Singular value decomposition for genome-wide expression data processing and modeling

Orly Alter*†, Patrick O. Brown‡, and David Botstein*

Departments of *Genetics and ‡Biochemistry, Stanford University, Stanford, CA 94305

Contributed by David Botstein, June 15, 2000

We describe the use of singular value decomposition in transforming genome-wide expression data from genes × arrays space to reduced diagonalized "eigengenes" × "eigenarrays" space, where the eigengenes (or eigenarrays) are unique orthonormal superpositions of the genes (or arrays). Normalizing the data by filtering out the eigengenes (and eigenarrays) that are inferred to represent noise or experimental artifacts enables meaningful comparison of the expression of different genes across different arrays in different experiments. Sorting the data according to the eigengenes and eigenarrays gives a global picture of the dynamics of gene expression, in which individual genes and arrays appear to be classified into groups of similar regulation and function, or similar cellular state and biological phenotype, respectively. After normalization and sorting, the significant eigengenes and eigenarrays can be associated with observed genome-wide effects of regulators, or with measured samples, in which these regulators are overactive or underactive, respectively.

DNA microarray technology (1, 2) and genome sequencing have advanced to the point that it is now possible to monitor gene expression levels on a genomic scale (3). These new data promise to enhance fundamental understanding of life on the molecular level, from regulation of gene expression and gene function to cellular mechanisms, and may prove useful in medical diagnosis, treatment, and drug design. Analysis of these new data requires mathematical tools that are adaptable to the large quantities of data, while reducing the complexity of the data to make them comprehensible. Analysis so far has been limited to identification of genes and arrays with similar expression patterns by using clustering methods (4–9).

We describe the use of singular value decomposition (SVD) (10) in analyzing genome-wide expression data. SVD is also known as Karhunen–Loève expansion in pattern recognition (11) and as principal-component analysis in statistics (12). SVD is a linear transformation of the expression data from the genes × arrays space to the reduced "eigengenes" × "eigenarrays" space. In this space the data are diagonalized, such that each eigengene is expressed only in the corresponding eigenarray, with the corresponding "eigenexpression" level indicating their relative significance. The eigengenes and eigenarrays are unique, and therefore also data-driven, orthonormal superpositions of the genes and arrays, respectively.

We show that several significant eigengenes and the corresponding eigenarrays capture most of the expression information. Normalizing the data by filtering out the eigengenes (and the corresponding eigenarrays) that are inferred to represent noise or experimental artifacts enables meaningful comparison of the expression of different genes across different arrays in different experiments. Such normalization may improve any further analysis of the expression data. Sorting the data according to the correlations of the genes (and arrays) with eigengenes (and eigenarrays) gives a global picture of the dynamics of gene expression, in which individual genes and arrays appear to be classified into groups of similar regulation and function, or similar cellular state and biological phenotype, respectively. These groups of genes (or arrays) are not defined by overall similarity in expression, but only by similarity in the expression of any chosen subset of eigengenes (or eigenarrays). Upon comparing two or more similar experiments, with a regulator being overactive or underactive in one but normally expressed in the others, the expression pattern of one of the significant eigengenes may be correlated with the expression patterns of this regulator and its targets. This eigengene, therefore, can be associated with the observed genome-wide effect of the regulator. The expression pattern of the corresponding eigenarray is correlated with the expression patterns observed in samples in which the regulator is overactive or underactive. This eigenarray, therefore, can be associated with these samples.

We conclude that SVD provides a useful mathematical framework for processing and modeling genome-wide expression data, in which both the mathematical variables and operations may be assigned biological meaning.

Mathematical Framework: Singular Value Decomposition

The relative expression levels of N genes of a model organism, which may constitute almost the entire genome of this organism, in a single sample, are probed simultaneously by a single microarray. A series of M arrays, which are almost identical physically, probe the genome-wide expression levels in M different samples—i.e., under M different experimental conditions. Let the matrix $\hat{e}$, of size N-genes × M-arrays, tabulate the full expression data. Each element of $\hat{e}$ satisfies $\langle n|\hat{e}|m\rangle \equiv e_{nm}$ for all $1 \leq n \leq N$ and $1 \leq m \leq M$, where e_{nm} is the relative expression level of the nth gene in the mth sample as measured by the mth array.§ The vector in the nth row of the matrix $\hat{e}$, $\langle g_n| \equiv \langle n|\hat{e}$, lists the relative expression of the nth gene across the different samples which correspond to the different arrays. The vector in the mth column of the matrix $\hat{e}$, $|a_m\rangle \equiv \hat{e}|m\rangle$, lists the genome-wide relative expression measured by the mth array.

SVD (10) is then linear transformation of the expression data from the N-genes × M-arrays space to the reduced L-"eigenarrays" × L-"eigengenes" space, where $L = \min\{M, N\}$ (see Fig. 7 in supplemental material at www.pnas.org),

$$\hat{e} = \hat{u}\hat{\varepsilon}\hat{v}^T. \qquad [1]$$

In this space the data are represented by the diagonal nonnegative matrix $\hat{e}$, of size L-eigengenes × L-eigenarrays, which satisfies $\langle k|\hat{\varepsilon}|l\rangle \equiv \varepsilon_l\delta_{kl} \geq 0$ for all $1 \leq k,l \leq L$, such that the lth eigengene is expressed only in the corresponding lth eigenarray, with the corresponding "eigenexpression" level ε_l. Therefore, the expression of each eigengene (or eigenarray) is decoupled

Abbreviation: SVD, singular value decomposition.

†To whom reprint requests should be addressed. E-mail: orly@genome.stanford.edu.

§In this report, $\hat{m}$ 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.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

from that of all other eigengenes (or eigenarrays). The "fraction of eigenexpression,"

$$p_l = \varepsilon_l^2 / \sum_{k=1}^{L} \varepsilon_k^2, \qquad [2]$$

indicates the relative significance of the lth eigengene and eigenarray in terms of the fraction of the overall expression that they capture. Assume also that the eigenexpression levels are arranged in decreasing order of significance, such that $\varepsilon_1 \geq \varepsilon_2 \geq \ldots \geq \varepsilon_L \geq 0$. "Shannon entropy" of a dataset,

$$0 \leq d = \frac{-1}{\log(L)} \sum_{k=1}^{L} p_k \log(p_k) \leq 1, \qquad [3]$$

measures the complexity of the data from the distribution of the overall expression between the different eigengenes (and eigenarrays), where $d = 0$ corresponds to an ordered and redundant dataset in which all expression is captured by a single eigengene (and 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 $\times$ L-eigenarrays and the L-eigengenes $\times$ M-arrays basis sets, respectively. The vector in the lth row of the matrix $\hat{v}^T$, $\langle\gamma_l| \equiv \langle l|\hat{v}^T$, lists the expression of the lth eigengene across the different arrays. The vector in the lth column of the matrix $\hat{u}$, $|\alpha_l\rangle \equiv \hat{u}|l\rangle$, lists the genome-wide expression in the lth eigenarray. The eigengenes and eigenarrays are orthonormal superpositions of the genes and arrays, such that the transformation matrices $\hat{u}$ and $\hat{v}$ are both orthogonal

$$\hat{u}^T\hat{u} = \hat{v}^T\hat{v} = \hat{I}, \qquad [4]$$

where $\hat{I}$ is the identity matrix. Therefore, the expression of each eigengene (or eigenarray) is not only decoupled but also decorrelated from that of all other eigengenes (or eigenarrays). The eigengenes and eigenarrays are unique, except in degenerate subspaces, defined by subsets of equal eigenexpression levels, and except for a phase factor of ± 1, such that each eigengene (or eigenarray) captures both parallel and antiparallel gene (or array) expression patterns. Therefore, SVD is data-driven, except in degenerate subspaces.

SVD Calculation. According to Eqs. **1** and **4**, the M-arrays $\times$ M-arrays symmetric correlation matrix $\hat{a} = \hat{e}^T\hat{e} = \hat{v}\hat{\varepsilon}^2\hat{v}^T$ is represented in the L-eigengenes $\times$ L-eigengenes space by the diagonal matrix $\hat{\varepsilon}^2$. The N-genes $\times$ N-genes correlation matrix $\hat{g} = \hat{e}\hat{e}^T = \hat{u}\hat{\varepsilon}^2\hat{u}^T$ is represented in the L-eigenarrays $\times$ L-eigenarrays space also by $\hat{\varepsilon}^2$, where for $L = \min\{M, N\} = M$, $\hat{g}$ has a null subspace of at least $N - M$ null eigenvalues. We, therefore, calculate the SVD of a dataset $\hat{e}$, with $M \ll N$, by diagonalizing $\hat{a}$, and then projecting the resulting $\hat{v}$ and $\hat{\varepsilon}$ onto $\hat{e}$ to obtain $\hat{u} = \hat{e}\hat{v}\hat{\varepsilon}^{-1}$.

Pattern Inference. The decorrelation of the eigengenes (and eigenarrays) suggests the possibility that some of the eigengenes (and the corresponding eigenarrays) represent independent regulatory programs or processes (and corresponding cellular states). We infer that an eigengene $|\gamma_l\rangle$ represents a regulatory program or process from its expression pattern across all arrays, when this pattern is biologically interpretable. This inference may be supported by a corresponding coherent biological theme reflected in the functions of the genes, whose expression patterns correlate or anticorrelate with the pattern of this eigengene. With this we assume that the corresponding eigenarray $|\alpha_l\rangle$ (which lists the amplitude of this eigengene pattern in the expression of each gene $|g_n\rangle$ relative to all other genes $\langle n|\alpha_l\rangle = \langle g_n|\gamma_l\rangle/\varepsilon_l$) represents the cellular state which corresponds to this process. We infer that the eigenarray $|\alpha_l\rangle$ represents a cellular state from the arrays whose expression patterns correlate or anticorrelate with the pattern of this eigenarray. Upon sorting of the genes, this inference may be supported by the expression pattern of this eigenarray across all genes, when this pattern is biologically interpretable.

Data Normalization. The decoupling of the eigengenes and eigenarrays allows filtering the data without eliminating genes or arrays from the dataset. We filter any of the eigengenes $|\gamma_l\rangle$ (and the corresponding eigenarray $|\alpha_l\rangle$) $\hat{e} \rightarrow \hat{e} - \varepsilon_l|\alpha_l\rangle\langle\gamma_l|$, by substituting zero for the eigenexpression level $\varepsilon_l = 0$ in the diagonal matrix $\hat{\varepsilon}$ and reconstructing the data according to Eq. **1**. We normalize the data by filtering out those eigengenes (and eigenarrays) that are inferred to represent noise or experimental artifacts.

Degenerate Subspace Rotation. The uniqueness of the eigengenes and eigenarrays does not hold in a degenerate subspace, defined by equal eigenexpression levels. We approximate significant similar eigenexpression levels $\varepsilon_l \approx \varepsilon_{l+1} \approx \ldots \approx \varepsilon_m$ with $\varepsilon_l = \ldots = \varepsilon_m = \sqrt{\sum_{k=1}^{m} \varepsilon_k^2/(m - l + 1)}$. Therefore, Eqs. **1–4** remain valid upon rotation of the corresponding eigengenes $\{(|\gamma_l\rangle, \ldots, |\gamma_m\rangle) \rightarrow \hat{R}(|\gamma_l\rangle, \ldots, |\gamma_m\rangle)\}$, and eigenarrays $\{(|\alpha_l\rangle, \ldots, |\alpha_m\rangle) \rightarrow \hat{R}(|\alpha_l\rangle, \ldots, |\alpha_m\rangle)\}$, for all orthogonal $\hat{R}$, $\hat{R}^T\hat{R} = \hat{I}$. We choose a unique rotation $\hat{R}$ by subjecting the rotated eigengenes to $m - l$ constraints, such that these constrained eigengenes may be advantageous in interpreting and presenting the expression data.

Data Sorting. Inferring that eigengenes (and eigenarrays) represent independent processes (and cellular states) allows sorting the data by similarity in the expression of any chosen subset of these eigengenes (and eigenarrays), rather than by overall similarity in expression. Given two eigengenes $|\gamma_k\rangle$ and $|\gamma_l\rangle$ (or eigenarrays $|\alpha_k\rangle$ and $|\alpha_l\rangle$), we plot the correlation of $|\gamma_k\rangle$ with each gene $|g_n\rangle$, $\langle\gamma_k|g_n\rangle/\langle g_n|g_n\rangle$ (or $|\alpha_k\rangle$ with each array $|a_m\rangle$) along the y-axis, vs. that of $|\gamma_l\rangle$ (or $|\alpha_l\rangle$) along the x-axis. In this plot, the distance of each gene (or array) from the origin is its amplitude of expression in the subspace spanned by $|\gamma_k\rangle$ and $|\gamma_l\rangle$ (or $|\alpha_k\rangle$ and $|\alpha_l\rangle$), relative to its overall expression $r_n \equiv \langle g_n|g_n\rangle^{-1}\sqrt{|\langle\gamma_k|g_n\rangle|^2 + |\langle\gamma_l|g_n\rangle|^2}$ (or $r_m \equiv \langle a_m|a_m\rangle^{-1}\sqrt{|\langle\alpha_k|a_m\rangle|^2 + |\langle\alpha_l|a_m\rangle|^2}$). The angular distance of each gene (or array) from the x-axis is its phase in the transition from the expression pattern $|\gamma_l\rangle$ to $|\gamma_k\rangle$ and back to $|\gamma_l\rangle$ (or $|\alpha_l\rangle$ to $|\alpha_k\rangle$ and back to $|\alpha_l\rangle$) $\tan\phi_n \equiv \langle\gamma_k|g_n\rangle/\langle\gamma_l|g_n\rangle$, (or $\tan\phi_m \equiv \langle\alpha_k|a_n\rangle/\langle\alpha_l|a_m\rangle$). We sort the genes (or arrays) according to ϕ_n (or ϕ_m).

Biological Data Analysis: Elutriation-Synchronized Cell Cycle

Spellman *et al.* (3) monitored genome-wide mRNA levels, for 6,108 ORFs of the budding yeast *Saccharomyces cerevisiae* simultaneously, over approximately one cell cycle period, $T \approx 390$ min, in a yeast culture synchronized by elutriation, relative to a reference mRNA from an asynchronous yeast culture, at 30-min intervals. The elutriation dataset we analyze (see supplemental data and Mathematica notebook at www.pnas.org and at http://genome-www.stanford.edu/SVD/) tabulates the measured ratios of gene expression levels for the $N = 5{,}981$ genes, 784 of which were classified by Spellman *et al.* as cell cycle regulated, with no missing data in the $M = 14$ arrays.

Pattern Inference. Consider the 14 eigengenes of the elutriation dataset. The first and most significant eigengene $|\gamma_1\rangle$, which describes time invariant relative expression during the cell cycle (Fig. 8*a* at www.pnas.org), captures more than 90% of the overall relative expression in this experiment (Fig. 8*b*). The entropy of

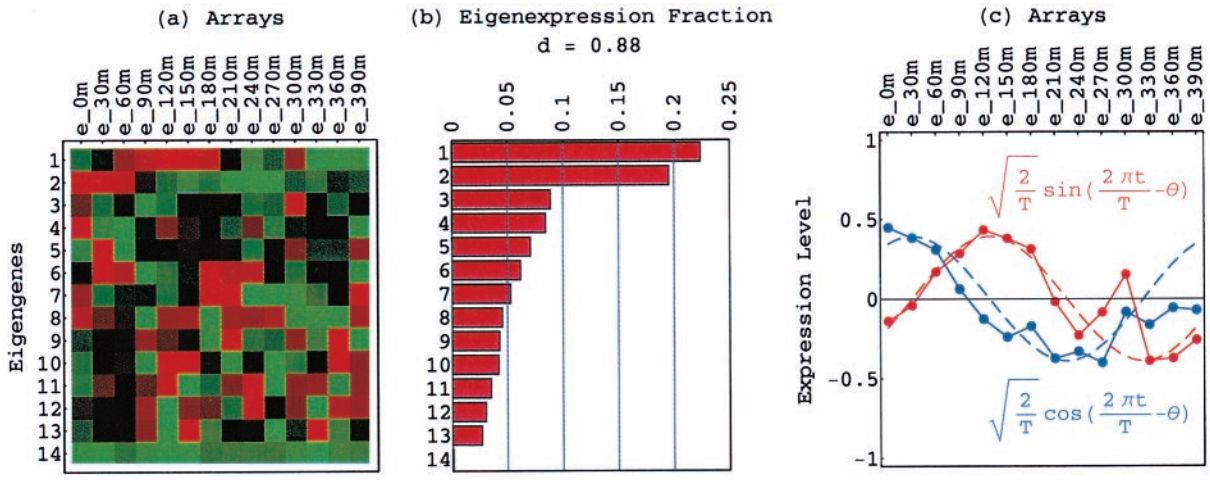

Fig. 1. Normalized elutriation eigengenes. (*a*) Raster display of $\hat{v}_N^T$, the expression of 14 eigengenes in 14 arrays. (*b*) Bar chart of the fractions of eigenexpression, showing that $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ capture about 20% of the overall normalized expression each, and a high entropy $d = 0.88$. (*c*) Line-joined graphs of the expression levels of $|\gamma_1\rangle_N$ (red) and $|\gamma_2\rangle_N$ (blue) in the 14 arrays fit dashed graphs of normalized sine (red) and cosine (blue) of period $T = 390$ min and phase $\theta = 2\pi/13$, respectively.

the dataset, therefore, is low $d = 0.14 \ll 1$. This suggests that the underlying processes are manifested by weak perturbations of a steady state of expression. This also suggests that time-invariant additive constants due to uncontrolled experimental variables may be superimposed on the data. We infer that $|\gamma_1\rangle$ represents experimental additive constants superimposed on a steady gene expression state, and assume that $|\alpha_1\rangle$ represents the corresponding steady cellular state. The second, third, and fourth eigengenes, which show oscillations during the cell cycle (Fig. 8*c*), capture about 3%, 1%, and 0.5% of the overall relative expression, respectively. The time variation of $|\gamma_3\rangle$ fits a normalized sine function of period T, $\sqrt{2/T}\sin(2\pi t/T)$. We infer that $|\gamma_3\rangle$ represents expression oscillation, which is consistent with gene expression oscillations during a cell cycle. The time variations of the second and fourth eigengenes fit a cosine function of period T with $\sqrt{1/2}$ the amplitude of a normalized cosine with this period, $\sqrt{1/T}\cos 2\pi t/T$. However, while $|\gamma_2\rangle$ shows decreasing expression on transition from $t = 0$ to 30 min, $|\gamma_4\rangle$ shows increasing expression. We infer that $|\gamma_2\rangle$ and $|\gamma_4\rangle$ represent initial transient increase and decrease in expression in response to the elutriation, respectively, superimposed on expression oscillation during the cell cycle.

Data Normalization. We filter out the first eigengene and eigenarray of the elutriation dataset, $\hat{e} \to \hat{e}_C = \hat{e} - \varepsilon_1|\alpha_1\rangle\langle\gamma_1|$, removing the steady state of expression. Each of the elements of the dataset $\hat{e}_C$, $\langle n|\hat{e}_C|m\rangle \equiv e_{C,nm}$, is the difference of the measured expression of the nth gene in the mth array from the steady-state levels of expression for these gene and array as calculated by SVD. Therefore, $e^2_{C,nm}$ is the variance in the measured expression of the nth gene in the mth array. Let $\hat{e}_{LV}$ tabulate the natural logarithm of the variances in elutriation expression, such that each element of $\hat{e}_{LV}$ satisfies $\langle n|\hat{e}_{LV}|m\rangle \equiv \log(e^2_{C,nm})$ for all $1 \le n \le N$ and $1 \le m \le M$, and consider the eigengenes of $\hat{e}_{LV}$ (Fig. 9*a* in supplemental material at www.pnas.org). The first eigengene $|\gamma_1\rangle_{LV}$, which captures more than 80% of the overall information in this dataset (Fig. 9*b*), describes a weak initial transient increase superimposed on a time-invariant scale of expression variance. The initial transient increase in the scale of expression variance may be a response to the elutriation. The time-invariant scale of expression variance suggests that a steady scale of experimental as well as biological uncertainty is associated with the expression data. This also suggests that time-invariant multiplicative constants due to uncontrolled experimental variables may be superimposed on the data. We filter out $|\gamma_1\rangle_{LV}$, removing the steady scale of expression variance, $\hat{e}_{LV} \to \hat{e}_{CLV} = \hat{e}_{LV} - \varepsilon_{1,LV}|\alpha_1\rangle_{LV\,LV}\langle\gamma_1|$.

The normalized elutriation dataset $\hat{e}_N$, where each of its elements satisfies $\langle n|\hat{e}_N|m\rangle \equiv \text{sign}(e_{C,nm})\sqrt{\exp(e_{CLV,nm})}$, tabulates for each gene and array expression patterns that are approximately centered at the steady-state expression level (i.e., of approximately zero arithmetic means), with variances which are approximately normalized by the steady scale of expression variance (i.e., of approximately unit geometric means). The first and second eigengenes,

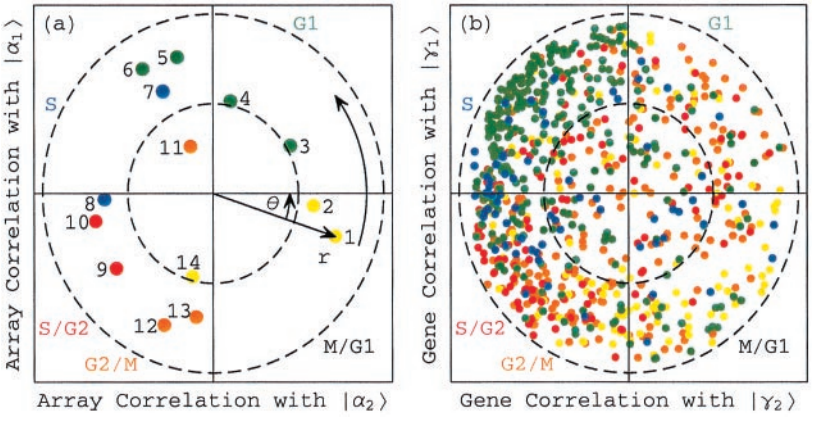

Fig. 2. Normalized elutriation expression in the subspace associated with the cell cycle. (*a*) Array correlation with $|\alpha_1\rangle_N$ along the *y*-axis vs. that with $|\alpha_2\rangle_N$ along the *x*-axis, color-coded according to the classification of the arrays into the five cell cycle stages, 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 outline 100% and 25% of overall normalized array expression in the $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ subspace. (*b*) Correlation of each gene with $|\gamma_1\rangle_N$ vs. that with $|\gamma_2\rangle_N$, for 784 cell cycle regulated genes, color-coded according to the classification by Spellman *et al.* (3).

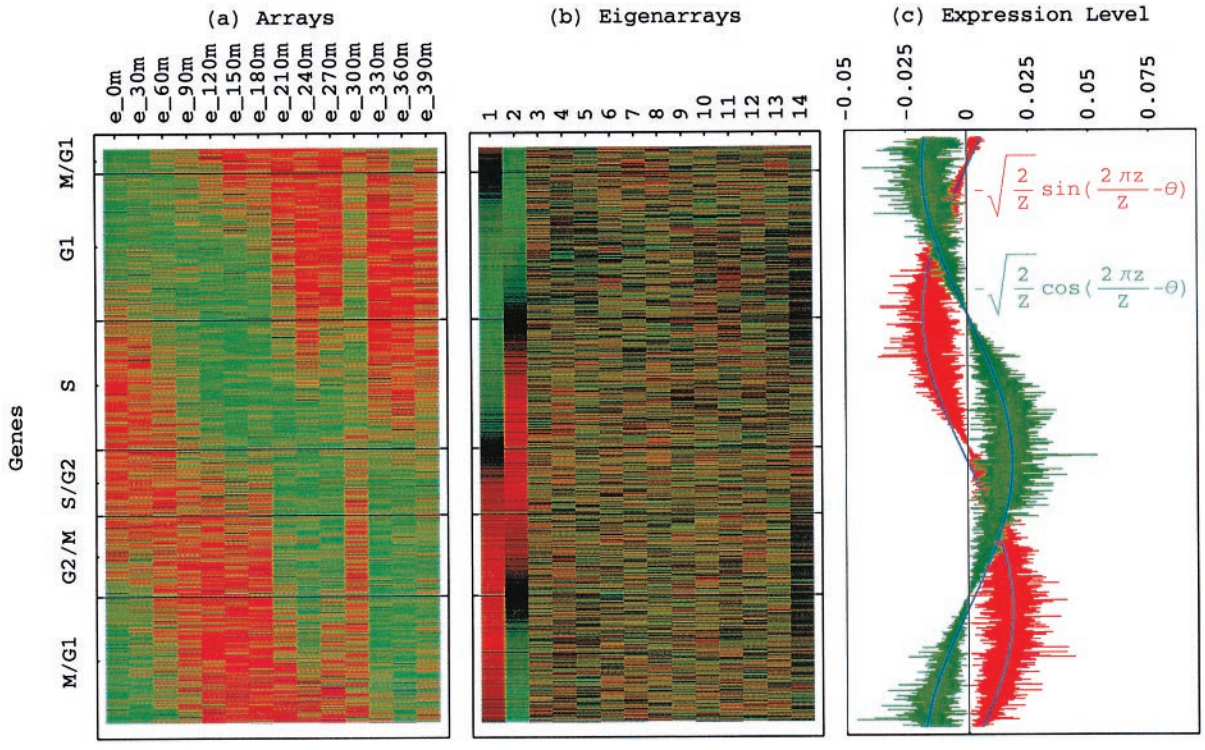

Fig. 3. Genes sorted by relative correlation with $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ of normalized elutriation. (*a*) Normalized elutriation expression of the sorted 5,981 genes in the 14 arrays, showing traveling wave of expression. (*b*) Eigenarrays expression; the expression of $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$, the eigenarrays corresponding to $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$, displays the sorting. (*c*) Expression levels of $|\alpha_1\rangle_N$ (red) and $|\alpha_2\rangle_N$ (green) fit normalized sine and cosine functions of period $Z \equiv N - 1 = 5{,}980$ and phase $\theta \approx 2\pi/13$ (blue), respectively.

$|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$, of $\hat{e}_N$ (Fig. 1*a*), which are of similar significance, capture together more than 40% of the overall normalized expression (Fig. 1*b*). The time variations of $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ fit normalized sine and cosine functions of period T and initial phase $\theta \approx 2\pi/13$, $\sqrt{2/T}\sin(2\pi t/T - \theta)$ and $\sqrt{2/T}\cos(2\pi t/T - \theta)$, respectively (Fig. 1*c*). We infer that $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ represent cell cycle expression oscillations, and assume that the corresponding eigenarrays $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ represent the corresponding cell cycle cellular states. Upon sorting of the genes (and arrays) according to $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ (and $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$), the initial phase $\theta \approx 2\pi/13$ can be interpreted as a delay of 30 min between the start of the experiment and that of the cell cycle stage G_1. The decay to zero in the time variation of $|\gamma_2\rangle_N$ at $t = 360$ and 390 min can be interpreted as dephasing in time of the initially synchronized yeast culture.

Data Sorting. Consider the normalized expression of the 14 elutriation arrays $\{|a_m\rangle\}$ in the subspace spanned by $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$, which is assumed to approximately represent all cell cycle cellular states (Fig. 2*a*). All arrays have at least 25% of their normalized expression in this subspace, with their distances from the origin satisfying $0.5 \leq r_m < 1$, except for the eleventh array $|a_{11}\rangle$. This suggests that $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ are sufficient to approximate the elutriation array expression. The sorting of the arrays according to their phases $\{\phi_m\}$, which describes the transition from the expression pattern $|\alpha_2\rangle_N$ to $|\alpha_1\rangle_N$ and back to $|\alpha_2\rangle_N$, gives an array order which is similar to that of the cell cycle time points measured by the arrays, an order that describes the progress of the cell cycle expression from the M/G_1 stage through G_1, S, S/G_2, and G_2/M and back to M/G_1.

Because $|\alpha_1\rangle_N$ is correlated with the arrays $|a_4\rangle$, $|a_5\rangle$, $|a_6\rangle$, and $|a_7\rangle$ and is anticorrelated with $|a_{13}\rangle$ and $|a_{14}\rangle$, we associate $|\alpha_1\rangle_N$ with the cell cycle cellular state of transition from G_1 to S, and $-|\alpha_1\rangle_N$ with the transition from G_2/M to M/G_1. Similarly, $|\alpha_2\rangle_N$ is correlated with $|a_2\rangle$ and $|a_3\rangle$, and therefore we associate $|\alpha_2\rangle_N$ with the transition from M/G_1 to G_1. Also, $|\alpha_2\rangle_N$ is anticorrelated with $|a_8\rangle$ and $|a_{10}\rangle$, and therefore we associate $-|\alpha_2\rangle_N$ with the transition from S to S/G_2. With these associations the phase of $|a_1\rangle$, $\phi_1 = -\theta \approx -2\pi/13$, corresponds to the 30-min delay between the start of the experiment and that of the cell cycle stage G_1, which is also present in the inferred cell cycle expression oscillations $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$.

Consider also the expression of the 5,981 genes $\{|g_n\rangle\}$ in the subspace spanned by $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$, which is inferred to approximately represent all cell cycle expression oscillations (Fig. 10 in supplemental material at www.pnas.org). One may expect that genes that have almost all of their normalized expression in this subspace with $r_n \approx 1$ are cell cycle regulated, and that genes that have almost no expression in this subspace with $r_n \approx 0$, are not regulated by the cell cycle at all. Indeed, of the 784 genes that were classified by Spellman *et al.* (3) as cell cycle regulated, 641 have more than 25% of their normalized expression in this subspace (Fig. 2*b*). We sort all 5,981 genes according to their phases $\{\phi_n\}$, to describe the transition from the expression pattern $|\gamma_2\rangle_N$ to that of $|\gamma_1\rangle_N$ and back to $|\gamma_2\rangle_N$, starting at $\phi_1 \approx -2\pi/13$. One may expect this to order the genes according to the stages in the cell cycle in which their expression patterns peak. However, for the 784 cell cycle regulated genes this sorting gives a classification of the genes into the five cell cycle stages, which is somewhat different than the classification by Spellman *et al.* This may be due to the poor quality of the elutriation expression data, as synchronization by elutriation was not very effective in this experiment. For the α factor-synchronized cell cycle expression there is much better agreement between the two classifications (Fig. 5*b*).

With all 5,981 genes sorted, the gene variations of $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ fit normalized sine and cosine functions of period $Z \equiv N - 1 = 5{,}980$ and initial phase $\theta \approx 2\pi/13$, $-\sqrt{2/Z}\sin(2\pi z/Z - \theta)$ and $\sqrt{2/Z}\cos(2\pi z/Z - \theta)$, respectively, where $z \equiv n - 1$ (Fig. 3 *b* and *c*). The sorted and normalized elutriation expression fit approximately a traveling wave of expression, varying

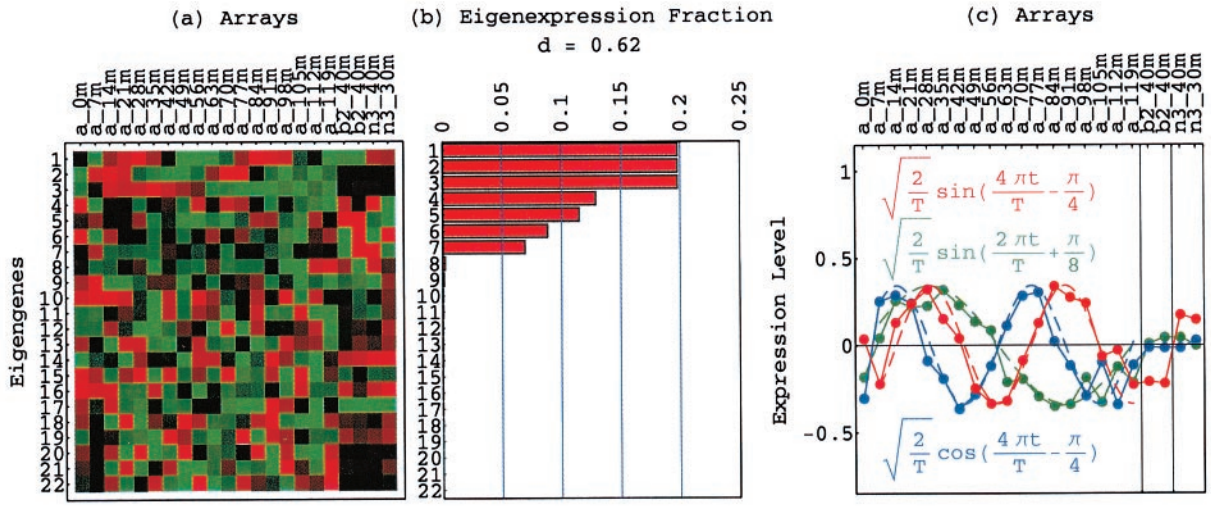

Fig. 4. Rotated normalized α factor, *CLB2*, and *CLN3* eigengenes. (*a*) Raster display of $\hat{v}^T_{RN}$, *where* $|\gamma_1\rangle_{RN} = \hat{R}_2\hat{R}_1|\gamma_1\rangle_N$, $|\gamma_2\rangle_{RN} = \hat{R}_1|\gamma_2\rangle_N$, and $|\gamma_3\rangle_{RN} = \hat{R}_2|\gamma_3\rangle_N$. (*b*) $|\gamma_1\rangle_{RN}$, $|\gamma_2\rangle_{RN}$ and $|\gamma_3\rangle_{RN}$ capture 20% of the overall normalized expression each. (*c*) Expression levels of $|\gamma_1\rangle_{RN}$ (red) and $|\gamma_2\rangle_{RN}$ (blue) fit dashed graphs of normalized sine (red) and cosine (blue) of period $T/2 = 66$ min and phase $\pi/4$, respectively, and $|\gamma_3\rangle_{RN}$ (green) fits dashed graph of normalized sine of period $T = 112$ min and phase $-\pi/8$, from $t = 7$ to $t = 119$ min during the cell cycle.

sinusoidally across both genes and arrays, such that the expression of the nth gene in the mth array satisfies $\langle n|\hat{e}_N|m\rangle \propto -2\cos[2\pi(t/T - z/Z)]/\sqrt{ZT}$ (Fig. 3*a*).

Biological Data Analysis: α Factor-Synchronized Cell Cycle and *CLB2* and *CLN3* Overactivations

Spellman *et al.* (3) also monitored genome-wide mRNA levels, for 6,108 yeast ORFs simultaneously, over approximately two cell cycle periods, in a yeast culture synchronized by α factor, relative to a reference mRNA from an asynchronous yeast culture, at 7-min intervals for 119 min. They also measured, in two independent experiments, mRNA levels of yeast strain cultures with overactivated *CLB2*, which encodes a G_2/M cyclin, both at $t = 40$ min relative to their levels at the start of overactivation at $t = 0$. Two additional independent experiments measured mRNA levels of strain cultures with overactivated *CLN3*, which encodes a G_1/S cyclin, at $t = 30$ and 40 min relative to their levels at the start of overactivation at $t = 0$. The dataset for the α factor, *CLB2*, and *CLN3* experiments we analyze (see supplemental data and Mathematica notebook at www.pnas.org) tabulates the ratios of gene expression levels for the $N = 4{,}579$ genes, 638 of which were classified by Spellman *et al.* as cell cycle regulated, with no missing data in the $M = 22$ arrays.

After data normalization and degenerate subspace rotation (see *Appendix* in supplemental material at www.pnas.org), the time variations of $|\gamma_1\rangle_{RN}$ and $|\gamma_2\rangle_{RN}$ fit normalized sine and cosine functions of two 66-min periods during the cell cycle, from $t = 7$ to 119 min, and initial phase $\theta \approx \pi/4$, respectively (Fig. 4*c*). While $|\gamma_2\rangle_{RN}$ describes steady-state expression in the *CLB2*- and *CLN3*-overactive arrays, $|\gamma_1\rangle_{RN}$ describes underexpression in the *CLB2*-overactive arrays and overexpression in the *CLN3*-overactive arrays.

Upon sorting the 4,579 genes in the subspace spanned by $|\gamma_1\rangle_{RN}$ and $|\gamma_2\rangle_{RN}$ (Fig. 5*b*), $|\gamma_1\rangle_{RN}$ is correlated with genes that peak late in the cell cycle stage G_1 and early in S, among them *CLN3*, and we associate $|\gamma_1\rangle_{RN}$ with the cell cycle expression oscillations that start at the transition from G_1 to S and are dependent on *CLN3*, which encodes a G_1/S cyclin. Also, $|\gamma_1\rangle_{RN}$ is anticorrelated with genes that peak late in G_2/M and early in M/G_1, among them *CLB2*, and therefore we associate $-|\gamma_1\rangle_{RN}$ with the oscillations that start at the transition from G_2/M to M/G_1 and are dependent on *CLB2*, which encodes a G_2/M cyclin. Similarly, $|\gamma_2\rangle_{RN}$ is correlated with genes that peak late in M/G_1 and early in G_1, anticorrelated with genes that peak late in S and early in S/G_2, and uncorrelated with *CLB2* and *CLN3*. We, therefore, associate $|\gamma_2\rangle_{RN}$ with the oscillations that start at the transition from M/G_1 to G_1 (and appear to be *CLB2*- and *CLN3*-independent), and $-|\gamma_2\rangle_{RN}$ with the oscillations that start at the transition from S to S/G_2 (and appear to be *CLB2*- and *CLN3*-independent).

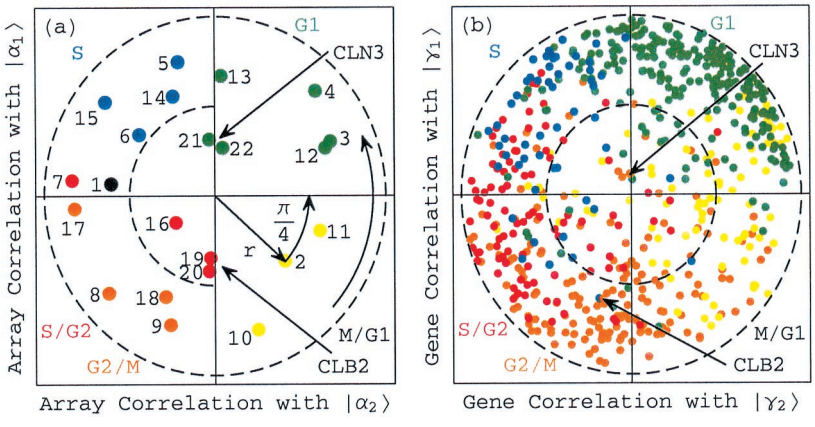

Fig. 5. Rotated normalized α factor, *CLB2*, and *CLN3* expression in the subspace associated with the cell cycle. (*a*) Array correlation with $|\alpha_1\rangle_{RN}$ along the *y*-axis vs. that with $|\alpha_2\rangle_{RN}$ along the *x*-axis, color-coded according to the classification of the arrays 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 outline 100% and 25% of overall normalized array expression in the $|\alpha_1\rangle_{RN}$ and $|\alpha_2\rangle_{RN}$ subspace. (*b*) Correlation of each gene with $|\gamma_1\rangle_{RN}$ vs. that with $|\gamma_2\rangle_{RN}$, for 638 cell cycle regulated genes, color-coded according to the classification by Spellman *et al.* (3).

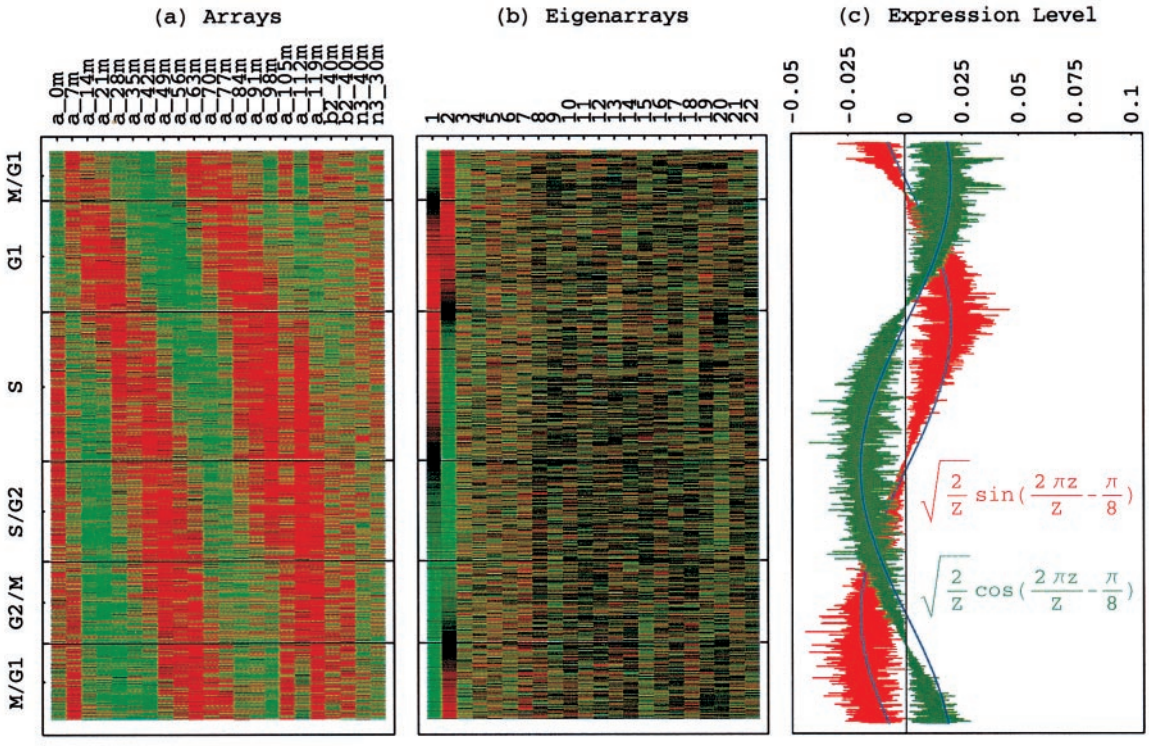

Fig. 6. Genes sorted by relative correlation with $|\gamma_1\rangle_{RN}$ and $|\gamma_2\rangle_{RN}$ of rotated normalized α factor, *CLB2*, and *CLN3*. (*a*) Normalized expression of the sorted 4,579 genes in the 22 arrays, showing traveling wave of expression from $t = 0$ to 119 min during the cell cycle and standing waves of expression in the *CLB2*- and *CLN3*-overactive arrays. (*b*) Eigenarrays expression; the expression of $|\alpha_1\rangle_{RN}$ and $|\alpha_2\rangle_{RN}$, the eigenarrays corresponding to $|\gamma_1\rangle_{RN}$ and $|\gamma_2\rangle_{RN}$, displays the sorting. (*c*) Expression levels of $|\alpha_1\rangle_{RN}$ (red) and $|\alpha_2\rangle_{RN}$ (green) fit normalized sine and cosine functions of period $Z \equiv N - 1 = 4{,}578$ and phase $\pi/8$ (blue), respectively.

Upon sorting the 22 arrays in the subspace spanned by $|\alpha_1\rangle_{RN}$ and $|\alpha_2\rangle_{RN}$ (Fig. 5*a*), $|\alpha_1\rangle_{RN}$ is correlated with the arrays $|a_{13}\rangle$ and $|a_{14}\rangle$, as well as with $|a_{21}\rangle$ and $|a_{22}\rangle$, which measure the *CLN3*-overactive samples. We therefore associate $|\alpha_1\rangle_{RN}$ with the cell cycle cellular state of transition from G_1 to S, which is simulated by *CLN3* overactivation. Also, $|\alpha_1\rangle_{RN}$ is anticorrelated with the arrays $|a_9\rangle$ and $|a_{10}\rangle$, as well as with $|a_{19}\rangle$ and $|a_{20}\rangle$, which measure the *CLB2*-overactive samples. We associate $-|\alpha_1\rangle_{RN}$ with the cellular transition from G_2/M to M/G_1, which is simulated by *CLB2* overactivation. Similarly, $|\alpha_2\rangle_{RN}$ appears to be correlated with $|a_2\rangle$, $|a_3\rangle$, $|a_{11}\rangle$, and $|a_{12}\rangle$, anticorrelated with $|a_6\rangle$, $|a_7\rangle$, $|a_{16}\rangle$, and $|a_{17}\rangle$, and uncorrelated with $|a_{19}\rangle$, $|a_{20}\rangle$, $|a_{21}\rangle$, or $|a_{22}\rangle$. We therefore associate $|\alpha_2\rangle_{RN}$ with the cellular transition from M/G_1 to G_1 (which appears to be *CLB2*- and *CLN3*-independent), and $-|\alpha_2\rangle_{RN}$ with the cellular transition from S to S/G_2 (which also appears to be *CLB2*- and *CLN3*-independent).

With all 4,579 genes sorted the gene variations of $|\alpha_1\rangle_{RN}$ and $|\alpha_2\rangle_{RN}$ fit normalized sine and cosine functions of period $Z \equiv N - 1 = 4{,}578$ and initial phase $\pi/8$, respectively (Fig. 6 *b* and *c*). The normalized and sorted cell cycle expression approximately fits a traveling wave, varying sinusoidally across both genes and arrays. The normalized and sorted expression in the *CLB2*- and *CLN3*-overactive arrays approximately fits standing waves, constant across the arrays and varying sinusoidally across genes only, which appear similar to $-|\alpha_1\rangle_{RN}$ and $|\alpha_1\rangle_{RN}$, respectively (Fig. 6*a*).

Conclusions

We have shown that SVD provides a useful mathematical framework for processing and modeling genome-wide expression data, in which both the mathematical variables and operations may be assigned biological meaning.

We thank S. Kim for insightful discussions, G. Sherlock for technical assistance and careful reading, and J. Doyle and P. Green for thoughtful reviews of this manuscript. This work was supported by a grant from the National Cancer Institute (National Institutes of Health, CA77097). O.A. is an Alfred P. Sloan and U.S. Department of Energy Postdoctoral Fellow in Computational Molecular Biology, and a National Human Genome Research Institute Individual Mentored Research Scientist Development Awardee in Genomic Research and Analysis (National Institutes of Health, 1 K01 HG00038-01). P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute.

1. Fodor, S. P., Rava, R. P., Huang, X. C., Pease, A. C., Holmes, C. P. & Adams, C. L. (1993) *Nature (London)* **364,** 555–556.
2. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270,** 467–470.
3. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998) *Mol. Biol. Cell* **9,** 3273–3297.
4. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. (1998) *Nat. Biotechnol.* **16,** 939–945.
5. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 14863–14868.
6. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. & Levine, A. J. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 6745–6750.
7. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. & Golub, T. R. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 2907–2912.
8. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) *Nat. Genet.* **22,** 281–285.
9. Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., Jr., & Haussler, D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 262–267.
10. Golub, G. H. & Van Loan, C. F. (1996) *Matrix Computation* (Johns Hopkins Univ. Press, Baltimore), 3rd Ed.
11. Mallat, S. G. (1999) *A Wavelet Tour of Signal Processing* (Academic, San Diego), 2nd Ed.
12. Anderson, T. W. (1984) *Introduction to Multivariate Statistical Analysis* (Wiley, New York), 2nd Ed.