Genomic Signal Processing:

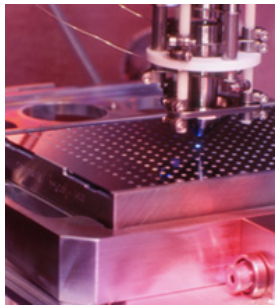
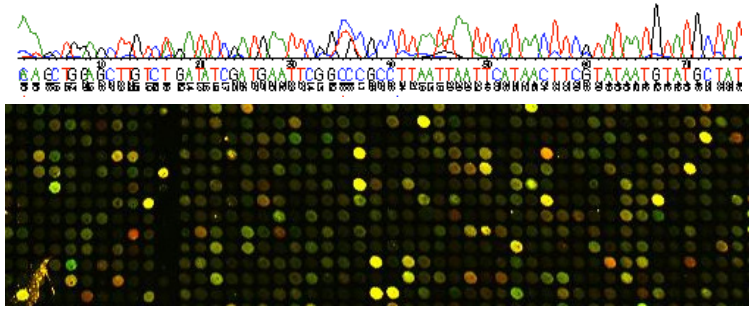
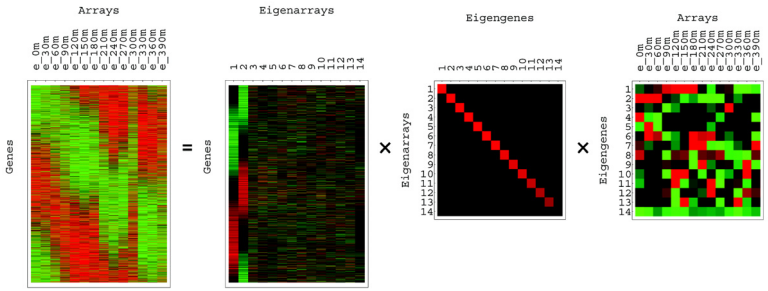
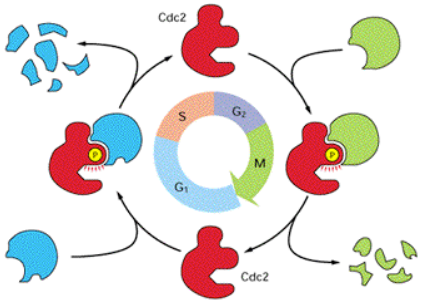
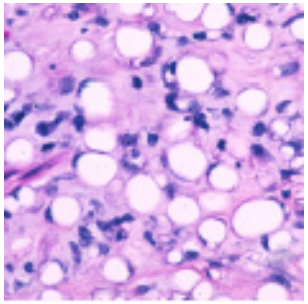
Large-Scale Data, Matrix (and Tensor) Algebra and Basic Biological Principles

Orly Alter

Department of Biomedical Engineering &
Institute of Cellular and Molecular Biology

University of Texas at Austin

NHGRI Individual Development Award in Genomic Research and Analysis

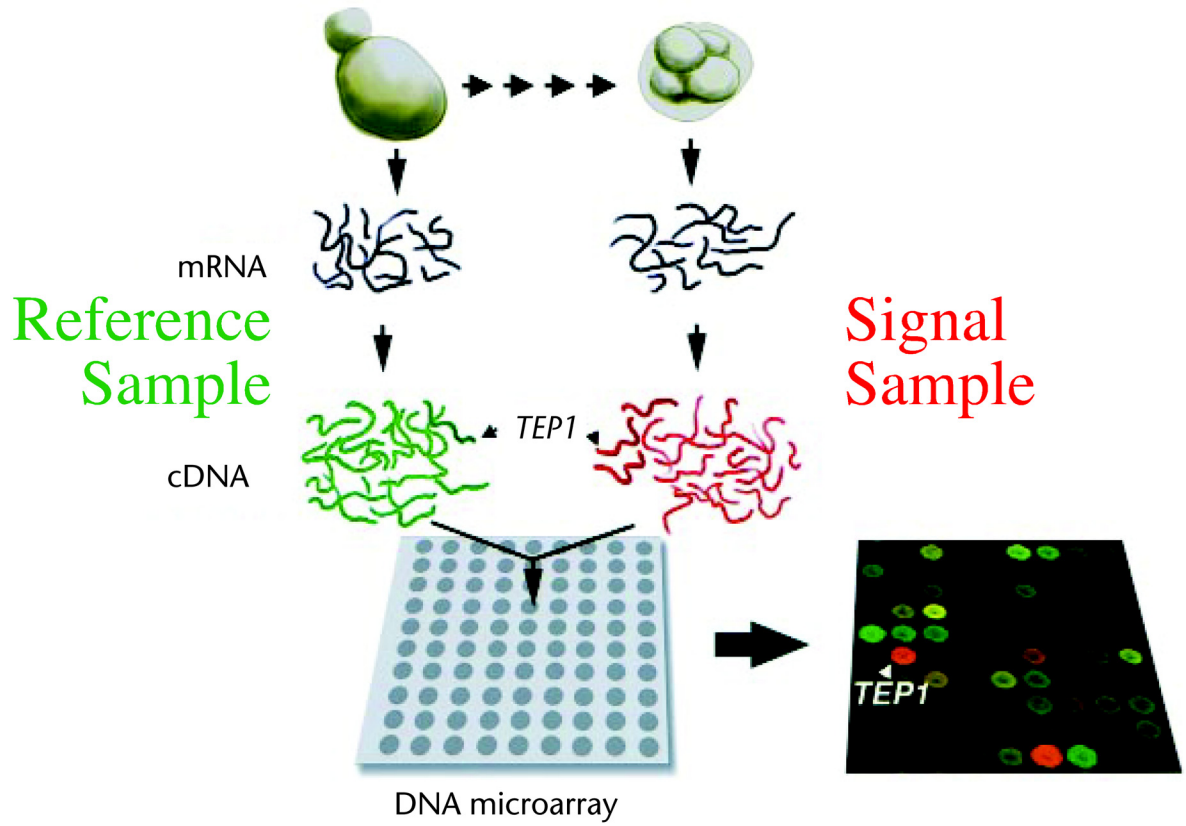
	Astronomy	Molecular Biology
Technology	Galileo	
Large-Scale Data	Brahe	
Mathematical Modeling	Kepler	
Basic Principles	Newton	 
Technology	NASA	Control of Cellular Mechanisms

New High Throughput Technologies

Sequencing of the human genome, and the genomes of other model organisms, such as yeast, is completed or well underway.

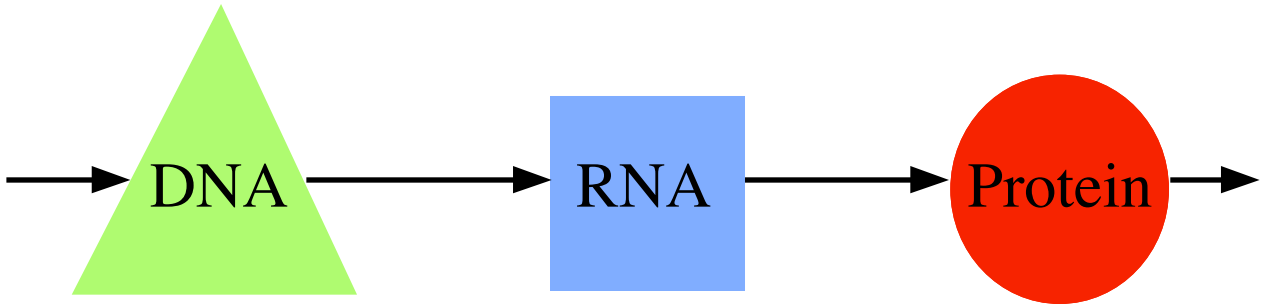
DNA microarray hybridization technology allows probing fluorescently tagged cDNA copies of mRNA from a single sample with thousands of DNA targets or synthetic oligonucleotides simultaneously.

Gene expression levels can now be monitored on a genomic scale.



Brown & Botstein, *Nature Genetics* 21, 33 (1999);
Lipshutz, Fodor, Gingeras & Lockhart, *Nature Genetics* 21, 20 (1999).

New Biology



These new data promise to enhance fundamental understanding of life processes on the molecular level, and may prove useful in medical diagnosis, treatment and drug design.

Clustering analyses already proved useful in assigning function to novel genes, identifying new promoters and classifying tumor tissues.

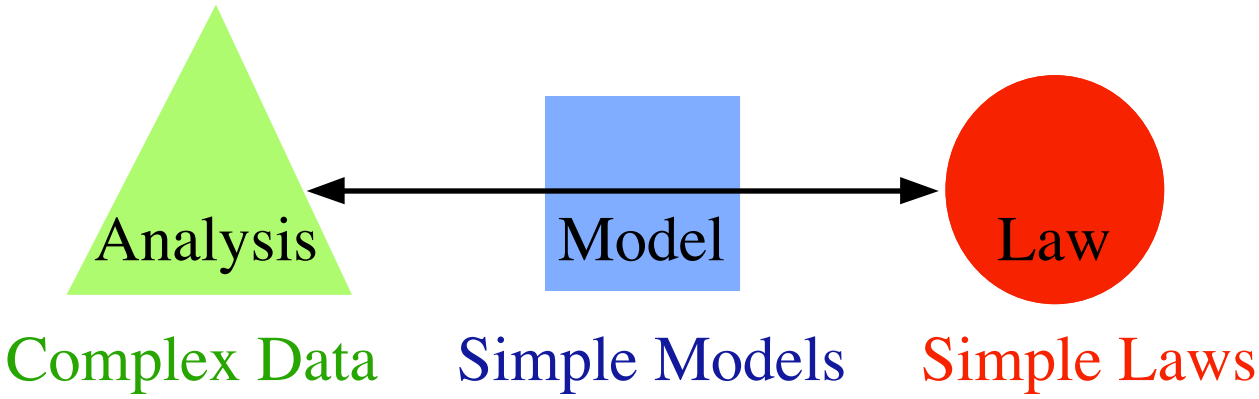
Genomic Signal Processing

Alter, in preparation for the *Wiley Encyclopedia of Biomedical Engineering*.

The data are in large quantities.
Artifacts are superimposed on the data.
Additional data need to be understood in light of the genomic data.
Existing small genetic networks models, when applied to genomic data, appear inconsistent.

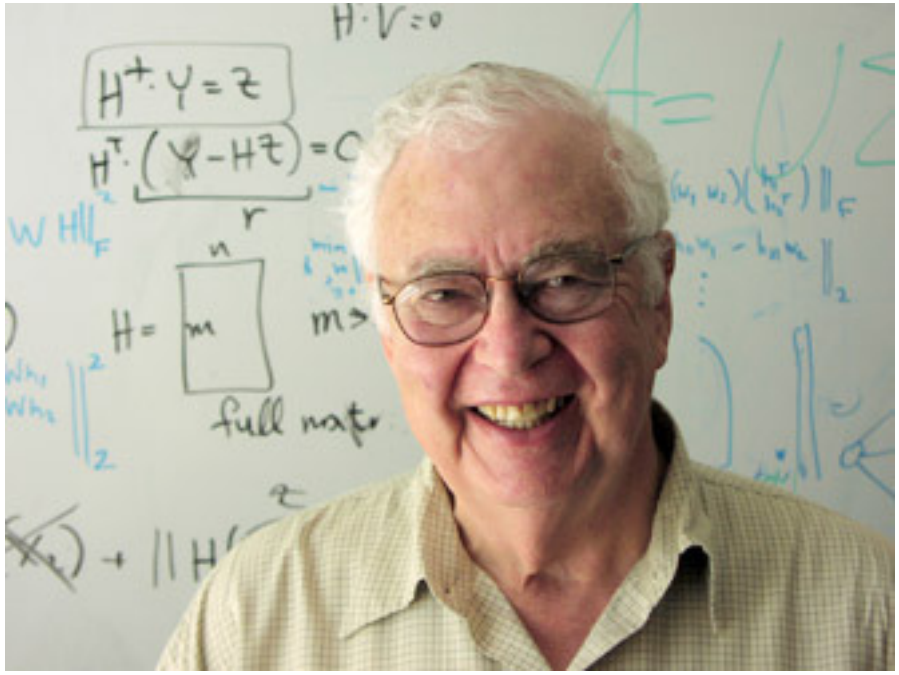
Data-Driven Models for Genomic Data

Mathematical frameworks for the description of the data, in which the mathematical variables and operations may represent some biological reality.



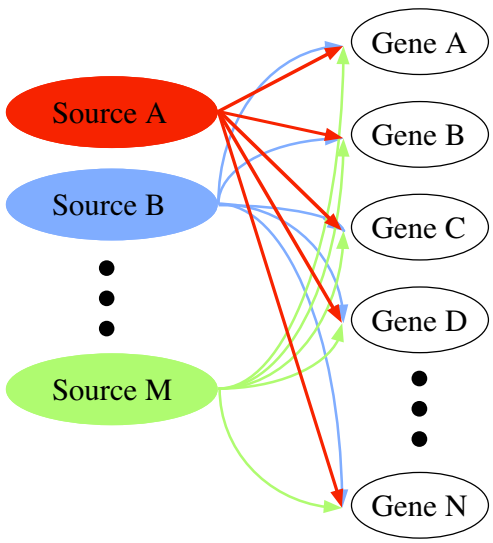
Analogous Problem: Machine Vision

Genomic signals appear complex, easily understood by the biological system → may be governed by simple laws.



Outline

SVD Modeling

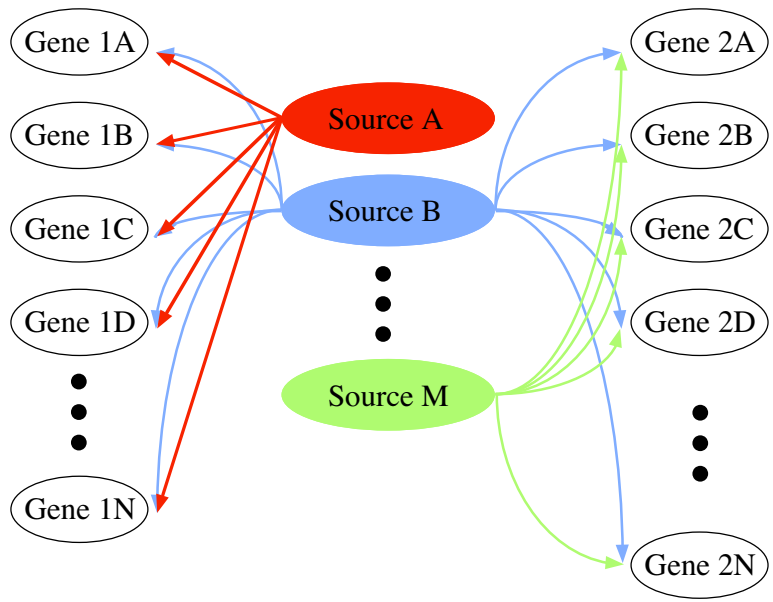


Time Courses

Tumor Samples

Transcript Size Distribution

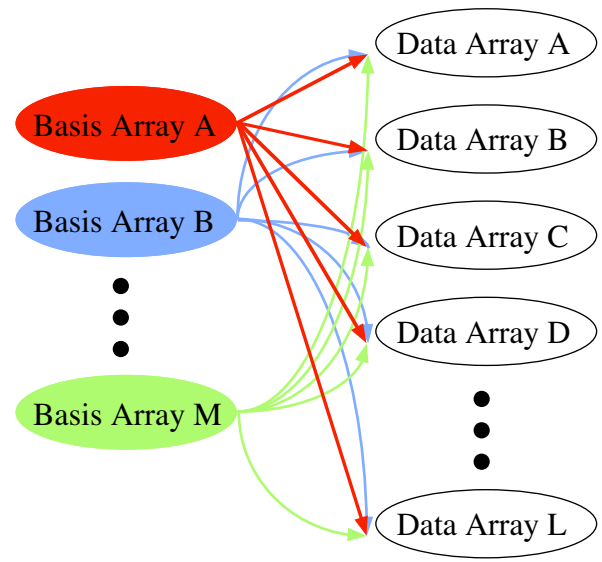
GSVD Comparative Modeling



Time Courses of Two Different Organisms

DNA Copy Number and RNA Expression in Tumors

Pseudoinverse Integrative Modeling

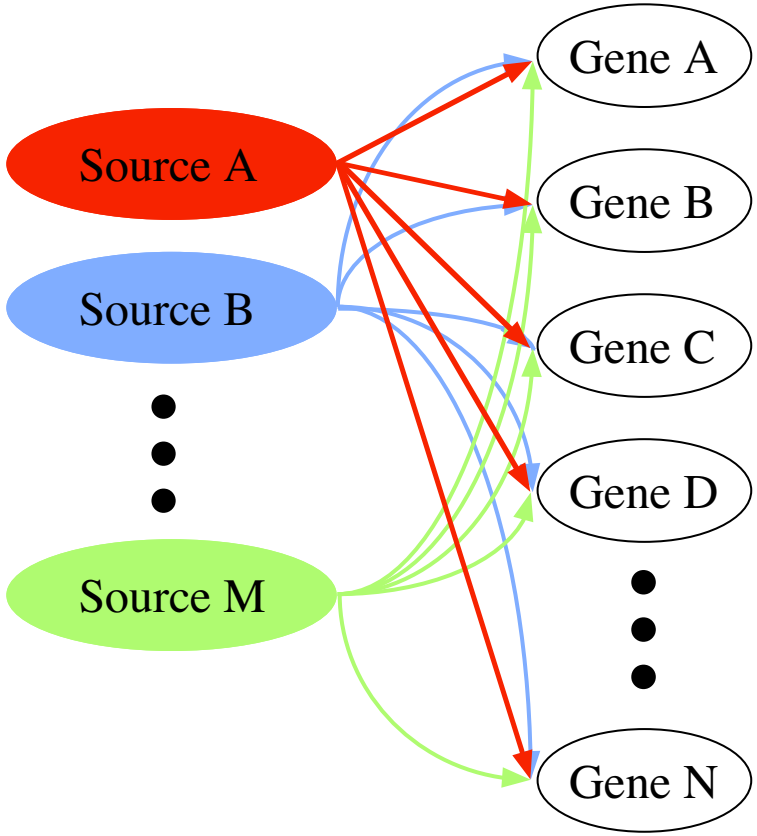


RNA Expression and Proteins' DNA-Binding

SVD Modeling of Genome-Wide Expression Data

Alter, Brown & Botstein, *PNAS* 97, 10101 (2000);
Alter, Brown & Botstein, *Proc. SPIE* 4266, 171 (2001);
http://genome-www.stanford.edu/SVD/.

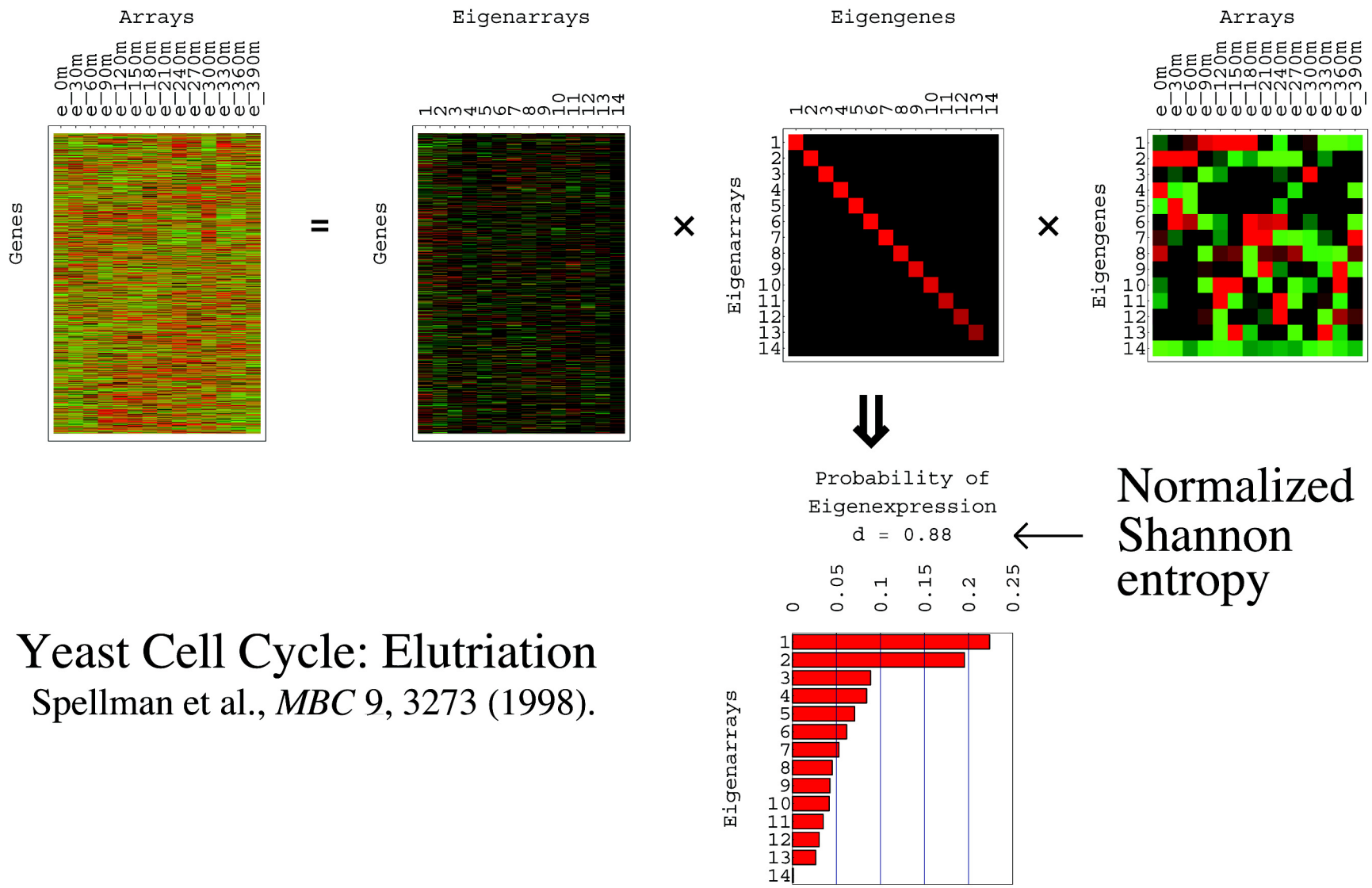
SVD formulates genome-wide expression as a superposition of the genome-wide effects of several independent sources of expression, such as regulatory programs, biological processes and experimental artifacts.



→ Data Normalization
→ Data Classification
→ Additional Data Incorporation

Singular Value Decomposition (I)

Linear transformation of gene expression data from genes × arrays space to reduced diagonalized "eigengenes" × "eigenarrays" space.



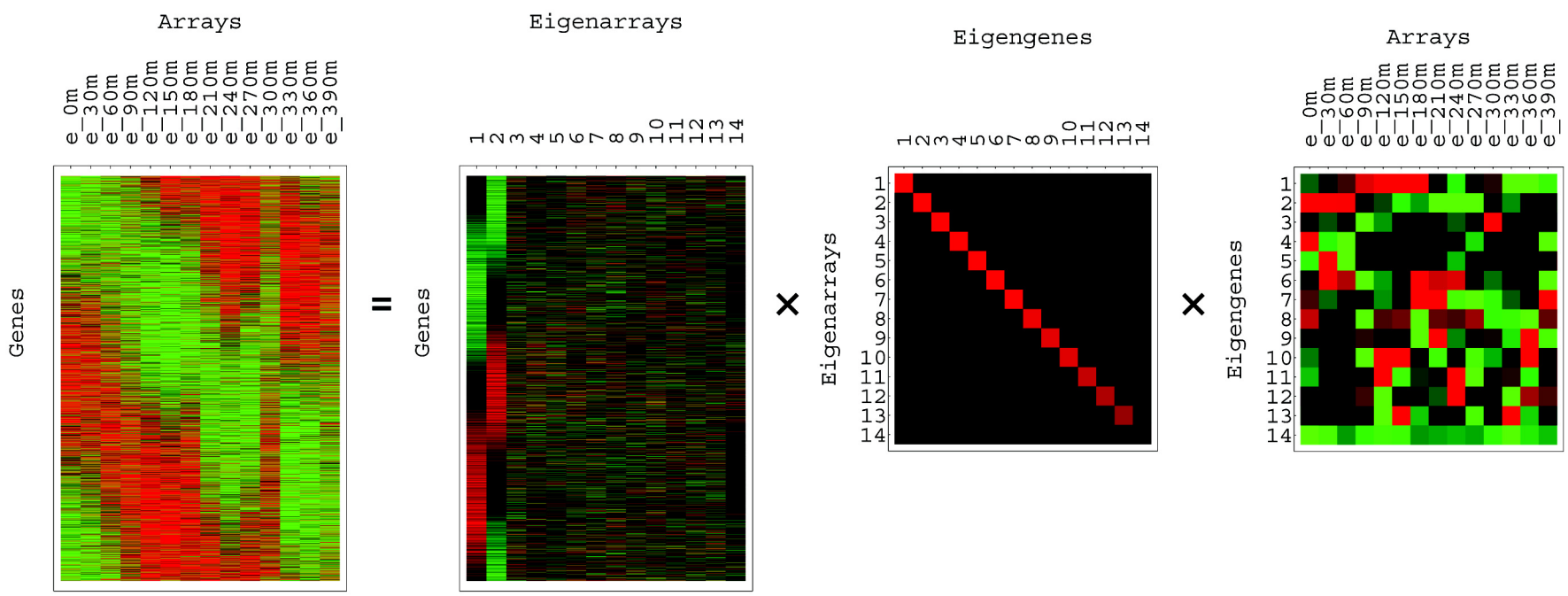
Normalized Shannon entropy

Yeast Cell Cycle: Elutriation

Spellman et al., *MBC* 9, 3273 (1998).

Singular Value Decomposition (II)

The eigengenes and eigenarrays are data-driven unique (except for a phase of ±1 and in degenerate subspaces), orthonormal (decorrelated and normalized), and decoupled superpositions of genes and arrays.



Does the decomposition of the genes (and arrays) expression to eigengenes (and eigenarrays) unravel the biological generation of the expression signal as a superposition of several cellular processes, biological and experimental (and the corresponding cellular states)?

SVD is linear transformation of the expression data from the N-genes $\times$ M-arrays space to the reduced L-"eigenarrays" $\times$ L-"eigengenes" space, where $L = \min\{M, N\}$,

$$\hat{e} = \hat{u}\,\hat{\epsilon}\,\hat{v}^T.$$

The "probability of eigenexpression" indicates the relative significance of the lth eigengene and eigenarray in terms of the fraction of the overall expression that they capture

$$p_l = \epsilon_l^2 / \sum_{k=1}^{L} \epsilon_k^2.$$

The "normalized Shannon entropy" measures the complexity of the data from the distribution of the overall expression between the different eigengenes (and eigenarrays),

$$0 \leq \; d = \frac{-1}{\log(L)} \sum_{k=1}^{L} p_k \log(p_k) \; \leq 1.$$

The transformation matrices $\hat{u}$ and $\hat{v}$ are both orthogonal,

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

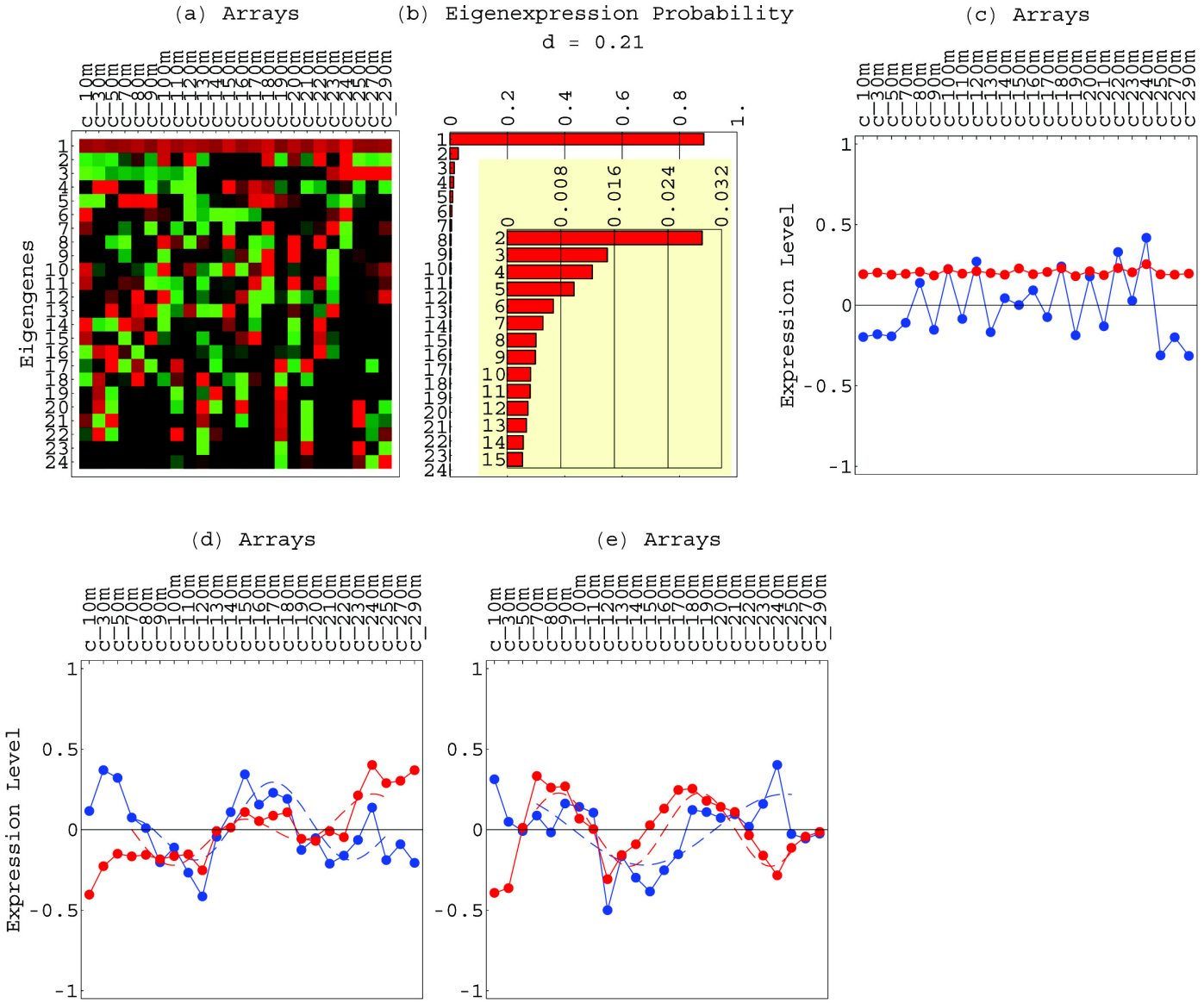
where $\hat{I}$ is the identity matrix.

Math Variables → Biology (I)

Significant eigengenes → independent biological processes and experimental artifacts:

90% of expression is steady state,
2.5% is day-of-hybridization artifact,
less than 7.5% is periodic →

Weak Signal Detection



Yeast Cell Cycle: Cdc15 Spellman et al., *MBC* 9, 3273 (1998).

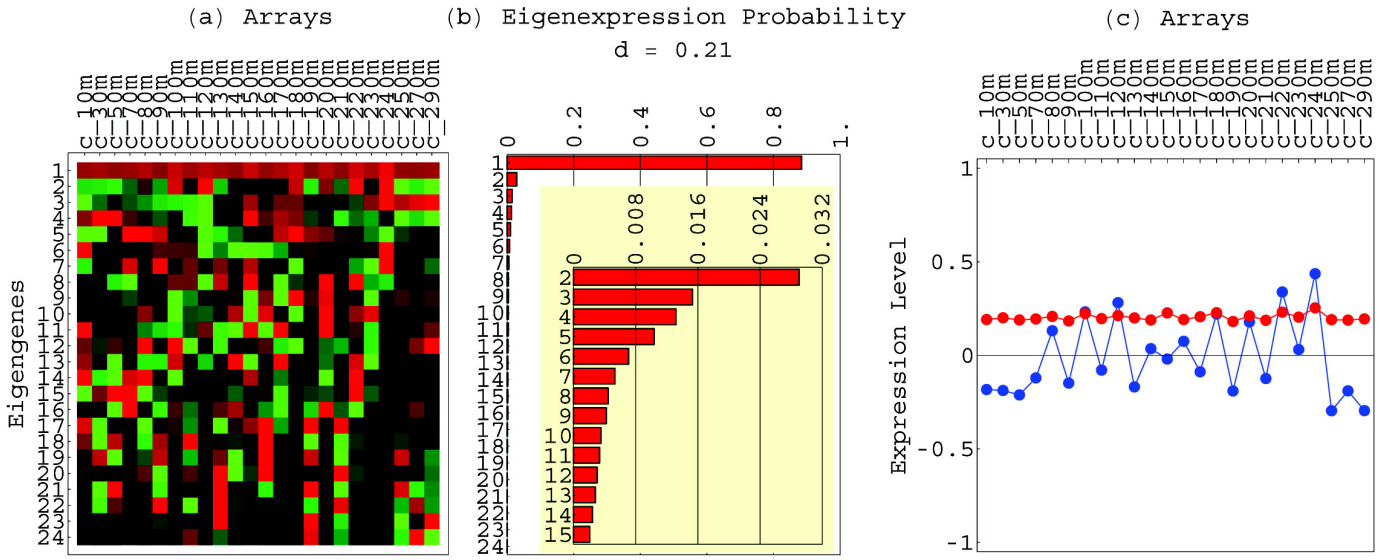
Math Operations → Biology (I)

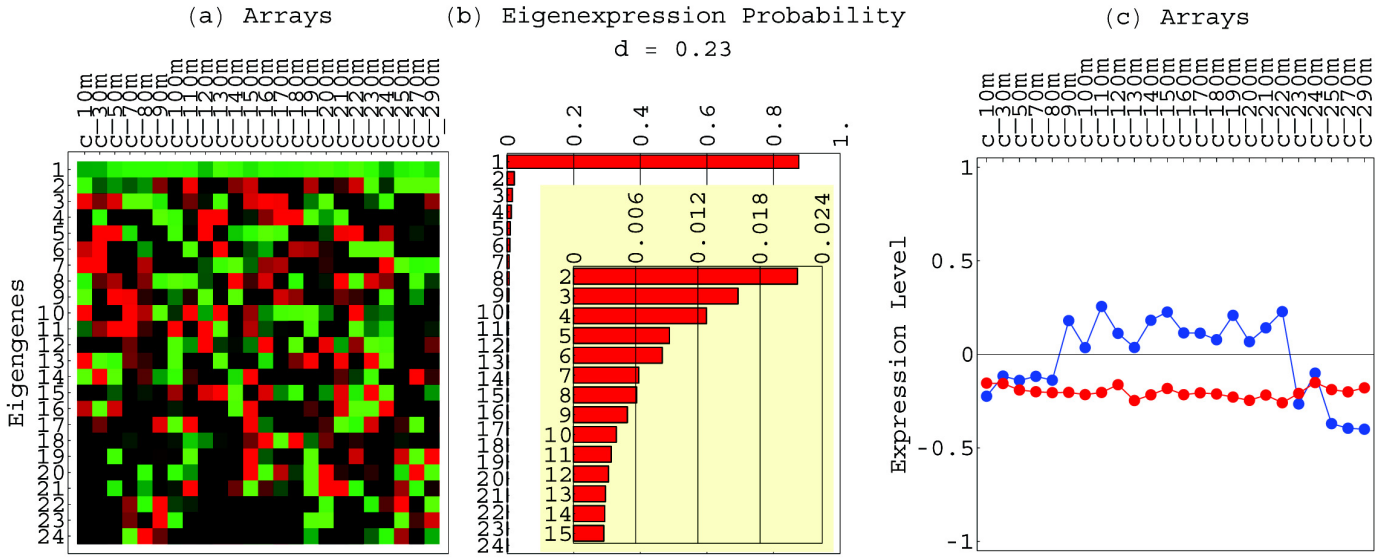
Detection of artifacts →

Filtering data without eliminating genes or arrays:

Normalization

Center data at additive steady state (and filter out additive day-of-hybridization artifact) …





… and normalize by multiplicative steady variance (and filter out multiplicative day-of-hybridization artifact).

Filter out the 1st and 2nd eigengenes (and eigenarrays) of the *CDC15* dataset, removing the steady state of expression and day-of-hybridization additive artifacts,

$$\hat{e} \rightarrow \hat{e}_C = \hat{e} - \epsilon_1 |\alpha_1\rangle\langle\gamma_1| - \epsilon_2 |\alpha_2\rangle\langle\gamma_2|.$$

Let $\hat{e}_{LV}$ tabulate the natural logarithm of the variances in expression, such that each element of $\hat{e}_{LV}$ satisfies for all $1 \leq n \leq N$ and $1 \leq m \leq M$,

$$\langle n|\hat{e}_{LV}|m\rangle \equiv \log(\langle n|\hat{e}_C|m\rangle^2).$$

Filter out the 1st and 2nd eigengenes of $\hat{e}_{LV}$, removing the steady scale of expression variance and day-of-hybridization multiplicative artifacts,

$$\begin{aligned}\hat{e}_{LV} \rightarrow \hat{e}_{CLV} &= \hat{e}_{LV} \\ &- \epsilon_{1,LV} |\alpha_1\rangle_{LV}\, {}_{LV}\langle\gamma_1| - \epsilon_{2,LV} |\alpha_2\rangle_{LV}\, {}_{LV}\langle\gamma_2|.\end{aligned}$$

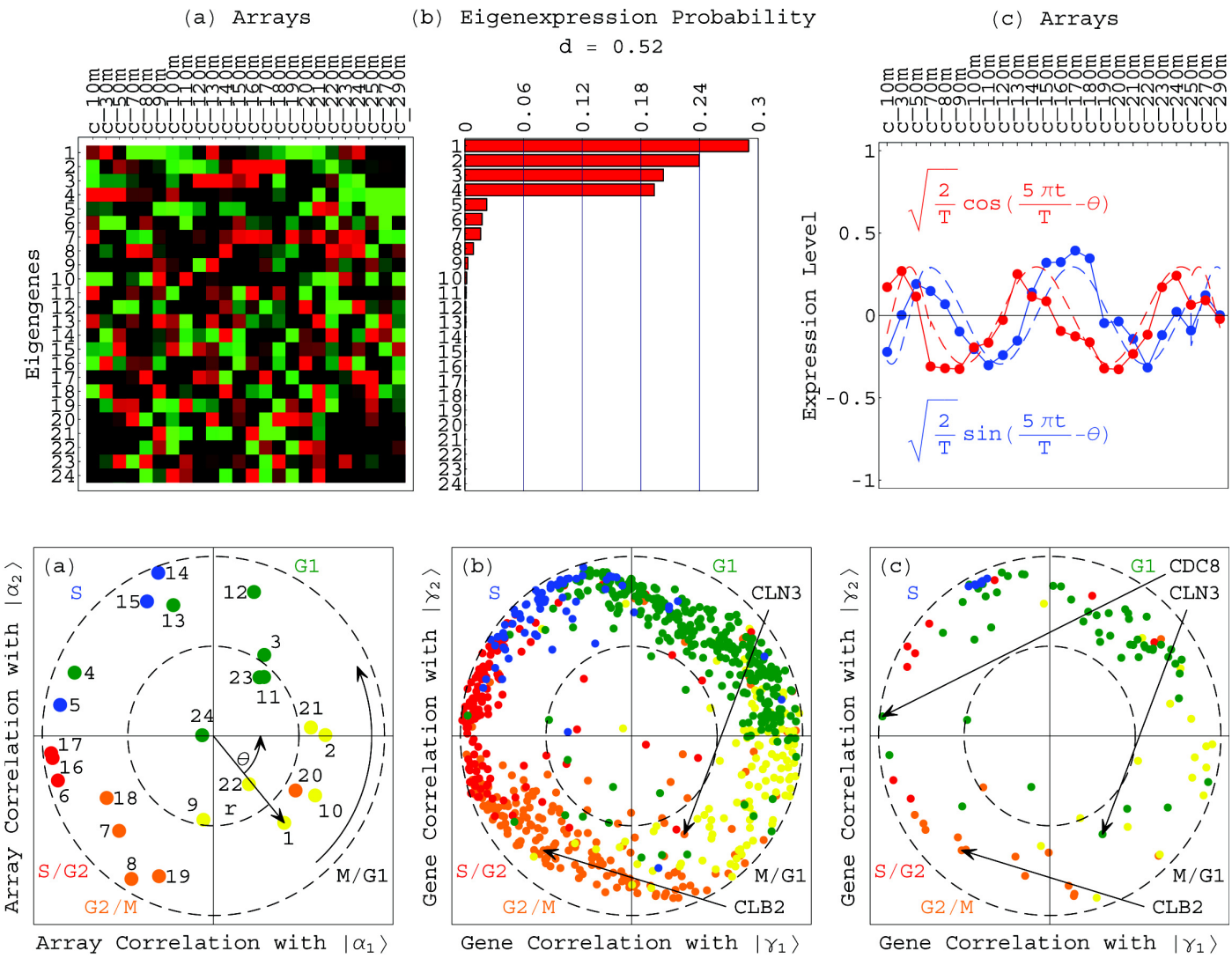
Each element in the normalized dataset $\hat{e}_N$ tabulates for each gene and array expression patterns that are of approximately zero arithmetic means and of approximately unit geometric means,

$$\langle n|\hat{e}_N|m\rangle \equiv \mathrm{sign}(e_{C,nm})\sqrt{\exp(e_{CLV,nm})}.$$

Math Variables → Biology (II)

Two dominant and periodic eigengenes of similar significance, and corresponding eigenarrays, span the

Cell Cycle Subspace:

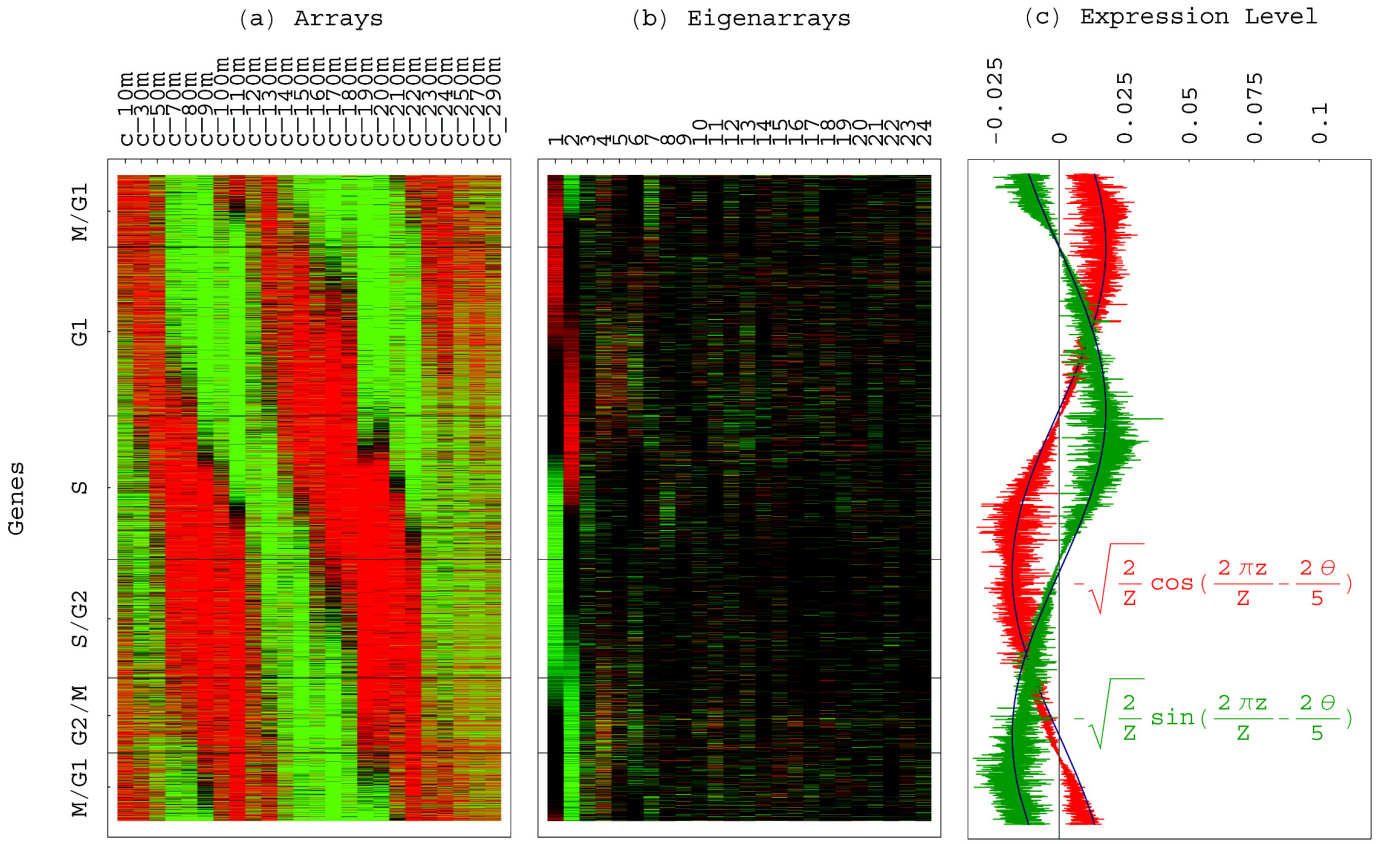


Math Operations → Biology (II)

Detection of biological signals →
sorting the data according to the eigengenes and eigenarrays, rather than overall expression:

Classification

Traveling Wave of Expression (I)



Consistent model for the expression of almost the full yeast genome during cell cycle.

→ Which yeast genes exhibit periodic expression during the cell cycle?

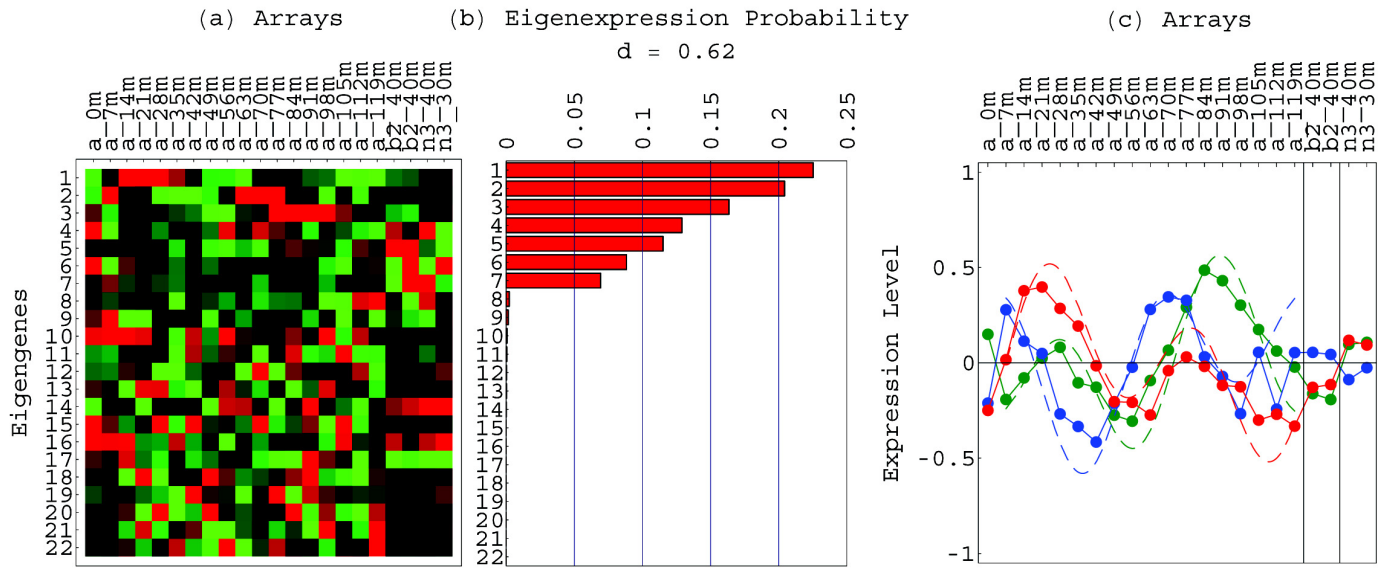
→ Is there a relation between DNA replication and RNA transcription?

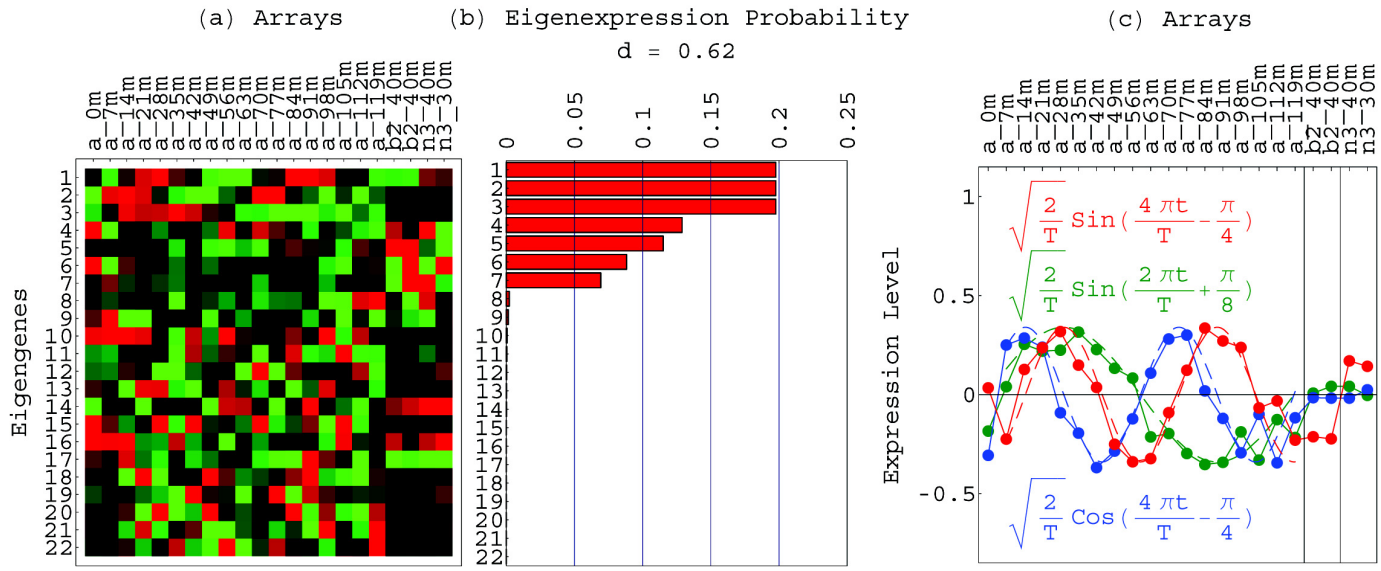
Math Operations → Biology (III)

Degeneracy of eigengenes (eigenarrays) subspace →
Unique rotation of eigengenes (eigenarrays) for better data interpretation and presentation:

Additional Data Incorporation

Cln3, Clb2 genome–wide effects = ± first eigengene





Yeast Cell Cycle: Alpha Factor, Clb2 & Cln3

Spellman et al., *MBC* 9, 3273 (1998).

Approximate the 1st, 2nd and 3rd eigenexpression levels of the α factor, *CLB2*, and *CLN3* dataset with

$$\varepsilon_{1,RN} = \varepsilon_{2,RN} = \varepsilon_{3,RN} = \sqrt{(\varepsilon_{1,N}^2 + \varepsilon_{2,N}^2 + \varepsilon_{3,N}^2)/3}.$$

First, rotate the eigengenes (and corresponding eigenarrays) requiring the rotated 2nd eigengene to describe equal expression in the *CLB2*- and *CLN3*-overactive arrays $|a_{20}\rangle$ and $|a_{21}\rangle$,

$$|\gamma_1\rangle_N \to \hat{R}_1|\gamma_1\rangle_N = \cos\rho_1|\gamma_1\rangle_N + \sin\rho_1|\gamma_2\rangle_N,$$
$$|\gamma_2\rangle_N \to \hat{R}_1|\gamma_2\rangle_N = -\sin\rho_1|\gamma_1\rangle_N + \cos\rho_1|\gamma_2\rangle_N.$$

Second, rotate the eigengenes requiring the rotated 3rd eigengene $\hat{R}_2|\gamma_3\rangle_N$ to describe equal expression in these arrays,

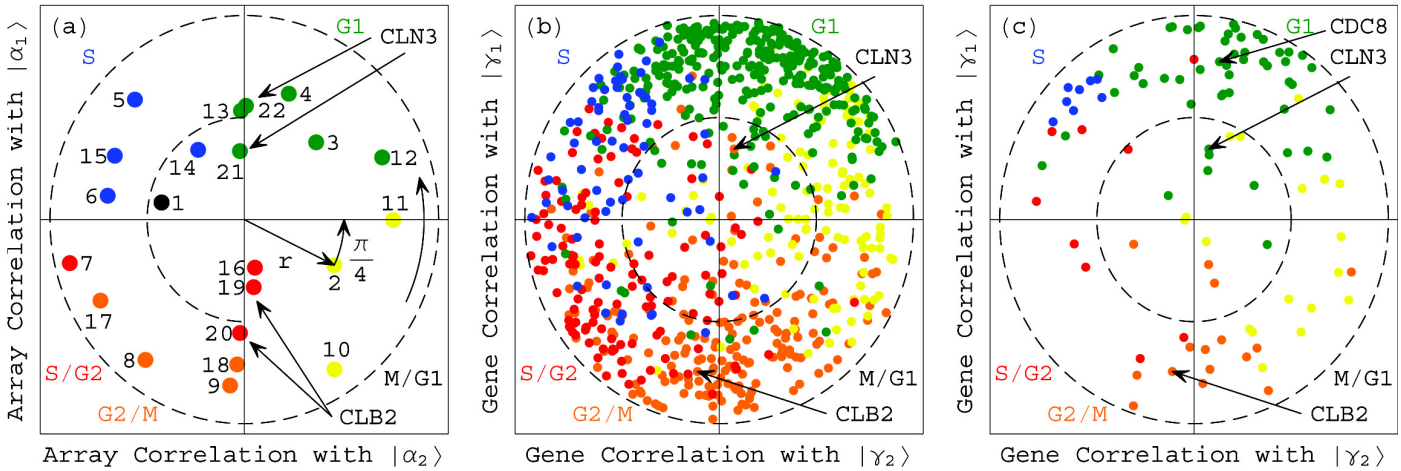
$$\hat{R}_1|\gamma_1\rangle_N \to \hat{R}_2\hat{R}_1|\gamma_1\rangle_N = \cos\rho_2\hat{R}_1|\gamma_1\rangle_N + \sin\rho_2|\gamma_3\rangle_N$$
$$|\gamma_3\rangle_N \to \hat{R}_2|\gamma_3\rangle_N = \sin\rho_2\hat{R}_1|\gamma_1\rangle_N - \cos\rho_2|\gamma_3\rangle_N$$

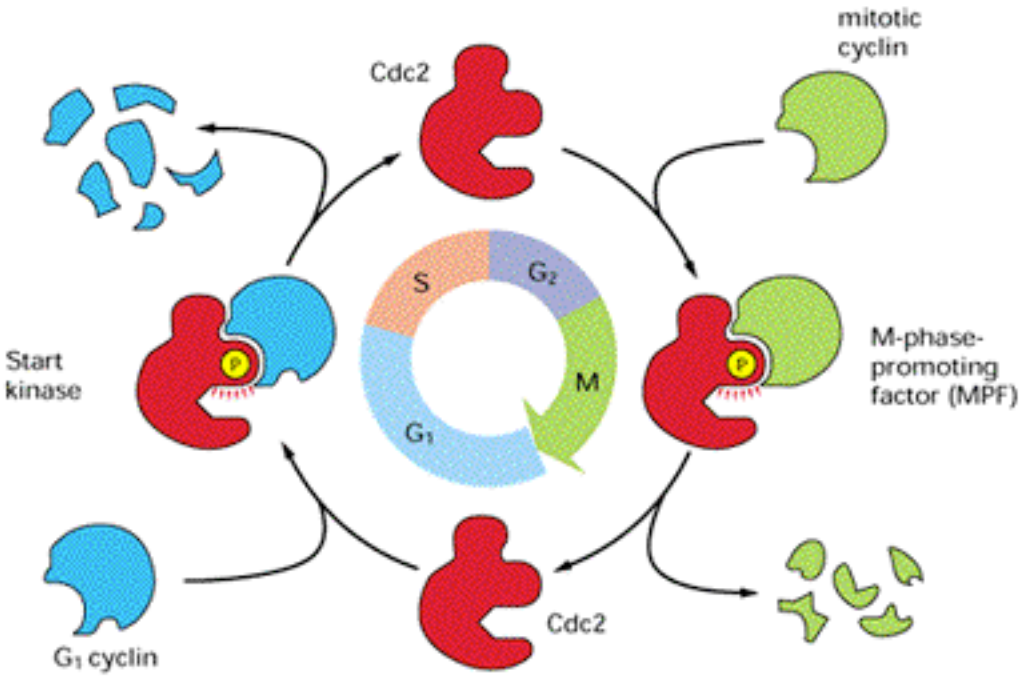
Math Variables → Biology (III)

Significant eigengenes and eigenarrays → genome-wide effects of regulators, and samples in which these regulators are overactive, respectively:

Cln3, Clb2 genome–wide effects = ± first eigengene

Cln3, Clb2 overactive samples = ± first eigenarray

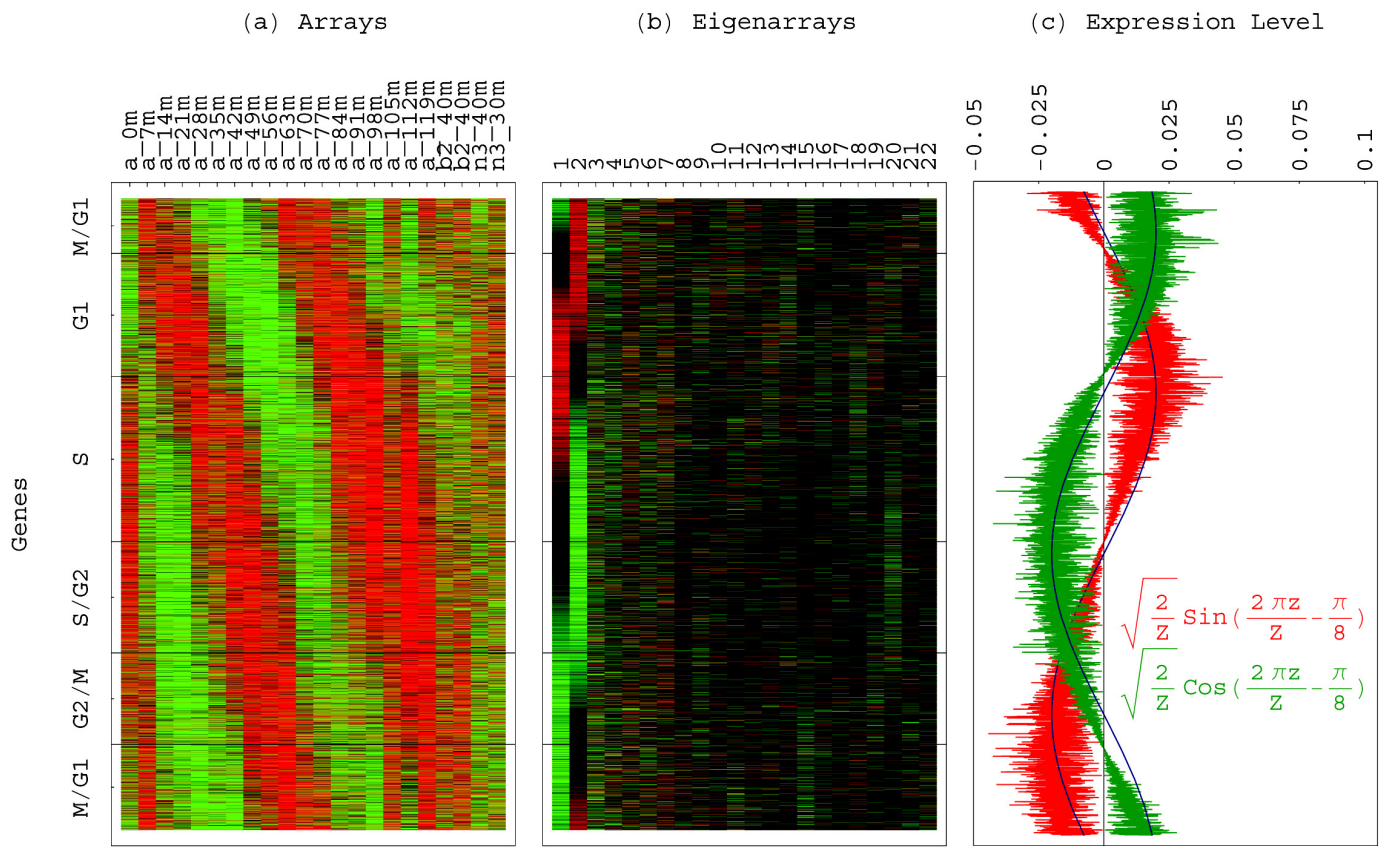




Alberts et al., *Molecular Biology of the Cell* (1994).

Traveling Wave of Expression (II)

Cln3, Clb2 overactive samples = ± first eigenarray

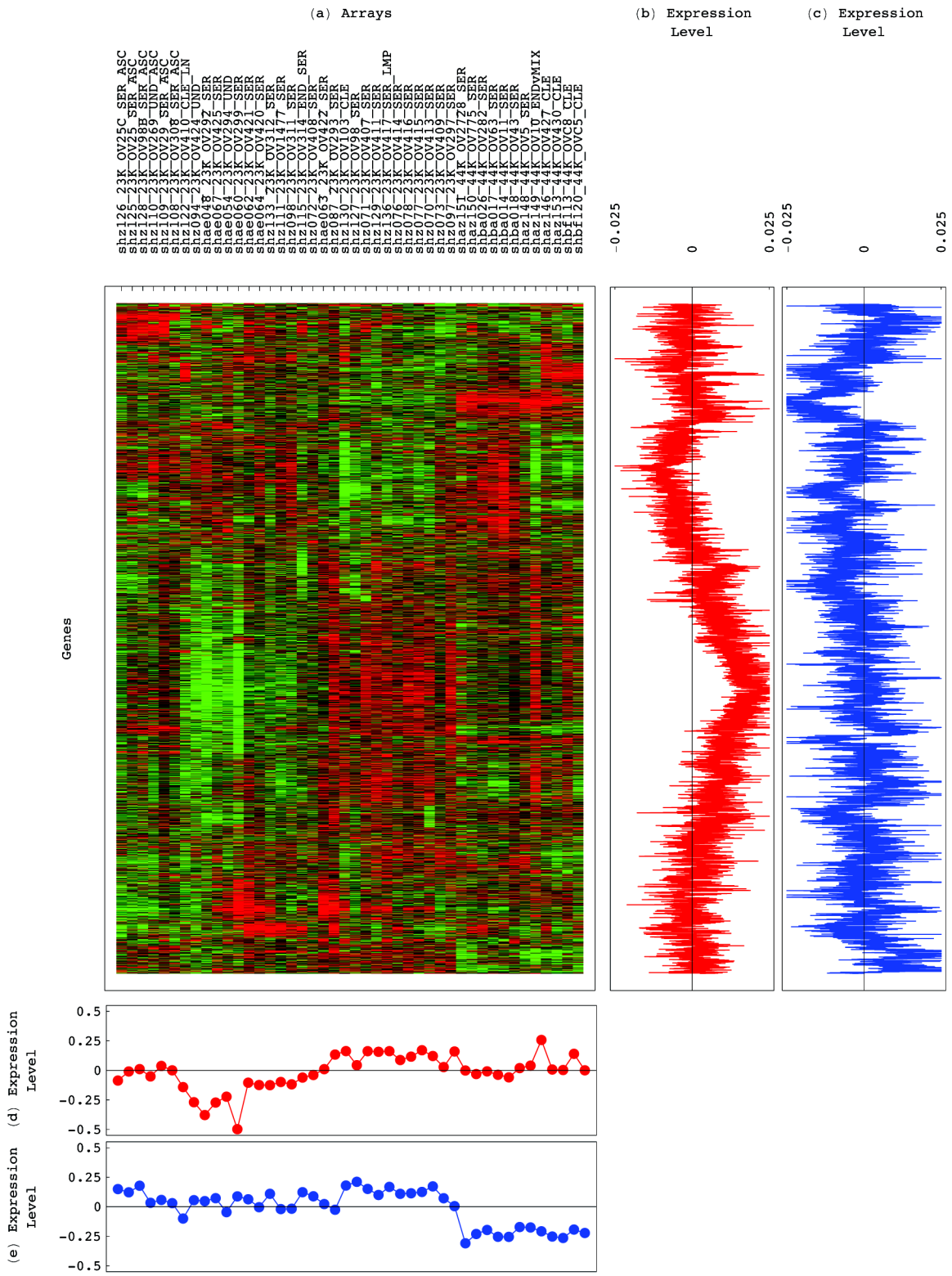


Consistent model for the expression of almost the full yeast genome during cell cycle, in a subspace spanned by only two eigengenes and corresponding eigenarrays.

→ Are there only two cellular modules that drive the yeast cell cycle?

→ Can we design a synthetic genetic network analogous to the LC circuit, which would simulate the yeast cell cycle?

SVD Normalization of Tumor Data



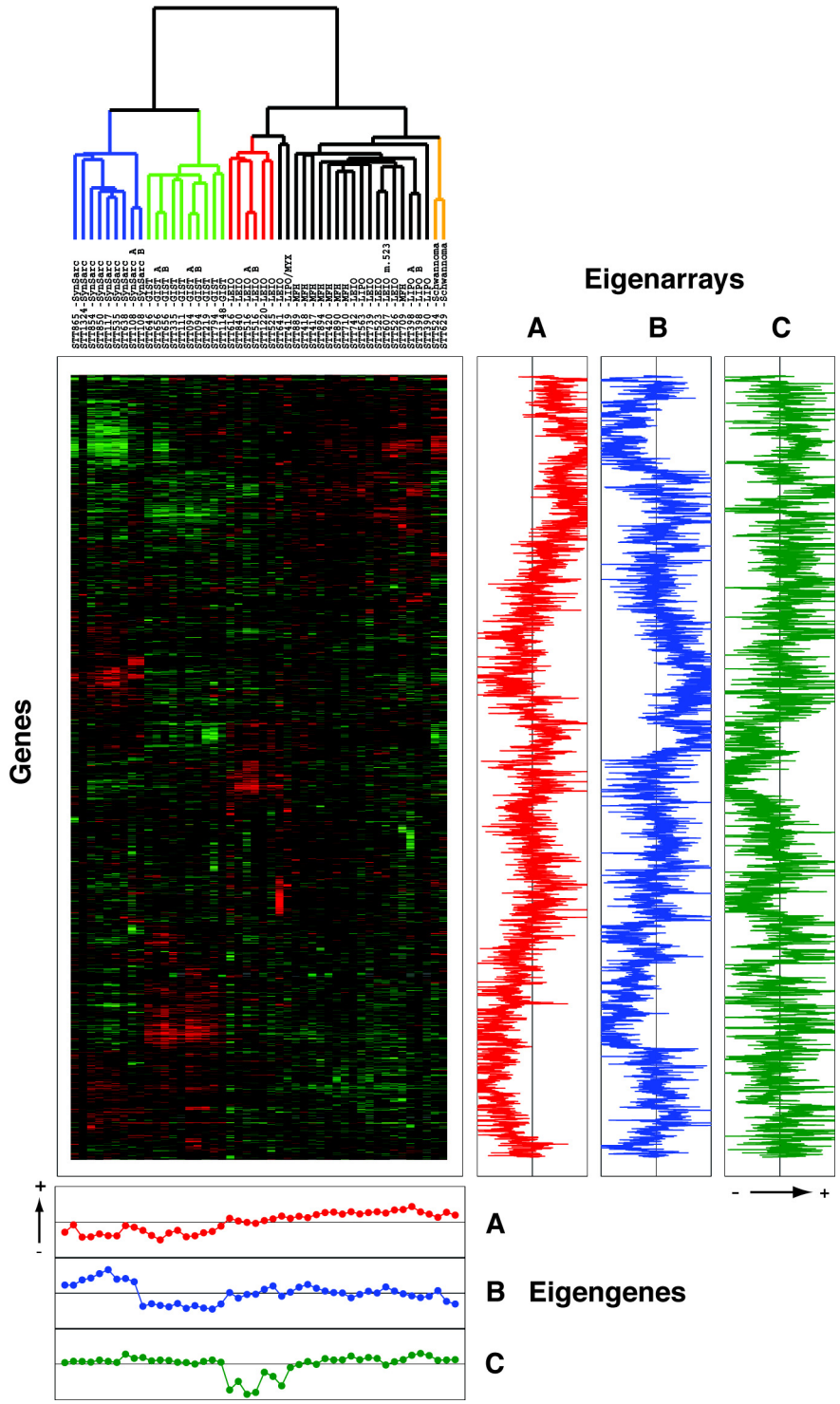
A: ↓ array print labeled "shae"

B: ↑ 23K array prints ↓ 44K array prints

Schaner et al., in preparation.

SVD Classification of Tumor Data

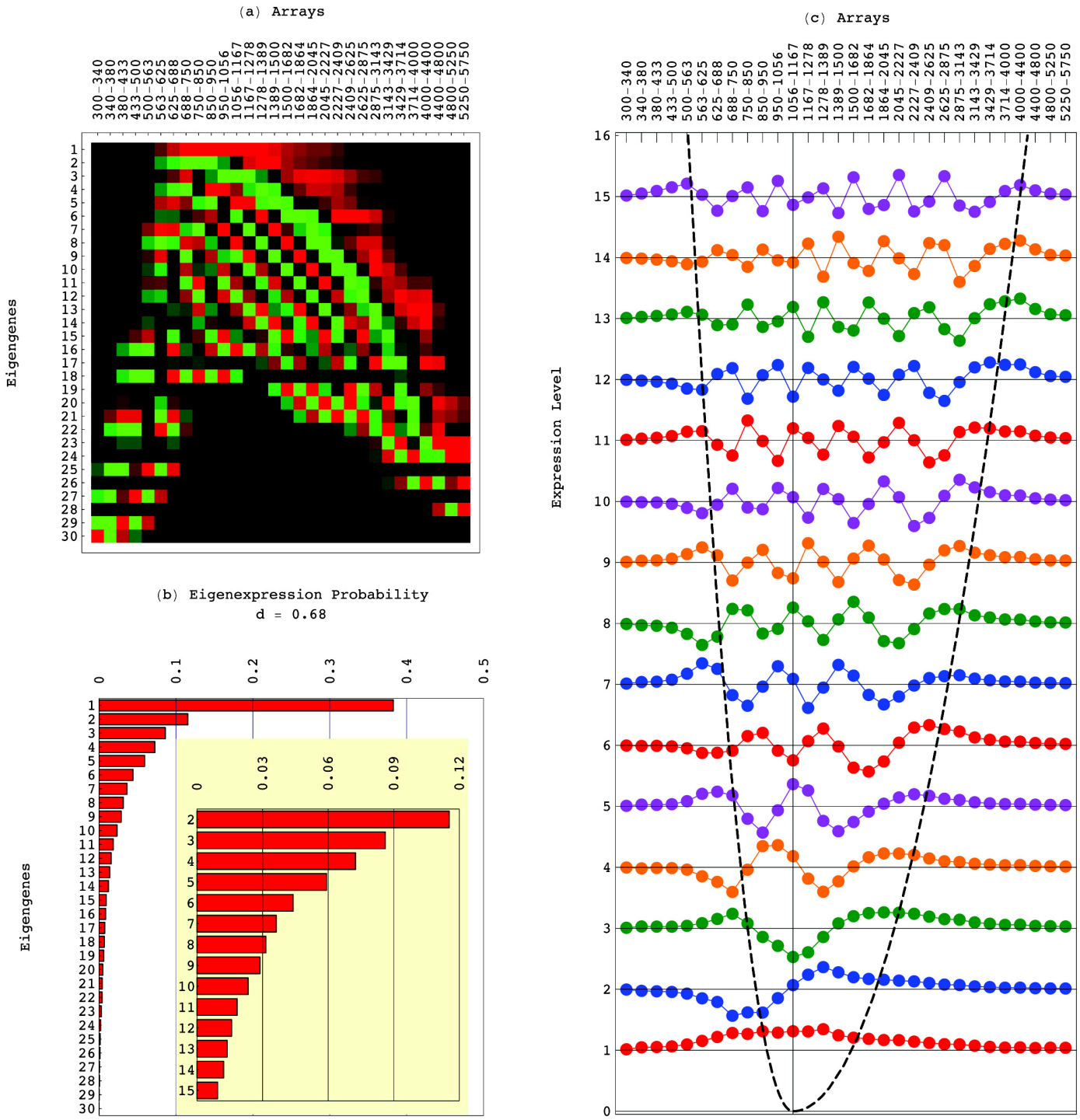
Nielsen et al., *Lancet* 359, 1301 (2002);
Alter, in preparation for *CPMB*.



A: ↓ synovial sarcomas and gastrointestinal stromal tumors (GISTs)
B: ↑ synovial sarcomas ↓ GISTs
C: ↓ leiomyosarcomas that express a group of muscle genes

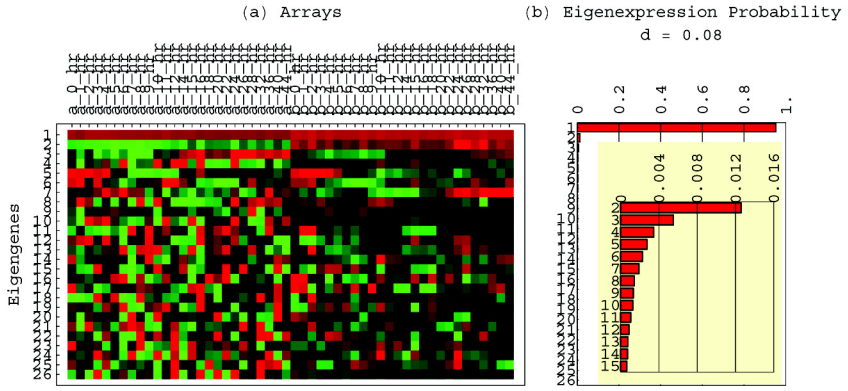
SVD Modeling of Genome-Wide Transcript Size Distribution Data:

Harmonic Oscillator



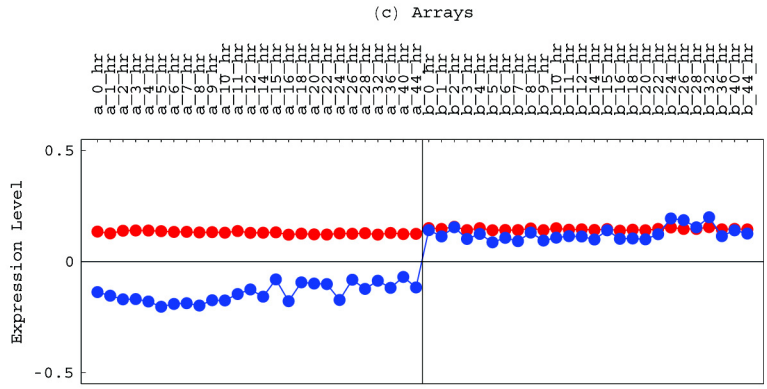
Hurowitz et al., in preparation.

SVD Comparison of Time Courses

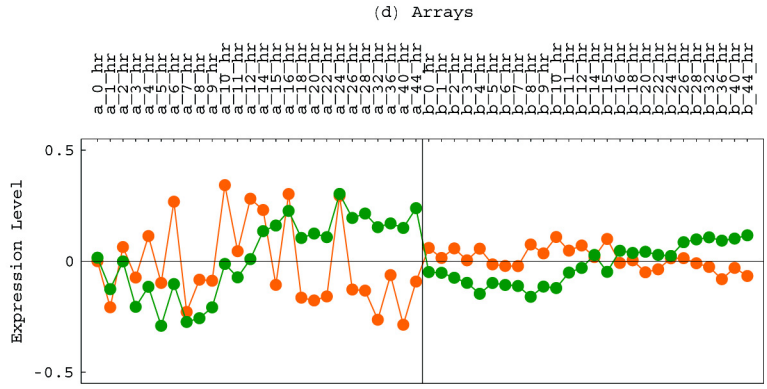


Human Cell Cycle: Double Thymidine Block

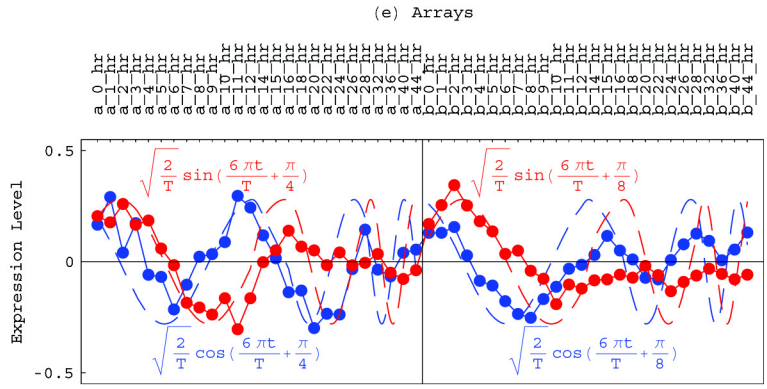
Whitfield et al., *MBC* 13, 1977 (2002).



Discovering artifacts



Comparing experimental protocols

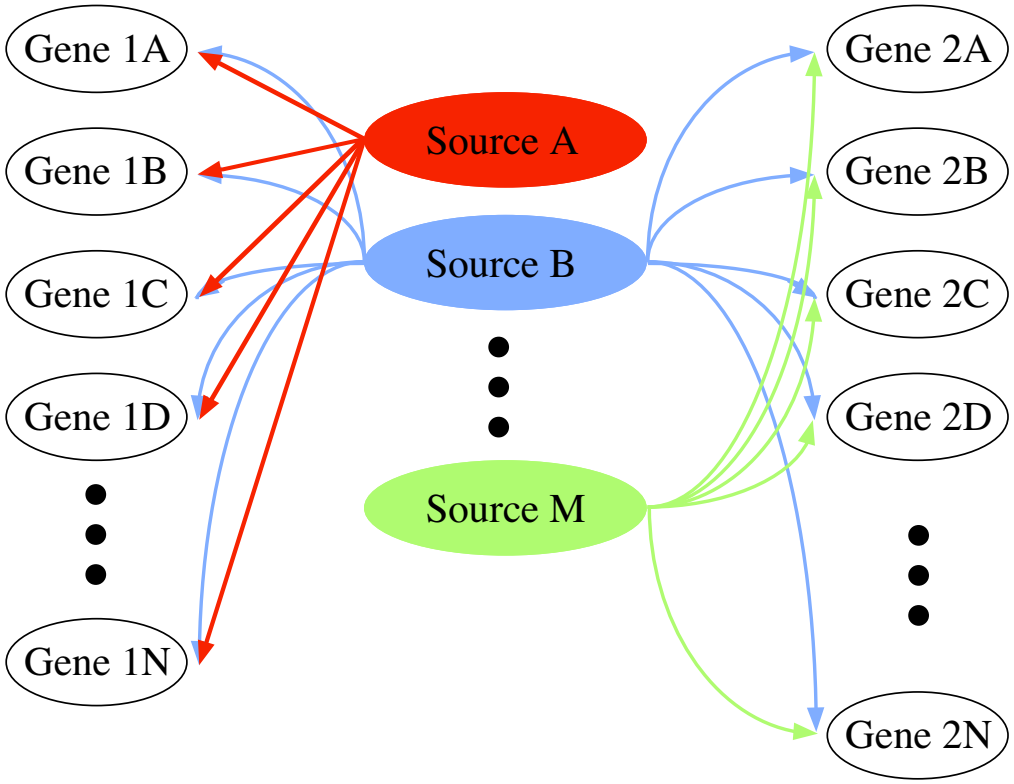


Comparing biological processes

GSVD for Comparative Modeling of Two Genome-Scale Datasets

Alter, Brown & Botstein, *PNAS* 100, 3351 (2003);
http://genome-www.stanford.edu/GSVD/.

GSVD formulates genome-scale expression as a superposition of the effects of several independent sources, such as regulatory programs, biological processes and experimental artifacts, that are common to both datasets, and several that are exclusive to one of the datasets or the other.



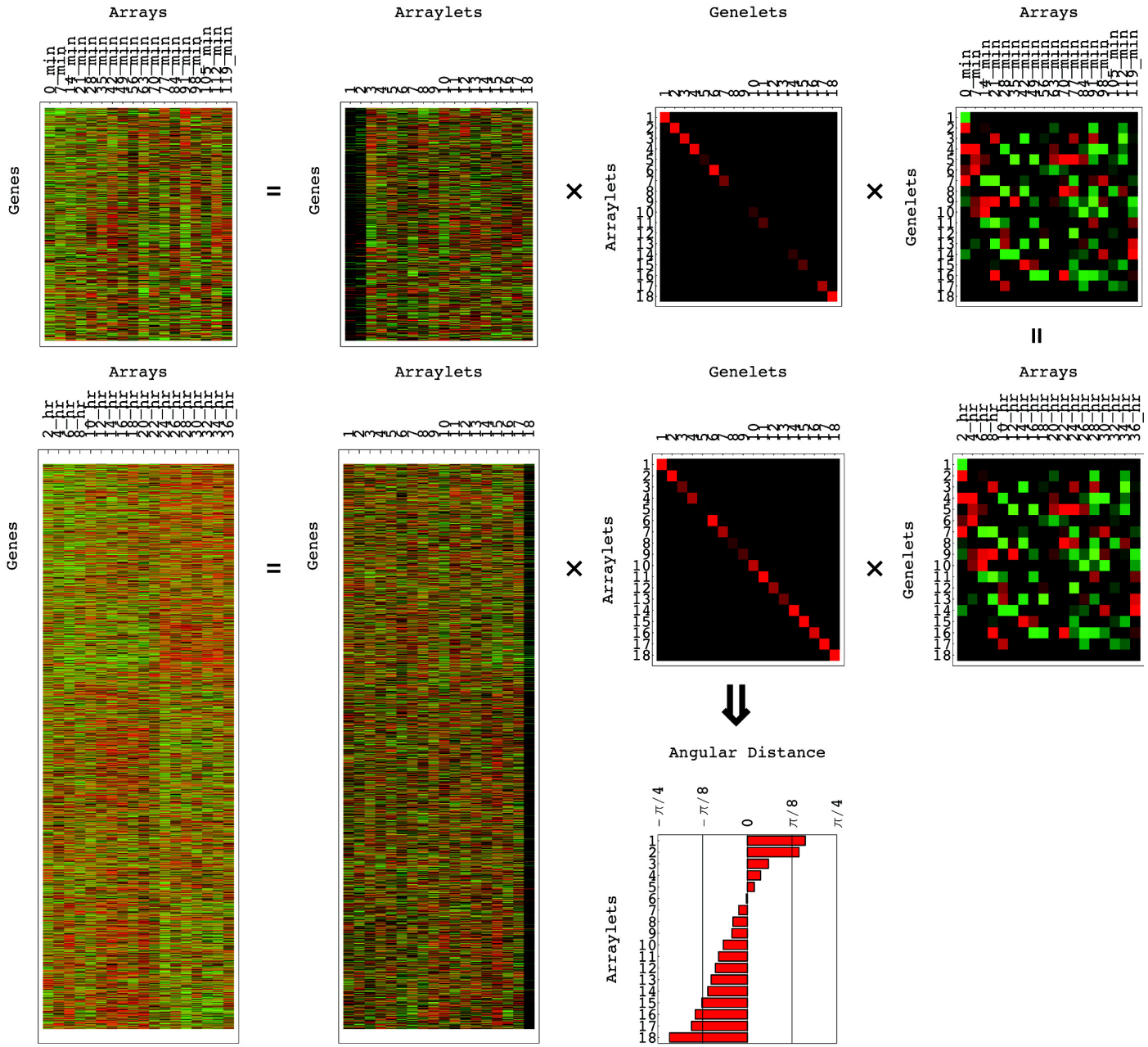
→ Comparative Classification
→ Comparative Data Reconstruction

Generalized SVD (I)

Linear transformation of two datasets from two genes × arrays spaces to two reduced diagonalized "genelets" × "arraylets" spaces.

Yeast Cell Cycle: Alpha Factor

Spellman et al., *MBC* 9, 3273 (1998).

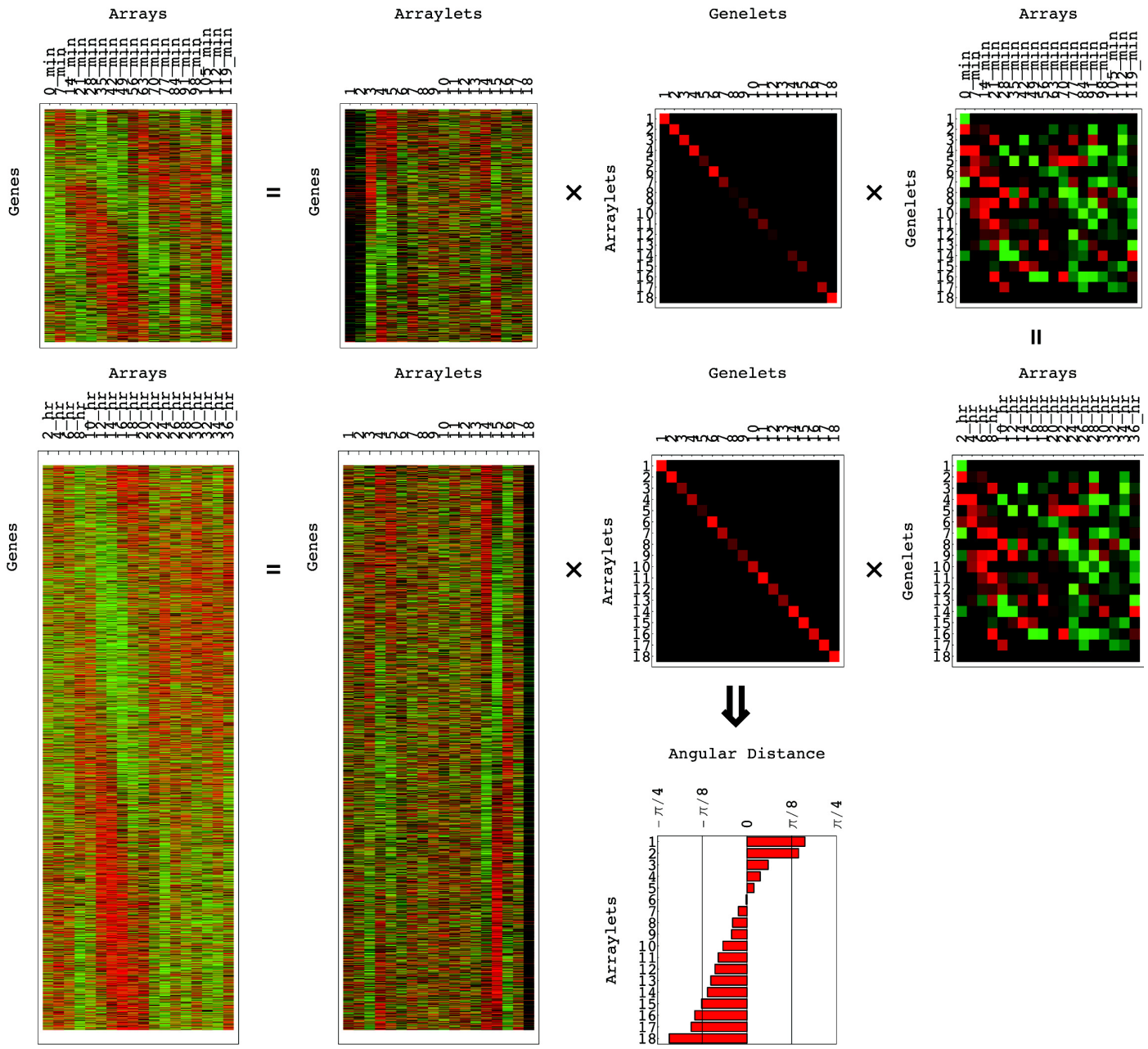


Human Cell Cycle: Double Thymidine Block

Whitfield et al., *MBC* 13, 1977 (2002).

Generalized SVD (II)

The genelets and arraylets are data-driven decoupled superpositions of genes and arrays. The arraylets are orthogonal (decorrelated) projections of the datasets onto the space spanned by the non-orthogonal genelets, that are shared by both datasets, with the "angular distances" indicating the significance of each genelet in one dataset relative to the other.



GSVD is simultaneous linear transformation of the two expression datasets from the two N_1-genes $\times$ M-arrays and N_2-genes $\times$ M-arrays spaces to the two reduced M-"genelets" $\times$ M-"arraylets" spaces,

$$\hat{e}_1 = \hat{u}_1 \, \hat{\epsilon}_1 \, \hat{x}^{-1},$$
$$\hat{e}_2 = \hat{u}_2 \, \hat{\epsilon}_2 \, \hat{x}^{-1}.$$

The antisymmetric "angular distance" between the datasets indicates the relative significance of the mth genelet,

$$\theta_m = \arctan(\epsilon_{1,m}/\epsilon_{2,m}) - \pi/4.$$

The transformation matrices $\hat{u}_1$ and $\hat{u}_2$ are both orthogonal, while, in general, $\hat{x}^{-1}$ is nonorthogonal,

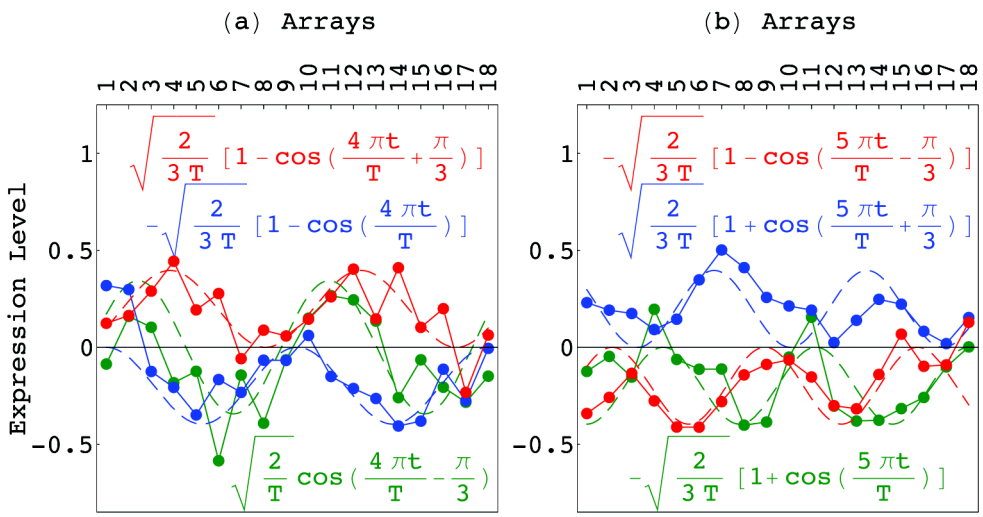
$$\hat{x}^{-1}(\hat{x}^{-1})^T \neq \hat{I} = \hat{u}_1^T \hat{u}_1 = \hat{u}_2^T \hat{u}_2,$$

where $\hat{I}$ is the identity matrix.

Math Variables → Biology (I)

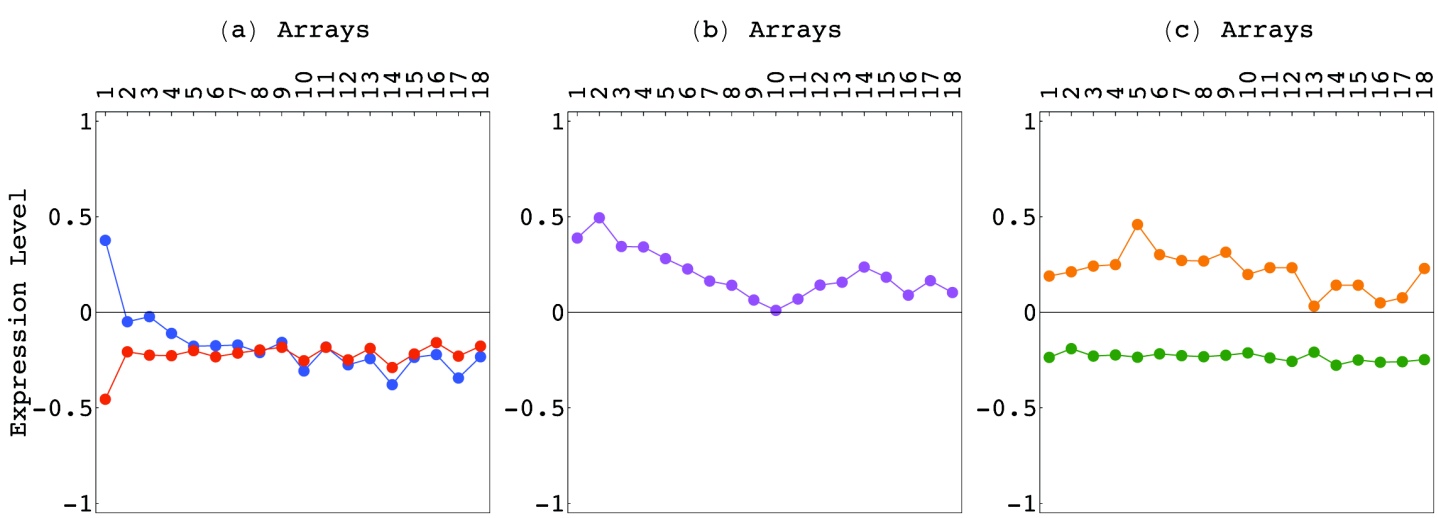
Genelets of almost equal significance in both datasets → processes common to both genomes:

Common Cell Cycle Subspace



Genelets of almost no significance in one dataset relative to the other → genome exclusive processes:

Exclusive Synchronization Responses Subspaces



← *Saccharomyces cerevisiae* Human →

	Classification	Dataset	Genelet and arraylet	Most likely parallel association	P-value of parallel association	Most likely antiparallel association	P-value of antiparallel association
a	Yeast	Microarray	3	G_1	2.1×10^{-49}	G_2/M	1.6×10^{-18}
			4	G_2/M	2.9×10^{-15}	G_1	1.1×10^{-36}
			5	M/G_1	1.3×10^{-36}	S/G_2	7.2×10^{-8}
			14	G_2/M	8.8×10^{-8}	G_1	2.6×10^{-13}
			15	S/G_2	5.9×10^{-7}	G_1	3.3×10^{-14}
			16	M/G_1	6.6×10^{-9}	S	7.5×10^{-3}
b		Traditional	3	G_1	1.7×10^{-12}	G_2/M	1.9×10^{-4}
			4	M/G_1	8.2×10^{-6}	G_1	2.6×10^{-22}
			5	M/G_1	1.2×10^{-10}	S	5.4×10^{-4}
			14	G_2/M	1.9×10^{-4}	G_1	2.2×10^{-8}
			15	G_2/M	3.2×10^{-3}	G_1	5.4×10^{-14}
			16	M/G_1	2.6×10^{-7}	S	1.5×10^{-5}
c	Human	Microarray	3	G_2/M	5.6×10^{-3}	G_1/S	7.9×10^{-2}
			4	S	8.2×10^{-21}	M/G_1	1.3×10^{-3}
			5	G_2/M	6.9×10^{-8}	G_1/S	6.4×10^{-5}
			14	G_2	3.3×10^{-34}	M/G_1	1.3×10^{-3}
			15	G_1/S	4.0×10^{-37}	G_2	1.9×10^{-37}
			16	G_2/M	2.0×10^{-33}	G_1/S	3.0×10^{-10}
d		Traditional	3	G_2/M	1.0×10^{-1}	S	3.2×10^{-3}
			4	S	1.5×10^{-8}	G_1/S	7.6×10^{-3}
			5	G_2	4.9×10^{-2}	G_1/S	1.2×10^{-1}
			14	G_2	6.6×10^{-8}	None	5.4×10^{-1}
			15	G_1/S	2.1×10^{-13}	G_2/M	2.1×10^{-14}
			16	G_2/M	9.0×10^{-17}	S	1.1×10^{-5}

Estimate the probability for such a coherent biological theme to be reflected in the annotations of a random unordered group of $n_i \leq N_i$ genes with the largest either positive or negative coefficients of the genelet, selected from the group of all N_i genes in either dataset without replacements, where $K_i \leq N_i$ of them are annotated to be of a particular molecular function or biological process, and where $k_i \leq K_i$ of them are in the group of the n_i genes, using combinatorics:

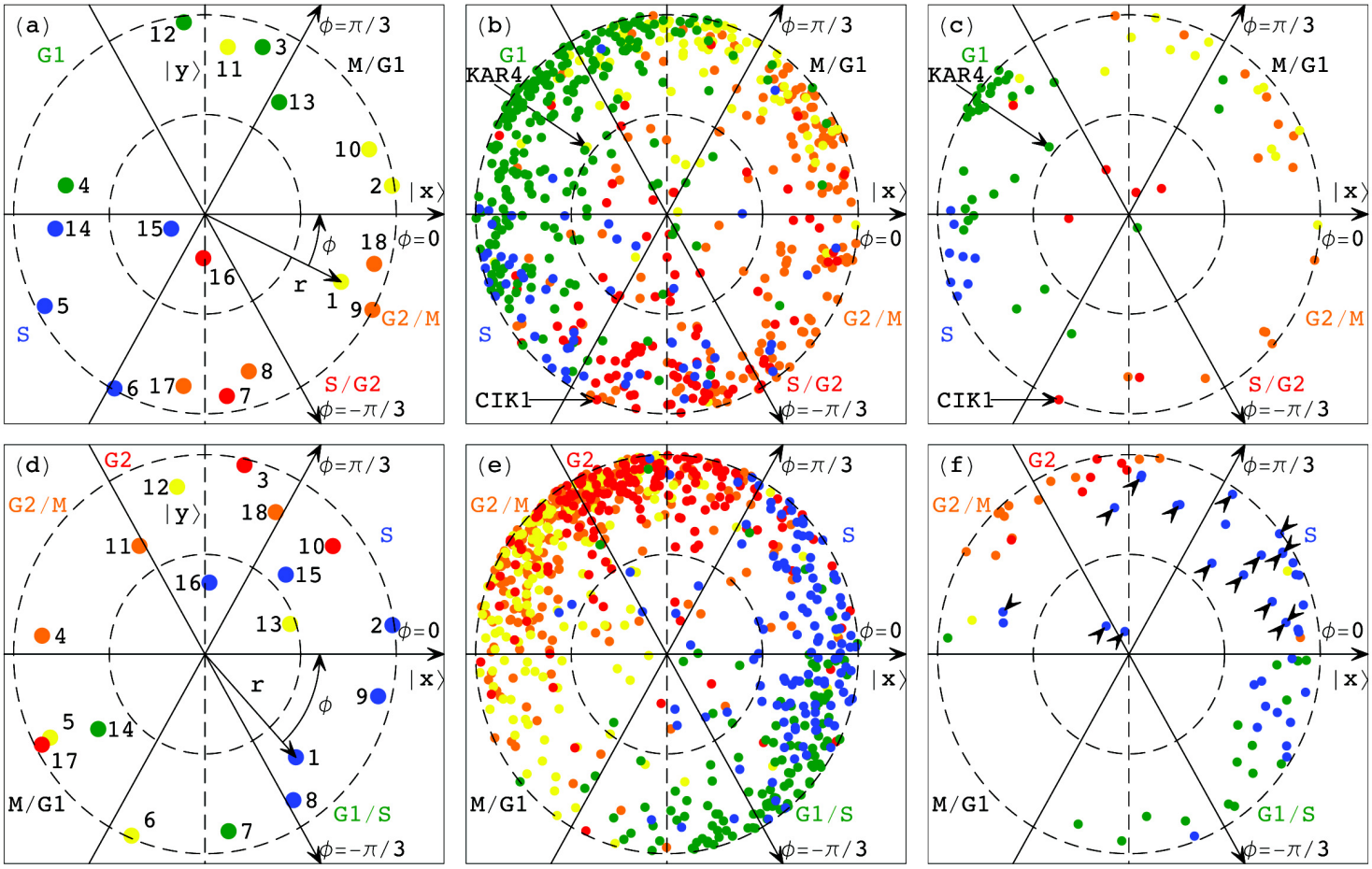
$$P(k_i; n_i, N_i, K_i) = \binom{N_i}{n_i}^{-1} \sum_{k=k_i}^{n_i} \binom{K_i}{k} \binom{N_i - K_i}{n_i - k}.$$

Math Operations → Biology (I)

Reconstruction of gene (array) expression in a subset of genelets (arraylets) → experimental observation of the cellular program (state) these genelets (arraylets) represent, with no other processes (states) present:

Simultaneous Classification in Common Cell Cycle Subspace …

Saccharomyces cerevisiae



Human

… outlines a correspondence between the groups of yeast genes and those of human genes.

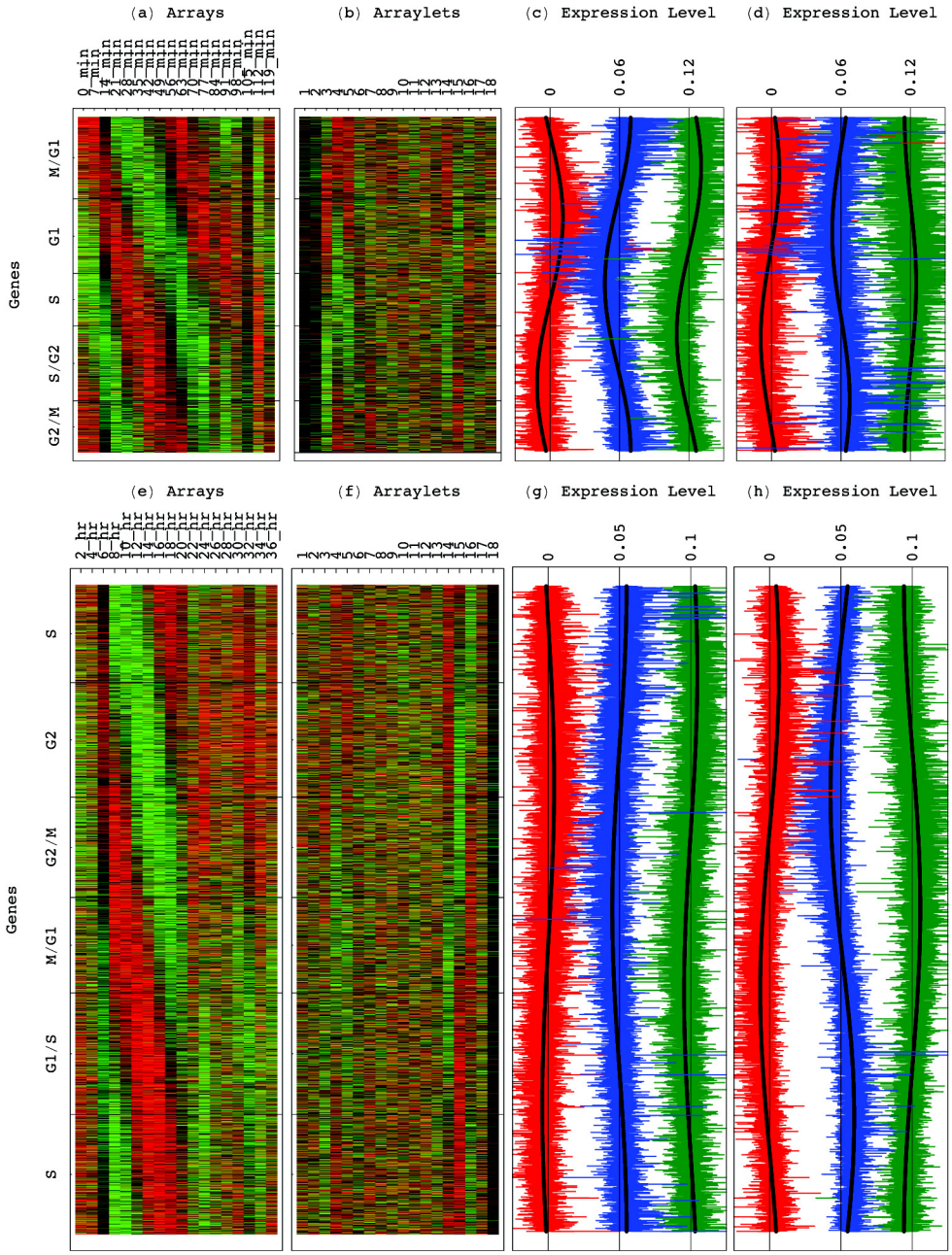
Math Variables → Biology (II)

Arraylets equally significant in both datasets

→ cellular states of common processes

Simultaneous Classification in Common Cell Cycle Subspace

Saccharomyces cerevisiae



Human

Comparative reconstruction of either dataset $\hat{e}_i$, where $i = 1, 2$, in a given subspace of K genelets and corresponding arraylets without eliminating genes or arrays:

$$\hat{e}_i \rightarrow \sum_{k=1}^{K} \epsilon_{i,k} |\alpha_{i,k}\rangle\langle\gamma_k|.$$

Comparative classification of either dataset $\hat{e}_i$ in a given subspace of $K > 2$ genelets: First, least squares-approximate the genelets' subspace with that spanned by the two orthonormal vectors $|x\rangle$ and $|y\rangle$, which maximize

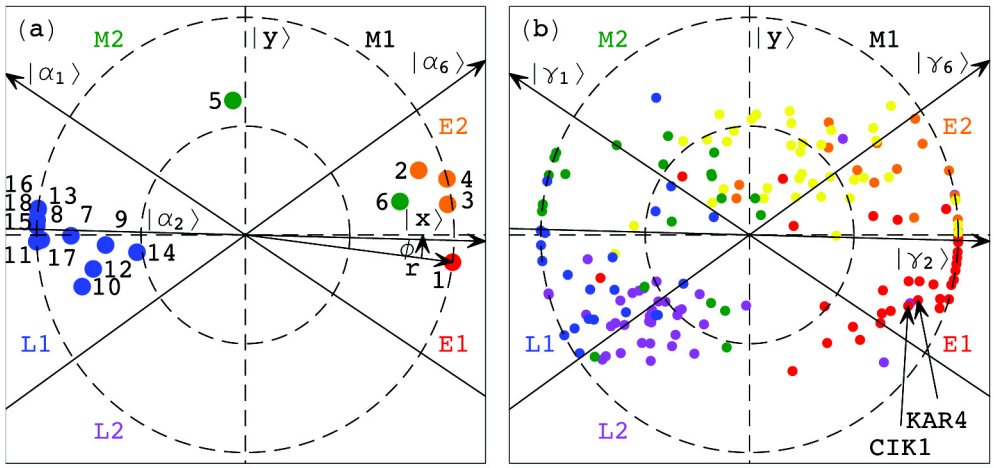
$$\sum_{k=1}^{K} \langle\gamma_k|\big(|x\rangle\langle x| + |y\rangle\langle y|\big)|\gamma_k\rangle.$$

Second, plot the projection of each gene of either dataset from the K-genelets subspace onto $|y\rangle$ along the y-axis vs. that onto $|x\rangle$ along the x-axis, normalized by its ideal amplitude, where the contribution of each genelet to the overall projected expression of the gene adds up rather than cancel out,

$$N_{i,n}^2 = \sum_{k=1}^{K} \sum_{l=1}^{K} \epsilon_{i,k}\epsilon_{i,l} \times \\ |\langle n|\alpha_{i,k}\rangle\langle\alpha_{i,l}|n\rangle\langle\gamma_k|\big(|x\rangle\langle x| + |y\rangle\langle y|\big)|\gamma_l\rangle|.$$

Third, plot the projection of each array of either dataset from the K-arraylets subspace onto $\Sigma_{k=1}^{K} |\alpha_{i,k}\rangle\langle\gamma_k|y\rangle$ along the y-axis vs. that onto $\Sigma_{k=1}^{K} |\alpha_{i,k}\rangle\langle\gamma_k|x\rangle$ along the x-axis, normalized by its ideal amplitude.

Yeast Exclusive Synchronization Response Subspace



signal transduction of mating signal

alpha factor response
cell wall integrity I (SLT2), agglutination
cell cycle arrest, M/G1: CLN3, PCL2, SWI4
mating, adaptation to mating signal

filamentous growth (RAS2)
cell wall integrity II (RHOs)
cell polarity (ACT1)
pseudohyphal growth

ATP synthesis, ARP2/3 protein complex, histones
G1/S: CLB4, CLB5

RSC, SWI/SNF and general chromatin modeling

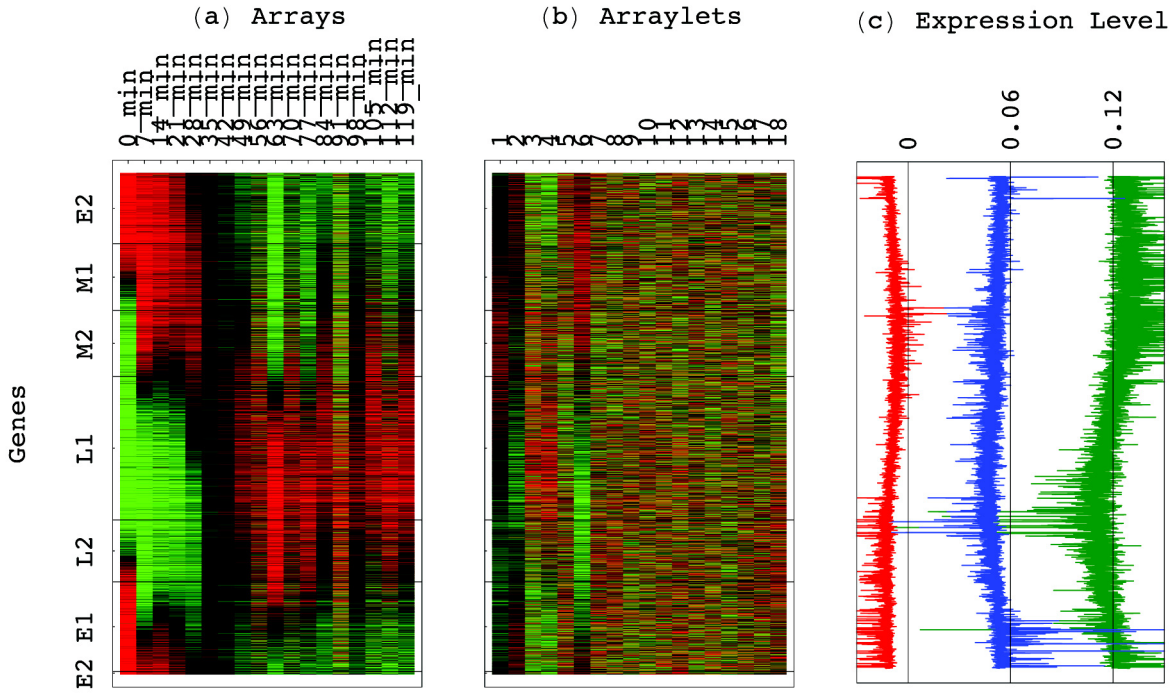
chromatin binding (MCMs) and architecture
STE20 and similar and downregulated genes

phosphate, iron transport
cell cycle control: CDC28, CKS1, BCK2

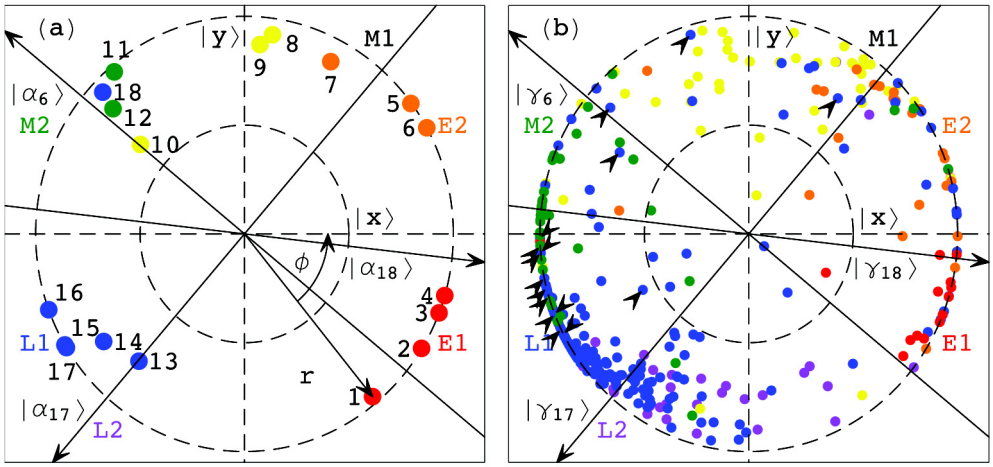
Math Variables → Biology (II)

Arraylets insignificant in one dataset relative to other → cellular state of an exclusive process:

Classification in Yeast Exclusive Pheromone Synchronization Response Subspace

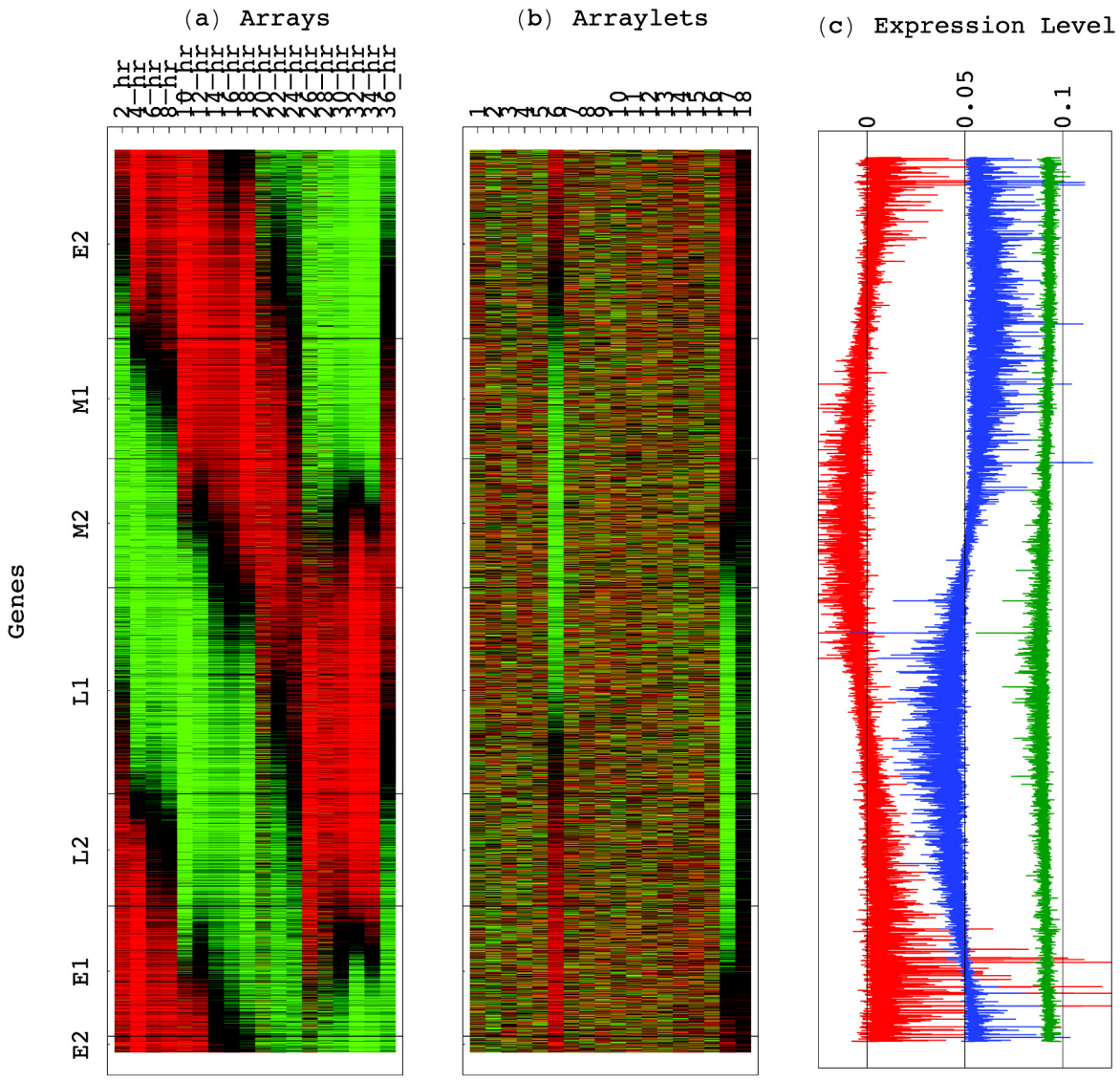


Human Exclusive Synchronization Stress Response Subspace



serum response
phosphatases, ubiquilins
interleukins, integrins, nexins
MCMs
ribosomal proteins, histones, tubulins
cytochromes, golgins, NADHs

Classification in Human Exclusive Synchronization Stress Response Subspace



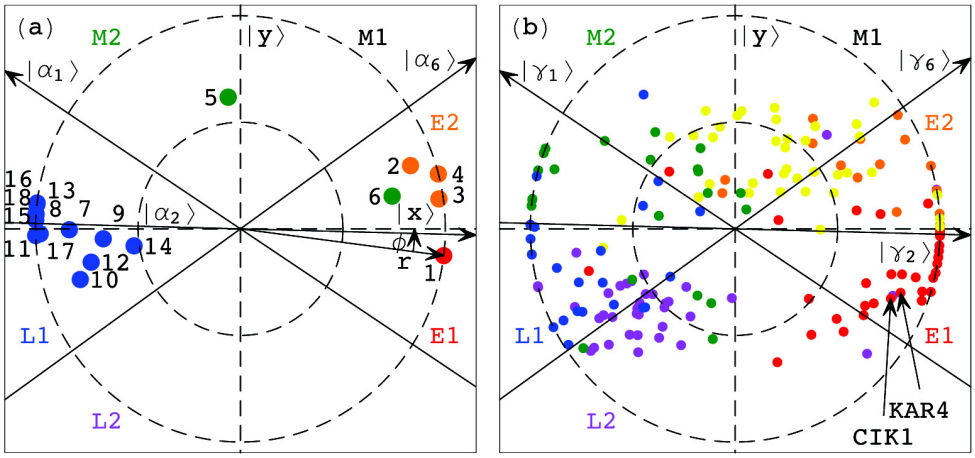
Math Operations → Biology (III)

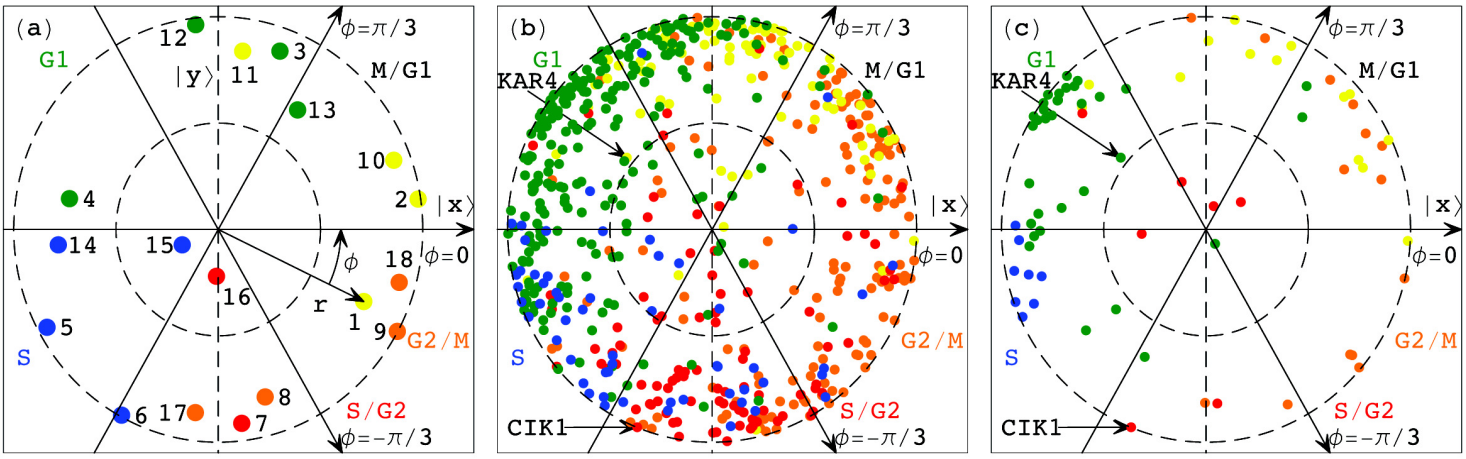
Data reconstruction in two subspaces → experimental observation of differential expression of a genome in the two cellular programs these subspaces represent:

Differential Expression in Yeast During Mating and Cell Cycle

Pheromone Synchronization Response Subspace:

KAR4 is required for CIK1 induction during mating*



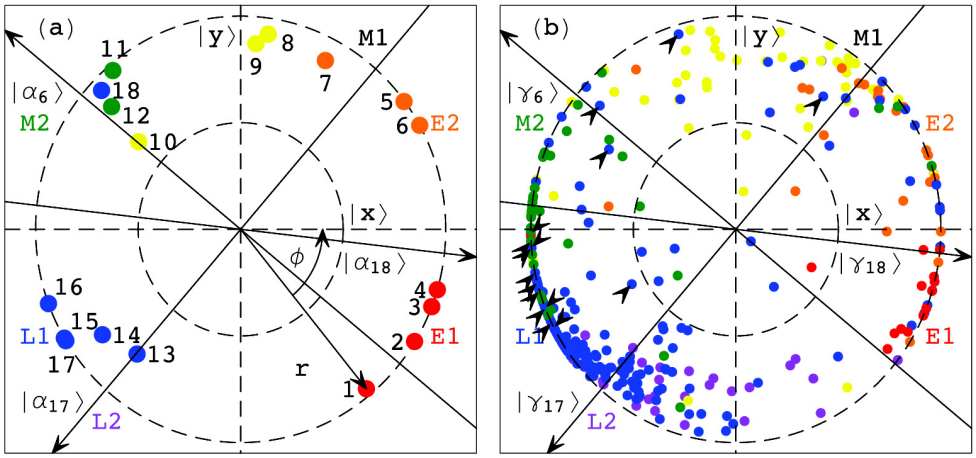


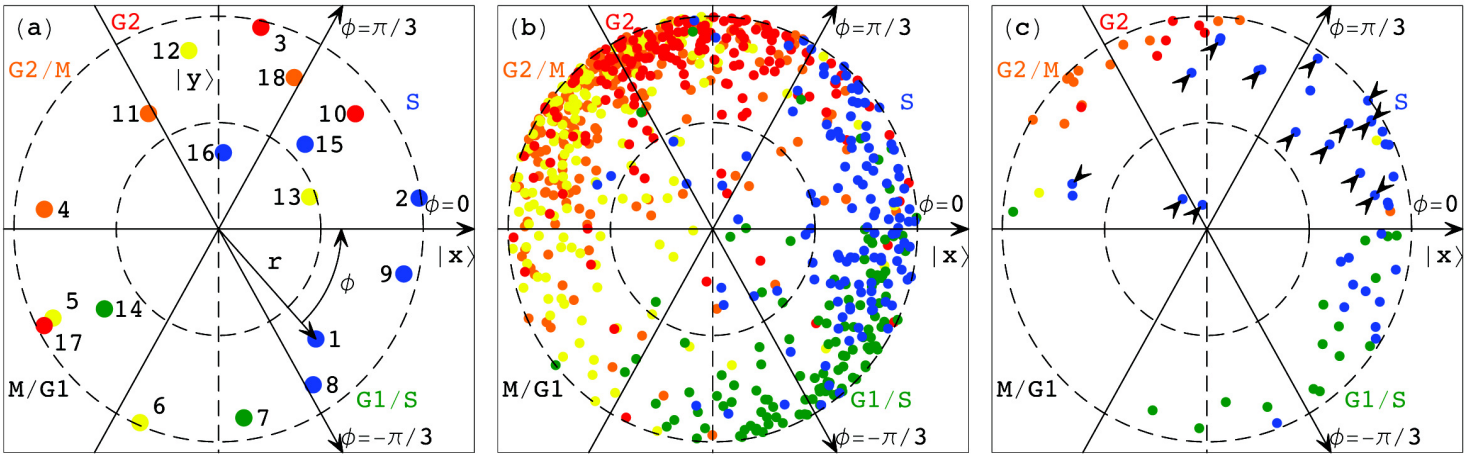
Common Cell Cycle Subspace: Mitotic expression of CIK1 during S/G2 is independent of KAR4*

*Kurihara, Stewart, Gammie & Rose, *MCB* 16, 3990 (1996).

Differential Expression in Human During Stress Synchronization Response and Cell Cycle

Synchronization Stress Response Subspace:
Histones reach expression minima early in time course



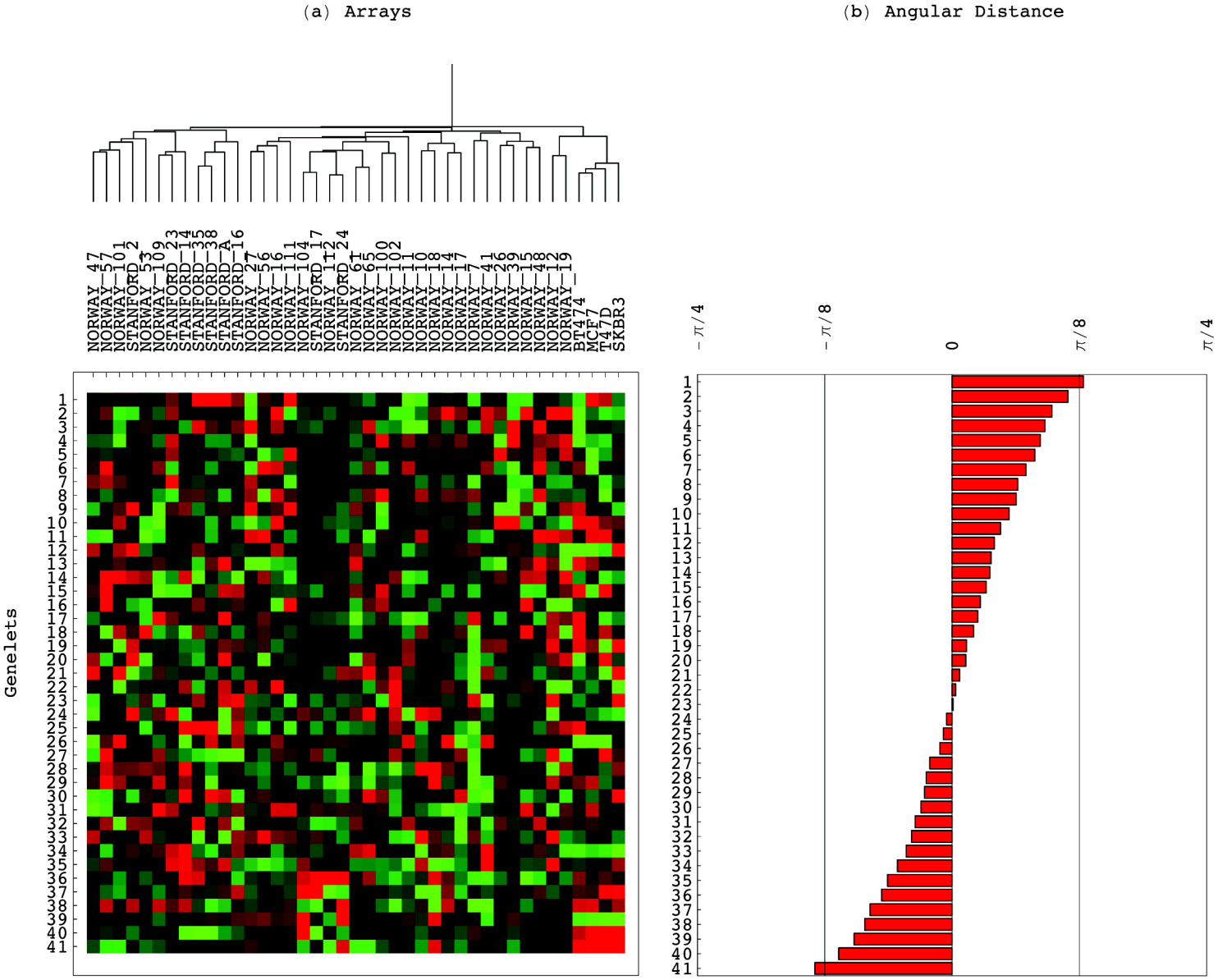


Common Cell Cycle Subspace:
Histones expression peaks early in the time course

- → Which yeast and human genes exhibit periodic expression during the cell cycle?
- → Is there a relation between DNA replication and RNA transcription that is manifested in both the yeast and human genomes?
- → Are there only three cellular modules that drive both the yeast and human cell cycles?

- → Can we design a synthetic genetic network analogous to the 3-transistor ring oscillator circuit, which would simulate the yeast and human cell cycles?
- → Can we identify new orthologs or new promoter motifs by comparing the sequence information for the groups of yeast and human genes that GSVD-correspond to one another?
- → Can we predict differential expression of yeast and human genes that would be validated by experiment?

GSVD Comparison of DNA Copy Number and RNA Expression



Pollack et al., *PNAS* 99, 12963 (2002).

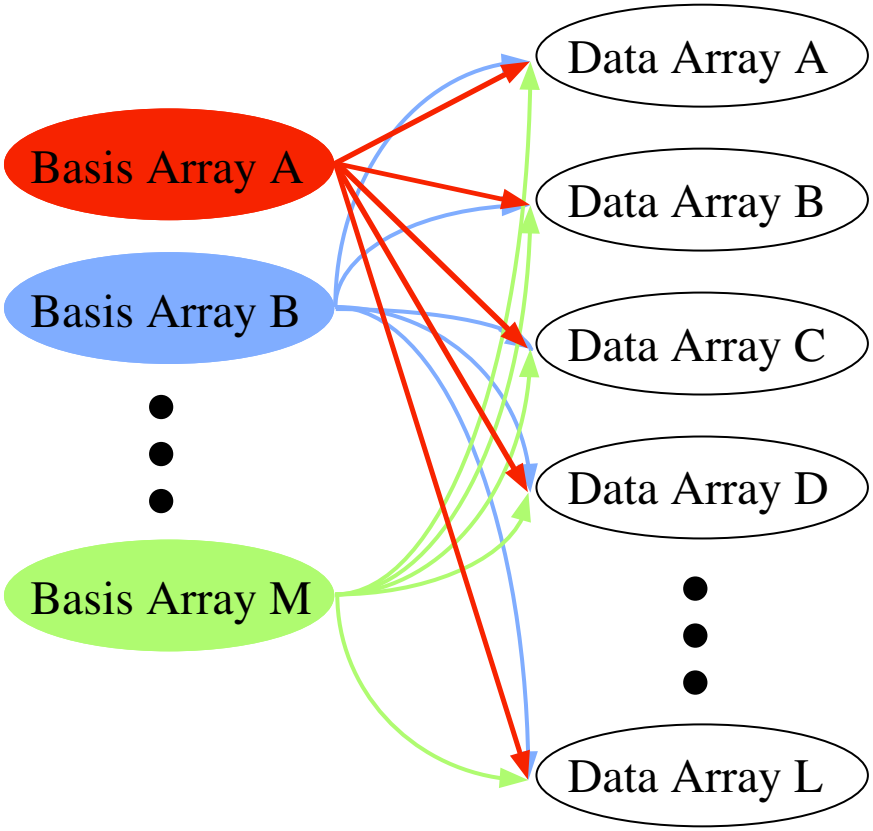
Breast cancer cell lines show higher expression of proliferation genes and lower expression of genes that represent tissue diversity than breast cancer tumors.

This quantitatively significant difference in expression results from an insignificant difference in DNA copy number, if at all.

Pseudoinverse Integrative Modeling of Genome-Scale Datasets

Alter, Golub, Brown & Botstein, *Proc. MNBWS* 15 (2004),
http://www.med.miami.edu/mnbws/Alter-.pdf;
Alter & Golub, in preparation.

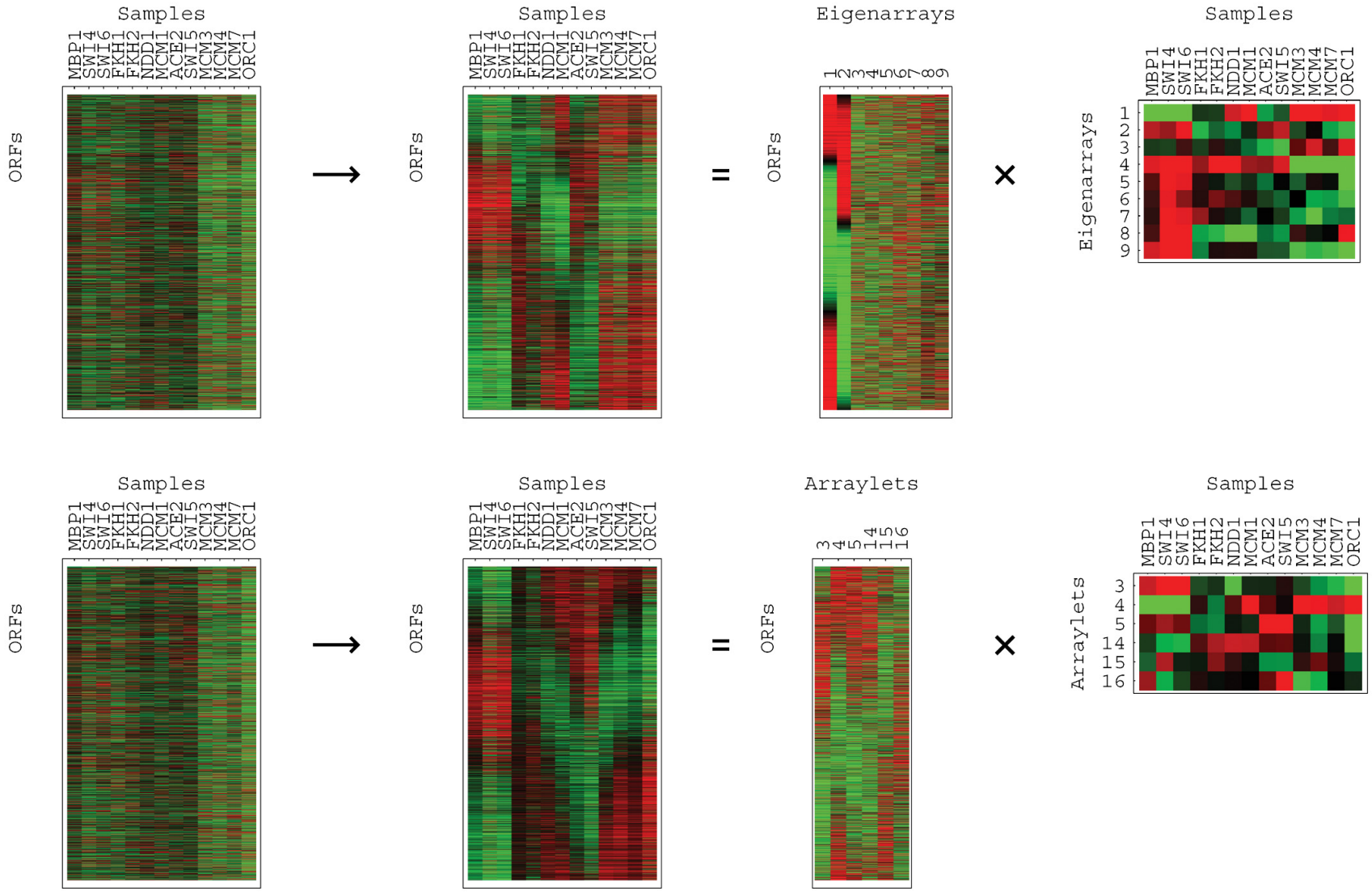
Pseudoinverse approximates the arrays of any number of genome-scale datasets as a superposition of the arrays of one dataset, designated the "basis" set.



→ Integrative Data Reconstruction
→ Integrative Data Classification

Pseudoinverse Projection

Linear transformation of the genome-scale data from ORFs × data arrays space to reduced basis arrays × data arrays space.



Moore-Penrose pseudoinverse projection of the data matrix $\hat{d}$ onto the basis matrix $\hat{b}$ is then linear transformation of the data $\hat{d}$ from the N-ORFs $\times$ L-data samples space to the M-basis samples $\times$ L-data samples space,

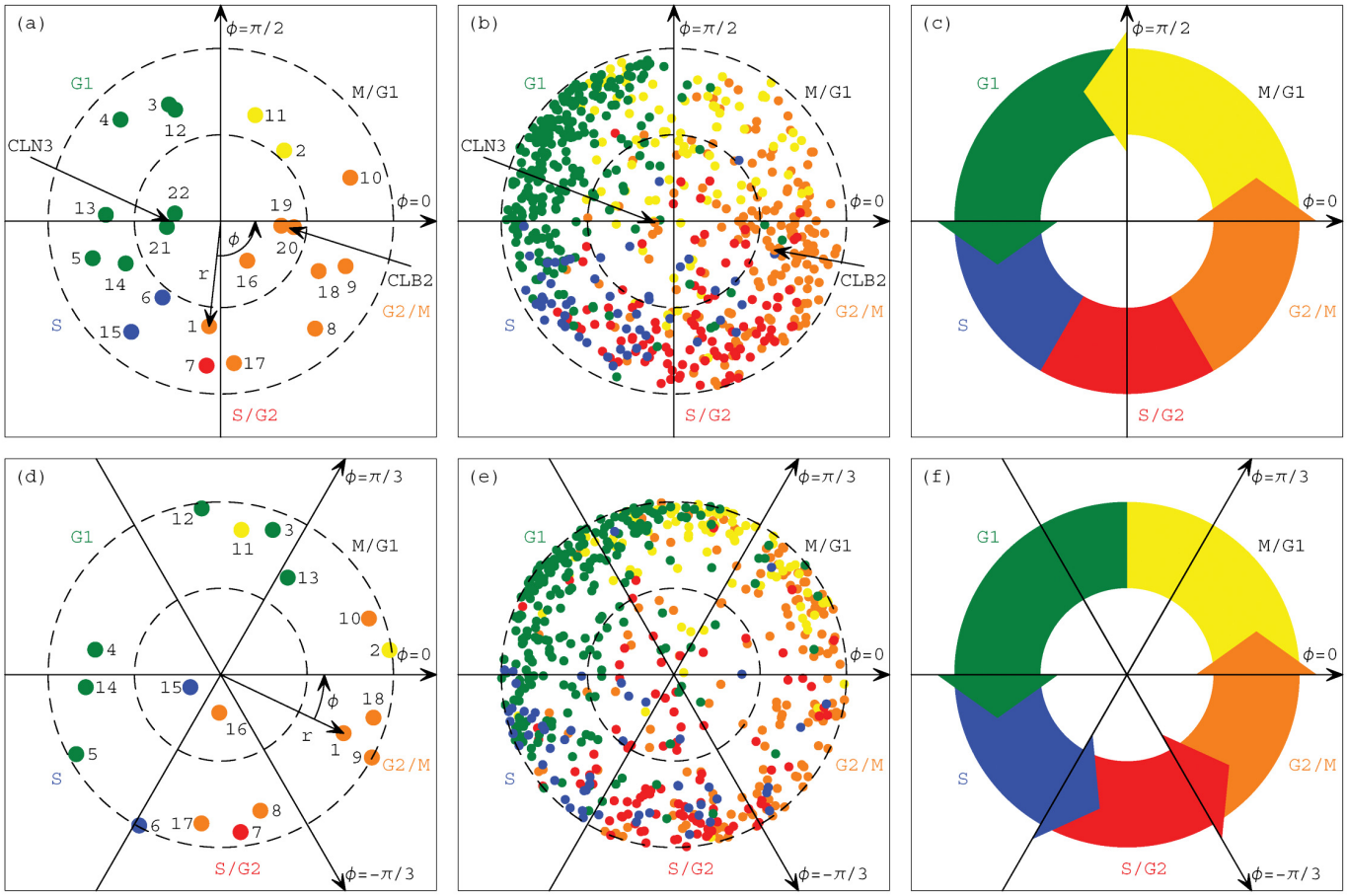
$$\hat{d} \rightarrow \hat{b}\hat{c},$$
$$\hat{b}^{\dagger}\hat{d} \equiv \hat{c},$$

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

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

such that the transformation matrices $\hat{b}\hat{b}^{\dagger}$ and $\hat{b}^{\dagger}\hat{b}$ are orthogonal projection matrices.

Pseudoinverse Projection Integrative Analysis of Cell Cycle RNA Transcription …



… and Proteins' DNA-Binding

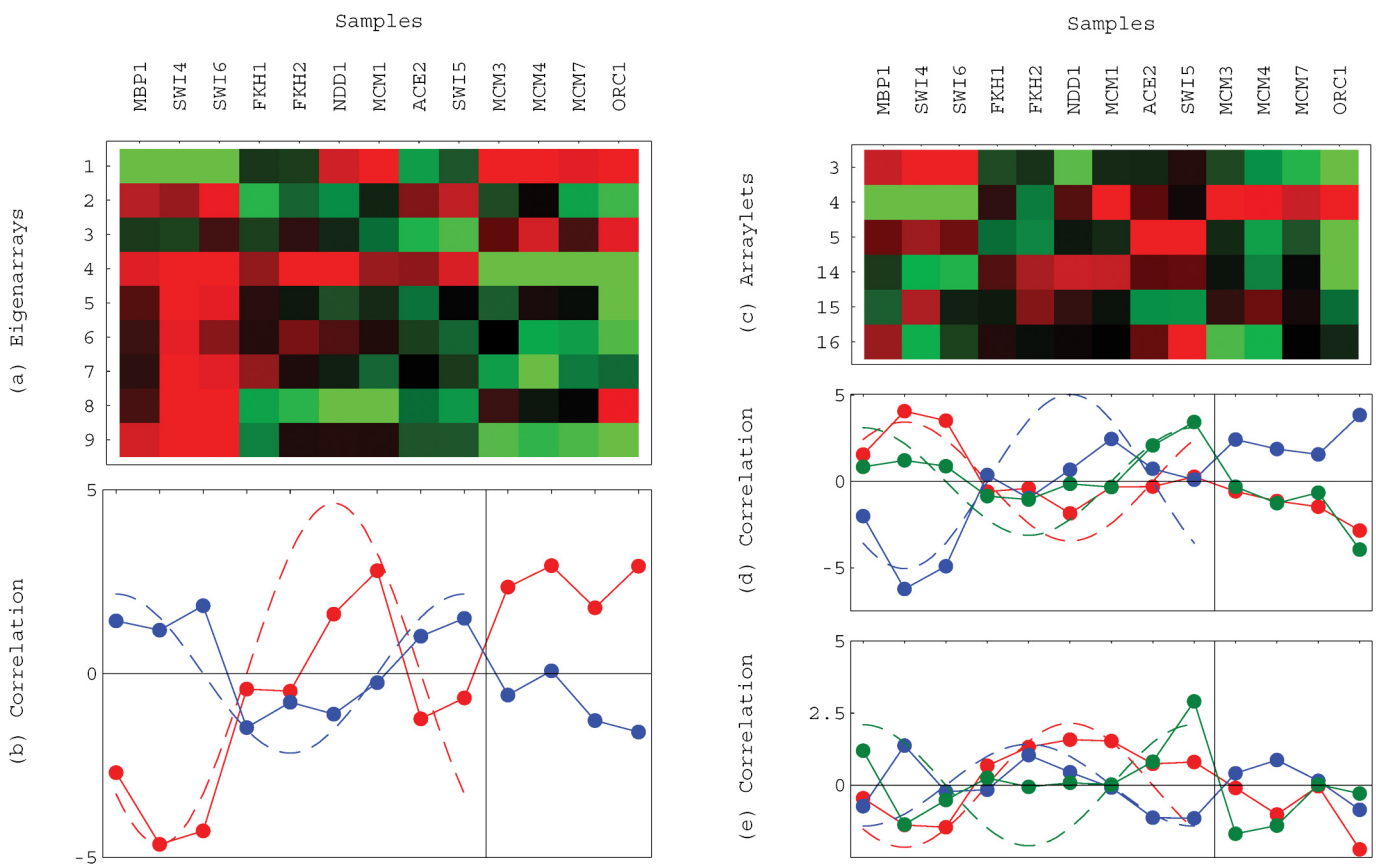
Binding of Transcription Factors:
Mbp1, Swi4, Swi6, Fkh1, Fkh2, Ndd1, Mcm1, Ace2, Swi5
Simon et al., *Cell* 106, 697 (2001).

Binding of Replication Initiation Proteins:
Orc1, Mcm3, Mcm4, Mcm7
Wyrick et al., *Science* 294, 2397 (2001).

Math Variables → Biology (I)

Pseudoinverse correlations define the transformation of variables → transformation of the patterns of the data into the patterns of the basis:

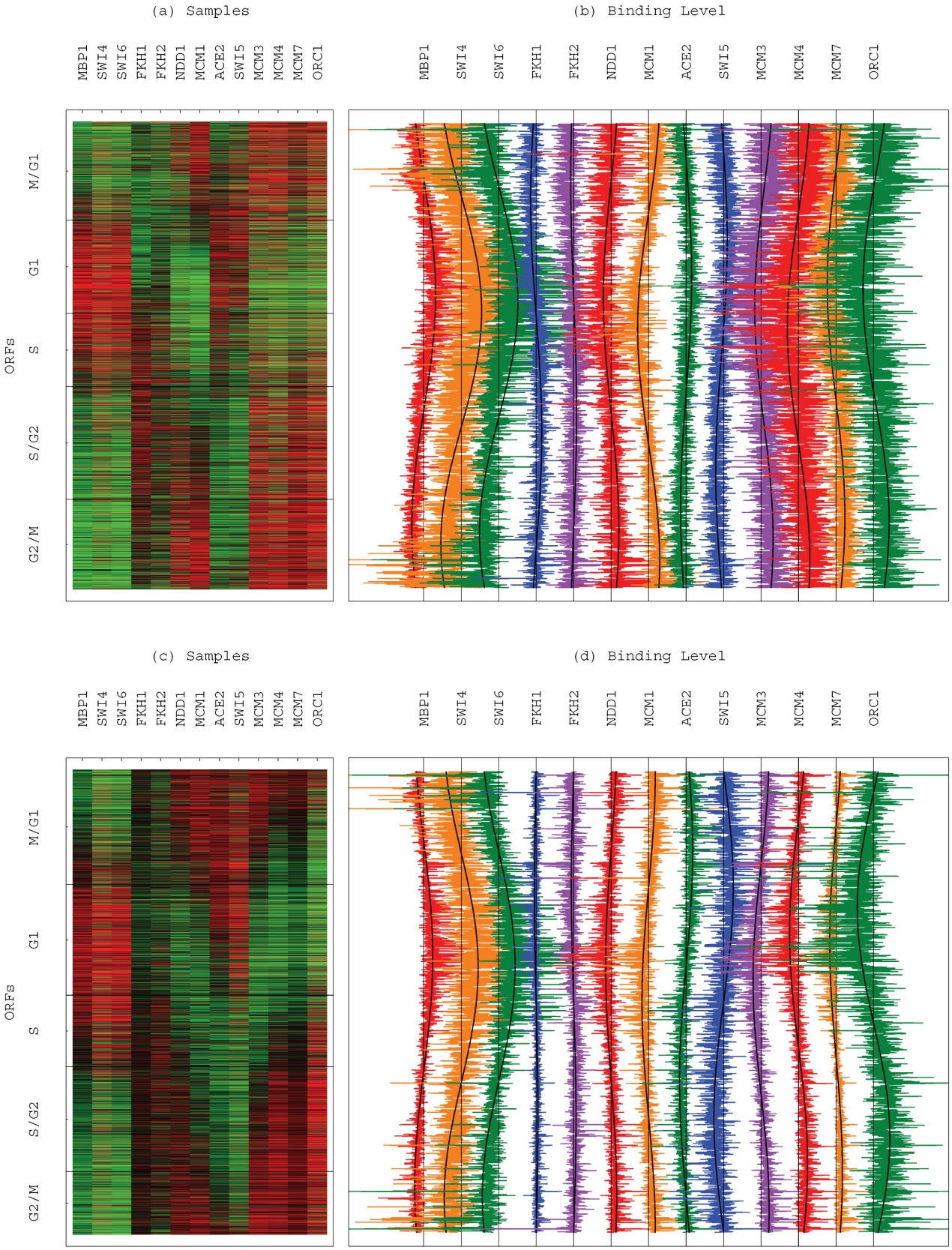
Pseudoinverse Correlations



Math Variables → Biology (II)

Pseudoinverse data reconstruction → experimental observation of only the cellular states manifested in the data that correspond to the states manifested in the basis:

Pseudoinverse Projecion Integrative Data Reconstruction



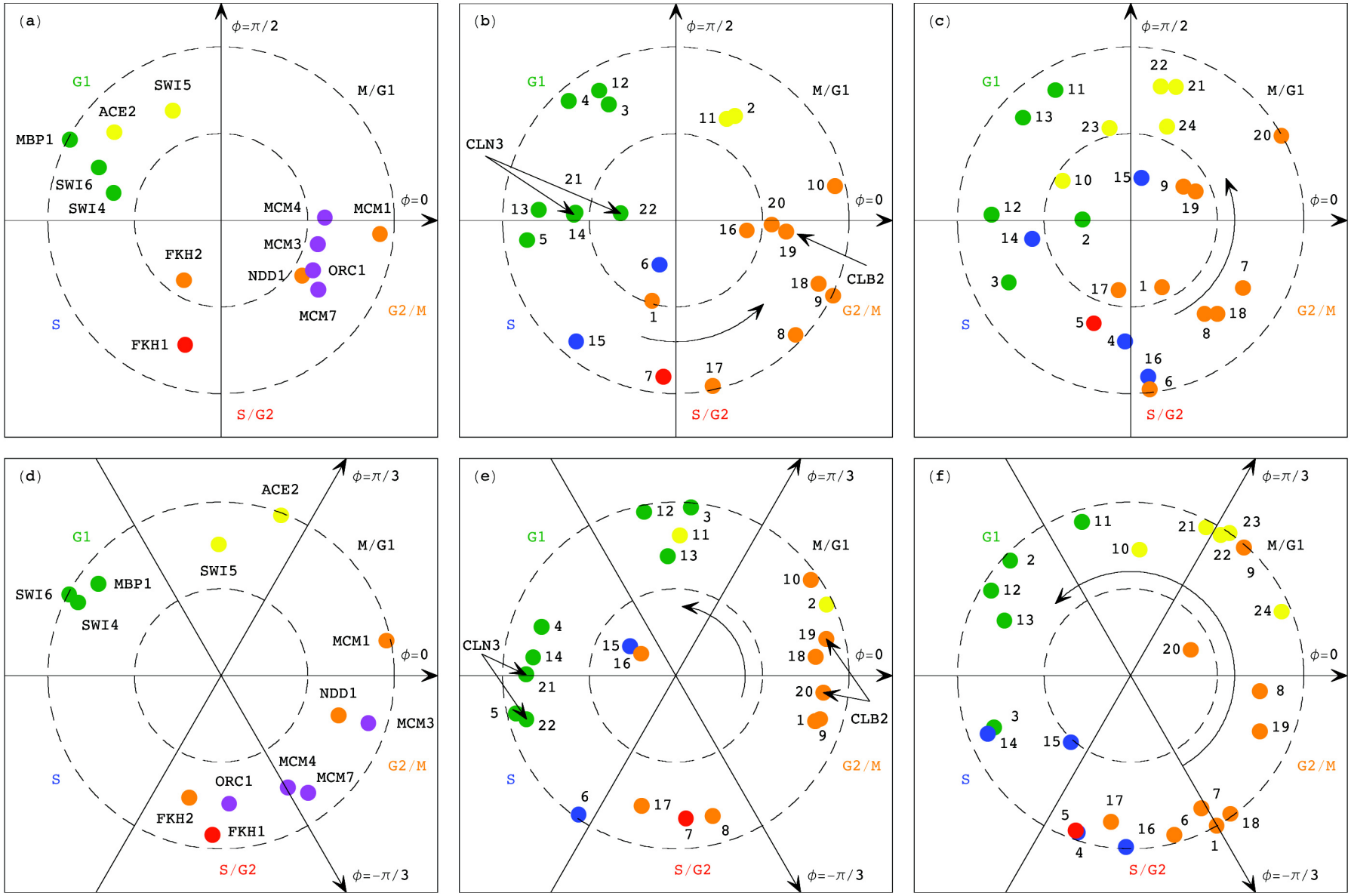
Integrative reconstruction of any one of a number of datasets $\hat{d}$ in the basis set $\hat{b}$ without eliminating ORFs or samples,

$$|d_l\rangle \rightarrow \sum_{m=0}^{M} c_{ml} \, |b_m\rangle,$$

that least-squares-approximates each of the samples that constitute the data matrix $\hat{d}$ by a linear superposition of the samples that constitute the basis matrix $\hat{b}$.

Integrative classification of the reconstructed data samples by similarity in the contribution of each of the basis samples to their overall profile rather than by their overall profile alone.

Pseudoinverse Integrative Mapping
Novel Correlation: DNA ↔ RNA



Probabilistic Associations by Annotations

				Microarray annotation				Traditional annotation			
				Parallel association		Antiparallel association		Parallel association		Antiparallel association	
	Dataset	Array		Most likely	P-value of	Most likely	P-value of	Most likely	P-value of	Most likely	P-value of
a	Binding of cell cycle transcription factors		*MBP1*	G_1	1.6×10^{-14}	None	4.0×10^{-3}	G_1	2.7×10^{-10}	None	9.3×10^{-2}
			SWI4	G_1	1.5×10^{-17}	None	1.2×10^{-1}	G_1	2.7×10^{-7}	None	9.3×10^{-2}
			SWI6	G_1	4.7×10^{-32}	G_2/M	7.3×10^{-2}	G_1	4.8×10^{-19}	G_2/M	4.4×10^{-2}
			FKH1	S/G_2	7.2×10^{-4}	None	3.5×10^{-1}	S/G_2	4.0×10^{-2}	S	3.9×10^{-1}
			FKH2	G_2/M	3.9×10^{-11}	None	8.3×10^{-2}	G_2/M	3.7×10^{-6}	None	2.7×10^{-2}
			NDD1	G_2/M	2.0×10^{-19}	G_1	9.5×10^{-2}	G_2/M	5.0×10^{-9}	M/G_1	3.3×10^{-1}
			MCM1	G_2/M	1.2×10^{-12}	G_1	4.0×10^{-3}	G_2/M	1.6×10^{-7}	G_1	3.3×10^{-2}
			ACE2	M/G_1	1.1×10^{-3}	G_2/M	8.4×10^{-3}	M/G_1	1.1×10^{-1}	S	7.8×10^{-2}
			SWI5	M/G_1	1.3×10^{-15}	G_1	4.5×10^{-5}	M/G_1	6.2×10^{-4}	G_2/M	6.2×10^{-5}
b	Binding of DNA replication initiation proteins		*ORC1*	None	4.0×10^{-3}	G_1	4.3×10^{-13}	None	2.2×10^{-1}	G_1	5.0×10^{-4}
			MCM3	None	4.5×10^{-4}	G_1	7.9×10^{-10}	None	2.7×10^{-2}	G_1	5.0×10^{-4}
			MCM4	None	1.3×10^{-2}	G_1	1.2×10^{-8}	None	4.0×10^{-3}	G_1	2.4×10^{-3}
			MCM7	None	1.3×10^{-2}	G_1	7.9×10^{-10}	None	2.7×10^{-2}	G_1	5.0×10^{-4}
c	α factor cell cycle expression time course	1	0 min	G_2/M	3.2×10^{-6}	G_1	4.9×10^{-27}	M/G_1	4.4×10^{-6}	G_1	7.0×10^{-14}
		2	7 min	M/G_1	5.7×10^{-4}	S	1.3×10^{-6}	M/G_1	1.3×10^{-2}	S	3.4×10^{-6}
		3	14 min	G_1	4.3×10^{-26}	G_2/M	3.2×10^{-6}	G_1	4.2×10^{-7}	S	3.4×10^{-6}
		4	21 min	G_1	3.9×10^{-57}	G_2/M	1.1×10^{-18}	G_1	7.0×10^{-14}	M/G_1	1.7×10^{-7}
		5	28 min	G_1	4.5×10^{-19}	G_2/M	6.2×10^{-16}	G_1	2.0×10^{-11}	M/G_1	4.5×10^{-9}
		6	35 min	S	2.1×10^{-10}	M/G_1	2.1×10^{-20}	S	3.4×10^{-6}	M/G_1	4.4×10^{-6}
		7	42 min	S/G_2	1.2×10^{-11}	M/G_1	4.8×10^{-25}	S	1.0×10^{-2}	M/G_1	1.1×10^{-12}
		8	49 min	G_2/M	6.2×10^{-16}	M/G_1	4.7×10^{-30}	G_2/M	7.6×10^{-3}	M/G_1	1.1×10^{-12}
		9	56 min	G_2/M	3.1×10^{-31}	G_1	6.8×10^{-51}	G_2/M	2.7×10^{-8}	G_1	3.5×10^{-15}
		10	63 min	G_2/M	6.2×10^{-16}	G_1	6.7×10^{-16}	G_2/M	5.7×10^{-4}	G_1	4.2×10^{-8}
		11	70 min	M/G_1	1.6×10^{-21}	S/G_2	4.1×10^{-9}	M/G_1	1.7×10^{-7}	S	3.4×10^{-6}
		12	77 min	G_1	5.1×10^{-61}	S/G_2	1.4×10^{-7}	G_1	2.3×10^{-22}	S/G_2	5.5×10^{-3}
		13	84 min	G_1	5.7×10^{-34}	G_2/M	1.4×10^{-21}	G_1	1.6×10^{-16}	G_2/M	1.1×10^{-6}
		14	91 min	G_1	1.8×10^{-8}	G_2/M	8.4×10^{-9}	S	3.4×10^{-6}	G_2/M	6.6×10^{-2}
		15	98 min	S/G_2	3.3×10^{-3}	M/G_1	2.0×10^{-3}	S	3.4×10^{-6}	M/G_1	1.3×10^{-2}
		16	105 min	G_2/M	4.8×10^{-4}	M/G_1	3.3×10^{-18}	G_2/M	7.6×10^{-3}	M/G_1	8.8×10^{-5}
		17	112 min	G_2/M	3.1×10^{-10}	M/G_1	3.3×10^{-18}	S/G_2	5.5×10^{-2}	G_1	3.8×10^{-6}
		18	119 min	G_2/M	3.3×10^{-14}	G_1	9.6×10^{-21}	G_2/M	3.0×10^{-5}	G_1	4.2×10^{-8}
d	Overexpression of cell cycle regulators	19	*CLB2*	G_2/M	2.1×10^{-67}	G_1	7.3×10^{-26}	G_2/M	7.0×10^{-14}	G_1	9.2×10^{-9}
		20	*CLB2*	G_2/M	1.0×10^{-55}	G_1	1.2×10^{-31}	G_2/M	7.0×10^{-14}	G_1	2.0×10^{-14}
		21	*CLN3*	G_1	1.2×10^{-61}	M/G_1	6.5×10^{-13}	G_1	5.2×10^{-27}	G_2/M	1.0×10^{-4}
		22	*CLN3*	G_1	5.1×10^{-48}	G_2/M	4.5×10^{-18}	G_1	1.1×10^{-15}	G_2/M	1.0×10^{-4}

				Microarray annotation				Traditional annotation			
				Parallel association		Antiparallel association		Parallel association		Antiparallel association	
	Dataset	Array		Most likely	*P*-value of	Most likely	*P*-value of	Most likely	*P*-value of	Most likely	*P*-value of
e	*CDC15*	1	10 min	M/G_1	1.1×10^{-3}	G_2/M	5.2×10^{-8}	G_2/M	1.4×10^{-2}	S	2.1×10^{-2}
	cell cycle	2	30 min	G_1	1.4×10^{-12}	G_2/M	4.1×10^{-21}	G_1	2.9×10^{-5}	G_2/M	1.4×10^{-3}
	expression	3	50 min	G_1	2.3×10^{-29}	G_2/M	1.5×10^{-30}	G_1	4.9×10^{-14}	G_2/M	1.6×10^{-7}
	time course	4	70 min	S	2.5×10^{-8}	M/G_1	2.8×10^{-19}	S	2.6×10^{-5}	M/G_1	1.6×10^{-5}
		5	80 min	S/G_2	2.2×10^{-9}	M/G_1	9.6×10^{-15}	S	1.0×10^{-3}	M/G_1	7.6×10^{-7}
		6	90 min	G_2/M	3.8×10^{-20}	G_1	3.0×10^{-20}	G_2/M	9.3×10^{-5}	G_1	3.2×10^{-9}
		7	100 min	G_2/M	3.8×10^{-20}	G_1	2.3×10^{-29}	G_2/M	9.3×10^{-5}	G_1	3.2×10^{-9}
		8	110 min	G_2/M	2.0×10^{-29}	G_1	1.9×10^{-27}	G_2/M	3.9×10^{-9}	G_1	2.4×10^{-10}
		9	120 min	G_2/M	9.3×10^{-15}	G_1	6.8×10^{-12}	G_2/M	1.4×10^{-3}	S	2.6×10^{-5}
		10	130 min	M/G_1	4.2×10^{-18}	S	4.5×10^{-3}	M/G_1	7.6×10^{-7}	S	1.0×10^{-3}
		11	140 min	G_1	3.0×10^{-20}	S/G_2	8.6×10^{-8}	G_1	3.7×10^{-6}	G_2/M	1.4×10^{-3}
		12	150 min	G_1	1.3×10^{-18}	G_2/M	4.7×10^{-5}	G_1	3.8×10^{-8}	G_2/M	1.4×10^{-2}
		13	160 min	G_1	6.8×10^{-12}	G_2/M	4.7×10^{-5}	G_1	4.0×10^{-7}	G_2/M	1.4×10^{-3}
		14	170 min	G_1	1.2×10^{-4}	M/G_1	2.7×10^{-4}	S	2.6×10^{-5}	G_2/M	1.4×10^{-2}
		15	180 min	S	1.1×10^{-3}	M/G_1	5.9×10^{-5}	S	1.0×10^{-3}	None	3.5×10^{-1}
		16	190 min	G_2/M	4.7×10^{-4}	G_1	1.0×10^{-14}	S	1.0×10^{-3}	M/G_1	2.7×10^{-3}
		17	200 min	G_2/M	4.6×10^{-10}	G_1	1.5×10^{-5}	G_2/M	1.4×10^{-3}	M/G_1	2.2×10^{-2}
		18	210 min	G_2/M	1.5×10^{-11}	G_1	3.0×10^{-20}	G_2/M	1.4×10^{-3}	G_1	3.2×10^{-9}
		19	220 min	G_2/M	2.3×10^{-17}	G_1	1.5×10^{-5}	G_2/M	1.4×10^{-3}	G_1	8.7×10^{-2}
		20	230 min	G_2/M	1.4×10^{-5}	G_1	5.0×10^{-6}	G_2/M	9.3×10^{-5}	G_1	8.7×10^{-2}
		21	240 min	G_2/M	1.1×10^{-8}	S	4.9×10^{-2}	M/G_1	2.4×10^{-4}	S	2.2×10^{-1}
		22	250 min	M/G_1	1.2×10^{-11}	S/G_2	1.2×10^{-1}	M/G_1	2.8×10^{-8}	S/G_2	1.3×10^{-2}
		23	270 min	M/G_1	1.2×10^{-5}	G_2/M	3.5×10^{-3}	M/G_1	1.6×10^{-5}	G_2/M	1.4×10^{-2}
		24	290 min	M/G_1	5.9×10^{-5}	G_2/M	4.7×10^{-4}	M/G_1	2.4×10^{-4}	G_2/M	1.4×10^{-2}

The genome-scale binding profiles of *ORC1*, *MCM3*, *MCM4*, and *MCM7* are correlated with transcription minima during the cell cycle stage G1.

Novel Genome-Scale Correlation Replication ↔ Transcription Predicted By Data-Driven Models

Replication may regulate transcription:

The binding of ORC and MCM proteins, which is known to be required for initiation of replication at origins across the yeast genome, represses, and possibly inhibits the transcription of genes that are located near the origins.

Transcription may regulate replication:

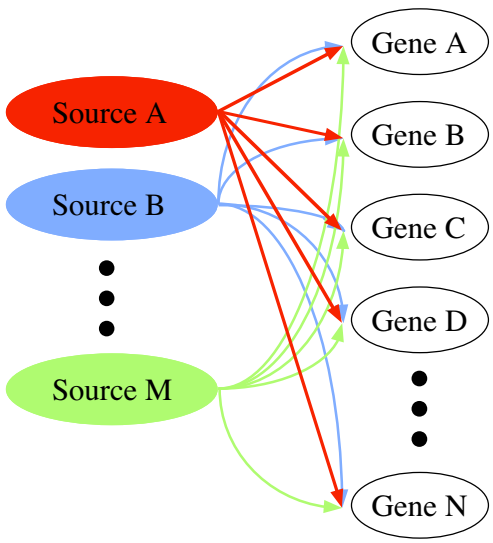
The transcription of genes at G1 reduces the efficiency of origins that are located near the transcribed genes.

→ This is the first time that a data-driven mathematical model has been used to predict a genome-scale biological principle.

→ This is the first time that the mathematical tool of pseudoinverse projection has been used to understand one genomic dataset in light of another.

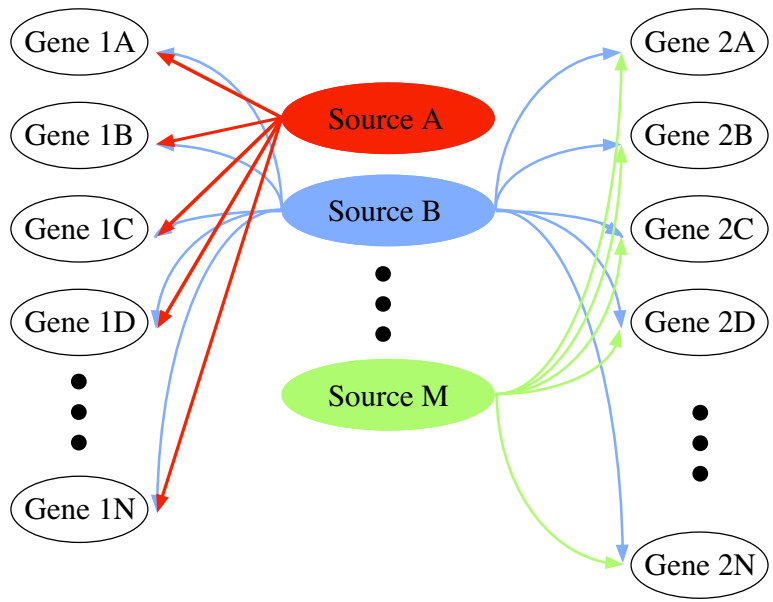
Conclusions

SVD Modeling



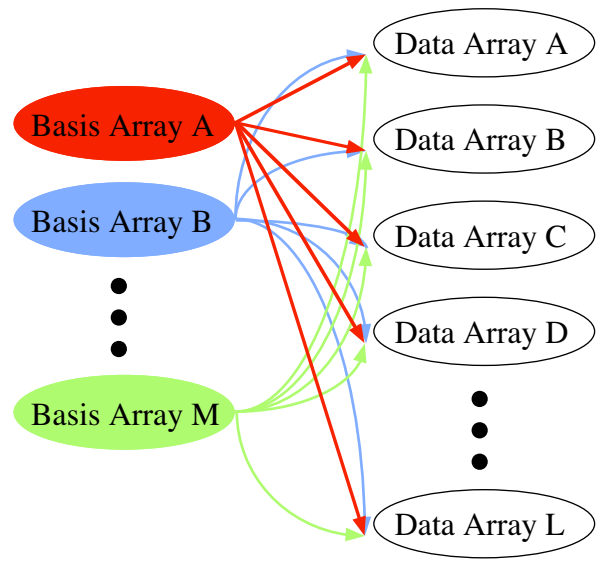
→ Data Normalization
→ Data Classification
→ Data Incorporation

GSVD Comparative Modeling

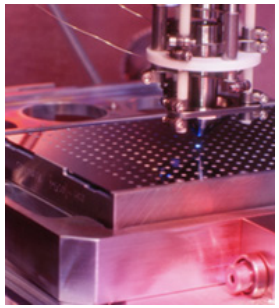
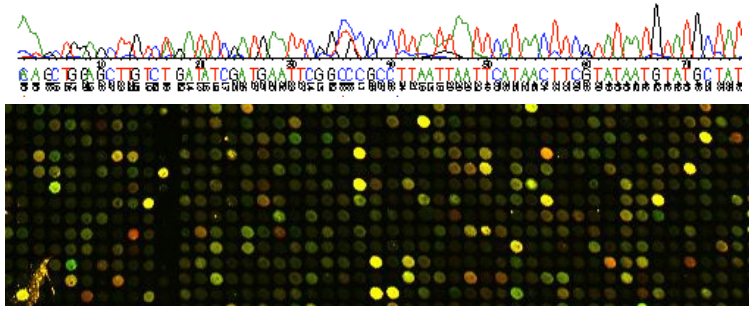
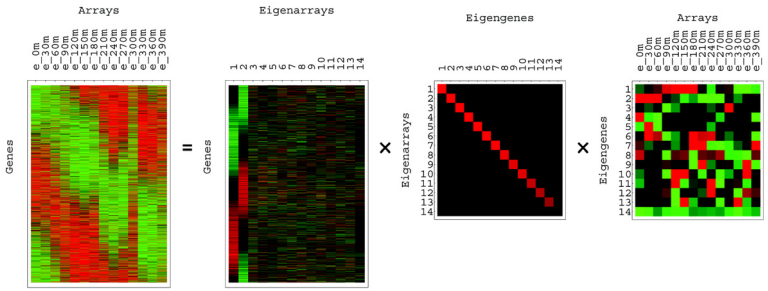
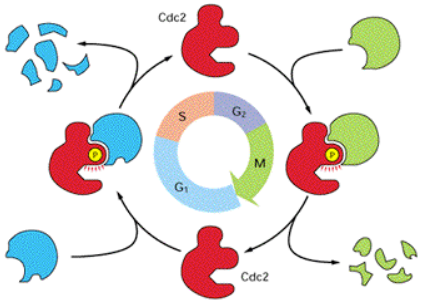
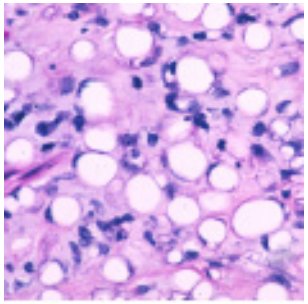


→ Comparative Data Reconstruction
→ Comparative Classification

Pseudoinverse Integrative Modeling

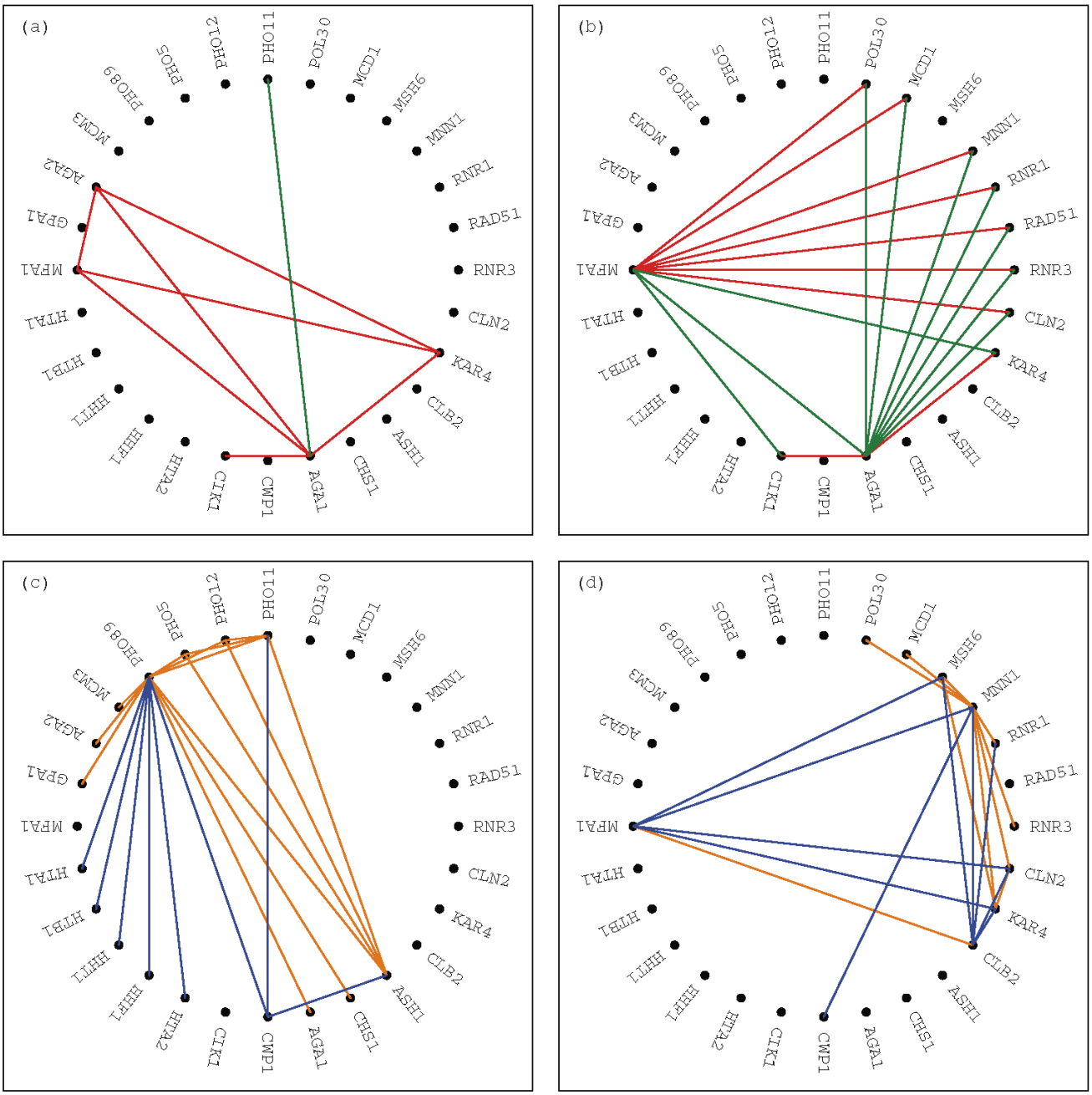


→ Integrative Reconstruction
→ Integrative Classification

	Astronomy	Molecular Biology
Technology	Galileo	
Large-Scale Data	Brahe	
Mathematical Modeling	Kepler	
Basic Principles	Newton	 
Technology	NASA	Control of Cellular Mechanisms

New Biology ↔ Innovative Math

Network Decomposition: From Systems to Pathways



Thanks!!!

Gene Golub

David Botstein

Pat Brown

Matt van de Rijn

Stanford Microarray Database

And, thank you!!!