

Reconstructing the pathways of a cellular system from genome-scale signals by using matrix and tensor computations

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, October 17, 2005

We describe the use of the matrix eigenvalue decomposition (EVD) and pseudoinverse projection and a tensor higher-order EVD (HOEVD) in reconstructing the pathways that compose a cellular system from genome-scale nondirectional networks of correlations among the genes of the system. The EVD formulates a genes \times genes network as a linear superposition of genes \times genes decorrelated and decoupled rank-1 subnetworks, which can be associated with functionally independent pathways. The integrative pseudoinverse projection of a network computed from a "data" signal onto a designated "basis" signal approximates the network as a linear superposition of only the subnetworks that are common to both signals and simulates observation of only the pathways that are manifest in both experiments. We define a comparative HOEVD that formulates a series of networks as linear superpositions of decorrelated rank-1 subnetworks and the rank-2 couplings among these subnetworks, which can be associated with independent pathways and the transitions among them common to all networks in the series or exclusive to a subset of the networks. Boolean functions of the discretized subnetworks and couplings highlight differential, i.e., pathway-dependent, relations among genes. We illustrate the EVD, pseudoinverse projection, and HOEVD of genome-scale networks with analyses of yeast DNA microarray data.

DNA microarrays | eigenvalue decomposition | higher-order eigenvalue decomposition | pseudoinverse projection | yeast *Saccharomyces cerevisiae* cell cycle and mating

DNA microarrays make it possible to record the complete genomic signals, such as mRNA expression (e.g., refs. 1 and 2) and DNA-bound proteins' occupancy levels (e.g., ref. 3), that are generated and sensed by cellular systems. The underlying genome-scale networks of relations among all genes of the cellular systems can be computed from these signals (e.g., refs. 4–6). These relations among the activities of genes, not only the activities of the genes alone, are known to be pathway-dependent, i.e., conditioned by the biological and experimental settings in which they are observed (e.g., ref. 7). For example, the mRNA expression patterns of the yeast *Saccharomyces cerevisiae* genes *KAR4* and *CIK1* are correlated during mating yet anticorrelated during cell-cycle progression (8). A single genome-scale nondirectional network of correlations cannot describe the pathway-dependent differences in relations, such as those between the expression patterns of *KAR4* and *CIK1*.

Recently, we showed that the matrix singular-value decomposition (SVD), generalized SVD, and pseudoinverse projection separate genome-scale signals, i.e., gene and array patterns of, e.g., mRNA expression and proteins' DNA binding, into mathematically defined patterns that correlate with the independent biological and experimental processes and cellular states that compose the signals (9–12). For example, the comparative generalized SVD of yeast and human mRNA expression during their cell cycles formulates the yeast expression as a linear superposition of cell-cycle oscillations, which are common to the

yeast and human, and response to synchronization by the mating pheromone, which is exclusive to the yeast, and describes a differential relation in the expression of genes such as *KAR4* and *CIK1* that is in agreement with their pathway-dependent activities (11).

Now, we describe the use of the matrix eigenvalue decomposition (EVD) and pseudoinverse projection and a tensor higher-order EVD (HOEVD) in reconstructing the pathways, or genome-scale pathway-dependent relations among the genes of a cellular system, from nondirectional networks of correlations, which are computed from measured genomic signals and tabulated in symmetric matrices. The EVD formulates a genes \times genes network, which is computed from a "data" signal, as a linear superposition of genes \times genes decorrelated and decoupled rank-1 subnetworks. We show that significant EVD subnetworks might represent functionally independent pathways that are manifest in the data signal. The integrative pseudoinverse projection of a network, computed from a data signal, onto a designated "basis" signal approximates the network as a linear superposition of only the subnetworks that are common to both signals, i.e., pseudoinverse projection filters off the network the subnetworks that are exclusive to the data signal. We show that the pseudoinverse-projected network simulates observation of only the pathways that are manifest under both sets of the biological and experimental conditions where the data and basis signals are measured. We define a comparative HOEVD that formulates a series of networks computed from a series of signals as linear superpositions of decorrelated rank-1 subnetworks and the rank-2 couplings among these subnetworks. We show that significant HOEVD subnetworks and couplings might represent independent pathways or transitions among them common to all or exclusive to a subset of the signals. Boolean functions of the discretized subnetworks and couplings highlight known as well as previously unknown differential, i.e., pathway-dependent relations between genes. We illustrate the EVD, pseudoinverse projection, and HOEVD of genome-scale networks with analyses of mRNA expression data from the yeast *Saccharomyces cerevisiae* during its cell cycle (1) and DNA-binding data of yeast transcription factors that are involved in cell-cycle, development, and biosynthesis programs (3).

Mathematical Methods: EVD, Pseudoinverse Projection, and HOEVD of Networks

Eigenvalue Decomposition. Let the symmetric matrix \hat{a}_1 of size N -genes \times N -genes tabulate the genome-scale nondirectional

Conflict of interest statement: No conflicts declared.

Abbreviations: EVD, eigenvalue decomposition; HOEVD, higher-order EVD; SVD, singular-value decomposition.

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

© 2005 by The National Academy of Sciences of the USA

network of correlations among the genes of a cellular system.[¶] The network \hat{a}_1 is computed from a genome-scale signal, designated the data signal, of, e.g., mRNA expression levels measured in a set of M_1 samples of the cellular system using M_1 DNA microarrays and tabulated in the N -genes \times M_1 -arrays matrix \hat{e}_1 , such that $\hat{a}_1 = \hat{e}_1 \hat{e}_1^T$. We compute the EVD of the network \hat{a}_1 ,

$$\hat{a}_1 = \hat{u}_1 \hat{\varepsilon}_1^2 \hat{u}_1^T, \quad [1]$$

from the SVD of the data signal $\hat{e}_1 = \hat{u}_1 \hat{\varepsilon}_1 \hat{v}_1^T$ (9, 10, 13). The M_1 -“eigenarrays” \times M_1 -“eigengenes” diagonal matrix $\hat{\varepsilon}_1$ defines the M_1 nonnegative “eigenexpression” levels, such that the expression of the m th eigen gene in the m th eigenarray is the m th eigenexpression level of \hat{e}_1 , $\varepsilon_{1,m} = \langle m | \hat{\varepsilon}_1 | m \rangle \geq 0$. The orthogonal transformation matrices \hat{u}_1 and \hat{v}_1^T define the N -genes \times M_1 -eigenarrays and the M_1 -eigen genes \times M_1 -arrays subspaces, respectively. The m th column of \hat{u}_1 , $|\alpha_{1,m}\rangle \equiv \hat{u}_1 |m\rangle$, lists the genome-scale expression of the m th eigenarray of \hat{e}_1 . The n th row of \hat{v}_1^T , $\langle n | \hat{v}_1^T$, lists the expression of the n th eigen gene.

EVD formulates the network \hat{a}_1 as a linear superposition of a series of M_1 rank-1 symmetric “subnetworks” of size N -genes \times N -genes each, where the m th subnetwork is the outer product of the m th eigenarray with its transpose $|\alpha_{1,m}\rangle \langle \alpha_{1,m}|$ (Fig. 5 in *Supporting Appendix*, which is published as supporting information on the PNAS web site),

$$\hat{a}_1 = \sum_{m=1}^{M_1} \varepsilon_{1,m}^2 |\alpha_{1,m}\rangle \langle \alpha_{1,m}|. \quad [2]$$

The significance of the m th subnetwork is indicated by the m th “fraction of eigenexpression” $p_{1,m} = \varepsilon_{1,m}^2 / (\sum_{m=1}^{M_1} \varepsilon_{1,m}^2)$, i.e., the expression correlation captured by the m th subnetwork relative to that captured by all subnetworks. Each subnetwork is decorrelated of all other subnetworks, i.e., $|\alpha_{1,m}\rangle \langle \alpha_{1,m}| \langle \alpha_{1,n}\rangle \langle \alpha_{1,n}| = 0$ for all $m \neq n$, since \hat{u}_1 is orthogonal. Each subnetwork is also decoupled of all other subnetworks, such that there are no contributions to the network \hat{a}_1 from the $M_1(M_1 - 1)/2$ rank-2 symmetric “couplings” among the subnetworks, i.e., $|\alpha_{1,m}\rangle \langle \alpha_{1,n}| + |\alpha_{1,n}\rangle \langle \alpha_{1,m}|$ for all $m \neq n$, since $\hat{\varepsilon}_1$ is diagonal. For a real data signal \hat{e}_1 , the eigenarrays are unique up to phase factors of ± 1 , and therefore the subnetworks are also unique, i.e., data-driven, except in degenerate subspaces defined by subsets of equal eigenexpression levels.

Pseudoinverse Projection. Let the matrix \hat{b} of size N -genes \times L -arrays tabulate the genome-scale signal, designated the “basis” signal, of, e.g., proteins’ DNA-binding occupancy levels measured in a set of L samples of the cellular system using L arrays. We compute the pseudoinverse projection (12, 13) of the network \hat{a}_1 onto the basis signal \hat{b} ,

$$\hat{a}_1 \rightarrow \hat{a}_2 \equiv (\hat{b} \hat{b}^\dagger) \hat{a}_1 (\hat{b} \hat{b}^\dagger), \quad [3]$$

from the projection of the data \hat{e}_1 onto the basis \hat{b} , $\hat{e}_1 \rightarrow \hat{e}_2 = \hat{b} \hat{b}^\dagger \hat{e}_1$, using the SVD of the basis $\hat{b} = \hat{U} \hat{\omega} \hat{V}^T$ to compute its pseudoinverse $\hat{b}^\dagger = \hat{V} \hat{\omega}^{-1} \hat{U}^T$. The l th column of \hat{U} , $|\beta_l\rangle \equiv \hat{U} |l\rangle$, lists the genome-scale binding of the l th eigenarray of \hat{b} . The pseudoinverse-projected network \hat{a}_2 is unique, i.e., data-driven. For a real basis signal \hat{b} , $\hat{b} \hat{b}^\dagger$ is an orthogonal projection matrix, and the projected network \hat{a}_2 is symmetric.

We compute the EVD of the projected network \hat{a}_2 ,

$$\hat{a}_2 = \hat{u}_2 \hat{\varepsilon}_2^2 \hat{u}_2^T = \sum_{m=1}^{M_2} \varepsilon_{2,m}^2 |\alpha_{2,m}\rangle \langle \alpha_{2,m}|, \quad [4]$$

where $M_2 = \min\{L, M_1\}$, from the SVD of the projected signal $\hat{e}_2 = \hat{u}_2 \hat{\varepsilon}_2 \hat{v}_2^T$, where the m th column of \hat{u}_2 , $|\alpha_{2,m}\rangle \equiv \hat{u}_2 |m\rangle$, lists the genome-scale expression of the m th eigenarray of \hat{e}_2 . In reconstructing \hat{a}_2 , the pseudoinverse projection filters out of \hat{a}_1 each of its subnetworks $|\alpha_{1,m}\rangle \langle \alpha_{1,m}|$, which is decorrelated of the series of L rank-1 symmetric subnetworks $|\beta_l\rangle \langle \beta_l|$ that compose the network $\hat{b} \hat{b}^\dagger$ computed from the basis signal \hat{b} , such that $|\beta_l\rangle \langle \beta_l| \langle \alpha_{1,m}\rangle \langle \alpha_{1,m}| = 0$ for all $l = 1, 2, \dots, L$ (Fig. 6 in *Supporting Appendix*).

Higher-Order EVD (HOEVD). Let the third-order tensor $\{\hat{a}_k\}$ of size K -networks \times N -genes \times N -genes tabulate a series of K genome-scale networks computed from a series of K genome-scale signals $\{\hat{e}_k\}$, of size N -genes \times M_k -arrays each, such that $\hat{a}_k = \hat{e}_k \hat{e}_k^T$ for all $k = 1, 2, \dots, K$. We define and compute a HOEVD of the tensor of networks $\{\hat{a}_k\}$,

$$\hat{a} \equiv \sum_{k=1}^K \hat{a}_k = \hat{u} \left(\sum_{k=1}^K \hat{\varepsilon}_k^2 \right) \hat{u}^T = \hat{u} \hat{\varepsilon}^2 \hat{u}^T, \quad [5]$$

using the SVD of the appended signals $\hat{e} \equiv (\hat{e}_1, \hat{e}_2, \dots, \hat{e}_K) = \hat{u} \hat{\varepsilon} \hat{v}^T$, where the m th column of \hat{u} , $|\alpha_m\rangle \equiv \hat{u} |m\rangle$, lists the genome-scale expression of the m th eigenarray of \hat{e} . Whereas the matrix EVD is equivalent to the matrix SVD for a symmetric nonnegative matrix, this tensor HOEVD is different from the tensor higher-order SVD (14–16) for the series of symmetric nonnegative matrices $\{\hat{a}_k\}$, where the higher-order SVD is computed from the SVD of the appended networks $(\hat{a}_1, \hat{a}_2, \dots, \hat{a}_K)$ rather than the appended signals. This HOEVD formulates the overall network computed from the appended signals $\hat{a} = \hat{e} \hat{e}^T$ as a linear superposition of a series of $M \equiv \sum_{k=1}^K M_k$ rank-1 symmetric “subnetworks” that are decorrelated of each other, $\hat{a} = \sum_{m=1}^M \varepsilon_m^2 |\alpha_m\rangle \langle \alpha_m|$. Each subnetwork is also decoupled of all other subnetworks in the overall network \hat{a} , since $\hat{\varepsilon}$ is diagonal.

This HOEVD formulates each individual network in the tensor $\{\hat{a}_k\}$ as a linear superposition of this series of M rank-1 symmetric decorrelated subnetworks and the series of $M(M-1)/2$ rank-2 symmetric couplings among these subnetworks (Fig. 7 in *Supporting Appendix*), such that

$$\hat{a}_k = \sum_{m=1}^M \varepsilon_{k,m}^2 |\alpha_m\rangle \langle \alpha_m| + \sum_{m=1}^M \sum_{l=m+1}^M \varepsilon_{k,lm}^2 (|\alpha_l\rangle \langle \alpha_m| + |\alpha_m\rangle \langle \alpha_l|), \quad [6]$$

for all $k = 1, 2, \dots, K$. The subnetworks are not decoupled in any one of the networks $\{\hat{a}_k\}$, since, in general, $\{\varepsilon_{k,m}^2\}$ are symmetric but not diagonal, such that $\varepsilon_{k,lm}^2 \equiv \langle l | \hat{\varepsilon}_k^2 | m \rangle = \langle m | \hat{\varepsilon}_k^2 | l \rangle \neq 0$. The significance of the m th subnetwork in the k th network is indicated by the m th fraction of eigenexpression of the k th network $p_{k,m} = \varepsilon_{k,m}^2 / (\sum_{m=1}^M \varepsilon_{k,m}^2) \geq 0$, i.e., the expression correlation captured by the m th subnetwork in the k th network relative to that captured by all subnetworks (and all couplings among them, where $\sum_{k=1}^K \varepsilon_{k,lm}^2 = 0$ for all $l \neq m$) in all networks. Similarly, the amplitude of the fraction $p_{k,lm} = \varepsilon_{k,lm}^2 / (\sum_{k=1}^K \sum_{m=1}^M \varepsilon_{k,m}^2)$ indicates the significance of the coupling between the l th and m th subnetworks in the k th network. The sign of this fraction indicates the direction of the coupling, such that $p_{k,lm} > 0$ corresponds to a transition from the l th to the m th subnetwork and $p_{k,lm} < 0$ corresponds to the transition from the m th to the

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

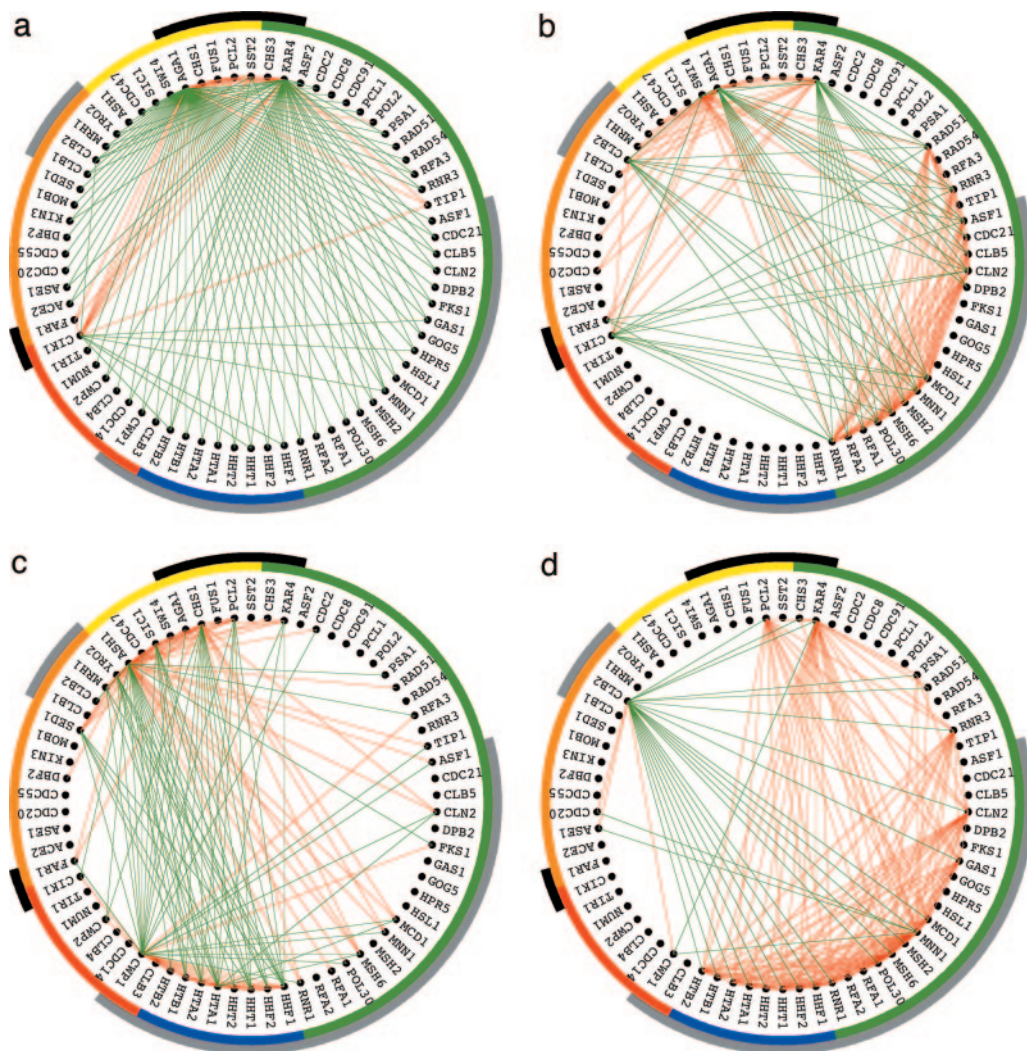


Fig. 1. Discretized significant EVD subnetworks of the network \hat{a}_1 in the subsets of 150 correlations (red) and anticorrelations (green) largest in amplitude among all traditionally classified cell-cycle genes of \hat{a}_1 , color-coded according to their cell-cycle classifications, M/G₁ (yellow), G₁ (green), S (blue), S/G₂ (red), and G₂/M (orange), and separately also according to their pheromone-response classifications, up-regulated (black) and down-regulated (gray). (a) The first subnetwork shows pheromone-response-dependent and cell-cycle-independent relations among the genes. (b) The second subnetwork shows pheromone-response- and cell-cycle-dependent relations. (c and d) The third and fourth subnetworks show cell-cycle-dependent relations that are orthogonal to each other.

lth. For real signals $\{\hat{e}_k\}$, the subnetworks are unique, and the couplings among them are unique up to phase factors of ± 1 , except in degenerate subspaces of \hat{e} .

Interpretation of the Subnetworks and Their Couplings. We parallel- and antiparallel-associate each subnetwork or coupling with most likely expression correlations, or none thereof, according to the annotations of the two groups of x pairs of genes each, with largest and smallest levels of correlations in this subnetwork or coupling among all $X = N(N - 1)/2$ pairs of genes, respectively. The P value of a given association by annotation is calculated by using combinatorics and assuming hypergeometric probability distribution of the Y pairs of annotations among the X pairs of genes, and of the subset of $y \subseteq Y$ pairs of annotations among the subset of $x \subseteq X$ pairs of genes, $P(x; y, Y, X) = \binom{X}{x}^{-1} \sum_{z=y}^X \binom{Y}{x-z} \binom{X-y}{x-z}$, where $\binom{X}{x} = X!x^{-1}(X - x)^{-1}$ is the binomial coefficient (17). The most likely association of a subnetwork with a pathway or of a coupling between two subnetworks with a transition between two pathways is that which corresponds to the smallest P value. Independently, we also parallel- and antiparallel-associate each eigenarray with most likely cellular states, or none thereof, assuming hypergeo-

metric distribution of the annotations among the N -genes and the subsets of $n \subseteq N$ genes with largest and smallest levels of expression in this eigenarray. The corresponding eigengene might be inferred to represent the corresponding biological process from its pattern of expression.

For visualization, we set the x correlations among the X pairs of genes largest in amplitude in each subnetwork and coupling equal to ± 1 , i.e., correlated or anticorrelated, respectively, according to their signs. The remaining correlations are set equal to 0, i.e., decorrelated. We compare the discretized subnetworks and couplings using Boolean functions (6).

Biological Results: Yeast Pathways from mRNA Expression and Proteins' DNA-Binding Signals

Significant EVD Subnetworks Are Associated with Functionally Independent Pathways. We compute the network \hat{a}_1 from the data signal \hat{e}_1 , which tabulates relative mRNA expression levels of $n = 4,153$ yeast genes with valid data in at least 15 of the $M = 18$ samples of a cell cycle time course of a culture synchronized by the mating pheromone α factor (1). The relative expression level of the n th gene in the m th sample is presumed valid when the

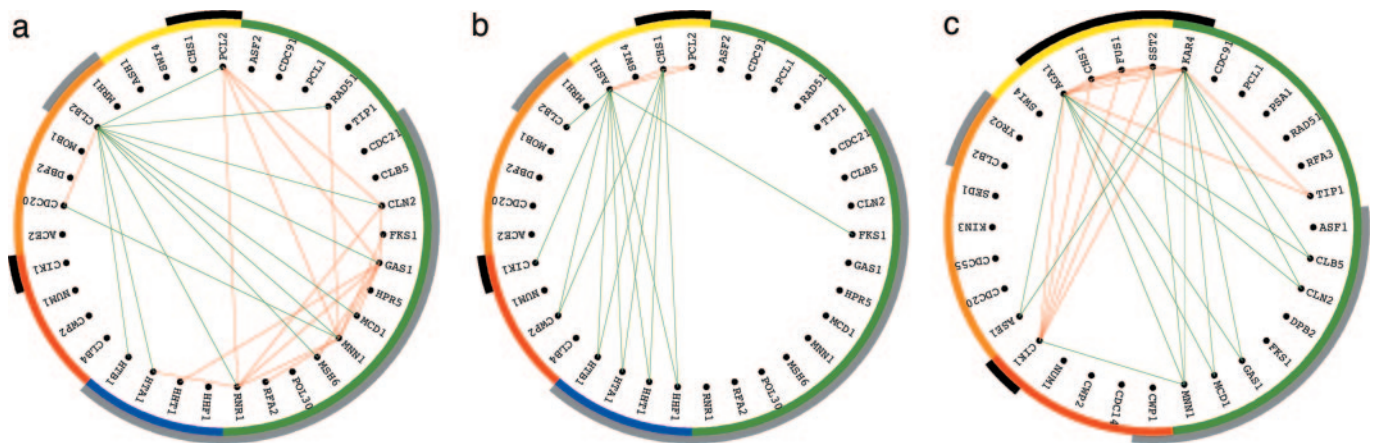


Fig. 2. Boolean AND intersections of the discretized EVD subnetworks of the pseudoinverse-projected \hat{a}_2 and \hat{a}_3 , in the subsets of 200 correlations largest in amplitude among all traditionally classified cell-cycle genes of \hat{a}_2 and \hat{a}_3 , respectively, with these of \hat{a}_1 . (a) The first subnetwork of \hat{a}_2 AND fourth subnetwork of \hat{a}_1 . (b) The second subnetwork of \hat{a}_2 AND third subnetwork of \hat{a}_1 . (c) The subnetwork of \hat{a}_3 AND first subnetwork of \hat{a}_1 .

ratio of the measured expression to the background signal is >1.5 for both the synchronized culture and asynchronous reference. Before computing \hat{a}_1 , we use SVD to estimate the missing data in \hat{e}_1 (10, 18) and to approximately center the expression pattern of each gene in \hat{e}_1 at its time-invariant level (*Supporting Appendix*).

EVD of the network \hat{a}_1 uncovers four significant subnetworks, which capture $>60\%$, 10% , 5% , and 5% , respectively, of the expression correlation of \hat{a}_1 . These subnetworks are associated with the independent pathways manifest in the data signal \hat{e}_1 , following the P values for the distribution of the $Y = 1,035$ pairs of the 46 genes that were microarray-classified as pheromone-regulated (2) among all $X = 2,926$ pairs of the 77 genes that were traditionally classified as cell-cycle-regulated (1), and among each of the subsets of $x = 150$ pairs of genes with largest and smallest levels, respectively, of expression correlation (Table 2 in *Supporting Appendix*). The associations of the EVD subnetworks of \hat{a}_1 are consistent with those of the corresponding SVD eigenarrays of \hat{e}_1 following the P values for the distribution of the 284 pheromone-regulated genes and that of the 574 genes, which were traditionally or microarray-classified as cell-cycle-regulated, among all 4,153 genes and among each of the subsets of 150 genes with largest and smallest levels, respectively, of expression (Table 1 in *Supporting Appendix*). The associations of the EVD subnetworks of \hat{a}_1 are also consistent with the patterns of expressions of the corresponding SVD eigengenes of \hat{e}_1 (Fig. 8 in *Supporting Appendix*). We visualize the discretized four subnetworks and their Boolean functions in the subset of 70 genes that constitute the $x = 150$ correlations in each subnetwork that are largest in amplitude among the $X = 2,926$ pairs of traditionally classified cell-cycle-regulated genes.

The first and most significant subnetwork is associated with the α factor signal-transduction pathway, where the relations among the genes depend only on their pheromone-response classifications. Genes that are up-regulated in response to pheromone, and separately also genes that are down-regulated, are correlated, even when these genes are classified into antipodal cell-cycle stages. Genes that are up-regulated in response to pheromone are anticorrelated with genes that are down-regulated, even when these genes are classified into the same cell-cycle stages. For example, *KAR4*, which is up-regulated in response to pheromone, is correlated with *CIK1*, which is also up-regulated, and anticorrelated with *CLN2*, which is down-regulated (Fig. 1a), even though the expression of both *KAR4* and *CLN2* peaks at the cell-cycle stage G_1 while the expression of *CIK1* peaks at the antipodal stage S/G_2 . In the second

subnetwork, which is associated with the exit from the α factor-induced cell-cycle arrest in M/G_1 and the entry into cell-cycle progression at G_1 , genes that are up-regulated in response to pheromone are correlated, independent of their cell-cycle classification. The relations among genes that are down-regulated, however, depend on their cell-cycle, rather than their pheromone-response, classification. For example, *CLN2* and *CLB2*, which encode cyclins of the antipodal stages G_1 and G_2/M , respectively, are anticorrelated, even though both are down-regulated in response to pheromone; and *SWI4*, which encodes a G_1 transcription factor, is correlated with *CLN2* and anticorrelated with *CLB2* (Fig. 1b). In the third and fourth subnetworks, which are associated with the two pathways of antipodal cell-cycle-expression oscillations that are orthogonal, i.e., $\pi/2$ out of phase relative to one another, the relations among genes depend only on their cell-cycle classifications. For example, in the third subnetwork, which is associated with the cell-cycle-expression oscillations at S vs. those at M, *KAR4* is anticorrelated with *CIK1*, where *KAR4* is correlated, and *CIK1* is anticorrelated with *ASH1* (Fig. 1c). In the fourth subnetwork, which is associated with expression at G_1 vs. that at G_2 , *KAR4* is correlated with *CLN2* (Fig. 1d).

Boolean functions of the discretized subnetworks highlight known pathway-dependent relations among genes, common to a subset of the subnetworks or antipodal across the subnetworks (Fig. 9 in *Supporting Appendix*).

Integrative Pseudoinverse-Projected Networks Simulate Observation of only the Pathways Manifest in both the Data and Basis Signals. We compute the network \hat{a}_2 by pseudoinverse-projecting the network \hat{a}_1 onto the basis signal, which tabulates the relative DNA-bound protein occupancy levels of the 2,120 genes with at least one valid data point in any one of $L = 12$ samples that correspond to 12 yeast-cell-cycle transcription factors (3). The relative binding occupancy level of the n th gene in the l th sample is presumed valid when the associated P value is <0.1 . Similarly, \hat{a}_3 is computed by projecting \hat{a}_1 onto the basis signal, which tabulates the occupancy levels of 2,476 genes in 12 samples of transcription factors involved in developmental programs, such as mating; and \hat{a}_4 is computed by projecting \hat{a}_1 onto the basis signal, which tabulates the occupancy levels of 2,943 genes in eight samples of factors involved in biosynthesis, such as DNA replication. Before computing \hat{a}_2 , \hat{a}_3 , and \hat{a}_4 for the 1,588, 1,827, and 2,254 genes at the intersections of \hat{a}_1 and the proteins' DNA-binding basis signals, we divide each gene measurement in each basis signal by the arithmetic mean of the measurements for that gene in that signal, thus converting the

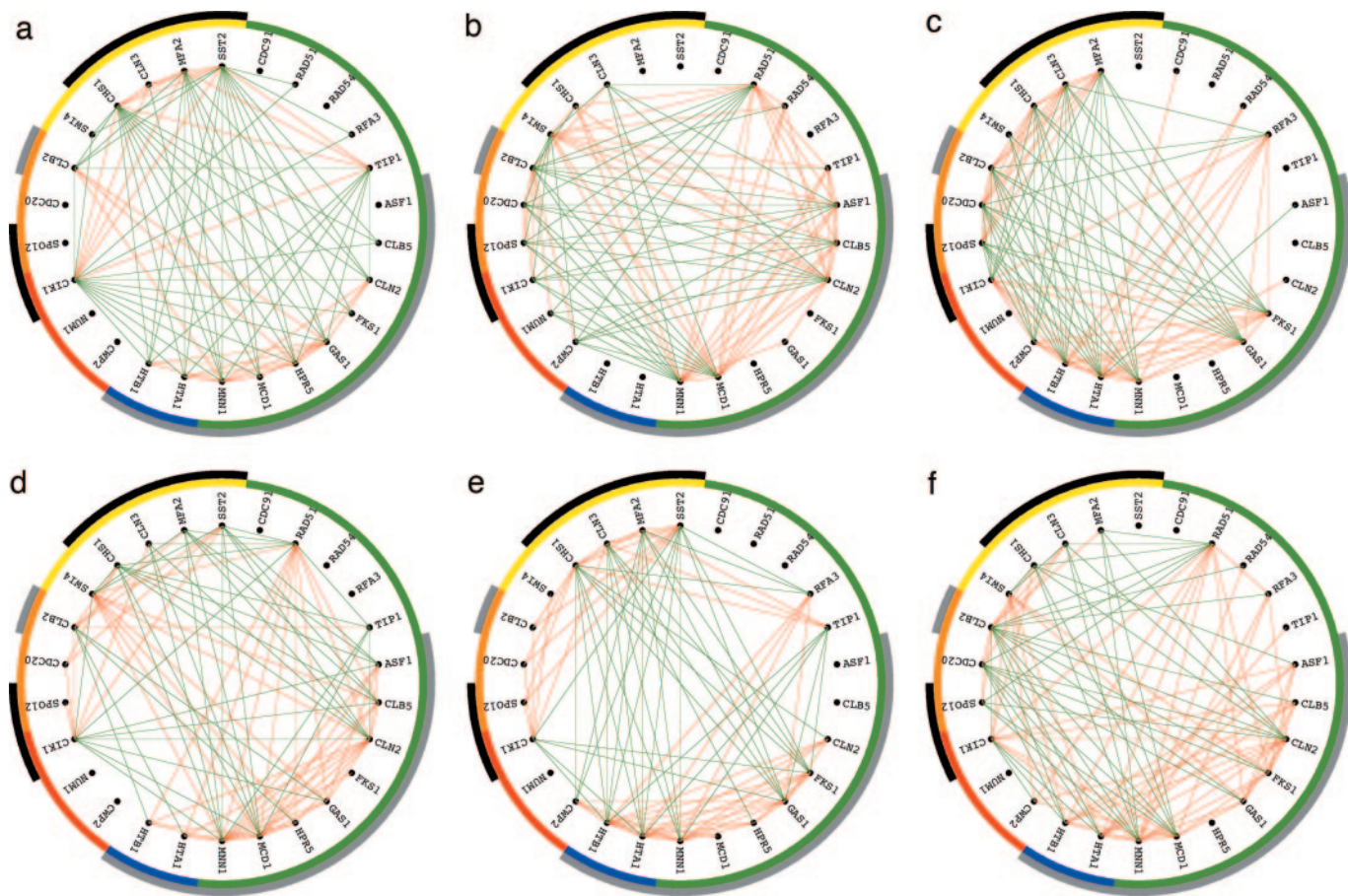


Fig. 3. Discretized significant HOEVD subnetworks of the series of networks $\{\hat{a}_1, \hat{a}_2, \hat{a}_3\}$ and their couplings, in the subsets of 100 correlations largest in amplitude among all traditionally classified cell-cycle genes of $\{\hat{a}_1, \hat{a}_2, \hat{a}_3\}$. (a) The first subnetwork shows pheromone-response-dependent only relations among the genes. (b and c) The second and third subnetworks show orthogonal cell-cycle-dependent relations. (d and e) The couplings between the first and second, and first and third subnetworks, respectively, both show pheromone-response- and cell-cycle-dependent relations. (f) The coupling between the second and third subnetworks shows cell-cycle-dependent only relations.

signals to DNA-binding levels of each transcription factor relative to those of all other factors. We also approximately center the binding pattern of each gene at its transcription factor-invariant level using SVD (*Supporting Appendix*).

EVD of the cell-cycle-projected network \hat{a}_2 uncovers only two significant subnetworks, which capture $\approx 55\%$ and 30% of the expression correlation of \hat{a}_2 , respectively, and are associated with the two pathways of antipodal cell-cycle-expression oscillations at

G_1 vs. those at G_2 and at S vs. M, respectively [Table 4 (row a) in *Supporting Appendix*]. Boolean AND intersection of the discretized first subnetwork of \hat{a}_2 , in the subset of 200 correlations largest in amplitude among all traditionally classified cell-cycle genes of \hat{a}_2 , with the discretized fourth subnetwork of \hat{a}_1 highlights correlations among traditionally classified M/ G_1 , G_1 , and S genes, and anticorrelations among these genes and G_2 /M genes, independent of their responses to pheromone (Fig. 2a). Boolean AND of the second

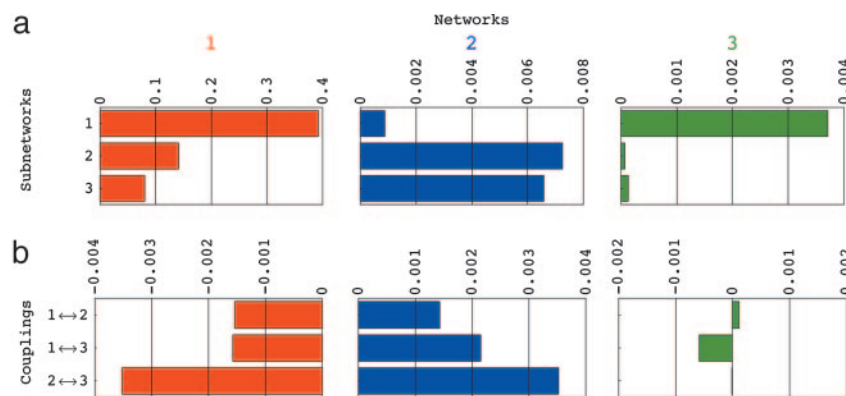


Fig. 4. Fractions of eigenexpression of the HOEVD subnetworks (a) and their couplings (b) in the individual networks \hat{a}_1 (red), \hat{a}_2 (blue), and \hat{a}_3 (green). The contributions of each coupling in each individual network cancel out in the overall network $\hat{a} \equiv \hat{a}_1 + \hat{a}_2 + \hat{a}_3$.

subnetwork of \hat{a}_2 with the third subnetwork of \hat{a}_1 highlights correlations among M/G₁ genes and their anticorrelations with S and S/G₂ genes (Fig. 2*b*). The α factor signal-transduction pathway that is manifest in the data but not in the basis signal is not associated with either one of the subnetworks of \hat{a}_2 . Similarly, EVD of the development-projected network \hat{a}_3 uncovers only one significant subnetwork, which captures >90% of the expression correlation of \hat{a}_3 and is associated with the α factor signal-transduction pathway [Table 4 (row b) in *Supporting Appendix*]. Boolean AND of the subnetwork of \hat{a}_3 with the first subnetwork of \hat{a}_1 highlights correlations among genes that are up-regulated in response to pheromone and their anticorrelations with down-regulated genes, independent of their cell-cycle classifications (Fig. 2*c*). The cell-cycle-expression oscillation pathways that are manifest in the data but not in the basis signal are not associated with either one of the subnetworks of \hat{a}_3 . EVD of the biosynthesis-projected network \hat{a}_4 uncovers three significant subnetworks, which capture together >90% of the expression correlation of \hat{a}_4 , all of which are associated with the activity of histones that peaks during DNA replication at the cell-cycle stage S [Table 4 (row c) and Fig. 13 in *Supporting Appendix*].

The associations of the EVD subnetworks of the projected networks \hat{a}_2 , \hat{a}_3 , and \hat{a}_4 are consistent with the associations of the corresponding SVD eigenarrays (Table 3 in *Supporting Appendix*) and eigengenes (Figs. 10–12 in *Supporting Appendix*) of the projected signals \hat{e}_2 , \hat{e}_3 , and \hat{e}_4 , respectively.

Comparative HOEVD Subnetworks and Their Couplings Are Associated with Pathways and the Transitions Among Them Common to the Series or Exclusive to a Subset of Networks. HOEVD of the series of networks $\{\hat{a}_1, \hat{a}_2, \hat{a}_3\}$ uncovers three significant subnetworks, which capture $\approx 40\%$, 15% , and 9% of the expression correlation of the overall network $\hat{a} \equiv \hat{a}_1 + \hat{a}_2 + \hat{a}_3$, respectively, and the three couplings among these subnetworks, which capture expression correlations only in the individual networks. The associations of the HOEVD subnetworks and couplings of $\{\hat{a}_1, \hat{a}_2, \hat{a}_3\}$ (Table 6 in *Supporting Appendix*) are consistent with the associations of the corresponding SVD eigenarrays (Table 5 in *Supporting Appendix*) and eigengenes (Fig. 14 in *Supporting Appendix*) of the appended signals $\hat{e} \equiv (\hat{e}_1, \hat{e}_2, \hat{e}_3)$, computed for the 868 genes at the intersection of \hat{e}_1 , \hat{e}_2 , and \hat{e}_3 .

The subnetworks are associated with the independent pathways that are manifest in the overall network as well as the individual networks. The first subnetwork, which is associated with the α factor signal-transduction pathway (Fig. 3*a*), contributes to the expression correlations of the network \hat{a}_1 as well as to the development-projected network \hat{a}_3 , but its contribution to the cell-cycle-projected network \hat{a}_2 is negligible (Fig. 4*a*). The second and third subnetworks, which are associated with the two pathways of antipodal

cell-cycle-expression oscillations at G₁ vs. that at G₂ and at S vs. that at M, respectively (Fig. 3*b* and *c*), contribute to \hat{a}_1 and \hat{a}_2 but not to \hat{a}_3 . The couplings are associated with the transitions among these independent pathways that are manifest in the individual networks only. The coupling between the first and second subnetworks is associated with the transition between the two pathways of response to pheromone and cell-cycle expression at G₁ vs. that at G₂, i.e., the exit from pheromone-induced arrest and entry into cell-cycle progression (Fig. 3*d*). The coupling between the first and third subnetworks is associated with cell-cycle expression at G₁/S vs. that at M (Fig. 3*e*). The coupling between the second and third subnetworks is associated with cell-cycle-expression oscillations at the two antipodal cell-cycle checkpoints of G₁/S vs. G₂/M (Fig. 3*f*). All these couplings contribute to the expression correlation of \hat{a}_2 . Their contributions to the expression correlations of \hat{a}_1 and \hat{a}_3 are negligible (Fig. 4*b*).

Boolean functions of the discretized subnetworks and couplings highlight known as well as previously unknown pathway-dependent relations among genes that are in agreement with current understanding of the cellular system of yeast (Fig. 15 in *Supporting Appendix*) (19).

Discussion

We have shown that the matrix EVD and pseudoinverse projection and a tensor HOEVD can separate genome-scale nondirectional networks of, e.g., mRNA expression and proteins' DNA-binding relations among genes into mathematically defined subnetworks and their couplings that can be associated with functionally independent pathways and the transitions among them. In analyses of genome-scale yeast networks, these subnetworks and couplings uncover coordinated differential relations among cell-cycle- and pheromone-regulated genes that are in agreement with reported pathway-dependent activities of these genes. Possible additional applications of EVD, pseudoinverse projection, and HOEVD include reconstruction of pathways and transitions among these pathways from nondirectional networks of correlations among sets of orthologous genes, which are computed from genome-scale signals of different types and from different organisms to elucidate organism, as well as pathway, dependence of relations among genes (e.g., refs. 6, 11, 20, and 21).

We thank T. G. Kolda and T. O. Yeates for thoughtful reviews of this manuscript; J. F. X. Diffley, V. R. Iyer, E. M. Marcotte, and B. K. Tye for helpful comments; and the American Institute of Mathematics in Palo Alto for hosting the 2004 Workshop on Tensor Decompositions where some of this work was done. This work was supported by National Science Foundation Grant CCR-0430617 (to G.H.G.) and National Human Genome Research Institute Individual Mentored Research Scientist Development Award in Genomic Research and Analysis 5 K01 HG00038 (to O.A.).

1. 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.
2. Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y. D., Dai, H., Walker, W. L., Hughes, T. R., et al. (2000) *Science* **287**, 873–880.
3. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., et al. (2002) *Science* **298**, 799–804.
4. Ihmels, J., Levy, R. & Barkai, N. (2002) *Nat. Biotechnol.* **22**, 86–92.
5. Balazsi, G., Barabasi, A. L. & Oltvai, Z. N. (2005) *Proc. Natl. Acad. Sci. USA* **102**, 7841–7846.
6. Bowers, P. M., Cokus, S. J., Eisenberg, D. & Yeates, T. O. (2004) *Science* **306**, 2246–2249.
7. Braun, E. & Brenner, N. (2004) *Phys. Biol.* **1**, 67–76.
8. Kurihara, L. J., Stewart, B. G., Gammie, A. E. & Rose, M. D. (1996) *Mol. Cell. Biol.* **16**, 3990–4002.
9. Alter, O., Brown, P. O. & Botstein, D. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 10101–10106.
10. 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.
11. Alter, O., Brown, P. O. & Botstein, D. (2003) *Proc. Natl. Acad. Sci. USA* **100**, 3351–3356.
12. Alter, O. & Golub, G. H. (2004) *Proc. Natl. Acad. Sci. USA* **101**, 16577–16582.
13. Golub, G. H. & Van Loan, C. F. (1996) *Matrix Computation* (Johns Hopkins Univ. Press, Baltimore), 3rd Ed.
14. De Lathauwer, L., De Moor, B. & Vandewalle, J. (2000) *SIAM J. Matrix Anal. Appl.* **21**, 1253–1278.
15. Kolda, T. G. (2001) *SIAM J. Matrix Anal. Appl.* **23**, 243–255.
16. Zhang, T. & Golub, G. H. (2000) *SIAM J. Matrix Anal. Appl.* **23**, 534–550.
17. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) *Nat. Genet.* **22**, 281–285.
18. Kim, H., Golub, G. H. & Park, H. (2005) *Bioinformatics* **21**, 187–198.
19. Caro, L. H., Smits, G. J., van Egmond, P., Chapman, J. W. & Klis, F. M. (1998) *FEMS Microbiol. Lett.* **161**, 345–349.
20. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. (2003) *Science* **302**, 249–255.
21. Bergmann, S., Ihmels, J. & Barkai, N. (2004) *PLoS Biol.* **2**, E9.