterns of the first and second eigenarrays, which correspond to the first and second eigengenes, respectively, display the sorting. **(C)** Line-joined graphs of the first (red) and second (green) rotated eigenarrays, fit normalized cosine and sine functions of one period across all genes.

It is still an open question whether all yeast genes or only a subset of the yeast genes, and if so, which subset, show periodic expression during the cell cycle.

This SVD model describes, to first order, the RNA expression of most of the yeast genome during the cell cycle program as being driven by the activities of two periodically oscillating cellular elements or modules, which are orthogonal, i.e., $\pi/2$ out of phase relative to one another. The underlying genetic network or circuit suggested by this model might be parallel in its design to the analog harmonic oscillator. This well-known oscillator design principle is at the foundations of numerous physical oscillators, including (1) the mechanical pendulum, the position and momentum of which oscillate periodically in time with a phase difference of $\pi/2$; (2) the electronic LC circuit, where the charge on the capacitor and the current flowing through the inductor oscillate periodically in time with a phase difference of $\pi/2$; and (3) the chemical Lotka-Volterra irreversible autocatalytic reaction model, where, far from thermodynamic equilibirum, the

concentrations of two intermediate reactants exhibit periodic oscillations in time that are $\pi/2$ out of phase relative to one another ***(35–37)***.

3. GSVD for Comparative Modeling of DNA Microarray Datasets

This section reviews the GSVD comparative model for DNA microarray datasets ***(19)***. GSVD is a comparative BSS algorithm that simultaneously decomposes two measured signals, i.e., the measured gene and array patterns of, e.g., RNA expression in two organisms, into mathematically decoupled "genelets" and two sets of "arraylets." The correspondence between these mathematical patterns uncovered in the measured signals and the similar and dissimilar among the biological programs that compose each of the two signals is illustrated with a comparative analysis of genome-scale RNA expression data from yeast ***(6)*** and human ***(7)*** during their cell cycle programs. One common picture of RNA expression oscillations during both the yeast and human cell cycles emerges from this analysis, which suggests an underlying eukaryotic genetic network or circuit that parallels the digital ring oscillator.

Comparisons of DNA sequence of entire genomes already give new insights into evolutionary, biochemical, and genetic pathways. Recent studies showed that the addition of RNA expression data to DNA sequence comparisons improves functional gene annotation and might expand the understanding of how gene expression and diversity evolved. For example, Stuart et al. ***(38)*** and independently also Bergmann, Ihmels, and Barkai ***(39)*** identified pairs of genes for which RNA coexpression is conserved, in addition to their DNA sequences, across several organisms. The evolutionary conservation of the coexpression of these gene pairs confers a selective advantage to the functional relations of these genes. The GSVD comparative model is not limited to genes of conserved DNA sequences, and as such it elucidates universality as well as specialization of molecular biological mechanisms that are truly on genomic scales. For example, the GSVD comparative model might be used to identify genes of common function across different organisms independently of the DNA sequence similarity among these genes, and therefore also to study nonorthologous gene displacement ***(40)***.

3.1. Mathematical Framework of GSVD

Let the matrix $\hat{e}_1$ of size N_1-genes $\times$ M_1-arrays tabulate the genome-scale signal, e.g., RNA expression levels, measured in a set of M_1 samples using M_1 DNA microarrays. As before, the mth column vector in the matrix $\hat{e}_1$, $|a_{1,m}\rangle$, lists the expression signal measured in the mth sample by the mth array across all N_1 genes simultaneously. The nth row vector in the matrix $\hat{e}_1$, $\langle g_{1,n}|$, lists the signal measured for the nth gene across the different arrays, which correspond to the different samples. Let the matrix $\hat{e}_2$ of size N_2-genes $\times$ M_2-arrays tabulate the genome-scale signal, e.g., RNA expression levels, measured in a set of M_2 samples under M_2 experimental conditions that correspond one-to-one to the M_1 conditions

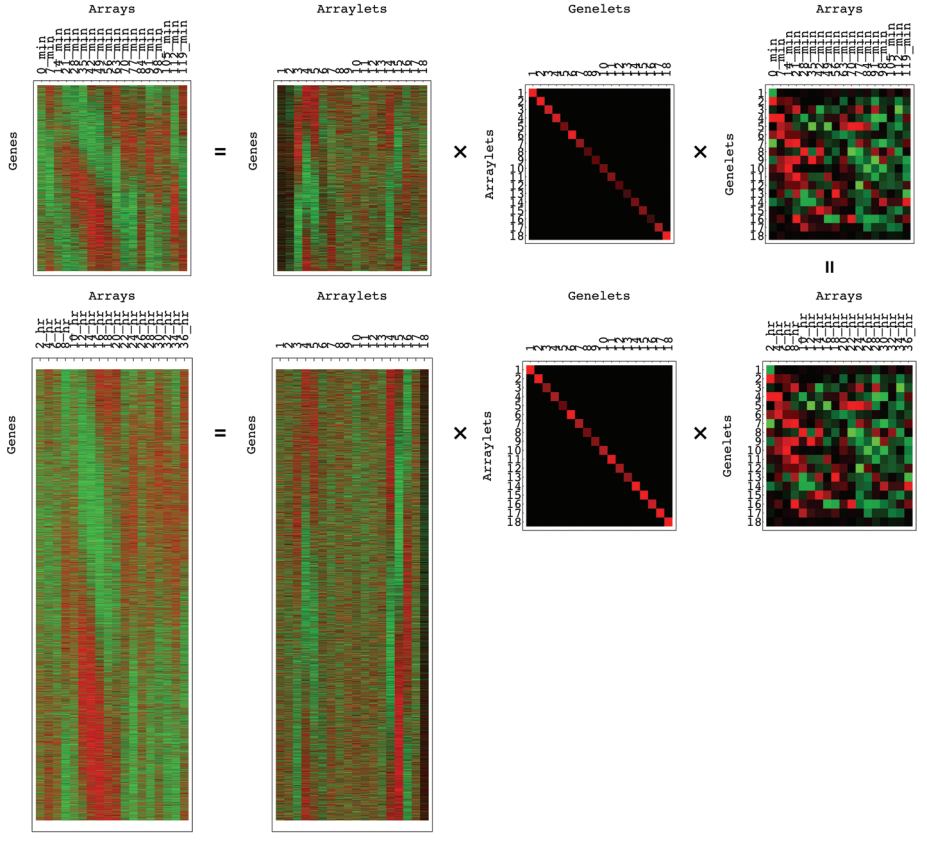


Fig. 8. Raster display of the GSVD of the yeast and human cell cycle RNA expression datasets, with overexpression (red), no change in expression (black), and underexpression (green) centered at the gene- and array-invariant expression of the 4523 yeast and 12,056 human genes. GSVD is a linear transformation of the yeast and human data from the 4523-yeast and 12,056-human genes × 18-arrays spaces to the reduced diagonalized 18-arraylets × 18-genelets spaces, which are spanned by the 4523- and 12,056-genes × 18-arraylets bases, respectively, and by the 18-genelets × 18-arrays shared basis.

underlying $\hat{e}_1$, such that $M_2 = M_1 \equiv M < \max\{N_1, N_2\}$. This one-to-one correspondence between the two sets of conditions is at the foundation of the GSVD comparative analysis of the two datasets, and should be mapped out carefully.

GSVD is a simultaneous linear transformation of the two expression datasets $\hat{e}_1$ and $\hat{e}_2$ from the two N_1-genes × M-arrays and N_2-genes × M-arrays spaces to the two reduced M-arraylets × M-genelets spaces **(Fig. 8)**,

$$\begin{aligned} \hat{e}_1 &= \hat{u}_1 \hat{\varepsilon}_1 \hat{x}^{-1}, \\ \hat{e}_2 &= \hat{u}_2 \hat{\varepsilon}_2 \hat{x}^{-1}. \end{aligned} \tag{5}$$

In these spaces the data are represented by the diagonal nonnegative matrices ε_1 and ε_2. Their diagonality means that each genelet is decoupled of all other genelets in both datasets simultaneously, such that each genelet is expressed only in the two corresponding arraylets, each of which is associated with one of the two datasets.

The antisymmetric "angular distances" between the datasets $\{\theta_m\}$ are calculated from the "generalized eigenexpression levels" $\{\varepsilon_{1,l}\}$ and $\{\varepsilon_{2,l}\}$, which are listed in the diagonals of ε_1 and ε_2,

$$-\pi/4 \le \theta_m = \arctan(\varepsilon_{1,m}/\varepsilon_{2,m}) - \pi/4 \le \pi/4 . \tag{6}$$

These angular distances indicate the relative significance of each genelet, i.e., its significance in the first dataset relative to that in the second dataset, in terms of the ratio of expression information captured by this genelet in the first dataset to that in the second. An angular distance of 0 indicates a genelet of equal significance in both datasets, with $\varepsilon_{1,m} = \varepsilon_{2,m}$. An angular distance of $\pm\pi/4$ indicates no significance in the second dataset relative to the first, with $\varepsilon_{1,m} \gg \varepsilon_{2,m}$, or in the first dataset relative to the second, with $\varepsilon_{1,m} \ll \varepsilon_{2,m}$, respectively.

The transformation matrix $\hat{x}^{-1}$ defines the M-genelets $\times$ M-arrays basis set, which is shared by both datasets. The transformation matrices $\hat{u}_1$ and $\hat{u}_2$ define the N_1-genes $\times$ M-arraylets and N_2-genes $\times$ M-arraylets basis sets, that correspond to the first and second datasets, respectively. The mth row vector in $\hat{x}^{-1}$, $\langle\gamma_m| \equiv \langle m|\hat{x}^{-1}$, lists the expression signal of the mth genelet across the different arrays in both datasets simultaneously. The mth column vector in $\hat{u}_1$ or $\hat{u}_2$, $|\alpha_{1,m}\rangle \equiv \hat{u}_1|m\rangle$ or $|\alpha_{2,m}\rangle \equiv \hat{u}_2|m\rangle$, lists the genome-scale signal of the mth arraylet of either the first or the second dataset, respectively. The genelets are normalized, but not necessarily orthogonal, superpositions of the genes of the first dataset and, at the same time, also the second dataset. The arraylets of the first or the second datasets are orthonormal superpositions of the arrays of the first and second datasets, respectively. In general, $\hat{x}^{-1}$ is nonorthogonal, while $\hat{u}_1$ and $\hat{u}_2$ are both orthogonal,

$$\hat{x}^{-1}\hat{x} \ne \hat{u}_1^T\hat{u}_1 = \hat{u}_2^T\hat{u}_2 = \hat{I}, \tag{7}$$

where $\hat{I}$ is the identity matrix. The expression of each arraylet of either dataset is, therefore, not only decoupled but also decorrelated from that of all other arraylets of this dataset. The genelets and arraylets are unique up to phase factors of ± 1 for real data matrices $\hat{e}_1$ and $\hat{e}_2$, such that each genelet and arraylet capture both parallel and antiparallel gene and array expression patterns, except in degenerate subspaces, defined by subsets of equal angular distances. GSVD is, therefore, data driven, except in degenerate subspaces.

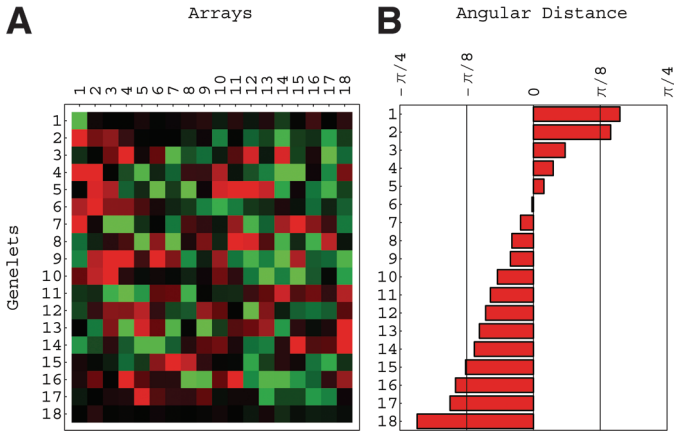


Fig. 9. The genelets of the yeast and human cell cycles RNA expression datasets. **(A)** Raster display of the expression of 18 genelets in the 18 yeast and 18 human arrays, simultaneously, centered at their array-invariant levels. **(B)** Bar chart of the angular distances, showing the first and second genelets highly significant in the yeast data relative to the human data, the third through the sixth and the 14th through the 16th almost equally significant in both datasets, and the 17th and 18th genelets highly significant in the human data relative to the yeast data. All other genelets are neither significant in the yeast data nor in the human data ***(19)***.

3.2. GSVD Comparative Analysis of Yeast and Human Cell Cycle RNA Expression Data

In this example, GSVD is applied to two datasets, which tabulate RNA expression of 4523 yeast genes and 12,056 human genes in 18 samples each of time courses of α factor-synchronized yeast culture ***(6)*** and double thymidine block-synchronized HeLa cell line culture ***(7)***, respectively. The yeast and human time courses span more than two and less than two and a half periods in the yeast and human cell cycles, respectively. Both yeast and human time courses are sampled at equal time intervals.

3.2.1. Common Genelets and Corresponding Arraylets Span the Common Yeast and Human Cell Cycle Subspace

Consider the 18 genelets of the yeast and human cell cycle datasets **(Fig. 9A)**. Six genelets are almost equally significant in the yeast and human datasets **(Fig. 9B)**: The third, fourth, and fifth genelets are slightly more significant in the yeast dataset than in the human dataset, with $0 < \theta_3 < \theta_4 < \theta_5 < \pi/16$. The 14th, 15th and 16th genelets are slightly more significant in the human dataset, with $-\pi/6 < \theta_{14} < \theta_{15} < \theta_{16} < 0$. The time-, i.e., array variations of the third, fourth

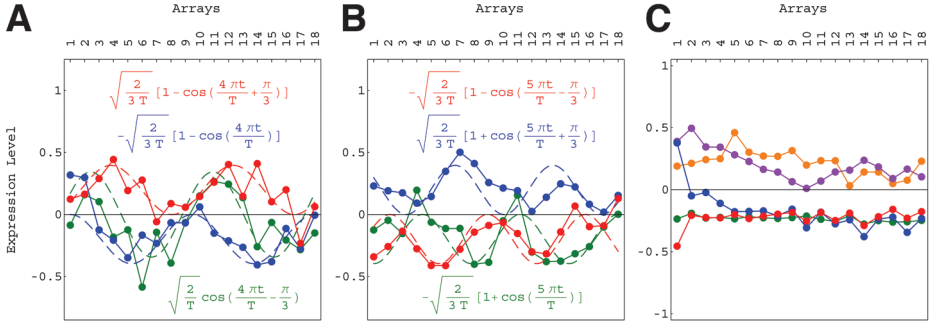


Fig. 10. Line-joined graphs of the expression levels of the significant genelets. **(A)** The third (red), fourth (blue), and fifth (green) genelets, which are associated with the common yeast and human cell cycle gene expression oscillations, fit dashed graphs of normalized cosines of two periods and initial phases of π/3 (red), 0 (blue) and –π/3 (green), respectively. **(B)** The 14th (red), 15th (blue) and 16th (green) genelets, which are also associated with cell cycle gene expression oscillations, fit dashed graphs of normalized cosines of two and a half periods and initial phases of –π/3 (red), π/3 (blue) and 0 (green), respectively. **(C)** The first (red) and second (blue) genelets are associated with the exclusive yeast response to the pheromone α factor, the 17th (orange) and 18th (green) are associated with the exclusive human stress response, and the sixth (violet) is associated with both the yeast and human transitions from synchronization responses into the cell cycle.

and fifth genelets fit normalized cosine functions of two periods and initial phases of π/3, 0 and –π/3, respectively, superimposed on time-invariant expression **(Fig. 10A)**. The 14th, 15th and 16th genelets fit normalized cosines of two and a half periods and initial phases of –π/3, π/3, and 0, respectively **(Fig. 10B)**. The time variations of the six common genelets suggest that they span the cell cycle subspace, which is common to both the yeast and human genomes, and is manifested in both datasets.

The corresponding six yeast and six human arraylets are associated by annotation with the corresponding yeast and human cell cycle cellular states, following the *p*-values for the distribution of the 604 yeast genes and 750 human genes, that were microarray-classified, and the 77 yeast genes and 73 human genes that were traditionally classified as cell cycle regulated, among all 4523 yeast and 12,056 human genes and among each of the subsets of 100 genes with largest and smallest levels of expression in each of the arraylets. The associations of the yeast and human arraylets are in agreement with the expression patterns of the genelets, taking into account the initial synchronization of the yeast culture in the cell cycle stage M/G_1 and that of the human culture in S. For example, the expression pattern of the fourth genelet is of 0 initial phase, suggesting that this genelet is correlated with the yeast cell cycle expression oscillations that peak at the stage M/G_1 and the human cell cycle expression

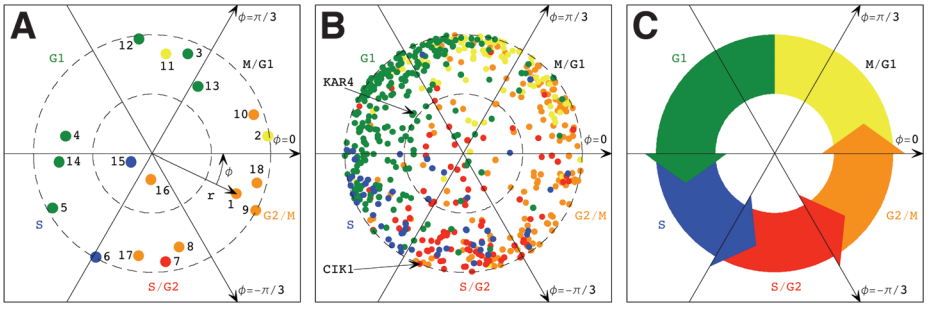


Fig. 11. Reconstructed yeast RNA expression in the GSVD common cell cycle subspace. **(A)** Projections of the expression of each of the 18 arrays, after reconstruction in the six-dimensional GSVD cell cycle subspace, onto the two-dimensional subspace that least-squares approximates it. The arrays are color coded according to their classification into the five cell cycle stages: M/G_1 (yellow), G_1 (green), S (blue), S/G_2 (red), and G_2/M (orange). The dashed unit and half-unit circles outline 100% and 50% of added up, rather than cancelled out, contributions of the six arraylets to the overall projected expression. The arrows describe the projections of the $-\pi/3$-, 0-, and $\pi/3$-phase arraylets. **(B)** Projections of the expression of each of the 612 cell cycle-regulated genes, reconstructed in the six-dimensional GSVD subspace, onto the two-dimensional subspace that approximates it. The genes are color coded according to either the traditional or microarray classifications. The expression patterns of *KAR4* and *CIK1* are anticorrelated. **(C)** The GSVD picture of the yeast cell cycle.

oscillations that peak at S. Following the traditional classifications, the corresponding yeast arraylet, i.e., the fourth yeast arraylet, is associated in parallel with the yeast cell cycle stage M/G_1, while the fourth human arraylet is associated in parallel with the human cell cycle stage S.

3.2.2. Simultaneous Reconstruction and Classification of the Yeast and Human Data in the Common Subspace Outlines the Biological Similarity in the Regulation of the Yeast and Human Cell Cycle Programs

The six-dimensional genelets subspace that represents the common yeast and human cell cycle expression oscillations is least squares-approximated with a two-dimensional subspace that is spanned by two orthonormal vectors $\langle x|$ and $\langle y|$. Projecting the expression of the 18 yeast arrays from the corresponding six-dimensional yeast arraylets subspace onto the corresponding approximate two-dimensional subspace **(Fig. 11A)** reveals that 50% or more of the contributions of the six arraylets add up, rather than cancel out, in the overall expression of 16 of the arrays. Sorting the arrays in this subspace gives an array order similar to that of the cell cycle time-points measured by the arrays. This order of the arrays describes the yeast cell cycle progression from the M/G_1 stage through G_1, S,

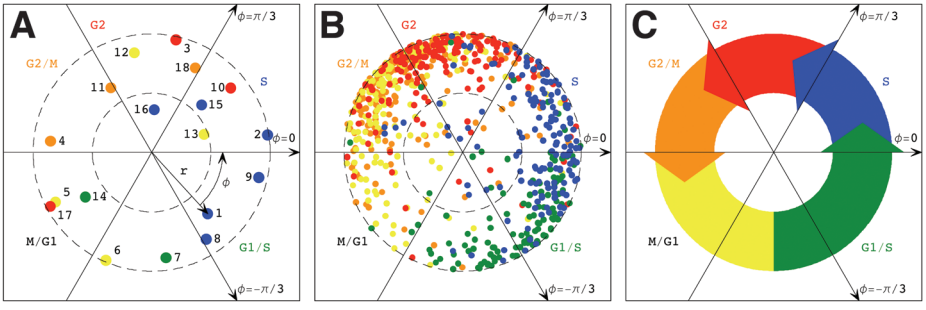


Fig. 12. Reconstructed human RNA expression in the GSVD common cell cycle subspace. **(A)** Projections of the expression of each of the 18 arrays, after reconstruction in the six-dimensional GSVD cell cycle subspace, onto the two-dimensional subspace that approximates it. The arrays are color coded according to their classification into the five cell cycle stages. The dashed unit and half-unit circles outline 100% and 50% of added up, rather than cancelled out, contributions of the six arraylets to the overall projected expression. The arrows describe the projections of –π/3-, 0- and π/3-phase arraylets. **(B)** Projections of the expression of each of the 774 cell cycle-regulated genes, reconstructed in the six-dimensional GSVD subspace, onto the two-dimensional subspace that approximates it. The genes are color coded according to either the traditional or microarray classifications. **(C)** The GSVD picture of the human cell cycle.

S/G_2, G_2/M back to M/G_1 twice. Projecting the expression of the 18 human arrays from the six-dimensional human arraylets subspace onto the approximate two-dimensional subspace reveals that 50% or more of the contributions of the six arraylets add up in the expression of 16 of the arrays **(Fig. 12A)**. Sorting the arrays describes the human cell cycle progression from S through G_2, G_2/M, M/G_1, G_1/S back to S two and a half times. Note that, the fourth and 16th yeast arraylets, which correspond to the two 0-phase genelets, correlate with the cell cycle transition from G_2/M to M/G_1, in which the yeast culture is synchronized initially, and anticorrelate with that from G_1 to S. Consistently, the fourth and 16th human arraylets anticorrelate with the transition from G_2/M to M/G_1, and correlate with that from G_1 to S, in which the human culture is synchronized initially.

Projecting the expression of the yeast and human genes from the six-dimensional genelets subspace onto the two-dimensional subspace that least squares-approximates it reveals that 50% or more of the contributions of the six genelets add up in the overall expression of 552 of the 612 yeast and 731 of the 774 human genes that were traditionally or microarray-classified as cell cycle-regulated (**Figs. 11B** and **12B**). These genes include, for example, 14 of 16 human histones, which were not microarray-classified as cell cycle-regulated based on their overall expression ***(19)***. Simultaneous classification of the yeast and human genes into the five cell cycle stages describes the progression of yeast

and human cell cycles along the yeast and human genes, respectively, and is in good agreement with both yeast and human microarray and traditional classifications. Note that, the two 0-phase genelets, the fourth and 16th genelets, correlate with cell cycle expression oscillations, which peak at the initial stages of synchronization of both yeast and human genes.

Simultaneous reconstruction and classification of the yeast and human arrays and genes in the subspaces spanned by the six yeast and six human arraylets, and six shared genelets, respectively, gives a picture that resembles the traditional understanding of the biological similarity in the regulation of the yeast and human, and perhaps all eukaryotic, cell cycles ***(32)*** of two antipodal checkpoints, at the transition from G_1 to S and at that from G_2/M to M/G_1, that are regulated independently of other cell cycle events **(Figs. 11C** and **12C)**.

3.2.3. Exclusive Genelets and Corresponding Arraylets Span the Exclusive Yeast and Human Synchronization Responses Subspaces

The first and second genelets, which capture most of the expression information in the yeast dataset, yet very little of the expression information in the human dataset, with $\theta_1,\theta_2 > \pi/7$ **(Fig. 9B)**, describe initial transient increase and decrease in expression, respectively **(Fig. 10C)**. A theme of yeast response to pheromone synchronization emerges from the annotations of the genes with the largest and smallest levels of expression in the first and second yeast arraylets. The sixth genelet, equally significant in both datasets, with $\theta_6 \sim 0$, describes an initial transient increase in expression superimposed on cosinusidial variation. A theme of transition from the response to the pheromone α factor into cell cycle progression emerges from the annotations of the yeast genes with the largest and smallest expression levels in the sixth yeast arraylet. These three genelets and corresponding three yeast arraylets are associated with the pheromone response program, which is exclusive to the yeast genome. Classification of the yeast genes and arrays into stages in the pheromone response in the subspaces spanned by these genelets and arraylets, respectively **(Fig. 13)**, is in good agreement with the traditional understanding of this program ***(41)***.

The 17th and 18th genelets are insignificant in the yeast dataset relative to that of the human, with $\theta_{17},\theta_{18} < -\pi/6$. A theme of human synchronization stress response emerges from the annotations of the genes with the largest and smallest expression levels in the 17th and 18th genelets. Also, from the annotations of the human genes with the largest and smallest expression levels in the sixth human arraylet emerges a theme of transition from stress response into cell cycle progression. These three genelets and corresponding three human arraylets are associated with this human exclusive stress response. Classification of the human genes and arrays into stress response

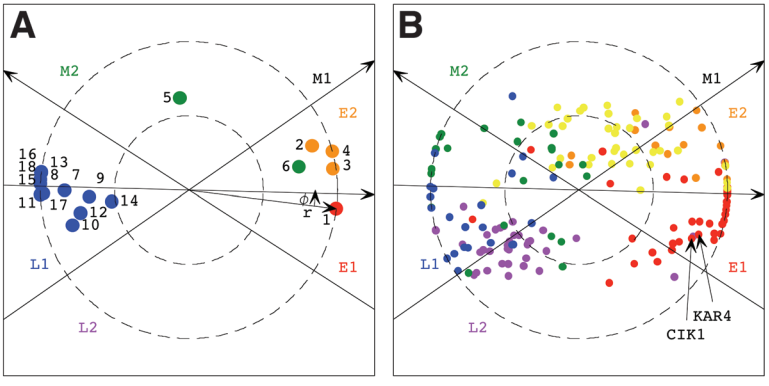


Fig. 13. Reconstructed yeast RNA expression in the GSVD yeast exclusive synchronization response subspace. **(A)** Projections of the expression of each of the 18 arrays, reconstructed in the three-dimensional GSVD synchronization response subspace, onto the two-dimensional subspace that least-squares approximates it. The arrays are color coded according to their classification into six stages in this response to synchronization program, which outlines the response to the pheromone α factor and the transition into cell cycle progression: early E_1 (red) and E_2 (orange), middle M_1 (yellow) and M_2 (green), and late L_1 (blue) and L_2 (violet). The dashed unit and half-unit circles outline 100% and 50% of added up, rather than cancelled out, contributions of the three arraylets to the overall projected expression. The arrows describe the projections of the three arraylets. **(B)** Projections of the expressions of 172 genes, reconstructed in the three-dimensional GSVD subspace, onto the two-dimensional subspace that approximates it. The genes are color coded according to the traditional understanding of the α factor synchronization response program. Genes that peak in E_1 are known to be involved in α factor response, mating, adaptation-to-mating signal, and cell cycle arrest; E_2 – filamentous and pseudohyphal growths and cell polarity; M_1 – ATP synthesis; M_2 – chromatin modeling; L_1 – chromatin binding and architecture; and L_2 – phosphate and iron transport. The expression patterns of *KAR4* and *CIK1* are correlated.

stages in the subspaces spanned by these genelets and arraylets, respectively ***(19)***, is in agreement with the current, somewhat limited, understanding of this program ***(7)***.

3.2.4. Data Reconstruction and Classification in the Common and Exclusive Subspaces Simulate Observation of Differential Expression in the Cell Cycle and Synchronization Response Programs

According to their expression in the yeast exclusive pheromone response subspace, the RNA expression patterns of the yeast genes *KAR4* and *CIK1* are correlated: The expression of both genes peaks early in the time course together

with the expression of other genes known to be involved in the response to the α factor **(Fig. 13B)**. In the common cell cycle subspace *KAR4* and *CIK1* are anti-correlated: *KAR4* peaks at the G_1 cell cycle stage, whereas *CIK1* peaks almost half a cell cycle period later (and also earlier) at S/G_2 **(Fig. 13B)**. This difference in the relation of the expression patterns of *CIK1* and *KAR4* in the response to pheromone program as compared with that of the cell cycle is in agreement with the experimental observation of Kurihara et al. ***(42)*** that induction of *CIK1* depends on that of *KAR4* during mating, which is mediated by the α factor pheromone, and is independent of *KAR4* during the mitotic cell cycle.

In the human exclusive stress response subspace, most human histones reach their expression minima early. In the common cell cycle subspace, most histones peak early, together with other genes known to peak in the cell cycle stage S. This differential expression of most histones may explain why these histones do not appear to be cell cycle regulated based on their overall expression ***(7)***: The superposition of the expression of the histones during the cell cycle and that in response to the synchronization leads to an overall steady-state expression early in the time course ***(19)***.

GSVD uncovers the program-dependent variation in the expression patterns of the human histones, as well as the program-dependent variation in the relations between the expression patterns of the yeast genes *KAR4* and *CIK1*.

3.3.1. GSVD Comparative Model for Genome-Scale RNA Expression During the Yeast and Human Cell Cycles Parallels the Digital Ring Oscillator

With all 4523 yeast and 12,056 human genes sorted according to their phases in the GSVD common cell cycle subspace, the reconstructed yeast and human expressions approximately fit traveling waves of one period cosinusoidal variation across the genes, and of two or two and a half periods across the arrays, respectively **(Fig. 14A)**. The gene variations of the six yeast and six human arraylets approximately fit one period cosines of $\pi/3$, 0, and $-\pi/3$ initial phases, such that the initial phase of each arraylet is similar to that of its corresponding genelet **(Fig. 14B,C)**. In this picture, all 4523 yeast genes, about three-quarters of the yeast genome, as well as all 12,056 human genes, about two-thirds of the human genome according to current estimates ***(35)***, appear to exhibit periodic expression during the cell cycle.

This GSVD model describes, to first order, the RNA expression of most of the yeast and human genomes during their common cell cycle programs as being driven by the activities of three periodically oscillating cellular elements or modules, which are $\pi/3$ out of phase relative to one another. The underlying eukaryotic genetic network or circuit suggested by this model might be parallel in its design to the digital three-inverter ring oscillator. Elowitz and Leibler ***(44)***

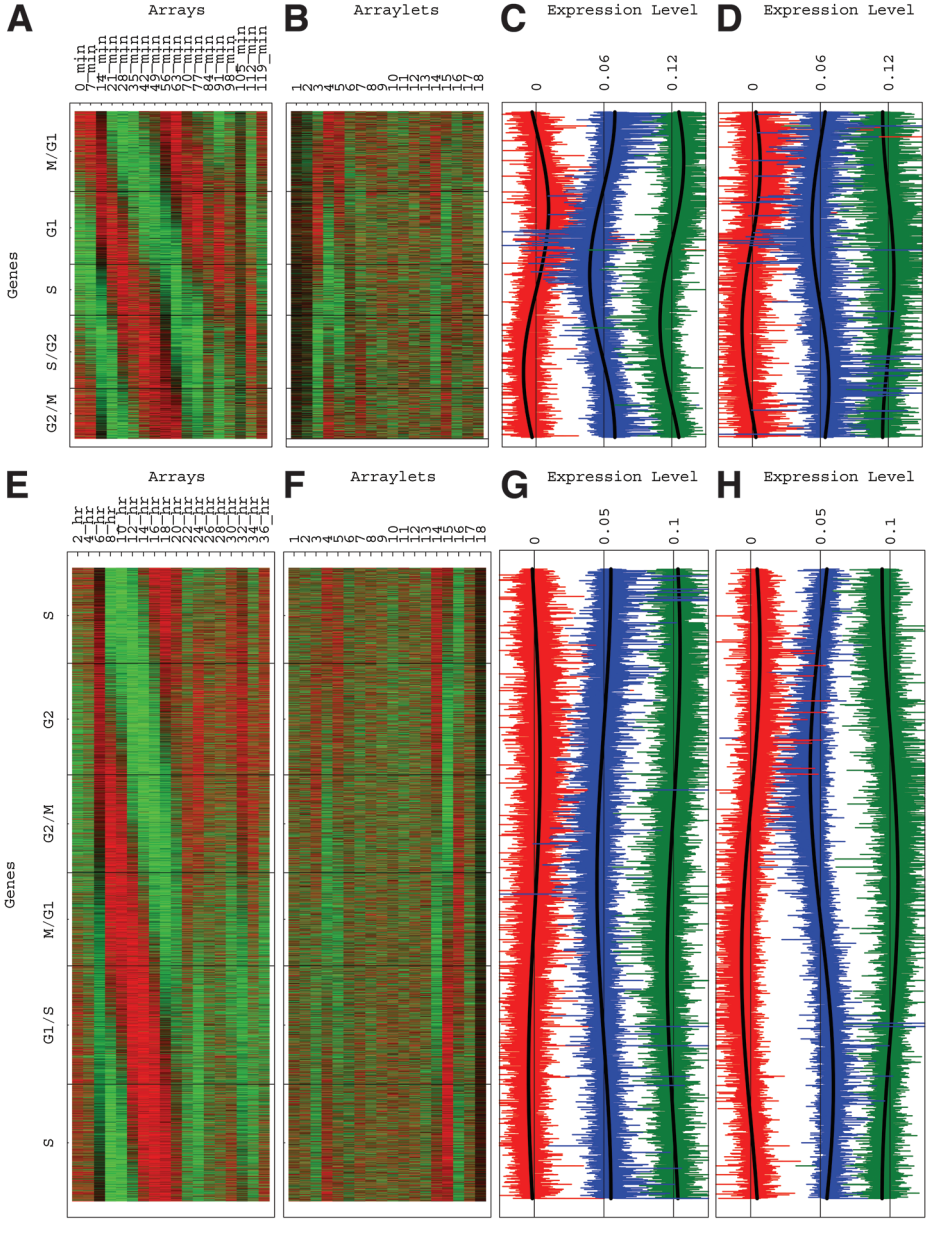


Fig. 14. Yeast and human cell cycles' RNA expression, reconstructed in the six-dimensional GSVD common subspace, with genes sorted according to their phases in the two-dimensional subspace that approximates it. **(A)** Yeast expression of the sorted 4523 genes in the 18 arrays, centered at their gene- and array-invariant levels, showing a traveling wave of expression. **(B)** Yeast expression of thesorted 4523 genes

recently demonstrated a synthetic genetic circuit analogous to this digital ring oscillator (*see also* Fung et al., **ref. *45***).

4. Pseudoinverse Projection for Integrative Modeling of DNA Microarray Datasets

Integrative analysis of different types of global signals, such as these measured by DNA microarrays from the same organism, promises to reveal global causal co-ordination of cellular activities. For example, Bussemaker, Li, and Siggia ***(46)*** predicted new regulatory motifs by linear regression of profiles of genome-scale RNA expression in yeast vs profiles of the abundance levels, or counts of DNA oligomer motifs in the promoter regions of the same yeast genes. Lu, Nakorchevskiy, and Marcotte ***(47)*** associated the knockout phenotype of individual yeast genes with cell cycle arrest by deconvolution of the RNA expression profiles measured in the corresponding yeast mutants into the RNA expression profiles measured during the cell cycle for all yeast genes that were microarray-classified as cell cycle regulated.

This section reviews the pseudoinverse projection integrative model for DNA microarray datasets and other large-scale molecular biological signals ***(20,21)***. Pseudoinverse projection is an integrative BSS algorithm that decomposes the measured gene patterns of any given "data" signal of, e.g., proteins' DNA-binding into mathematically least squares-optimal pseudoinverse correlations with the measured gene patterns of a chosen "basis" signal of, e.g., RNA expression, in a different set of samples from the same organism. The measured array patterns of the data signal are least squares-approximated with a decomposition into the measured array patterns of the basis. The correspondence between these mathematical patterns that are uncovered in the measured signals and the independent

---

Fig. 14. *(Continued)* in the 18 arraylets, centered at their array-invariant levels. The expression patterns of the third through fifth and 14th through 16th arraylets display the sorting. **(C)** The third (red), fourth (blue), and fifth (green) yeast arraylets fit one period cosines of $\pi/3$ (red), 0 (blue) and $-\pi/3$ (green) initial phases. **(D)** The 14th (red), 15th (blue), and 16th (green) yeast arraylets fit one period cosines of $-\pi/3$- (red), $\pi/3$- (blue), and 0- (green) phases. **(E)** Human expression of the sorted 12,056 genes in the 18 arrays, centered at their gene- and array-invariant levels, showing a traveling wave of expression. **(F)** Human expression of the sorted 12,056 genes in the 18 arraylets, centered at their array-invariant levels. The expression patterns of the third through fifth and 14th through 16th arraylets display the sorting. **(G)** The third (red), fourth (blue), and fifth (green) human arraylets fit one period cosines of $\pi/3$- (red), 0- (blue), and $-\pi/3$- (green) phases. **(H)** The 14th (red), 15th (blue) and 16th (green) human arraylets fit one period cosines of $-\pi/3$- (red), $\pi/3$- (blue) and 0- (green) phases.

activities of cellular elements that compose the signals is illustrated with an integration of yeast genome-scale DNA-binding occupancy of cell cycle transcription factors ***(8)*** and DNA replication initiation proteins ***(9)*** with RNA expression during the cell cycle, using as basis sets the eigenarrays and arraylets determined by SVD and GSVD, respectively. One consistent picture emerges that predicts novel correlation between DNA replication initiation and RNA transcription during the yeast cell cycle. This novel correlation, which might be due to a previously unknown mechanism of regulation, demonstrates the power of the SVD, GSVD, and pseudoinverse projection models to predict previously unknown biological principles.

4.1. Mathematical Framework of Pseudoinverse Projection

Let the basis matrix $\hat{b}$ of size N-genomic sites or open reading frames (ORFs) $\times$ M-basis profiles tabulate M genome-scale molecular biological profiles of, e.g., RNA expression, measured from a set of M samples or extracted mathematically from a set of M or more measured samples. As before, the mth column vector in the matrix $\hat{b}$, $|b_m\rangle \equiv \hat{b}|m\rangle$, lists the signal measured in the mth sample by the mth array across all N ORFs simultaneously. The nth row vector in the matrix $\hat{b}$, $\langle n|\hat{b}$, lists the signal measured in the nth ORF across the different arrays, which correspond to the different samples. Let the data matrix $\hat{d}$ of size N-ORFs $\times$ L-data samples tabulate L genome-scale molecular biological profiles of, e.g., proteins' DNA binding, measured for the same ORFs in L samples from the same organism. The lth column vector in the matrix $\hat{d}$, $|d_l\rangle \equiv \hat{d}|l\rangle$, lists the signal measured in the lth sample across all N ORFs simultaneously.

Moore–Penrose pseudoinverse projection of the data matrix $\hat{d}$ onto the basis matrix $\hat{b}$ is a linear transformation of the data $\hat{d}$ from the N-ORFs $\times$ L-data samples space to the M-basis profiles $\times$ L-data samples space **(Fig. 15)**,

$$\begin{aligned} \hat{d} &\rightarrow \hat{b}\hat{c}, \\ \hat{b}^{\dagger}\hat{d} &\equiv \hat{c}, \end{aligned} \tag{8}$$

where the matrix $\hat{b}^{\dagger}$, that is, the pseudoinverse of $\hat{b}$, satisfies

$$\begin{aligned} \hat{b}\hat{b}^{\dagger}\hat{b} &= \hat{b}, \\ \hat{b}^{\dagger}\hat{b}\hat{b}^{\dagger} &= \hat{b}^{\dagger}, \\ (\hat{b}\hat{b}^{\dagger})^{T} &= \hat{b}\hat{b}^{\dagger}, \\ (\hat{b}^{\dagger}\hat{b})^{T} &= \hat{b}^{\dagger}\hat{b}, \end{aligned} \tag{9}$$

such that the transformation matrices $\hat{b}\hat{b}^{\dagger}$ and $\hat{b}^{\dagger}\hat{b}$ are orthogonal projection matrices for a real basis matrix $\hat{b}$.

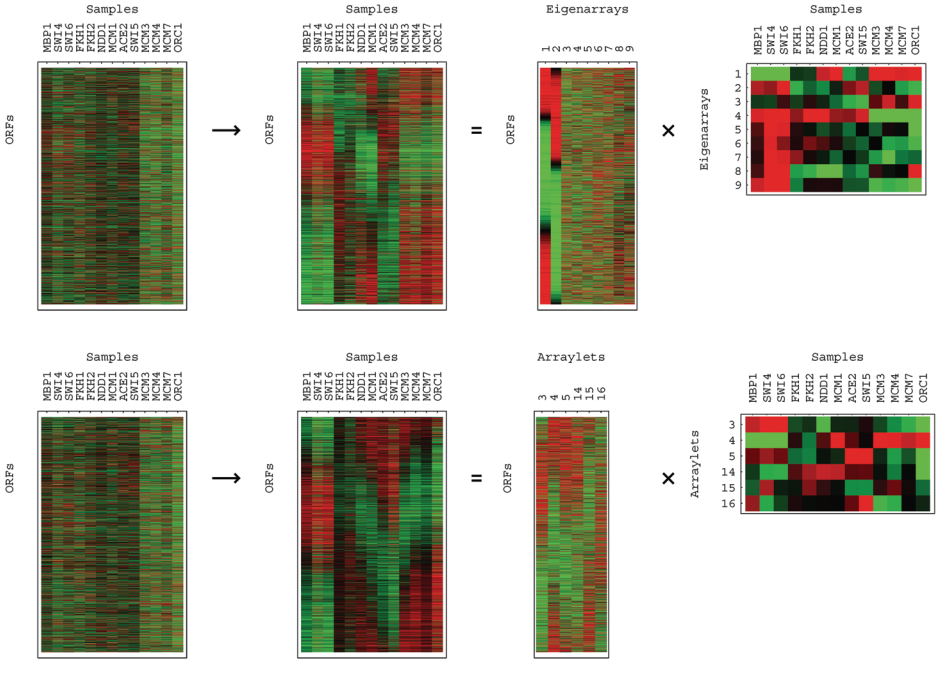


Fig. 15. Raster display of the pseudoinverse projection of the yeast cell cycle transcription factors and replication initiation proteins' DNA-binding data onto the SVD and GSVD cell cycle RNA expression bases, with overexpression (red), no change in expression (black) and underexpression (green) centered at ORF- and sample-invariant expression, and with the ORFs sorted according to their SVD and GSVD phases, respectively. Pseudoinverse projection is a linear transformation of the proteins' binding data from the 2227 ORFs × 13-data samples space to the nine eigenarrays of the SVD basis × 13-data samples space ***(upper)***, and also of the proteins' binding data from the 2139 ORFs × 13-data samples space to the six arraylets of the GSVD basis × 13-data samples space ***(lower)***.

In this space the data matrix $\hat{d}$ is represented by the pseudoinverse correlations matrix $\hat{c}$. The vector in the mth row of the matrix $\hat{c}$, $\langle c_m| \equiv \langle m|\hat{c}$, lists the pseudoinverse correlations of the L data profiles with the mth basis profile. The pseudoinverse correlations matrix $\hat{c}$ is unique, i.e., data driven.

4.2. Pseudoinverse Projection Integrative Analysis of Yeast Cell Cycle RNA Expression and Proteins' DNA-Binding Data

In this example, a data matrix that tabulates DNA-binding occupancy levels of nine yeast cell cycle transcription factors ***(8)*** and four yeast replication initiation proteins ***(9)*** across 2928 yeast ORFs is pseudoinverse projected onto (1)

the SVD cell cycle RNA expression basis matrix, which tabulates the expression of the nine most significant eigenarrays of the α factor, *CLB2*, and *CLN3* dataset, including the two eigenarrays that span the SVD cell cycle subspace, across 4579 ORFs, 2227 of which are present in the data matrix; and (2) the GSVD cell cycle RNA expression basis matrix, which tabulates the expression of the six arraylets that span the GSVD cell cycle subspace across 4523 ORFs, 2139 of which are present in the data matrix.

4.2.1. Pseudoinverse Correlations Uncovered in the Data Correspond to Reported Functions of Transcription Factors

The nine transcription factors are ordered, following Simon et al. ***(8)***, from these that have been reported to function in the cell cycle stage G_1, through these that have been reported to function in S, S/G_2, G_2/M, and M/G_1: Mbp1, Swi4, Swi6, Fkh1, Fkh2, Ndd1, Mcm1, Ace2, and Swi5. With this order, the SVD- and GSVD-pseudoinverse correlations approximately fit cosine functions of one period and of varying initial phases across the nine transcription factors' samples and are approximately invariant across the four samples of the replication initiation proteins, Mcm3, Mcm4, Mcm7, and Orc1 **(Fig. 16)**. Transcription factors that have been reported to function in antipodal cell cycle stages, such as Mbp1, Swi4, and Swi6 that are known to function in G_1 and Mcm1 that is known to function in G_2/M, consistently exhibit anticorrelated levels of DNA-binding in all patterns of pseudoinverse correlations. Each pattern of pseudoinverse correlations $\langle c_m|$ represents the activity of the transcripition factors during the cell cycle stage that the corresponding basis profile $\langle b_m|$ correlates with. For example, the first SVD basis profile, i.e., the first eigenarray, correlates with RNA expression oscillations at the transition from the cell cycle stage G_2/M to M/G_1 and anticorrelates with oscillations at the transition from G_1 to S **(Fig. 6C)**. Correspondingly, the first pattern of SVD-pseudoinverse correlations describes enhanced activity of the transcription factor Mcm1 and reduced activity of Mbp1, Swi4, and Swi6 **(Fig. 16B)**.

4.2.2. Pseudoinverse Reconstruction of the Data in the Basis Simulates Experimental Observation of Only the Cellular States Manifest in the Data that Correspond to Those in the Basis

The proteins' DNA-binding data is SVD- and independently also GSVD-reconstructed using pseudoinverse projections in the intersections of the SVD and GSVD bases matrices with the data matrix **(Fig. 17)**. With the 2227 and 2139 ORFs sorted according to their SVD and GSVD cell cycle phases, respectively,

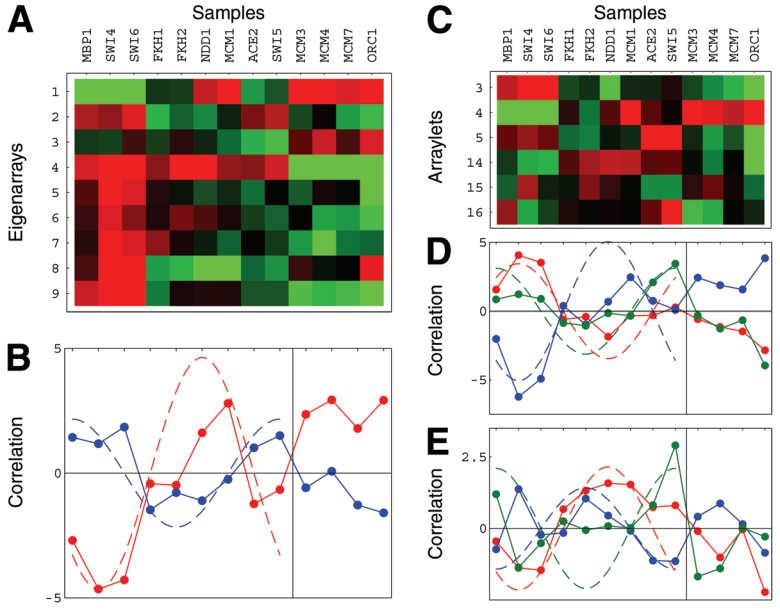


Fig. 16. Pseudoinverse correlations of the proteins' DNA-binding data with the SVD and GSVD cell cycle RNA expression. **(A)** Raster display of the correlations with the nine eigenarrays that span the SVD basis. **(B)** Line-joined graphs of the correlations with the first (red) and second (blue) most significant eigenarrays that span the SVD subspace. **(C)** Raster display of the correlations with the six arraylets that span the GSVD basis and the GSVD subspace. **(D)** Line-joined graphs of the correlations with third (red), fourth (blue), and fifth (green) arraylets, and **(E)** the 14th (red), 15th (blue), and 16th (green) arraylets.

the variations of the SVD- and GSVD-reconstructed binding profiles across the ORFs approximately fit cosine functions of one period and of varying initial phases.

The SVD- and GSVD-reconstructed transcription factors' data approximately fit traveling waves, cosinusoidally varying across the ORFs as well as the nine samples. Simon et al. ***(8)*** observed a similar traveling wave in the binding data from the nine transcription factors, ordered as in **Subheading 4.2.1.** above, across only 213 ORFs. These traveling waves are in agreement with current understanding of the progression of cell cycle transcription along the genes and in time as it is regulated by DNA binding of the transcription factors at the promoter regions of the transcribed genes. Pseudoinverse reconstruction of the data in both the SVD and GSVD bases, therefore, simulates experimental observation of only the proteins' DNA-binding cellular states that correspond to those of RNA expression during the cell cycle.

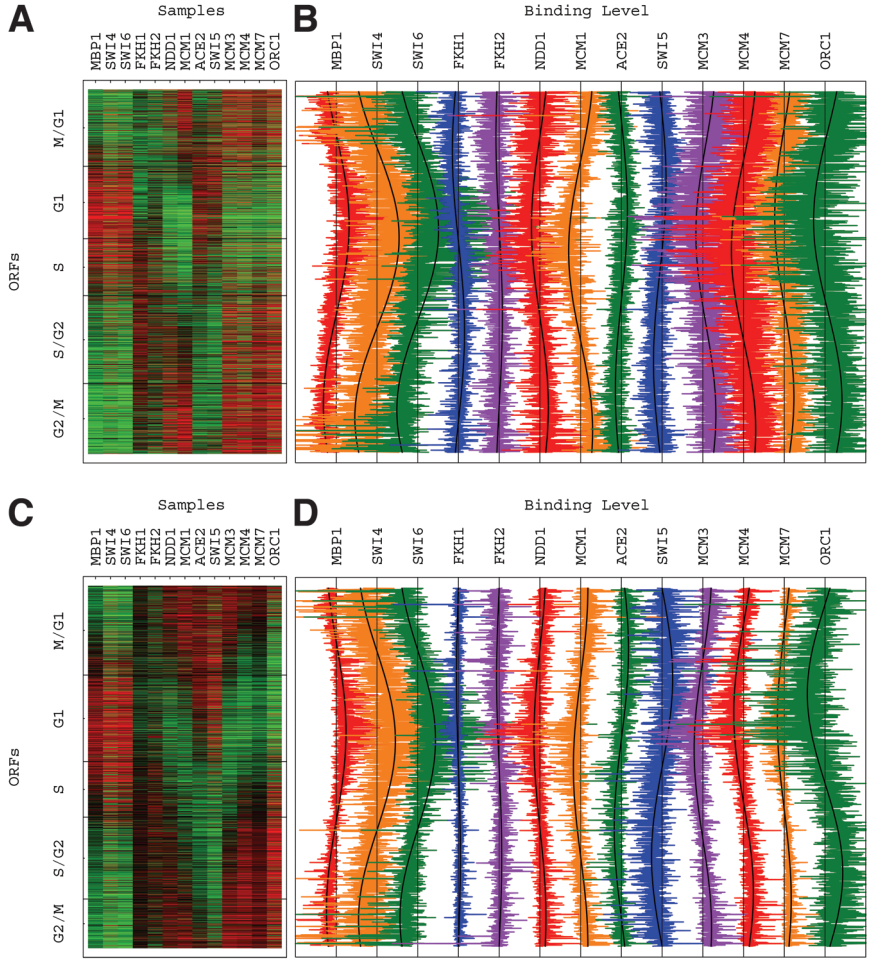


Fig. 17. Pseudoinverse reconstructions of the proteins' DNA-binding data in the SVD and GSVD cell cycle RNA expression bases, with the open reading frames sorted according to their SVD and GSVD phases, respectively, showing a traveling wave in the nine transcription factors and a standing wave in the four replication initiation proteins. **(A)** Raster display of the SVD-reconstructed data. **(B)** Line-joined graphs of the SVD-reconstructed data profiles. **(C)** Raster display of the GSVD-reconstructed data. **(D)** Line-joined graphs of the GSVD-reconstructed data profiles.

The SVD- and GSVD-reconstructed replication initiation proteins' data approximately fit a standing wave, cosinusoidally varying across the ORFs and constant across the four samples. These replication initiation proteins' reconstructed profiles are antiparallel to the reconstructed profiles of Mbp1, Swi4, and Swi6, and parallel to that of Mcm1.

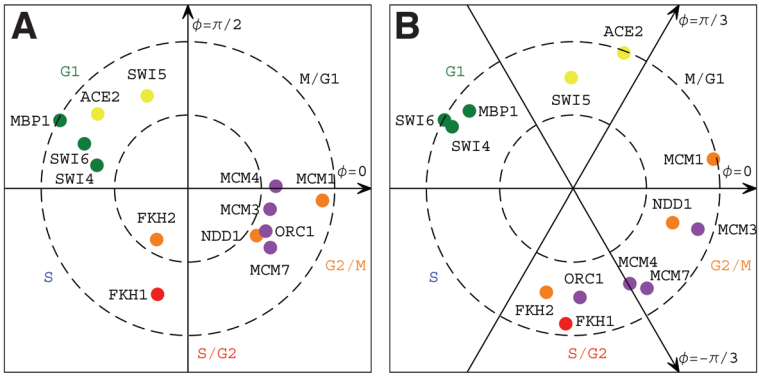


Fig. 18. Reconstructed yeast proteins' DNA-binding data in the RNA expression bases. **(A)** Correlations of the reconstructed binding of each of the 13 proteins with the first and second rotated eigenarrays along the *x*- and *y*-axes. The transcription factors are color coded according to their classification into the five cell cycle stages: M/G_1 (yellow), G_1 (green), S (blue), S/G_2 (red), and G_2/M (orange). The replication initiation proteins are colored violet. The dashed unit and half-unit circles outline 100% and 25% of overall normalized array expression in this subspace. **(B)** Projections of the binding of each of the nine transcription factors and four replication initiation proteins, after reconstruction in the six-dimensional GSVD cell cycle subspace, onto the two-dimensional subspace that least-squares approximates it. The dashed unit and half-unit circles outline 100% and 50% of added up, rather than cancelled out, contributions of the six arraylets to the overall projected reconstructed binding. The arrows describe the projections of the $-\pi/3$-, 0-, and $\pi/3$-phase arraylets.

4.2.3. Classification of the Basis-Reconstructed Data Samples Maps the Cellular States of the Data Onto Those of the Basis and Gives a Global Picture of Possible Causal Coordination of These States

Projected from the SVD basis, that is spanned by nine eigenarrays, onto the SVD cell cycle subspace, that is spanned by the two most significant of these eigenarrays, all SVD-reconstructed samples have at least 25% of their binding profiles in this subspace, except for Fkh2 **(Fig. 18A)**. Projected from the six-dimensional GSVD cell cycle subspace, that is spanned by six arraylets, onto the two-dimensional subspace that approximates it, 50% or more of the contributions of the six arraylets to each GSVD-reconstructed sample add up, rather than cancel out **(Fig. 18B)**.

Sorting the samples according to their SVD or GSVD phases gives an array order that is similar to that of Simon et al. ***(8)***, and describes the yeast cell cycle progression from the cellular state of Mbp1's binding through that of Swi5's. The SVD and GSVD mappings of the transcription factors' binding profiles

onto the expression subspaces are also in agreement with the current understanding of the cell cycle program. Mapping the binding of Mbp1, Swi4, and Swi6 onto the cell cycle expression stage G_1 corresponds to the biological coordination between the binding of these factors to the promoter regions of ORFs and the subsequent peak in transcription of these ORFs during G_1. The mapping of Mbp1, Swi4, and Swi6 onto G_1, which is antipodal to G_2/M, also corresponds to their binding to promoter regions of ORFs that exhibit transcription minima or shutdown during G_2/M, and to their minimal or lack of binding at promoter regions of ORFs which transcription peaks in G_2/M. Similarly, the mapping of Mcm1 onto G_2/M corresponds to its binding to the promoter regions of ORFs that are subsequently transcribed during the transition from G_2/M to M/G_1. The binding profiles of the replication initiation proteins are SVD- and GSVD-mapped onto the cell cycle stage that is antipodal to G_1. These SVD and GSVD mappings are consistent with the reconstructed profiles of Mcm3, Mcm4, Mcm7, and Orc1 being antiparallel to those of Mbp1, Swi4, and Swi6 and parallel to that of Mcm1.

The parallel and antiparallel associations by annotation of the proteins' DNA-binding profiles with the cellular states of RNA expression during the cell cycle are also consistent with the SVD and GSVD mappings. These associations follow the *p*-values for the distribution of the 400 and 377 ORFs that were microarray-classified and the 58 and 60 ORFs that were traditionally classified as cell cycle regulated among all 2227 and 2139 ORFs that are mapped onto the SVD and GSVD subspaces, respectively, and among each of the subsets of 200 ORFs with largest and smallest levels of binding occupancy in each of the profiles. Again, the binding profiles of all four DNA replication initiation proteins, Mcm3, Mcm4, Mcm7, and Orc1 are anticorrelated with RNA expression in the cell cycle stage G_1, together with the profile of the transcription factor Mcm1, whereas the profiles of the transcription factors Mbp1, Swi4, and Swi6 that are known to drive the cell cycle stage G_1, are correlated with RNA expression in this stage ***(20,21)***.

Thus, DNA-binding of Mcm3, Mcm4, Mcm7, and Orc1 adjacent to ORFs is pseudoinverse-correlated with minima or even shutdown of the transcription of these ORFs during the cell cycle stage G_1. This novel correlation suggests a previously unknown genome-scale coordination between DNA replication initiation and RNA transcription during the cell cycle in yeast.

The correlation between Mcm3, Mcm4, Mcm7, and Orc1 and the transcription factor Mcm1 suggests a genome-scale, or maybe even a genome-wide coordination in the activities of the DNA replication initiation proteins and Mcm1. One possible explanation of this correlation may be provided by the recent suggestion by Chang et al. ***(48***; *see also* Donato, Chang and Tye, **ref.** ***49)*** that Mcm1 binds origins of replication, and thus functions as a replication initiation protein

in addition to its function as a transcription factor. However, this correlation does not necessarily mean that Mcm1 colocalizes with origins. It is the tendency of ORFs adjacent to Mcm1's binding sites to exhibit transcription minima during the cell cycle stage G_1, which correlates with a similar tendency of those ORFs that are adjacent to binding sites of the replication initiation proteins.

4.3. Pseudoinverse Projection Integrative Model for Genome-Scale RNA Transcription and DNA-Binding of Cell Cycle Transcription Factors and Replication Initiation Proteins in Yeast

One consistent picture emerges upon integrating the genome-scale proteins' DNA-binding data with the SVD and GSVD cell cycle RNA expression bases, which is in agreement with the current understanding of the yeast cell cycle program ***(50–53)***, and is supported by recent experimental results ***(49)***. This picture correlates for the first time the binding of replication initiation proteins with minima or shutdown of the transcription of adjacent ORFs during the cell cycle stage G_1, under the assumption that the measured cell cycle RNA expression levels are approximately proportional to cell cycle RNA transcription activity. It was shown by Diffley et al. ***(50)*** that replication initiation requires binding of Mcm3, Mcm4, Mcm7, and Orc1 at origins of replication across the yeast genome during G_1 (*see also* **ref. *51***). And, it was shown by Micklem et al. ***(52)*** that these replication initiation proteins are involved with transcriptional silencing at the yeast mating loci (*see also* **ref. *53***). Either one of at least two mechanisms of regulation may be underlying this novel genome-scale correlation between DNA replication initiation and RNA transcription during the yeast cell cycle: the transcription of genes may reduce the binding efficiency of adjacent origins. Or, the binding of replication initiation proteins to origins of replication may repress, or even shut down, the transcription of adjacent genes.

This is the first time that a data-driven mathematical model, where the mathematical variables and operations represent biological or experimental reality, has been used to predict a biological principle that is truly on a genome scale. The ORFs in either one of the basis or data matrices were selected based on data quality alone, and were not limited to ORFs that are traditionally or microarray-classified as cell cycle regulated, suggesting that the RNA transcription signatures of yeast cell cycle cellular states may span the whole yeast genome.

5. Are Genetic Networks Linear and Orthogonal?

The SVD model, the GSVD comparative model, and the pseudoinverse projection integrative model are all mathematically linear and orthogonal. These models formulate genome-scale molecular biological signals as linear superpositions of mathematical patterns, which correlate with activities of cellular elements, such as regulators or transcription factors, that drive the measured signal

and cellular states where these elements are active. These models associate the independent cellular states with orthogonal, i.e., decorrelated, mathematical profiles suggesting that the overlap or crosstalk between the genome-scale effects of the corresponding cellular elements or modules is negligible.

Recently, Ihmels, Levy, and Barkai ***(54)*** found evidence for linearity as well as orthogonality in the metabolic network in yeast. Integrating genome-scale RNA expression data with the structural description of this network, they showed that at the network's branchpoints, only distinct branches are coexpressed, and concluded that transcriptional regulation biases the metabolic flow toward linearity. They also showed that individual isozymes, i.e., chemically distinct but functionally similar enzymes, tend to be corregulated separately with distinct processes. They concluded that transcriptional regulation uses isozymes as means for reducing crosstalk between pathways that use a common chemical reaction.

Orthogonality of the cellular states that compose a genetic network suggests an efficient network design. With no redundant functionality in the activities of the independent cellular elements, the number of such elements needed to carry out a given set of biological processes is minimized. An efficient network, however, is fragile. The robustness of biological systems to diverse perturbations, e.g., phenotypic stability despite environmental changes and genetic variation, suggests functional redundancy in the activities of the cellular elements, and therefore also correlations among the corresponding cellular states. Carslon and Doyle ***(55)*** introduced the framework of "highly optimized tolerance" to study fundamental aspects of complexity in, among others, biological systems that appear to be naturally selected for efficiency as well as robustness. They showed that trade-offs between efficiency and robustness might explain the behavior of such complex systems, including occurrences of catastrophic failure events.

Linearity of a genetic network may seem counterintuitive in light of the nonlinearity of the chemical processes, which underlie the network. Arkin and Ross ***(56)*** showed that enzymatic reaction mechanisms can be thought to compute the mathematically nonlinear functions of logic gates on the molecular level. They also showed that the qualitative logic gate behavior of such a reaction mechanism may not change when situated within a model of the cellular program that uses the reaction. This program functions as a biological switch from one pathway to another in response to chemical signals, and thus computes a nonlinear logic gate function on the cellular scale. Another cellular program that can be thought to compute nonlinear functions is the well-known genetic switch in the bacteriophage λ, the program of decision between lysis and lysogeny ***(57)***. McAdams and Shapiro ***(58)*** modeled this program with a circuit of integrated

logic components. However, even if the kinetics of biochemical reactions are nonlinear, the mass balance constraints that govern these reactions are linear. Schilling and Palsson ***(59)*** showed that the underlying pathway structure of a biochemical network, and therefore also its functional capabilities, can be extracted from the linear set of mass balance constraints corresponding to the set of reactions that compose this network.

That genetic networks might be modeled with linear and orthogonal mathematical frameworks does not necessarily imply that these networks are linear and orthogonal(e.g., **refs. *60–62***). Dynamical systems, linear and nonlinear, are regularly studied with linear orthogonal transforms ***(63).*** For example, SVD might be used to reconstruct the phase-space description of a dynamical system from a series of observations of the time evolution of the coordinates of the system. In such a reconstruction, the experimental data are mapped onto a subspace spanned by selected patterns that are uncovered in the data by SVD. The phase-space description of linear systems, for which the time evolution, or "motion," of the coordinates is periodic, such as the analog harmonic oscillator, is the "limit cycle." The phase-space description of nonlinear systems, for which the coordinates' motion is chaotic, such as the chemical Lotka-Volterra irreversible autocatalytic reaction ***(35–37)***, is the "strange attractor." Broomhead and King ***(64)*** were the first to use SVD to reconstruct the strange attractor.

Although it is still an open question whether genetic networks are linear and orthogonal, linear and orthogonal mathematical frameworks have already proven successful in describing the physical world, in such diverse areas as mechanics and perception. It may not be surprising, therefore, that linear and orthogonal mathematical models for genome-scale molecular biological signals (1) provide mathematical descriptions of the genetic networks that generate and sense the measured data, where the mathematical variables and operations represent biological or experimental reality; (2) elucidate the design principles of cellular systems as well as guide the design of synthetic ones; and (3) predict previously unknown biological principles.

These models may become the foundation of a future in which biological systems are modeled as physical systems are today.

Acknowledgments

The author thanks D. Botstein and P. O. Brown for introducing her to genomics, G. H. Golub for introducing her to matrix and tensor computation and M. van de Rijn for introducing her to translational cancer research. The author also thanks T. M. Baer, G. M. Church, J. F. X. Diffley, J. Doyle, S. R. Eddy, P. Green, R. R. Klevecz, E. Rivas, and J. J. Wyrick for thoughtful and thorough reviews of parts of the work presented in this chapter. This work was supported

by a National Human Genome Research Institute Individual Mentored Research Scientist Development Award in Genomic Research and Analysis (K01 HG00038-05) and by a Sloan Foundation and Department of Energy Postdoctoral Fellowship in Computational Molecular Biology (DE-FG03-99ER62836).

References

1. Fodor, S. P., Rava, R. P., Huang, X. C., Pease, A. C., Holmes, C. P., and Adams, C. L. (1993) Multiplexed biochemical assays with biological chips. *Nature* **364,** 555–556.
2. Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470.
3. Brown, P. O., and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* **21,** 31–37.
4. Pollack, J. R., and Iyer, V. R. (2002) Characterizing the physical genome. *Nat. Genet.* **32,** 515–521.
5. Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., et al. (2001) The Stanford microarray database. *Nucleic Acids Res.* **29,** 152–155.
6. Spellman, P. T., Sherlock, G., Zhang, M. Q., et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* **9,** 3273–3297.
7. Whitfield, M. L., Sherlock, G., Saldanha, A., et al. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13,** 1977–2000.
8. Simon, I., Barnett, J., Hannett, N., et al. (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106,** 697–708.
9. Wyrick, J. J., Aparicio, J. G., Chen, T., et al. (2001) Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins. *Science* **294,** 2301–2304.
10. Newton, I. (1999) *The Principia: Mathematical Principles of Natural Philosophy.* (Cohen, I. B., and Whitman, A., trans.) University of California Press, Berkeley, CA.
11. Hubel, D. H., and Wiesel, T. N. (1968) Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195,** 215–243.
12. Barlow, H. B. (1972) Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* **1,** 371–394.
13. Olshausen, B. A., and Field, D. J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381,** 607–609.
14. Bell, A. J., and Sejnowski, T. J. (1997) The "independent components" of natural scenes are edge filters. *Vision Res.* **37,** 3327–3338.
15. Golub, G. H., and Van Loan, C. F. (1996) *Matrix Computation*, *3rd ed.*, Johns Hopkins University, Press, Baltimore, MD.
16. 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,** 10,101–10,106.

17. Alter, O., Brown, P. O., and Botstein, D. (2001) Processing and modeling genome-wide expression data using singular value decomposition. In: *Microarrays: Optical Technologies and Informatics, vol. 4266* (Bittner, M. L., Chen, Y., Dorsel, A. N., and Dougherty, E. R., eds.), Int. Soc. Optical Eng., Bellingham, WA, pp. 171–186.
18. Nielsen, T. O., West, R. B., Linn, S. C., et al. (2002) Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet* **359,** 1301–1307.
19. Alter, O., Brown, P. O., and Botstein, D. (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl. Acad. Sci. USA* **100,** 3351–3356.
20. Alter, O., Golub, G. H., Brown, P. O., and Botstein, D. (2004) Novel genome-scale correlation between DNA replication and RNA transcription during the cell cycle in yeast is predicted by data-driven models. In: *Proc. Miami Nat. Biotechnol. Winter Symp. on the Cell Cycle, Chromosomes and Cancer, vol. 15* (Deutscher, M. P., Black, S., Boehmer, P. E., et al., eds.), Univ. Miami Sch. Med., Miami, FL, www.med.miami.edu/mnbws/Alter-.pdf.
21. Alter, O. and Golub, G. H. (2004) Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. *Proc. Natl. Acad. Sci. USA* **101,** 16,577–16,582.
22. Alter, O., and Golub, G. H. (2005) Reconstructing the pathways of a cellular system from genome-scale signals using matrix and tensor computations. *Proc. Natl. Acad. Sci. USA* **102,** 17,559–17,564.
23. Alter, O., and Golub, G. H. (2006) Singular value decomposition of genome-scale mRNA lengths distribution reveals asymmetry in RNA gel electrophoresis band broadening. *Proc. Natl. Acad. Sci. USA* **103,** 11,828–11,833.
24. Alter, O. (2006) Discovery of principles of nature from mathematical modeling of DNA microarray data. *Proc. Natl. Acad. Sci. USA* **103,** 16,063–16,064.
25. Wigner, E. P. (1960) The unreasonable effectiveness of mathematics in the natural sciences. *Commun. Pure Appl. Math.* **13,** 1–14.
26. Hopfield, J. J. (1999) Odor space and olfactory processing: collective algorithms and neural implementation. *Proc. Natl. Acad. Sci. USA* **96,** 12,506–12,511.
27. Sirovich, L., and Kirby, M. (1987) Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc. Am. A* **4,** 519–524.
28. Turk, M., and Pentland, A. (1991) Eigenfaces for recognition. *J. Cogn. Neurosci.* **3,** 71–86.
29. Landau, L. D., and Lifshitz, E. M. (1976) *Mechanics*, *3rd ed.* (Sykes, J. B., and Bell, J. S., trans.), Butterworth-Heinemann, Oxford, UK.
30. 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.
31. Roberts, C. J., Nelson, B., Marton, M. J., et al. (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287,** 873–880.
32. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1994) *Molecular Biology of the Cell*, *3rd ed.*, Garland Pub., New York, NY.

33. Klevecz, R. R., Bolen, J., Forrest, G., and Murray, D. B. (2004) A genomewide oscillation in transcription gates DNA replication and cell cycle. *Proc. Natl. Acad. Sci. USA* **101,** 1200–1205.
34. Li, C. M., and Klevecz, R. R. (2006) A rapid genome-scale response of the transcriptional oscillator to perturbation reveals a period-doubling path to phenotypic change. *Proc. Natl. Acad. Sci. USA* **103,** 16,254–16,259.
35. Nicolis, G. and Prigogine, I. (1971) Fluctuations in nonequilibrium systems. *Proc. Natl. Acad. Sci. USA* **68,** 2102–2107.
36. Rössler O. E. (1976) An equation for continuous chaos. *Phys. Lett. A* **35,** 397–398.
37. Roux, J. -C., Simoyi, R. H., and Swinney, H. L. (1983) Observation of a strange attractor. *Physica D* **8,** 257–266.
38. Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302,** 249–255.
39. Bergmann, S., Ihmels, J., and Barkai, N. (2004) Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol* **2,** E9.
40. Mushegian, A. R., and Koonin, E. V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA* **93,** 10,268–10,273.
41. Dwight, S. S., Harris, M. A., Dolinski, K., et al. (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.* **30,** 69–72.
42. Kurihara, L. J., Stewart, B. G., Gammie, A. E., and Rose, M. D. (1996) Kar4p, a karyogamy-specific component of the yeast pheromone response pathway. *Mol. Cell. Biol.* **16,** 3990–4002.
43. Ewing, B. and Green, P. (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25,** 232–234.
44. Elowitz, M. B., and Leibler, S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature* **403,** 335–338.
45. Fung, E., Wong, W. W., Suen, J. K., Butler, T., Lee, S. G., and Liao, J. C. (2005) A synthetic gene-metabolic oscillator. *Nature* **435,** 118–122.
46. Bussemaker, H. J., Li, H., and Siggia, E. D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.* **27,** 167–171.
47. Lu, P., Nakorchevskiy, A., and Marcotte, E. M. (2003) Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proc. Natl. Acad. Sci. USA* **100,** 10,370–10,375.
48. Chang, V. K., Fitch, M. J., Donato, J. J., Christensen, T. W., Merchant, A. M., and Tye, B. K. (2003) Mcm1 binds replication origins. *J. Biol. Chem.* **278,** 6093–6100.
49. Donato, J. J., Chung, S. C., and Tye, B. K. (2006) Genome-wide hierarchy of replication origin usage in *Saccharomyces cerevisiae. PloS Genet.* **2,** E9.
50. Diffley, J. F. X., Cocker, J. H., Dowell, S. J., and Rowley, A. (1994) Two steps in the assembly of complexes at yeast replication origins in vivo. *Cell* **78,** 303–316.
51. Kelly, T. J. and Brown, G. W. (2000) Regulation of chromosome replication. *Annu. Rev. Biochem.* **69,** 829–880.

52. Micklem, G., Rowley, A., Harwood, J., Nasmyth, K., and Diffley, J. F. X. (1993) Yeast origin recognition complex is involved in DNA replication and transcriptional silencing. *Nature* **366,** 87–89.
53. Fox, C. A. and Rine, J. (1996) Influences of the cell cycle on silencing. *Curr. Opin. Cell Biol.* **8,** 354–357.
54. Ihmels, J., Levy, R., and Barkai, N. (2004) Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **60,** 86–92.
55. Carlson, J. M. and Doyle, J. (1999) Highly optimized tolerance: a mechanism for power laws in designed systems. *Phys. Rev. E* **60,** 1412–1427.
56. Arkin, A. P. and Ross, J. (1994) Computational functions in biochemical reaction networks. *Biophys. J.* **67,** 560–578.
57. Ptashne, M. (1992) *Genetic Switch: Phage Lambda and Higher Organisms*, *2nd ed.*, Blackwell Publishers, Oxford, UK.
58. McAdams, H. H. and Shapiro, L. (1995) Circuit simulation of genetic networks. *Science* **269,** 650–656.
59. Schilling, C. H. and Palsson, B. O. (1998) The underlying pathway structure of biochemical reaction networks. *Proc. Natl. Acad. Sci. USA* **95,** 4193–4198.
60. Yeung, M. K., Tegner, J., and Collins, J. J. (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA* **99,** 6163–6168.
61. Price, N. D., Reed, J. L., Papin, J. A., Famili, I., and Palsson, B. O. (2003) Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices. *Biophys. J.* **84,** 794–804.
62. Vlad, M. O., Arkin, A. P., and Ross, J. (2004) Response experiments for nonlinear systems with application to reaction kinetics and genetics. *Proc. Natl. Acad. Sci. USA* **101,** 7223–7228.
63. Doyle, J. and Stein, G. (1981) Multivariable feedback design: Concepts for a classical/modern synthesis. *IEEE Trans. Automat. Contr.* **26,** 4–16.
64. Broomhead, D. S. and King, G. P. (1986) Extracting qualitative dynamics from experimental-data. *Physica D* **20,** 217–236.