Singular value decomposition of genome-scale mRNA lengths distribution reveals asymmetry in RNA gel electrophoresis band broadening

Orly Alter*† and Gene H. Golub†‡

*Department of Biomedical Engineering, Institute for Cellular and Molecular Biology and Institute for Computational Engineering and Sciences, University of Texas, Austin, TX 78712; and ‡Scientific Computing and Computational Mathematics Program and Department of Computer Science, Stanford University, Stanford, CA 94305

Contributed by Gene H. Golub, June 7, 2006

We describe the singular value decomposition (SVD) of yeast genome-scale mRNA lengths distribution data measured by DNA microarrays. SVD uncovers in the mRNA abundance levels data matrix of genes × arrays, i.e., electrophoretic gel migration lengths or mRNA lengths, mathematically unique decorrelated and decoupled ''eigengenes.'' The eigengenes are the eigenvectors of the arrays × arrays correlation matrix, with the corresponding series of eigenvalues proportional to the series of the ''fractions of eigen abundance.'' Each fraction of eigen abundance indicates the significance of the corresponding eigengene relative to all others. We show that the eigengenes fit ''asymmetric Hermite functions,'' a generalization of the eigenfunctions of the quantum harmonic oscillator and the integral transform which kernel is a generalized coherent state. The fractions of eigen abundance fit a geometric series as do the eigenvalues of the integral transform which kernel is a generalized coherent state. The ''asymmetric generalized coherent state'' models the measured data, where the profiles of mRNA abundance levels of most genes as well as the distribution of the peaks of these profiles fit asymmetric Gaussians. We hypothesize that the asymmetry in the distribution of the peaks of the profiles is due to two competing evolutionary forces. We show that the asymmetry in the profiles of the genes might be due to a previously unknown asymmetry in the gel electrophoresis thermal broadening of a moving, rather than a stationary, band of RNA molecules.

DNA microarrays | yeast *Saccharomyces cerevisiae* | Hermite functions | generalized coherent states | evolutionary forces

Advances in sequencing technology (1), including DNA and RNA gel electrophoresis (2–6), fueled the Human Genome Project, promoted the resulting sequencing of numerous complete genomes, and stimulated the emergence of DNA microarray hybridization technology. This high-throughput technology makes it possible to assay the hybridization of DNA or RNA molecules, extracted from a single sample, with several thousands of probes simultaneously (7, 8). Different types of molecular biological signals, such as abundance levels of DNA, RNA, and DNA-bound proteins can now be measured on genomic scales (9, 10).

Recently, Hurowitz and Brown (11) described the use of DNA microarrays in the genome-scale measurement of the distribution of the lengths of mRNA gene transcripts in yeast. Electrophoresis was used to separate the transcripts by migration length in an agarose gel, where each migration length corresponds to a transcript length. The gel was cut into slices, and the relative abundance levels of the different yeast transcripts in each slice was measured with a DNA microarray. We describe the singular value decomposition (SVD) (12) of the mRNA abundance levels data matrix of genes × arrays, i.e., gel slices, electrophoretic migration lengths, or mRNA lengths. SVD separates the measured profiles of the genes into mathematically unique decorrelated and decoupled "eigengenes," which are the eigenvectors of the arrays × arrays correlation matrix. The corresponding series of eigenvalues are proportional to the series of "fractions of eigen abundance," each of which indicates the significance of the corresponding eigengene relative to all eigengenes. Recently, we illustrated the possible correspondence between significant eigengenes uncovered in DNA microarray data and the independent biological and experimental processes that compose the data with an analysis of genome-scale mRNA expression from yeast during its cell cycle program (ref. 13, and see also refs. 14 and 15).

Here we show that the eigengenes of the yeast mRNA lengths distribution data fit "asymmetric Hermite functions." Hermite functions are the eigenfunctions of the quantum harmonic oscillator (17, 18) and the integral transform which kernel is a generalized coherent state (19, 20). We show that the corresponding fractions of eigen abundance fit a geometric series, as do the eigenvalues of the integral transform which kernel is a generalized coherent state. We show that, as follows from the uniqueness of SVD, the "asymmetric generalized coherent state" model fits the measured genome-scale distribution of the lengths of mRNA gene transcripts in yeast, i.e., that (*i*) the profiles of mRNA abundance levels of most genes fit asymmetric Gaussians; and (*ii*) the distribution of the peaks of the profiles of the genes fits an asymmetric Gaussian that peaks approximately at the mRNA length of 1,000 ± 50 nucleotides. We hypothesize that the asymmetric Gaussian distribution of the peaks of the profiles of the genes is due to two competing evolutionary forces that balance at the peak of this distribution.

Recently, we predicted a previously unknown biological principle by modeling DNA microarray data. Integrating genome-scale proteins' DNA-binding data with cell cycle mRNA expression time course data from yeast, using pseudoinverse projection, we predicted a previously unknown correlation between DNA replication initiation and RNA transcription, which might be due to an undiscovered mechanism of regulation (21). Now we reveal a previously unknown physical principle by modeling DNA microarray data. We show that the asymmetry in the profiles of mRNA abundance levels of the genes across the arrays, i.e., gel slices, might be due to a previously unknown asymmetry in the thermal broadening of a moving band of mRNA molecules. The peak of the symmetric Gaussian profile of a stationary band shifts within the moving band, changing its profile into an asymmetric Gaussian. We conclude that the mathematical modeling of DNA microarray data might be used to uncover the physical as well as the biological principles that govern the activities of DNA and RNA.

SVD of Genome-Scale mRNA Lengths Distribution in Yeast

Hurowitz and Brown (11) measured relative mRNA abundance levels for 9,867 putative genes of the yeast *Saccharomyces cerevisiae*,

Conflict of interest statement: No conflicts declared.

Freely available online through the PNAS open access option.

Abbreviation: SVD, singular value decomposition.

†To whom correspondence may be addressed. E-mail: orlyal@mail.utexas.edu or golub@stanford.edu.

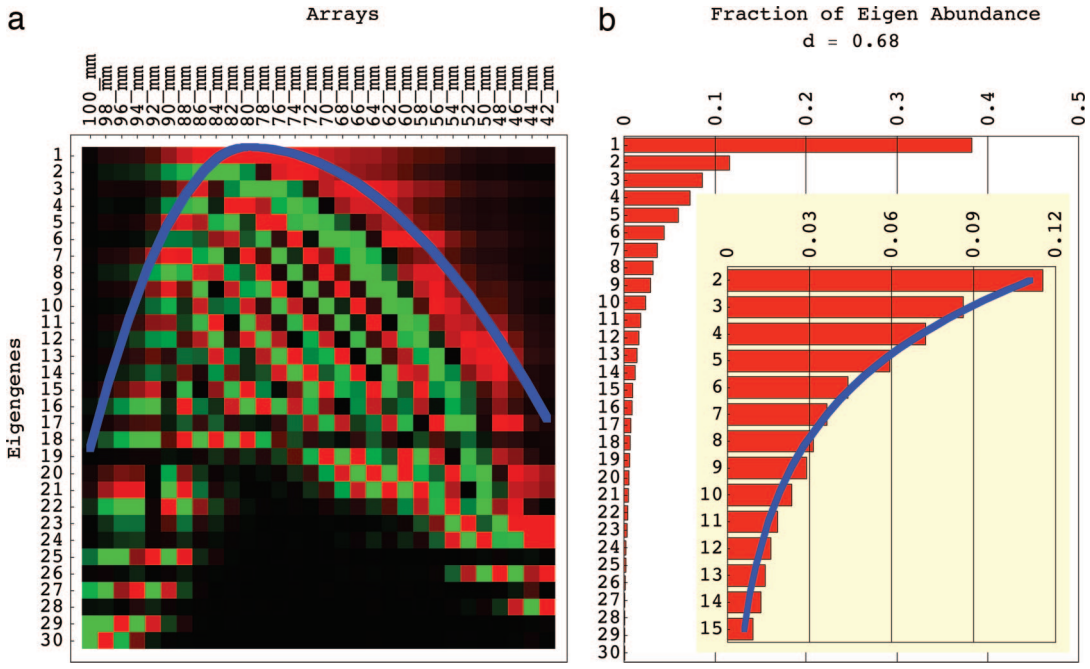

Fig. 1. Eigengenes of the yeast genome-scale mRNA lengths distribution data. (*a*) Raster display of $\hat{v}^T$, the abundance of $X = 30$ eigengenes in 30 arrays, corresponding to 30 gel slices, with overabundance (red), no change in abundance (black), and underabundance (green) around the "ground state" of abundance, which is captured by the first eigengene $\langle 1|\hat{v}^T$. The inflection points of the eigengenes approximately sample a graph of the asymmetric parabolic potential $kx^2/2$, where $k = k_1$ for $x \leq 0$ and $k = k_2 = k_1/4$ for $x \geq 0$ (blue) at unit intervals. (*b*) Bar chart of the 30 fractions of eigen abundance $\{\Omega_n\}$, which approximately fit a graph of the exponential function of n, $\{c\lambda^n\}$ (blue).

over a range of 60 mm of an agarose gel cut into 30 slices of 2 mm each, spanning the electrophoretic migration range of 42–100 mm and the corresponding mRNA lengths range of 4,400–300 nucleotides, after 2 h of separation by length in an electric field of 9 V/cm. The data set we analyze tabulates the ratios of mRNA abundance levels for the $P = 6{,}776$ genes with no missing data in any of the $X = 30$ arrays corresponding to the 30 slices of gel. Let the matrix $\hat{d}$ of size P-genes $\times$ X-arrays tabulate the genome-scale mRNA lengths distribution of yeast (see www.bme.utexas.edu/research/orly/harmonic_oscillator). The vector in the pth row of the matrix $\hat{d}$, $\langle p|\hat{d}$ lists the profile of relative abundance levels of the pth gene transcript across the different gel slices which correspond to the different arrays.[§] The vector in the xth column of the matrix $\hat{d}$, $\hat{d}|x\rangle$, lists the relative abundance levels of the P gene transcripts as measured in the xth gel slice by the xth array.

We compute the singular value decomposition (SVD) of the data matrix $\hat{d} = \hat{u}\hat{\omega}\hat{v}^T$ (12) (see www.bme.utexas.edu/research/orly/harmonic_oscillator). The nth row of the orthogonal transformation matrix $\hat{v}^T$ lists the nth "eigengene" $\langle n|\hat{v}^T$, which is unique up to a phase factor of ± 1 (13) (Fig. 1*a*). The diagonal nonnegative matrix $\hat{\omega}$ lists the "eigen abundance" levels $\{\langle n|\hat{\omega}|n\rangle\}$ (Fig. 1*b*). The significance of the nth eigengene is indicated by the nth "fraction of eigen abundance" $\Omega_n = \langle n|\hat{\omega}|n\rangle^2/(\Sigma_{n=1}^{N}\langle n|\hat{\omega}|n\rangle^2)$, i.e., the abundance captured by the nth eigengene relative to that captured by all eigengenes. Note that the eigengenes are the eigenvectors of the X-arrays $\times$ X-arrays correlation matrix $\hat{d}^T\hat{d}$ with the corresponding series of eigenvalues proportional to the series of the fractions of eigen abundance.

Eigengenes Fit "Asymmetric Hermite Functions"

The nth Hermite function,

$$h_n(\sqrt{k}x) = \left(\frac{k}{2^{2n}n!^2\pi}\right)^{1/4} \exp\left(-\frac{kx^2}{2}\right) H_n(\sqrt{k}x), \qquad [1]$$

where $H_n(\sqrt{k}x)$ is the nth Hermite polynomial,

$$H_n(\sqrt{k}x) = \frac{n!}{2\pi i}\oint z^{-n-1}\exp(-z^2 + 2z\sqrt{k}x)dz, \qquad [2]$$

is a solution of the differential equation that describes the generalized coordinate x of the quantum harmonic oscillator (17, 18) with the generalized Hooke's constant k,

$$\left(-\frac{1}{2k}\frac{d^2}{dx^2} + \frac{kx^2}{2}\right)h_n(\sqrt{k}x) = \left(n + \frac{1}{2}\right)h_n(\sqrt{k}x). \qquad [3]$$

Substituting $d^2[h_n(\sqrt{k}x)]/dx^2 = 0$ in Eq. **3**, it can be shown that the inflection points of the Hermite functions $h_n(\sqrt{k}x)$ sample the parabolic potential of the harmonic oscillator at unit intervals, where $kx^2/2 = (n + 1/2)$. The Hermite functions form an orthogonal basis in the range $x \in (-\infty, \infty)$, and are, therefore, eigenfunctions of the quantum harmonic oscillator in the generalized coordinate representation.

We fit the nth eigengene with the $(n-1)$th continuous "asymmetric Hermite function," where the generalized Hooke's constant is asymmetric with respect to the equilibrium $x = 0$ (Fig. 2),

$$k^{-1/4}h_{n-1}(\sqrt{k}x), \quad k = \begin{cases} k_1, & x \leq 0 \\ k_2, & x \geq 0 \end{cases}. \qquad [4]$$

These asymmetric Hermite functions form only an approximately orthogonal basis. The $(n-1)$th asymmetric Hermite function is, therefore, normalized after discretization by sampling at unit intervals in the range $x \in [-11, 18]$, where the equilibrium $x = 0$ is set at the gel migration length of 78 mm, and then fit to the nth eigengene, for $n = 1, \ldots, 10$. We find that the "asymmetric generalized Hooke's constant" is approximately $k = k_1 \approx 0.36$ for $x \leq 0$ and $k = k_2 \approx k_1/4 \approx 0.09$ for $x \geq 0$. The arithmetic mean of the correlations between the $(n-1)$th asymmetric Hermite function and the nth eigengene for $n = 1, \ldots, 10$ is $\approx$0.78. The

[§]In this manuscript, $\hat{m}$ 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.

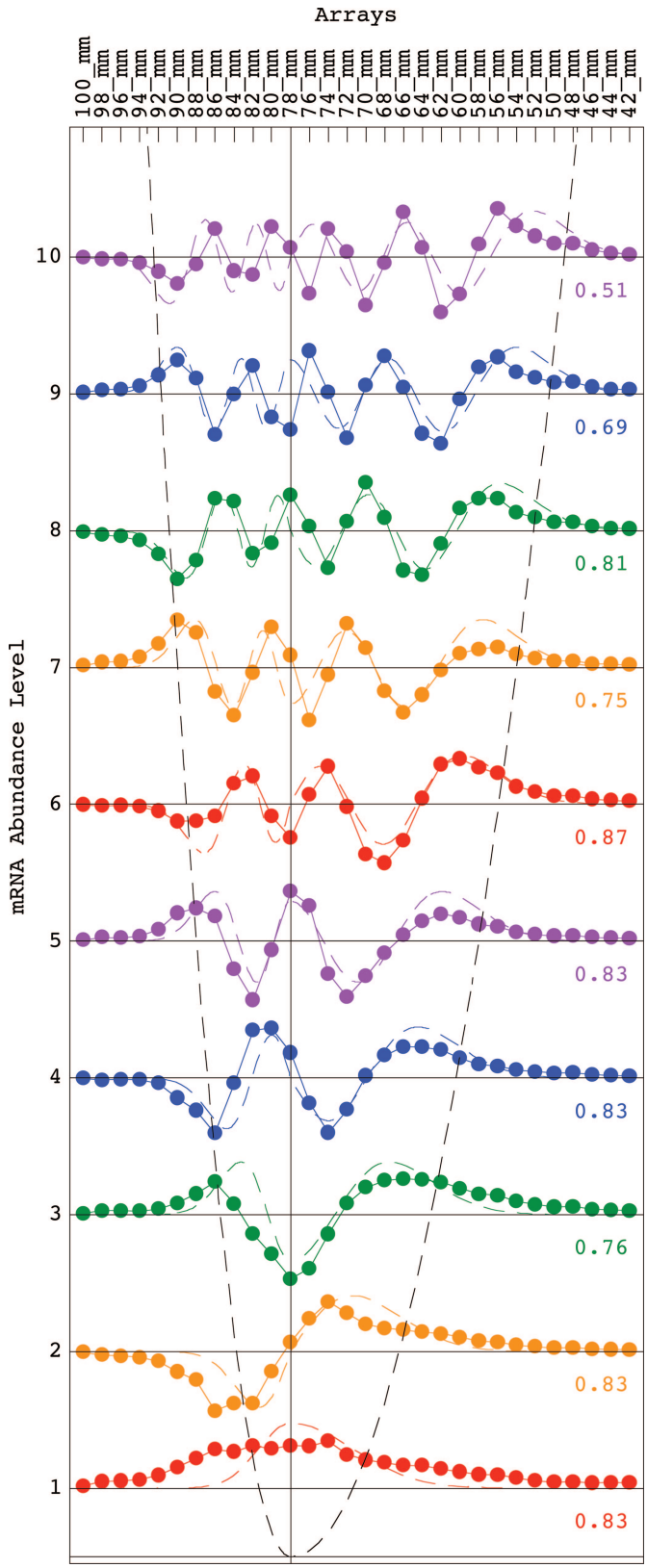

Fig. 2. Line-joined graphs of the abundance levels of the 1st (red) through 10th (violet) eigengenes of the yeast genome-scale mRNA lengths distribution data, $\{\langle n|\hat{v}^T\}$ for $n = 1, \ldots, 10$, approximately fit dashed graphs of the 0th (red) through 9th (violet) asymmetric Hermite functions of Eq. **4** with correlations ranging from 0.51 to 0.87. The inflection points of the eigengenes approximately sample a dashed graph of the asymmetric parabolic potential at unit intervals.

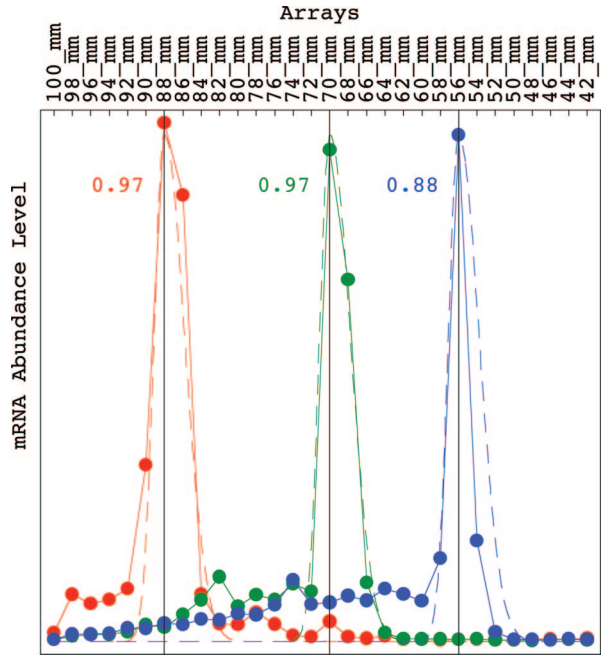

Fig. 3. Line-joined graphs of the measured profiles of abundance levels of the yeast genes *VMA7* (red), *ADE12* (green), and *HIS4* (blue) approximately fit dashed graphs of the asymmetric Gaussian $\exp[-a(x-p)^2]$, where $a = a_1$ for $x \leq p$ and $a = a_2 = a_1/4$ for $x \geq p$, and where the Gaussian peak is set at the gel migration lengths of 88, 70, and 56 mm, respectively, with the corresponding correlations of 0.97, 0.97, and 0.88.

inflection points of the eigengenes approximately sample the corresponding continuous "asymmetric parabolic potential" $kx^2/2$, where $k = k_1$ for $x \leq 0$ and $k = k_2 = k_1/4$ for $x \geq 0$, at unit intervals (Figs. 1*a* and 2).

Asymmetric Generalized Coherent State Model of the Genome-Scale mRNA Lengths Distribution

Assume that the profile of mRNA abundance levels measured for the pth gene across the X slices of gel approximately fits a Gaussian of the variable x which peaks at $x = p$ with the variance $\sigma_x^2 = 1/2a > 0$, $\exp[-a(x-p)^2]$. Assume also that the distribution of the peaks across the P genes fits a Gaussian of the variable p which peaks at the equilibrium $p = x = 0$ with the variance $\sigma_p^2 = 1/2b > 0$, where $\sigma_p^2 \gg \sigma_x^2$ such that $a \gg b$. With these assumptions, the abundance level of the pth gene transcript as measured by the xth array is proportional to the generalized coherent state (19)

$$f(x,p) = \exp[-a(x-p)^2 - bp^2]. \qquad [5]$$

The correlation of the mRNA abundance levels measured by the xth and yth arrays across the P genes, in the limit of $P \to \infty$, is then proportional to the generalized coherent state

$$g(x,y) = \exp[-\alpha(x^2+y^2) + 2\beta xy]$$

$$\propto \int_{-\infty}^{\infty} f(x,p)f(y,p)dp, \qquad [6]$$

where $\alpha = (a^2 + 2ab)/[2(a+b)]$ and $\beta = a^2/[2(a+b)]$. For $a \gg b$, $\alpha \approx a/2 + b$ and $\beta \approx a/2$. Using Eq. **2**, it can be shown that the Hermite functions $h_n(\sqrt{k}x)$ are the eigenfunctions of the integral transform which kernel is $g(x,y)$ in the range $x \in (-\infty, \infty)$ with the corresponding eigenvalues proportional to the geometric series $\{\lambda^n\}$ (20)

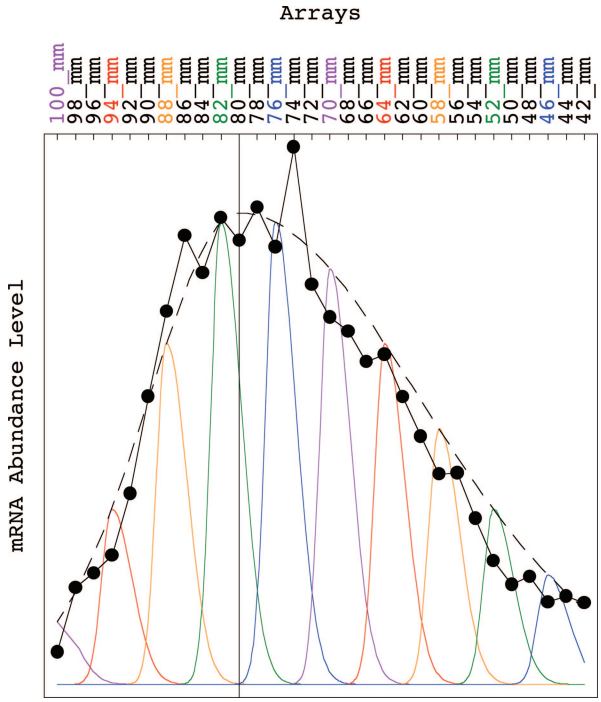

Fig. 4. Asymmetric generalized coherent state model of the yeast genome-scale mRNA lengths distribution data of Eq. **9**. Line-joined graph of the arithmetic mean of the profiles of mRNA abundance levels of the P genes of Eq. **8** approximately fits a dashed graph of the asymmetric Gaussian $\exp(-bp^2)$, which models the distribution of the peaks of the profiles of the genes, where $b = b_1$ for $x \leq 0$ and $b = b_2 = b_1/4$ for $x \geq 0$ and where the Gaussian peak is set at the gel migration length of 80 mm. Graphs of the asymmetric Gaussian $\exp[-a(x-p)^2]$ model the profiles of the genes, where the Gaussian peaks are set at 100 mm (violet) through 46 mm (blue), and where the Gaussian amplitudes are determined by the distribution of the peaks, which models the relative abundance of each Gaussian peak among all genes.

$$\int_{-\infty}^{\infty} g(x,y)h_n(\sqrt{k}y)dy \propto \lambda^n h_n(\sqrt{k}x), \qquad [7]$$

where $k = 2\sqrt{\alpha^2 - \beta^2}$ and $\lambda = \sqrt{(\alpha - k/2)/(\alpha + k/2)}$. For $a \gg b$, $k \approx 2\sqrt{ab}$ and $\lambda \approx 1 - \sqrt{b/a}$.

Fractions of Eigen Abundance Fit a Geometric Series. Fitting the fractions of eigen abundance $\{\Omega_n\}$ with the geometric series of Eq. **7** $\{c\lambda^n\}$ for $n = 2, \ldots, 10$, we find that $\lambda \approx 0.8$. The correlation between the fractions of eigen abundance and the geometric series for $n = 2, \ldots, 10$ is >0.99 (Fig. 1*b*). From $k = k_1 \approx 0.36$ for $x \leq 0$, $k = k_2 \approx k_1/4 \approx 0.09$ for $x \geq 0$ and $\lambda \approx 0.8$ for all x, we calculate that $a = a_1 \approx 1.6$ for $x \leq p$, and $a = a_2 \approx a_1/4 \approx 0.4$ for $x \geq p$, and that $b = b_1 \approx 0.02$ for $x \leq 0$, and $b = b_2 \approx b_1/4 \approx 0.005$ for $x \geq 0$. Note that $a_1/b_1 \approx a_2/b_2 \approx 80$, such that $a \gg b$.

Note also that, for the fractions of eigen abundance computed from the generalized coherent state of Eq. **5** after discretization to fit approximately the geometric series of Eq. **7**, it follows that $\lambda \approx 1 - \sqrt{b/a}$ is symmetric with respect to the equilibrium $p = x = 0$, whereas for the asymmetric Hermite functions of Eq. **4** to fit approximately the eigengenes computed from the generalized coherent state of Eq. **5** after discretization it follows that $k \approx 2\sqrt{ab}$ is asymmetric.

Profiles of mRNA Abundance Levels of Most Genes Fit Asymmetric Gaussians. Hurowitz and Brown (11) observed that, for most genes, the profile of mRNA abundance levels of each gene, $\langle p|\hat{d}$, peaks at only one of the X gel slices. We fit the profiles of mRNA abundance levels measured for the yeast genes *VMA7*, *ADE12* and *HIS4*, which were selected by Hurowitz and Brown as representative genes, with the asymmetric Gaussians $\exp[-a(x-p)^2]$, where $a = a_1$ for $x \leq p$ and $a = a_2 \approx a_1/4$ for $x \geq p$, sampled at unit intervals in the ranges $x \in [-7, 22]$, $[-16, 13]$, and $[-23, 6]$, such that their peaks are set at the gel migration lengths of 88, 70, and 56 mm, which according to Hurowitz and Brown correspond approximately to the mRNA lengths of 675 ± 35, 1,500 ± 60, and 2,575 ± 100 nucleotides,

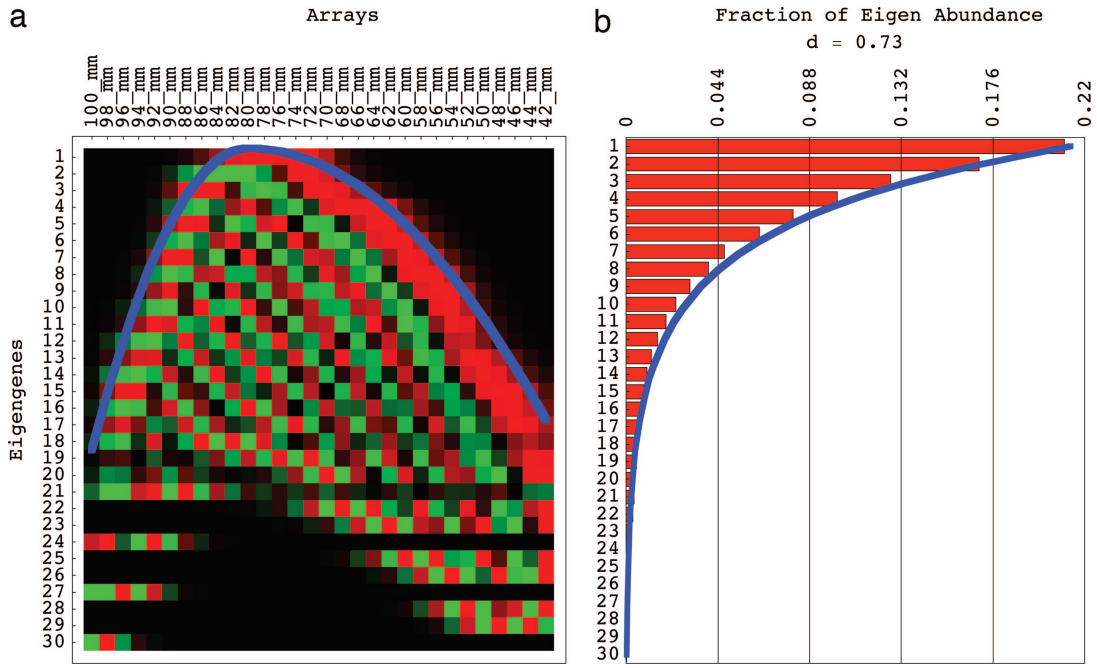

Fig. 5. Eigengenes of the discretized approximated asymmetric generalized coherent state of Eq. **10**. (*a*) Raster display of the 30 eigengenes in 30 arrays, the inflection points of which approximately sample a graph of the asymmetric parabolic potential (blue) at unit intervals. (*b*) Bar chart of the 30 fractions of eigen abundance, which approximately fit a graph of the exponential function of n (blue).

respectively. We find that the arithmetic mean of the correlations between the measured profile for each of these three genes and the corresponding discretized Gaussian is ≈0.94 (Fig. 3).

Distribution of the Peaks of the Profiles of the Genes Fits an Asymmetric Gaussian. From Eq. **5**, the arithmetic mean of the profiles of mRNA abundance levels of the P genes, in the limit of $P \to \infty$, is approximately proportional to the distribution of the peaks $x = p$ across the X gel slices for $a \gg b$

$$\frac{1}{P}\sum_{p=1}^{P}\langle p|\hat{d} \propto \int_{-\infty}^{\infty} f(x,p)dp$$

$$\propto \exp[-ab(x^2/(a+b)] \approx \exp(-bx^2). \qquad [8]$$

We fit the arithmetic mean of the profiles of mRNA abundance levels of the P genes with the asymmetric Gaussian $\exp(-bx^2)$, where $b = b_1$ for $x \le 0$ and $b = b_2 = b_1/4$ for $x \ge 0$, sampled at unit intervals in the range $x \in [-10, 19]$, where the equilibrium is set at the gel migration length of 80 mm. We find that the correlation is >0.99 (Fig. 4).

Genome-Scale mRNA Lengths Distribution Fits an Asymmetric Generalized Coherent State. We fit the mRNA lengths distribution data with the analytical continuous asymmetric generalized coherent state $f(x, p)$ of Eq. **5**, where the variances $\sigma_x^2 = 1/2a$ and $\sigma_p^2 = 1/2b$ are asymmetric with respect to the peaks $x = p$ of the profiles of the genes and the equilibrium $p = x = 0$, respectively (Fig. 4),

$$a \approx \begin{cases} k_1/(1-\lambda), & x \le p \\ k_2/(1-\lambda), & x \ge p \end{cases},$$
$$b \approx \begin{cases} (1-\lambda)k_1/2, & x \le 0 \\ (1-\lambda)k_2/2, & x \ge 0 \end{cases}. \qquad [9]$$

SVD of the asymmetric generalized coherent state is computed after discretization by sampling at unit intervals in the range $x \in [-10, 19]$, where the equilibrium is set at the gel migration length of 80 mm (Fig. 7, which is published as supporting information on the PNAS web site). We find that the arithmetic mean of the correlations between the nth eigengene computed for the discretized asymmetric generalized coherent state of Eq. **9** and the nth eigengene computed for the measured data for $n = 1, \ldots, 10$ is ≈0.73. The inflection points of the eigengenes computed for the discretized asymmetric generalized coherent state approximately sample the asymmetric parabolic potential $kx^2/2$, where $k = k_1$ for $x \le 0$ and $k = k_2 = k_1/4$ for $x \ge 0$ at unit intervals (Figs. 7*a* and 8, which are published as supporting information on the PNAS web site). We find that the correlation between the fractions of eigen abundance computed for the discretized asymmetric generalized coherent state and the geometric series $\{c\lambda^n\}$ for $n = 2, \ldots, 10$ is >0.99 (Fig. 7*b*).

Following Eq. **8**, we approximate the distribution of the peaks of the genes $\exp(-bp^2)$ in the asymmetric coherent state of Eq. **5** with the arithmetic mean of the profiles of the genes

$$\exp[-a(x-p)^2]\left(\frac{1}{P}\sum_{p=1}^{P}\langle p|\hat{d}\right). \qquad [10]$$

SVD is computed again after discretization of $\exp[-a(x-p)^2]$ by sampling at unit intervals in the range $x \in [-10, 19]$, where the equilibrium is set at the gel migration length of 80 mm (Fig. 9, which is published as supporting information on the PNAS web site). We find that the approximated asymmetric generalized coherent state of Eq. **10** fits the measured data better than the asymmetric generalized coherent state of Eqs. **5** and **9**: The arithmetic mean of

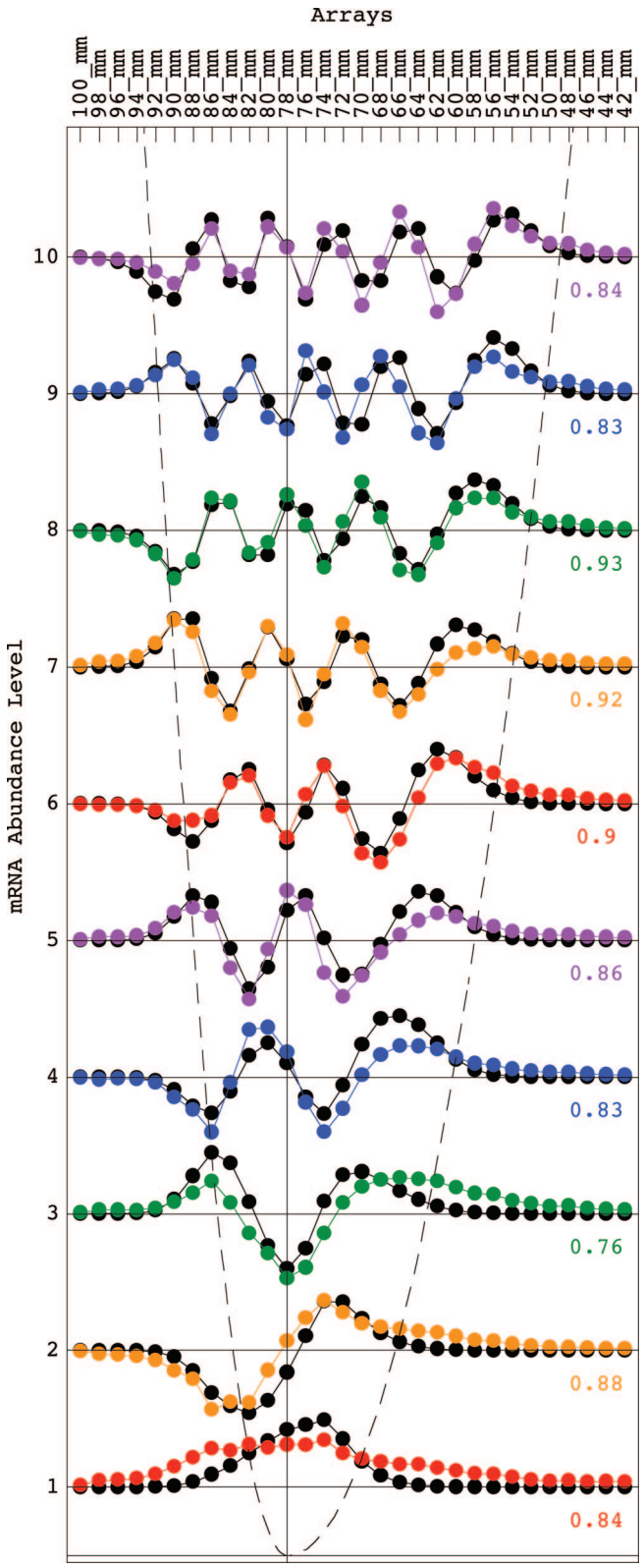

Fig. 6. Line-joined graphs of the abundance levels of the 1st (red) through 10th (violet) eigengenes of the yeast genome-scale mRNA lengths distribution data, $\{\langle n|\hat{v}^T\}$ for $n = 1, \ldots, 10$, approximately fit line-joined graphs of the abundance levels of the 1st through 10th eigengenes (black) of the discretized approximated asymmetric generalized coherent state of Eq. **10** with correlations ranging from 0.76 to 0.93. The inflection points of the eigengenes approximately sample a dashed graph of the asymmetric parabolic potential at unit intervals.

the correlations between the nth eigengene computed for the approximated and discretized asymmetric generalized coherent state of Eq. **10** and the nth eigengene computed for the measured data for $n = 1, \ldots, 10$ is ≈ 0.86 (Figs. 5 and 6). Note also the robustness of the eigengenes and the fractions of eigen abundance computed for the asymmetric generalized coherent state of Eq. **9** to the perturbation introduced by the approximation of Eq. **10**.

Why Does the Distribution of the Peaks of the Profiles of the Genes Fit an Asymmetric Gaussian?

Our hypothesis is that there are two competing evolutionary forces that determine the lengths of mRNA gene transcripts, and also the distribution of the peaks of the P profiles measured for the P genes. The first force acts to maximize the information content of the genes and their functional specificity, and therefore also their mRNA lengths. The second force acts to minimize the costs associated with the transcription as well as posttranscriptional processes, such as translation, and therefore also the lengths of mRNA gene transcripts. These forces balance at the peak of the distribution $\exp(-bp^2)$, i.e., the equilibrium $p = x = 0$. We find the equilibrium at the gel migration length of 80 mm, which according to Hurowitz and Brown (11) corresponds approximately to the mRNA length of 1,000 ± 50 nucleotides. Both forces are linearly proportional to and oppositely directed to the displacement of the gel migration length of a transcript from this equilibrium gel migration length in the manner of the restoring force of the harmonic oscillator. Note that the gel migration length of a transcript is approximately linearly proportional to the logarithm of the mRNA length of the transcript (3, 11). The proportionality constants are $b = b_1$ for the first force in the range of $x \leq 0$, and $b = b_2$ for the second force in the range of $x \geq 0$. We find that $b_2 \approx b_1/4$, suggesting that the first force, which acts to increase mRNA lengths that are shorter than the equilibrium length of 1,000 nucleotides, is larger than the second force, which acts to decrease mRNA lengths that are longer than the equilibrium length.

Why Do the Profiles of mRNA Abundance Levels of Most Genes Fit Asymmetric Gaussians?

Asymmetry in the gel electrophoresis thermal broadening of a moving, rather than a stationary, band of RNA molecules, along the axis of the electric field, might be underlying this previously unknown asymmetry in the profiles of mRNA abundance levels of the genes. When the electric field is turned off, the thermal broadening of a band of RNA molecules in an agarose gel is symmetric along the axis of the field, such that the distribution of the RNA molecules is a Gaussian profile, $\exp[-a(x - p)^2]$, which peaks at the position where the RNA molecules were loaded onto the gel, $x = p$. When the electric field is turned on, the RNA molecules migrate in the gel along the axis of the electric field in addition to their thermal diffusion (3–6). The electrophoretic velocity depends on the lengths of the RNA molecules. Molecules which are of the same lengths will migrate with the same velocity and form a moving band. In the thermal broadening of a moving band of RNA molecules the peak of the band is moving toward the front of the band and away from its back. As a result, the width of the band narrows in the direction of migration, $\sigma_{x,1} = \sqrt{1/2a_1}$ for $x \leq p$, and widens in the opposite direction, $\sigma_{x,2} = \sqrt{1/2a_2}$ for $x \geq p$, such that $\sigma_{x,2} > \sigma_{x,1}$, and the distribution of the RNA molecules is an asymmetric Gaussian profile. We find that $\sigma_{x,2}/\sigma_{x,1} = \sqrt{a_1/a_2} \approx 2$, and therefore that the peak of each Gaussian profile is approximately shifted by $(\sigma_{x,1} - \sigma_{x,2})/2$ within the band, which width is $\sigma_{x,1} + \sigma_{x,2}$. Substituting $a_1 \approx 0.16$ and $a_2 \approx 0.4$ and taking into account the 2-mm widths of the gel slices, we find that the width of the Gaussian profile is ≈3 mm, and that the peak of the Gaussian profile is shifted by ≈0.5 mm within the band.

Discussion

We have shown, using SVD, that the genome-scale distribution of yeast mRNA gene transcript lengths measured by DNA microarrays can be modeled mathematically as an asymmetric generalized coherent state, where the profiles of mRNA abundance levels of most genes as well as the distribution of the peaks of these profiles fit asymmetric Gaussians. We have hypothesized that the asymmetric Gaussian distribution of the peaks of the profiles of the genes is due to two competing evolutionary forces, which balance at the peak of this distribution, approximately at the mRNA length of 1,000 ± 50 nucleotides. SVD analyses of DNA microarray measured genome-scale distributions of the lengths of unspliced pre-mRNA, spliced mRNA exons and separately mRNA introns in different organisms (22), as well as different experimentally evolved strains of any one organism (23, 24), may be used to study the evolutionary forces which affect mRNA lengths.

We have shown that the asymmetry in the profiles of the genes might be due to a previously unknown asymmetry in the gel electrophoresis thermal broadening of a moving, rather than a stationary, band of RNA molecules. Previous simulations and measurements of DNA band broadening in gel electrophoresis have shown that the broadening of a moving band can be different from that of a stationary band, but have not suggested asymmetry (4–6). We conclude that the mathematical modeling of DNA microarray data might be used to uncover the physical as well as the biological principles which govern the activities of DNA and RNA.

We thank D. B. Oberman for recognizing that the eigengenes are Hermite functions, for recognizing that the asymmetric band broadening is due to a shift of the Gaussian peak within the moving band, and for many very insightful discussions. We also thank J. J. Collins, T. A. Duke, A. Goriely, J. Ross, and M. O. Vlad for thoughtful and thorough reviews of this manuscript and M. V. Berry, D. Botstein, P. O. Brown, G. M. Church, E. H. Hurowitz, V. R. Iyer, W. H. Press, G. W. Slater, and D. W. L. Sumners for helpful comments. This work was supported by National Science Foundation Grant CCR-0430617 (to G.H.G.) and a National Human Genome Research Institute Individual Mentored Research Scientist Development Award in Genomic Research and Analysis (K01 HG00038) (to O.A.).

1. Church, G. M. & Gilbert, W. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 1991–1995.
2. Southern, E. D. (1975) *J. Mol. Biol.* **98,** 503–517.
3. Lerman, L. S. & Frisch, H. L. (1982) *Biopolymers* **21,** 995–997.
4. Duke, T. A. & Viovy, J. L. (1992) *Phys. Rev. Lett.* **68,** 542–545.
5. Slater, G. W. (1993) *Electrophoresis* **14,** 1–7.
6. Tinland, B., Pernodet, N. & Pluen, A. (1998) *Biopolymers* **46,** 201–214.
7. Fodor, S. P., Rava, R. P., Huang, X. C., Pease, A. C., Holmes, C. P. & Adams, C. L. (1993) *Nature* **364,** 555–556.
8. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270,** 467–470.
9. Brown, P. O. & Botstein, D. (1999) *Nat. Genet.* **21,** 33–37.
10. Pollack, J. R. & Iyer, V. R. (2002) *Nat. Genet.* **32,** 515–521.
11. Hurowitz, E. H. & Brown, P. O. (2003) *Genome Biol.* **5,** R2.
12. Golub, G. H. & Van Loan, C. F. (1996) *Matrix Computation* (Johns Hopkins Univ. Press, Baltimore), 3rd Ed.
13. Alter, O., Brown, P. O. & Botstein, D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10101–10106.
14. Yeung, M. K., Tegner, J. & Collins, J. J. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6163–6168.
15. Vlad, M. O., Arkin, A. P. & Ross, J. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 7223–7228.
16. Alter, O. & Golub, G. H. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 17559–17564.
17. Schiff, L. I. (1968) *Quantum Mechanics* (McGraw–Hill, New York).
18. Alter, O. & Yamamoto, Y. (2001) *Quantum Measurement of a Single System* (Wiley, New York).
19. Glauber, R. J. (1963) *Phys. Rev.* **131,** 2766–2788.
20. Daubechies, I. (1988) *IEEE Trans. Inf. Theory* **34,** 605–612.
21. Alter, O. & Golub, G. H. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 16577–16582.
22. Yandell, M., Mungall, C. J., Smith, C., Prochnik, S., Kaminker, J., Hartzell, G., Lewis, S. & Rubin, G. M. (2006) *PLoS Comput. Biol.* **2,** E15.
23. Lenski, R. E. & Travisano, M. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 6808–6814.
24. Dunham, M. J., Badrane, H., Ferea, T., Adams, J., Brown, P. O., Rosenzweig, F. & Botstein, D. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 16144–16149.