Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription

Orly Alter*‡ and Gene H. Golub§

*Department of Biomedical Engineering and Institute for Cellular and Molecular Biology, 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, September 13, 2004

We describe an integrative data-driven mathematical framework that formulates any number of genome-scale molecular biological data sets in terms of one chosen set of data samples, or of profiles extracted mathematically from data samples, designated the ''basis'' set. By using pseudoinverse projection, the molecular biological profiles of the data samples are least-squares-approximated as superpositions of the basis profiles. Reconstruction of the data in the basis simulates experimental observation of only the cellular states manifest in the data that correspond to those of the basis. Classification of the data samples according to their reconstruction in the basis, rather than their overall measured profiles, maps the cellular states of the data onto those of the basis and gives a global picture of the correlations and possibly also causal coordination of these two sets of states. We illustrate this framework with an integration of yeast genome-scale proteins' DNA-binding data with cell cycle mRNA expression time course data. Novel correlation between DNA replication initiation and RNA transcription during the yeast cell cycle, which might be due to a previously unknown mechanism of regulation, is predicted.

singular value decomposition | generalized singular value decomposition | DNA microarrays | yeast *Saccharomyces cerevisiae* cell cycle

Recent advances in high-throughput technologies enable monitoring molecular biological signals, e.g., mRNA expression levels and proteins' DNA-binding occupancy levels, that correspond to activities of cellular systems, e.g., DNA replication, RNA transcription, and proteins' DNA-binding on a genomic scale. Integrative analysis of these global signals promises to give new insights into cellular mechanisms of regulation, i.e., global causal coordination of cellular activities. Integrative analysis of different types of large-scale molecular biological data requires mathematical tools that are able to formulate any number of large-scale data sets in terms of a common frame of reference, while reducing the complexity of the data to make them comprehensible (1, 2). These tools should provide data-driven models or mathematical frameworks for the description of the data, where the variables, i.e., the patterns that they uncover in the data, and operations, i.e., data reconstruction and classification in subspaces spanned by these patterns, may represent some biological reality.

Recently we showed that singular value decomposition (SVD) (3, 4) and generalized SVD (GSVD) (5) provide such data-driven frameworks for genome-scale molecular biological data. For example, the variables of SVD, "eigengenes" and corresponding "eigenarrays," in the analyses of yeast *Saccharomyces cerevisiae* cell cycle time course mRNA expression data (6), and those of GSVD, "genelets" and corresponding "arraylets," in the comparative analysis of yeast and human (7) cell cycle time course mRNA expression data, were shown to correlate with observed genome-scale effects of known cell cycle regulators and measured samples of the cell cycle stages that they regulate, respectively. Mathematical reconstruction of the yeast data in these subsets of eigengenes and corresponding eigenarrays, or genelets and corresponding arraylets, was shown to simulate approximately the experimental observation of the cell cycle progression alone, rather than the cell cycle progression together with concurrent biological processes and experimental artifacts. Mathematical classification of yeast genes and arrays according to their expression of these eigengenes and eigenarrays, or genelets and arraylets, rather than overall expression, mapped the data onto cell cycle stages and outlined the progression of the cell cycle along genes and in time, respectively.

Now we show that pseudoinverse projection (8) provides an integrative data-driven framework that formulates any number of genome-scale data sets in terms of a chosen set of data samples, or profiles extracted mathematically from data samples, which is designated the "basis" set. Pseudoinverse projection of a data set onto the basis set is a linear transformation of the data set from the open reading frames (ORFs) $\times$ data-samples space to the data-samples $\times$ basis-samples space, where each of the data samples is least-squares-approximated as a linear superposition of the basis profiles. We show that mathematical reconstruction of the data in the basis may simulate experimental observation of only the cellular states manifest in the data that correspond to those of the basis. Mathematical classification of the data samples according to their reconstruction in the basis, rather than their overall molecular biological profiles, maps the cellular states of the data onto those of the basis and gives a global picture of the correlations and possibly also coordination of these two sets of cellular states. Novel correlations between data samples and basis profiles might be due to previously unknown mechanisms of regulation.

We illustrate this framework with an integration of yeast genome-scale proteins' DNA-binding occupancy data (9) of nine cell cycle transcription factors (10) and four DNA replication initiation proteins (11) with the cell cycle time course mRNA expression data, using as basis sets the eigenarrays and arraylets determined by SVD and GSVD, respectively.¶

Mathematical Methods: Pseudoinverse Projection

Let the basis matrix $\hat{b}$, of size N-ORFs or genomic sites $\times$ M-basis profiles, tabulate M genome-scale molecular biological profiles, measured from a set of M samples or extracted mathematically from a set of M or more measured samples. The vector in the nth row of the matrix $\hat{b}$, $\langle n|\hat{b}$, lists the signal of the nth ORF across the different samples which correspond to the different arrays.‖ The vector in the mth column of the matrix $\hat{b}$, $|b_m\rangle \equiv \hat{b}|m\rangle$, lists the measured genome-scale signal levels of the mth basis sample. Let the data matrix $\hat{d}$, of size N-ORFs $\times$ L-data samples, tabulate a genome-scale molecular biological data set of a different type of data and

Abbreviations: SVD, singular value decomposition; GSVD, generalized SVD.

‡To whom correspondence should be addressed. E-mail: orlyal@mail.utexas.edu.

¶Alter, O., Golub, G. H., Brown, P. O. & Botstein, D., Miami Nature Biotechnology Winter Symposium: The Cell Cycle, Chromosomes and Cancer, Jan. 31–Feb. 4, 2004, Miami Beach, FL (www.med.miami.edu/mnbws/alter-.pdf).

‖In this article, $\hat{m}$ 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.

for the same ORFs in the same genome, measured in L samples. The vector in the lth column of the matrix $\hat{d}$, $|d_l\rangle \equiv \hat{d}|l\rangle$, lists the measured genome-scale signal levels of the lth data sample.

Moore–Penrose pseudoinverse projection (8) of the data matrix $\hat{d}$ onto the basis matrix $\hat{b}$ is then 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 (see Fig. 5 and *Appendix*, which are published as supporting information on the PNAS web site),

$$\hat{d} \to \hat{b}\hat{c}, \quad \hat{b}^\dagger\hat{d} \equiv \hat{c}, \qquad [1]$$

where the matrix $\hat{b}^\dagger$, i.e., the pseudoinverse of $\hat{b}$, satisfies

$$\hat{b}\hat{b}^\dagger\hat{b} = \hat{b}, \quad (\hat{b}\hat{b}^\dagger)^T = \hat{b}\hat{b}^\dagger, \qquad [2]$$

$$\hat{b}^\dagger\hat{b}\hat{b}^\dagger = \hat{b}^\dagger, \quad (\hat{b}^\dagger\hat{b})^T = \hat{b}^\dagger\hat{b},$$

such that the transformation matrices $\hat{b}\hat{b}^\dagger$ and $\hat{b}^\dagger\hat{b}$ are orthogonal projection matrices. The pseudoinverse of $\hat{b}$ is data-driven and unique. In this space the data matrix $\hat{d}$ is represented by the matrix $\hat{c}$, which tabulates the correlations between the M vectors that span the pseudoinverse $\hat{b}^\dagger$, $\{\langle\beta_m^\dagger|\} \equiv \{\langle m|\hat{b}^\dagger\}$, and the L profiles of the samples that span the data matrix $\hat{d}$, $\{|d_l\rangle\}$, such that $c_{ml} \equiv \langle m|\hat{c}|l\rangle = \langle\beta_m^\dagger|d_l\rangle$ for all $1 \le m \le M$ and $1 \le l \le L$.

Pseudoinverse Computation. We use the SVD of the basis matrix $\hat{b} = \hat{u}\hat{w}\hat{v}^T$, where $\hat{w}$ is a diagonal nonnegative matrix, $\hat{u}^T\hat{u} = \hat{v}^T\hat{v} = \hat{I}$, and $\hat{I}$ is the identity matrix, to compute the pseudoinverse $\hat{b}^\dagger = \hat{v}\hat{w}^{-1}\hat{u}^T$, such that Eq. **2** is satisfied, and $\hat{b}\hat{b}^\dagger = \hat{u}\hat{u}^T$ and $\hat{b}^\dagger\hat{b} = \hat{v}\hat{v}^T$ are orthogonal projection matrices. We then compute the pseudoinverse correlations $\hat{c}$ from $\hat{b}^\dagger$ and $\hat{d}$. We also compute the canonical correlation of each data profile with the basis, $0 \le \cos\theta_l = (\Sigma_{m=1}^M |\langle m|\hat{u}^T|d_l\rangle|^2/\langle d_l|d_l\rangle)^{-1/2} \le 1$.

Integrative Data Reconstruction. The pseudoinverse projection of $\hat{d}$ onto $\hat{b}$ allows reconstruction of the data, $\hat{d} \to \hat{b}\hat{b}^\dagger\hat{d}$, where each of the data samples is least-squares-approximated by a linear superposition of the basis profiles, $|d_l\rangle \to \Sigma_{m=1}^M c_{ml}|b_m\rangle$, without eliminating ORFs or samples. For reconstruction and visualization, we set the arithmetic mean of each ORF across the samples and that of each sample across the ORFs to zero, such that each ORF and sample in the reconstructed data set is centered at its sample- or ORF-invariant level, respectively.

Integrative Data Classification. The reconstructed data samples are classified by similarity in the contributions of the basis profiles to their overall measured profiles rather than by their overall measured profiles alone.

Consider a basis that is determined by SVD analysis of a set of measured samples, and is spanned by $M > 2$ eigenarrays, $\{|b_m\rangle\}$, two of which, $|b_1\rangle$ and $|b_2\rangle$, span a subspace of interest. We plot the correlation of $\langle\beta_2^\dagger|$ with each reconstructed data sample $|d_l\rangle$, $c_{2l}\langle d_l|\hat{b}\hat{b}^\dagger|d_l\rangle^{-1/2}$, along the y-axis vs. that of $\langle\beta_1^\dagger|$ along the x-axis. In this plot, the distance of each sample from the origin is its amplitude in the subspace spanned by $|b_1\rangle$ and $|b_2\rangle$ relative to its overall reconstructed amplitude, $r_l = (c_{1l}^2 + c_{2l}^2)^{1/2}\,\langle d_l|\hat{b}\hat{b}^\dagger|d_l\rangle^{-1/2}$. The angular distance of each sample from the x-axis is its phase in the transition from the profile $|b_1\rangle$ to $|b_2\rangle$ and back to $|b_1\rangle$, $\tan\phi_l = c_{2l}/c_{1l}$. We sort the reconstructed samples according to their angular distances from the x-axis, ϕ_l.

Consider also a basis that is determined by GSVD comparative analysis of two sets of measured samples and is spanned by $M > 2$ arraylets of one of these sets, $\{|b_m\rangle\}$. We approximate this basis with that spanned by the two vectors $\Sigma_{m=1}^M|b_m\rangle\langle\gamma_m|x\rangle$ and $\Sigma_{m=1}^M|b_m\rangle\langle\gamma_m|y\rangle$, where the vectors $|x\rangle$ and $|y\rangle$ least-squares-approximate the corresponding M-genelets subspace, $\{\langle\gamma_m|\}$, and maximize $\Sigma_{m=1}^M \langle\gamma_m|(|x\rangle\langle x| + |y\rangle\langle y|)|\gamma_m\rangle$. We plot the projection of each data sample, $|d_l\rangle$, from the M-arraylets subspace onto $\Sigma_{m=1}^M|b_m\rangle\langle\gamma_m|y\rangle$, that is N_l^{-1} $\Sigma_{m=1}^M c_{ml}\langle y|\gamma_m\rangle$ along the y-axis vs. that onto $\Sigma_{m=1}^M|b_m\rangle\langle\gamma_m|x\rangle$ along the x-axis, normalized by its ideal amplitude, where the contribution of each arraylet to the overall projected sample adds up rather than cancels out, $N_l^2 = \Sigma_{m=1}^M\Sigma_{k=1}^M c_{ml}c_{kl}|\langle\gamma_m|(|x\rangle\langle x| + |y\rangle\langle y|)|\gamma_k\rangle|$. In this plot, the distance of each sample from the origin, r_l, is the amplitude of its normalized projection. An amplitude of 1 indicates that the contributions of the arraylets add up, and an amplitude of 0 indicates that they cancel out. The angular distance of each sample from the x-axis, ϕ_l, is its phase in the transition from the profile $\Sigma_{m=1}^M|b_m\rangle\langle\gamma_m|x\rangle$ to $\Sigma_{m=1}^M|b_m\rangle\langle\gamma_m|y\rangle$ and back, going through the projections of all M arraylets in this subspace. Again, we sort the reconstructed samples according to ϕ_l.

Independently, we also parallel- and antiparallel-associate each data sample with most likely parallel and antiparallel cellular states, or none thereof, according to the annotations of the two groups of n ORFs each, with largest and smallest levels of biological signal in this sample among all N ORFs, respectively. The P value of a given association by annotation is calculated by using combinatorics and assuming hypergeometric probability distribution of the K annotations among the N ORFs, and of the subset of $k \subseteq K$ annotations among the subset of $n \subset N$ ORFs, $P(k; n, N, K) = \binom{N}{n}^{-1}\,\Sigma_{l=k}^n \binom{K}{l}\binom{N-K}{n-l}$, where $\binom{N}{n} = N!n!^{-1}(N-n)!^{-1}$ is the binomial coefficient (12). We define the most likely association of a data sample with a cellular state as the association which corresponds to the smallest P value.

Biological Results: Integrative Analysis of mRNA Expression and Proteins' DNA-Binding Data

Basis Sets. (i) *SVD cell cycle mRNA expression basis.* SVD analysis (3, 4) of relative mRNA expression levels of 4,579 ORFs in 22 yeast samples measured by Spellman *et al.* (6) determined two dominant orthogonal eigenarrays and corresponding eigengenes of similar significance that span the yeast cell cycle expression subspace (see Data Sets 1 and 2 and Mathematica Notebook 1, which are published as supporting information on the PNAS web site). The 22 samples correspond to 18 samples of a cell cycle time course of an α-factor-synchronized culture, and two samples each of strains with overexpressed *CLN*3 and *CLB*2, which encode G_1 and G_2/M cyclins, respectively. One eigenarray was shown to correlate and anticorrelate with the samples of overexpressed *CLN*3 and *CLB*2, respectively (Fig. 1*a*). The corresponding eigengene was shown to correlate with *CLN*3 and its targets, i.e., genes for which expression peaks at the transition from G_1 to S, and anticorrelate with *CLB*2 and its respective targets, for which expression peaks at that from G_2/M to M/G_1 (Fig. 1*b*). Classification of the yeast arrays and genes in the subspaces spanned by these two eigenarrays and eigengenes gives a picture that resembles the traditional understanding of yeast cell cycle regulation (13): 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. 1*c*). The SVD cell cycle mRNA expression basis we use is spanned by the $M = 9$ most significant eigenarrays across the $N = 4{,}579$ ORFs, including the two eigenarrays that span the SVD cell cycle expression subspace. (ii) *GSVD cell cycle mRNA expression basis.* GSVD comparative analysis (5) of mRNA expression of 4,523 yeast and 12,056 human ORFs in 18 samples each of time courses of α-factor-synchronized yeast culture (6), and double thymidine block-synchronized HeLa cell line culture measured by Whitfield *et al.* (7), determined six dominant genelets and corresponding six yeast and six human arraylets, at $-\pi/3$, 0 and $\pi/3$ initial phases, of similar significance in both data sets that span the yeast and human common cell cycle expression subspace (Data Sets 2, 3, and 4 and Mathematica Notebook 1, which are published as supporting information on the PNAS web site). The two 0-phase yeast arraylets were shown to correlate with 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 (Fig. 1*d*). The two 0-phase human arraylets were shown to anticorrelate with the

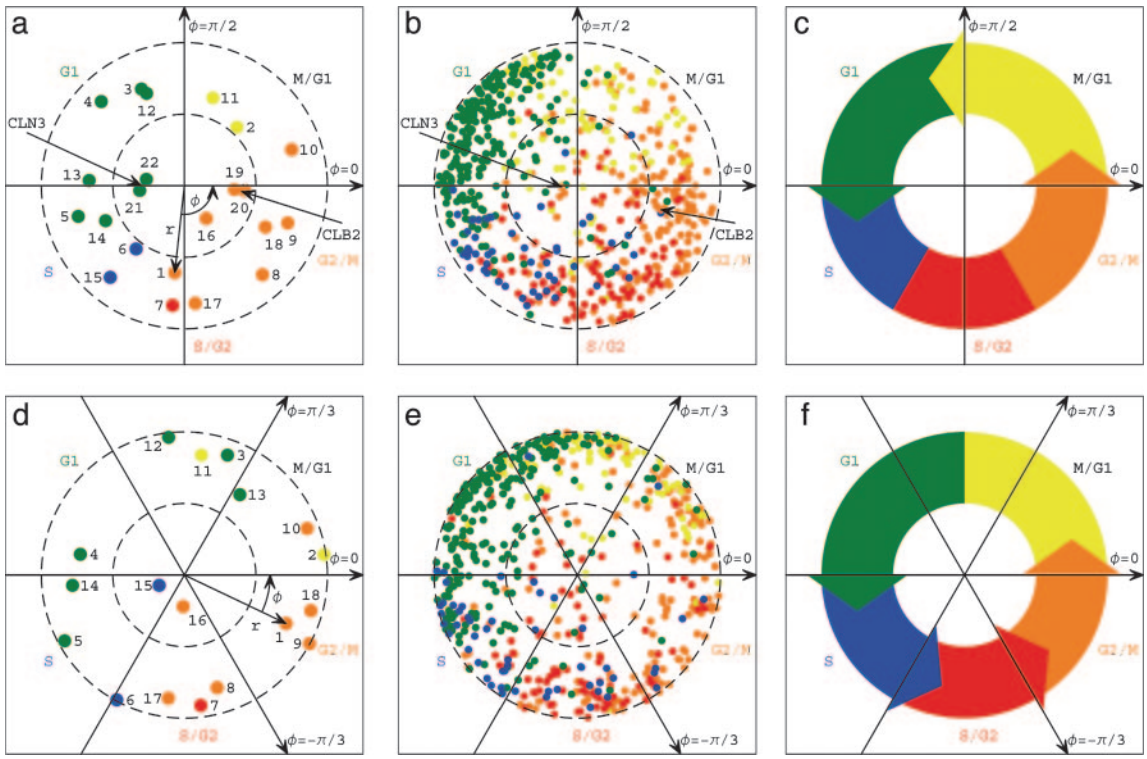

Fig. 1. The SVD (3, 4) and GSVD (5) cell cycle mRNA expression subspaces. (*a*) Normalized array correlation with the $\pi/2$-phase eigenarray along the *y*-axis vs. that with the 0-phase along the *x*-axis, color-coded according to the classification of the arrays into the five cell cycle stages by using combinatorics: 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 this subspace. (*b*) Normalized correlation of each of the 646 cell cycle-regulated genes with the two corresponding eigengenes, color-coded according to either the traditional or microarray classifications. (*c*) The SVD picture of the yeast cell cycle. (*d*) Array expression, projected from the six-arraylets GSVD subspace onto $\pi/2$-phase along the *y*-axis vs. that onto 0-phase along the *x*-axis. The dashed unit and half-unit circles outline 100% and 50% of added up (rather than canceled 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. (*e*) Expression of the 612 cell cycle-regulated genes, projected from the six-genelets GSVD subspace onto $\pi/2$-phase along the *y*-axis vs. that onto 0-phase along the *x*-axis. (*f*) The GSVD picture of the yeast cell cycle.

transition from G_2/M to M/G_1, and to correlate with that from G_1 to S, in which the human culture is synchronized initially. The two shared 0-phase genelets were shown to correlate with 0-phase oscillations of both yeast and human genes (Fig. 1*e*). Simultaneous 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 cell cycles (13), i.e., 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 (Fig. 1*f*). The GSVD cell cycle mRNA expression basis we use is spanned by the six yeast arraylets across the 4,523 ORFs.

Data Sets. (i) *Proteins' DNA-binding data.* This data set tabulates the relative DNA-bound protein occupancy levels of the $N = 2{,}928$ ORFs with at least one valid data point in any one of $L = 13$ samples, which correspond to the nine yeast cell cycle transcription factors measured by Simon *et al.* (10) and four yeast replication initiation proteins measured by Wyrick *et al.* (11) (Data Set 5, which is published as supporting information on the PNAS web site). The relative binding occupancy level of the nth ORF in the lth sample is presumed valid when the P value calculated by either Simon *et al.* or Wyrick *et al.* that is associated with the measured relative binding occupancy signal is <0.1. We divide each ORF measurement by the arithmetic mean of the measurements for that ORF, thus converting the data to binding levels of each protein relative to those of all other proteins. (ii) *α-Factor mRNA expression data.* This set tabulates the relative mRNA expression levels of the 4,636 ORFs with valid data in all of the 18 samples of a cell cycle time course of an α-factor-synchronized culture (6) (Data Set 6, which is published as supporting information on the PNAS web site). The relative expression level of the nth ORF in the lth sample is presumed valid when the ratio of the measured expression signal to that of the background is >1 for both the synchronized culture and the asynchronous reference. (iii) CLB2 *and* CLN3 *mRNA overexpression data.* This set tabulates mRNA expression of 5,840 ORFs with valid data in four samples, two samples each of strains with overexpressed *CLN*3 and *CLB*2, which encode G_1 and G_2/M cyclins, respectively (6) (Data Set 7, which is published as supporting information on the PNAS web site). (iv) CDC15 *mRNA expression data.* This set tabulates mRNA expression of 4,122 ORFs with valid data in all 24 samples of a cell cycle time course of a yeast *CDC*15 mutant culture synchronized by temperature change (6) (Data Set 8, which is published as supporting information on the PNAS web site).

Pseudoinverse Reconstruction of the Proteins' DNA-Binding Data in the mRNA Expression Bases. Of the 2,227 and 2,139 ORFs in the intersections of the 2,928 ORFs of the proteins' DNA-binding data set and the 4,579 and 4,523 ORFs of the SVD- and GSVD-cell cycle mRNA expression bases, 400 and 377 ORFs were microarray-classified, and 58 and 60 were traditionally classified as cell cycle-regulated, respectively. In these intersections, at least one canonical correlation of each binding profile with either the SVD or GSVD bases is >0.1 (see Fig. 6 and Mathematica Notebook 2, which are published as supporting information on the PNAS web site). We

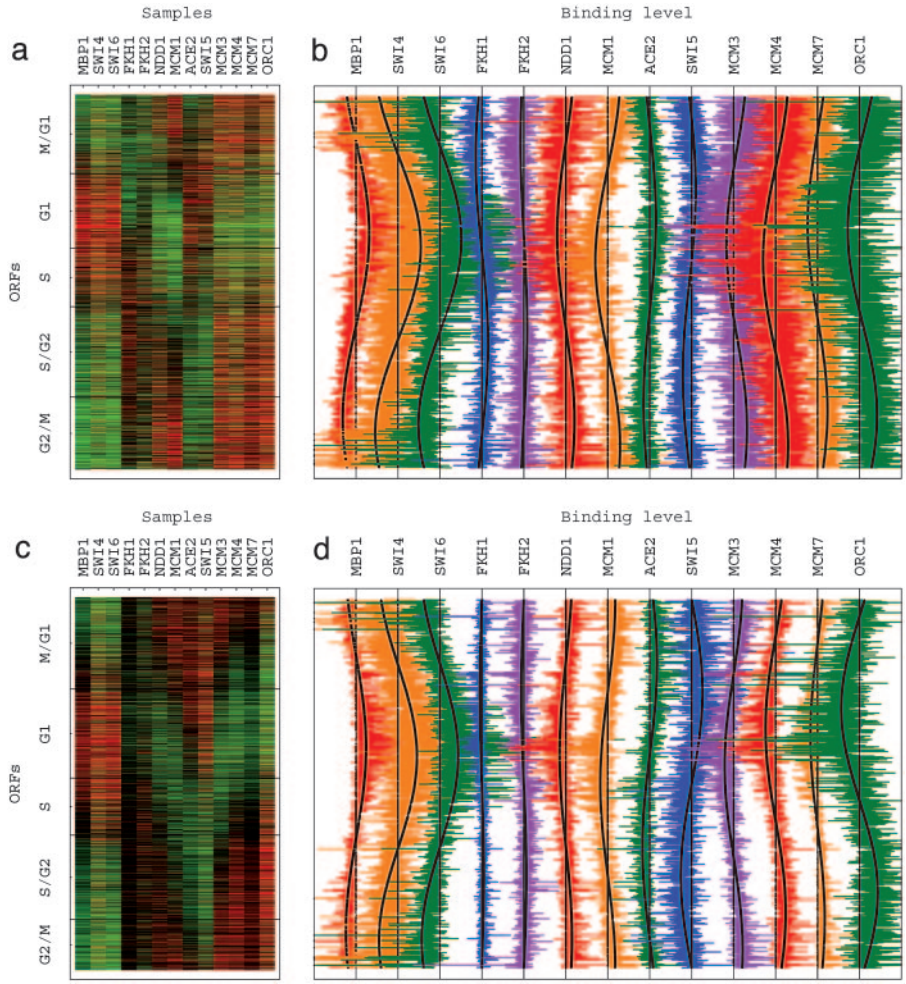

Fig. 2. Pseudoinverse reconstruction of the proteins' DNA-binding data in the SVD (*a* and *b*) and GSVD (*c* and *d*) cell cycle mRNA expression bases, with the ORFs sorted according to their SVD- and GSVD phases, respectively. Raster displays (*a* and *c*), with overexpression (red), no change in expression (black), and underexpression (green), and line-joined graphs (*b* and *d*) of the SVD- and GSVD-reconstructed 13 binding profiles along 2,227 and 2,139 ORFs, centered at their sample- and ORF-invariant levels, show a traveling wave in the nine transcription factors and a standing wave in the four replication initiation proteins.

reconstruct the proteins' DNA-binding data in the SVD and GSVD bases by using pseudoinverse projections in these intersections (Fig. 2). With the ORFs sorted according to their SVD and GSVD cell cycle phases, the ORF variations of the SVD- and GSVD-reconstructed binding profiles approximately fit cosine functions of one period and of varying initial phases. With the nine transcription factors ordered Mbp1, Swi4, Swi6, Fkh1, Fkh2, Ndd1, Mcm1, Ace2, and Swi5, following Simon *et al.* (10), the SVD- and GSVD-pseudoinverse correlations approximately fit cosine functions of one period and of varying initial phases across the nine samples, and are approximately invariant across the four samples of the replication initiation proteins, Mcm3, Mcm4, Mcm7, and Orc1 (Fig. 3).

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.* (10) observed a similar traveling wave in the binding data of the nine transcription factors, ordered as above, across only 213 ORFs in the intersection of ORFs with a P value <0.001 for at least one data point in any one of the nine samples, and ORFs that were microarray-classified as cell cycle-regulated, sorted according to their cell cycle phases as calculated by Spellman *et al.* (6). These traveling waves are in agreement with current understanding of the cell cycle's progression of 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 proteins' DNA-binding cellular states that correspond to those of mRNA expression during the cell cycle. The SVD- and GSVD-reconstructed replication initiation proteins' data approximately fit standing waves, cosinusoidally varying across the ORFs and constant across the four samples, that are antiparallel to the reconstructed profiles of Mbp1, Swi4, and Swi6, and parallel to that of Mcm1.

Pseudoinverse Mapping of the Proteins' DNA-Binding Data onto the Cell Cycle mRNA Expression Subspaces. We map the SVD- and GSVD-reconstructed proteins' DNA-binding data onto the SVD- and GSVD-cell cycle mRNA expression subspaces, respectively, associating with each binding profile cell cycle phase and amplitude (Fig. 4). Projected from the SVD basis, that is spanned by nine eigenarrays, onto the SVD-cell cycle subspace, which is spanned by two of these eigenarrays, all SVD-reconstructed samples have at least 25% of their binding profiles in this subspace, where their distances from the origin satisfy $0.5 \leq r_l < 1$, except for Fkh2. Projected from the six-dimensional GSVD-cell cycle subspace, which 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, where the distance of each array from the origin satisfies $0.5 \leq r_l < 1$.

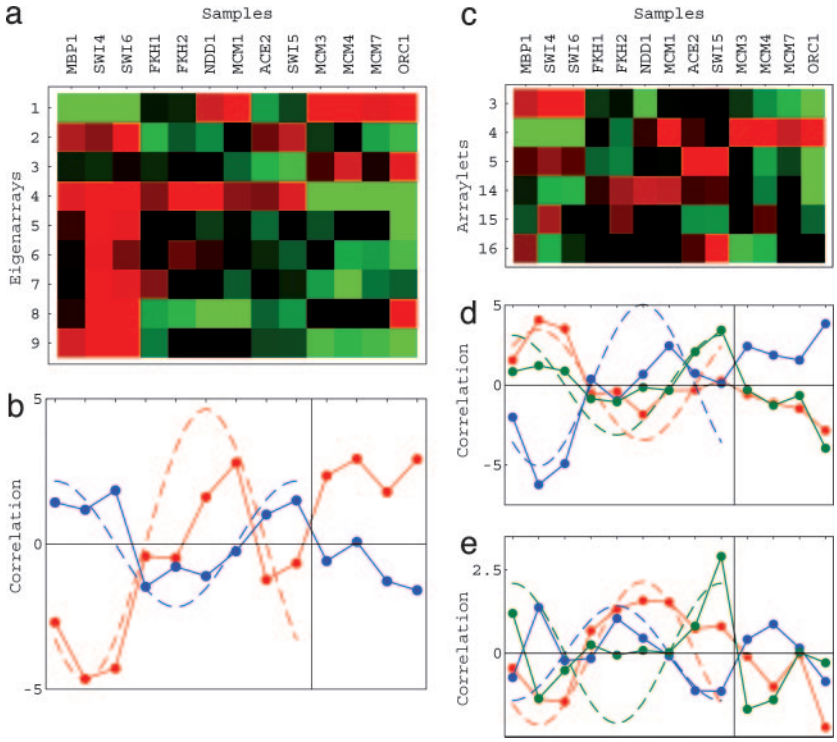

Fig. 3. Pseudoinverse correlations of the proteins' DNA-binding data with the SVD (*a* and *b*) and GSVD (*d* and *e*) cell cycle mRNA expression bases. Shown are raster displays of $\hat{c} \equiv \hat{b}^{+}\hat{d}$, the correlations of the 13 binding profiles with the nine eigenarrays (*a*) and six arraylets (*c*) that span the SVD and GSVD bases, respectively. Also shown are line-joined graphs of the pseudoinverse correlations with the first (red) and second (blue) eigenarrays that span the SVD-cell cycle expression subspace (*b*), the third (red), fourth (blue), and fifth (green) arraylets (*d*), and the 14th (red), 15th (blue), and 16th (green) arraylets that span the GSVD cell cycle expression subspace (*e*).

Sorting the samples according to their SVD or GSVD phases gives an array order that is similar to that of Simon *et al.* (10) 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 that have 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. This mapping is 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. Thus, DNA binding of Mcm3, Mcm4, Mcm7, and Orc1 adjacent to ORFs is shown to be correlated with minima or even shutdown of the transcription of these ORFs during the cell cycle stage G_1, suggest-

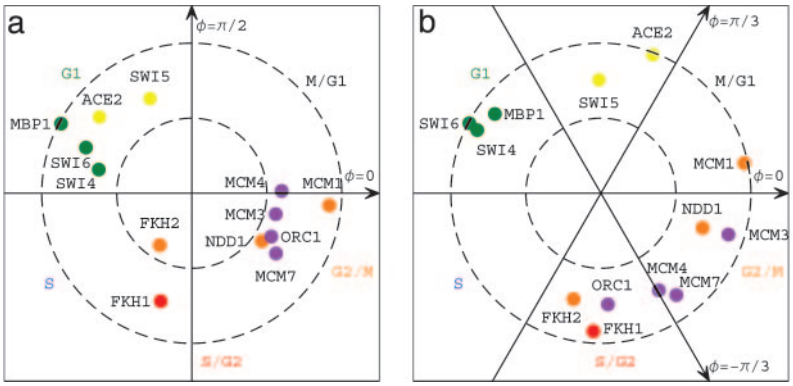

Fig. 4. Pseudoinverse mapping of the proteins' DNA-binding data onto the SVD (*a*) and GSVD (*b*) cell cycle mRNA expression subspaces. (*a*) Normalized sample correlation with the π/2-phase eigenarray along the *y*-axis vs. that with the 0-phase along the *x*-axis. (*b*) Sample binding projected from the six-arraylets GSVD subspace onto π/2-phase along the *y*-axis vs. that onto 0-phase along the *x*-axis.

ing a previously unknown genome-scale coordination between DNA replication initiation and RNA transcription during the cell cycle in yeast.

Independently, we also parallel- and antiparallel-associate each binding profile with most likely parallel and antiparallel cell cycle stages, or none thereof (Table 1, which is published as supporting information on the PNAS web site), by calculating the *P* value for the distribution of the 506 and 77 ORFs that were microarray and traditionally classified as cell cycle-regulated, respectively, among all 2,928 ORFs and among each of the subsets of 200 ORFs with largest and smallest levels of binding occupancy, respectively (Fig. 7, which is published as supporting information on the PNAS web site). At least one of the four *P* values for each profile, following either the microarray or traditional classification, for either parallel or antiparallel association, is <0.01. Most of the *P* values are $\ll 0.01$. Almost all parallel and antiparallel associations of each profile are consistently antipodal, i.e., half of a cell-cycle period apart. Also, almost all associations following the microarray classification are consistent with the associations following the traditional classification. For example, following both the microarray and traditional classifications, the profile of Mcm1 is associated in parallel with G_2/M and in antiparallel with G_1. The SVD and GSVD mappings of all of the binding profiles onto the cell cycle transcription subspaces are consistent with these probabilistic associations by ORF annotations.

Pseudoinverse Integration of the mRNA Expression Data with the mRNA Expression Bases. We integrate the α-factor cell cycle, *CLB*2 and *CLN*3 overexpression and *CDC*15 cell cycle mRNA expression data sets with the SVD- and GSVD-cell cycle mRNA expression bases by using pseudoinverse projections (see Figs. 8–18 and Tables 2 and 3, which are published as supporting information on the PNAS web site). The results are all consistent and in agreement with the current understanding of the cell cycle program.

Pseudoinverse Integration of the Replication Initiation Proteins' DNA-Binding Data with the Transcription Factors' DNA-Binding Basis. We integrate the replication initiation proteins' DNA-binding data with the transcription factors' DNA-binding data after reconstruction in either the SVD- or GSVD-cell cycle RNA transcription bases (see Figs. 19 and 20, which are published as supporting information on the PNAS web site). Again we find that the binding profiles of the replication initiation proteins, Mcm3, Mcm4, Mcm7, and Orc1, are anticorrelated with the profiles of Mbp1, Swi4, and Swi6 and correlated with the profile of Mcm1.

Discussion

We showed that pseudoinverse projection can be used for integrative analysis of different types of large-scale molecular biological data. One consistent picture emerges upon integrating genome-scale proteins' DNA-binding data with the SVD- and GSVD-cell cycle mRNA expression bases, which is in agreement with the current understanding of the yeast cell cycle program. This picture correlates 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 mRNA expression levels are approximately proportional to cell cycle RNA transcription activity. It is known that replication initiation requires binding of Mcm3, Mcm4, Mcm7, and Orc1 at origins of replication across the yeast genome during G_1 (14, 15) and that these replication initiation proteins are involved with transcriptional silencing at the yeast mating loci (16, 17). It was suggested recently that the transcription factor Mcm1 also binds origins of replication (18). 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. Thus a data-driven mathematical model, where the mathematical variables and operations represent biological 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 sets were selected on the basis of data quality alone and were not limited to ORFs that are microarray or traditionally classified as cell cycle-regulated, suggesting that the RNA transcription signatures of yeast cell cycle cellular states may span the whole yeast genome. This idea is in agreement with the recent observation that a genome-wide oscillation in transcription gates DNA replication and the cell cycle (19).

Possible additional applications of pseudoinverse projection include integrating additional data of different cellular programs, e.g., yeast meiosis or invasive growth, and of different type, e.g., DNA sequence motif abundance in ORFs' promoter regions, DNA copy number, mRNA expression, or proteins' DNA-binding levels, with the basis set of yeast cell cycle mRNA expression to elucidate the coordination of these programs in terms of their genomic signals.

We thank D. Botstein and P. O. Brown for introducing us to genomics; J. F. X. Diffley, P. Green, R. R. Klevecz, and J. J. Wyrick for thoughtful and thorough reviews of this manuscript; I. W. Dawes, V. R. Iyer, E. M. Marcotte, and K. Nasmyth for insightful discussions; and G. W. Brown, I. Haviv, I. Simon, and B. K. Tye for helpful comments. This work was supported by National Science Foundation Grant CCR-0430617 (to G.H.G.). O.A. is an Individual Mentored Research Scientist Development Awardee in Genomic Research and Analysis (5 K01 HG00038-01) of the National Human Genome Research Institute.

1. Bussemaker, H. J., Li, H. & Siggia, E. D. (2001) *Nat. Genet.* **27,** 167–171.
2. Lu, P., Nakorchevskiy, A. & Marcotte, E. M. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 10370–10375.
3. Alter, O., Brown, P. O. & Botstein, D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10101–10106.
4. Alter, O., Brown, P. O. & Botstein, D. (2001) in *Microarrays: Optical Technologies and Informatics*, eds. Bittner, M. L., Chen, Y., Dorsel, A. N. & Dougherty, E. R. (Int. Soc. Optical Eng., Bellingham, WA), Vol. 4266, pp. 171–186.
5. Alter, O., Brown, P. O. & Botstein, D. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 3351–3356.
6. 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.
7. Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., Brown, P. O. & Botstein, D. (2002) *Mol. Biol. Cell* **13,** 1977–2000.
8. Golub, G. H. & Van Loan, C. F. (1996) *Matrix Computation* (Johns Hopkins Univ. Press, Baltimore), 3rd Ed.
9. Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M. & Brown, P. O. (2001) *Nature* **409,** 533–538.
10. Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S. & Young, R. A. (2001) *Cell* **106,** 697–708.
11. Wyrick, J. J., Aparicio, J. G., Chen, T., Barnett, J. D., Jennings, E. G., Young, R. A., Bell, S. P. & Aparicio, O. M. (2001) *Science* **294,** 2357–2360.
12. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) *Nat. Genet.* **22,** 281–285.
13. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. D. (1994) *Molecular Biology of the Cell* (Garland, New York), 3rd Ed.
14. Diffley, J. F. X., Cocker, J. H., Dowell, S. J. & Rowley, A. (1994) *Cell* **78,** 303–316.
15. Kelly, T. J. & Brown, G. W. (2000) *Annu. Rev. Biochem.* **69,** 829–880.
16. Micklem, G., Rowley, A., Harwood, J., Nasmyth, K. & Diffley, J. F. X. (1993) *Nature* **366,** 87–89.
17. Fox, C. A. & Rine, J. (1996) *Curr. Opin. Cell Biol.* **8,** 354–357.
18. Chang, V. K., Fitch, M. J., Donato, J. J., Christensen, T. W., Merchant, A. M. & Tye, B. K. (2003) *J. Biol. Chem.* **278,** 6093–6100.
19. Klevecz, R. R., Bolen, J., Forrest, G. & Murray, D. B. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 1200–1205.