Tensor Computations
For Genomic Signal Processing:
From Data Patterns to Principles of Nature

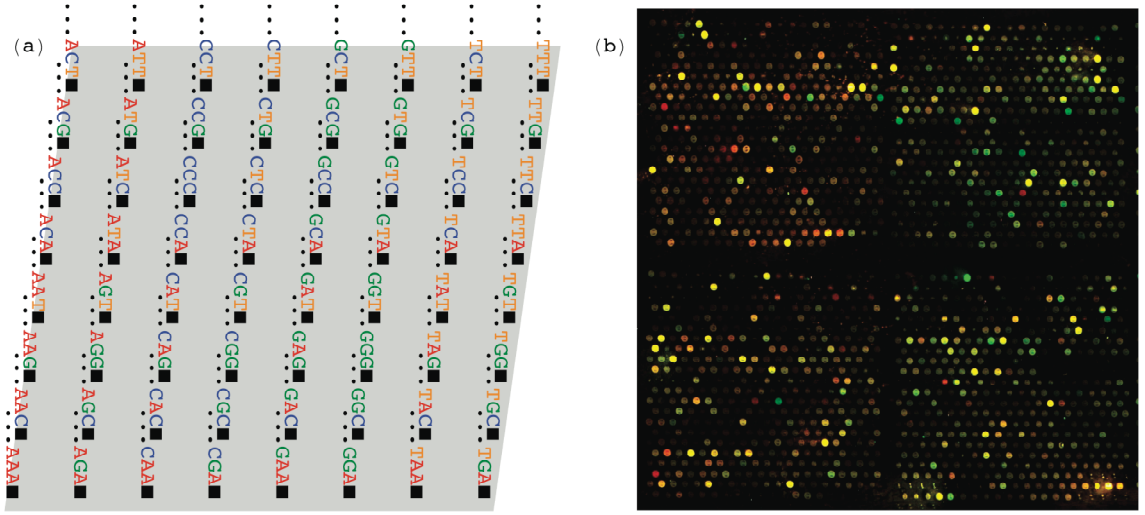
Orly Alter

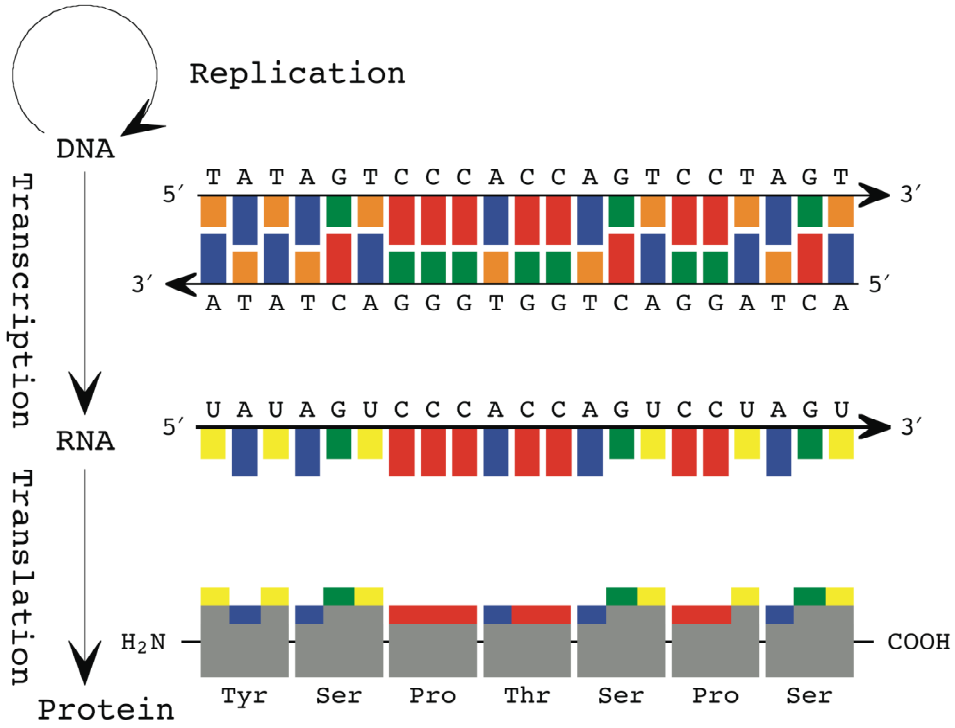
Department of Biomedical Engineering,
Institute for Cellular and Molecular Biology and
Institute for Computational Engineering and Sciences

University of Texas at Austin

DNA Microarrays Record Genomic Signals

DNA microarrays rely on hybridization to record the complete genomic signals that guide the progression of cellular processes, such as abundance levels of DNA, RNA and DNA-bound proteins on a genomic scale.

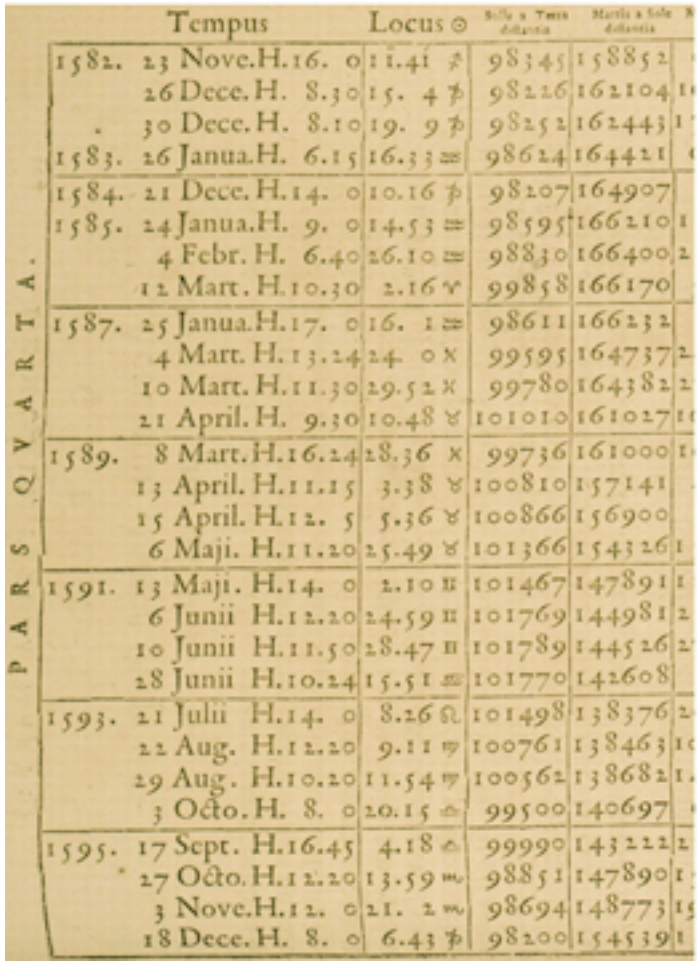




From Data Patterns to Principles of Nature

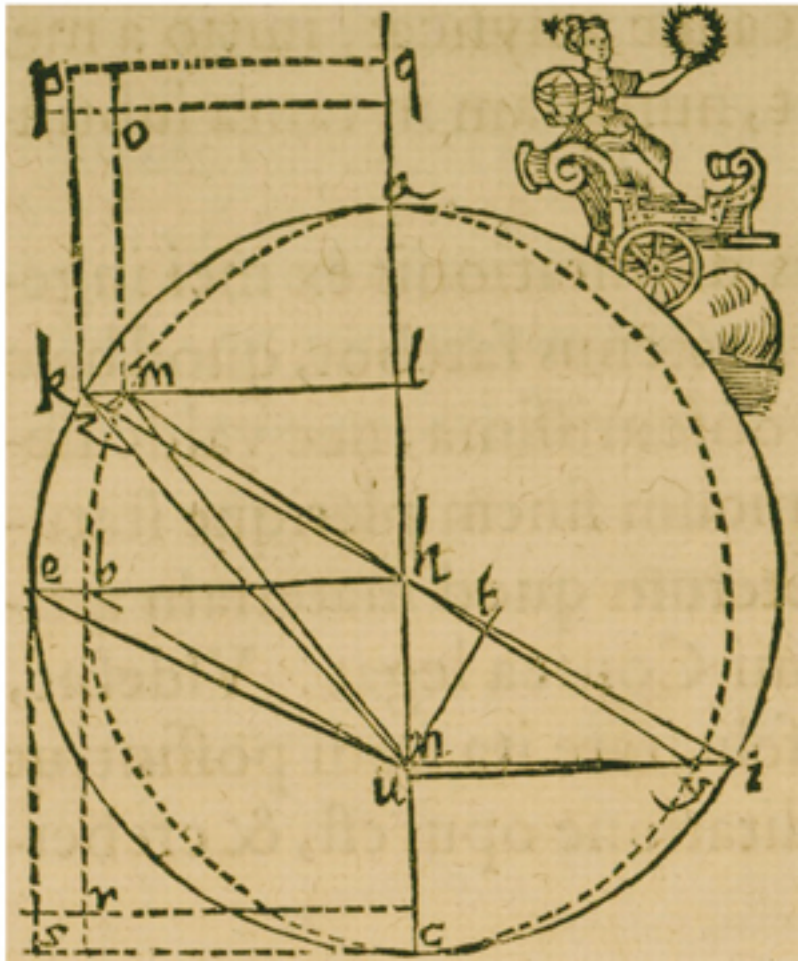
Alter, *PNAS* 103, 16063 (2006);
Alter, in *Microarray Data Analysis: Methods and Applications* (Humana Press, 2007), pp. 17–59.

Kepler's discovery of his first law of planetary motion from mathematical modeling of Brahe's astronomical data:



PARS QVARTA.

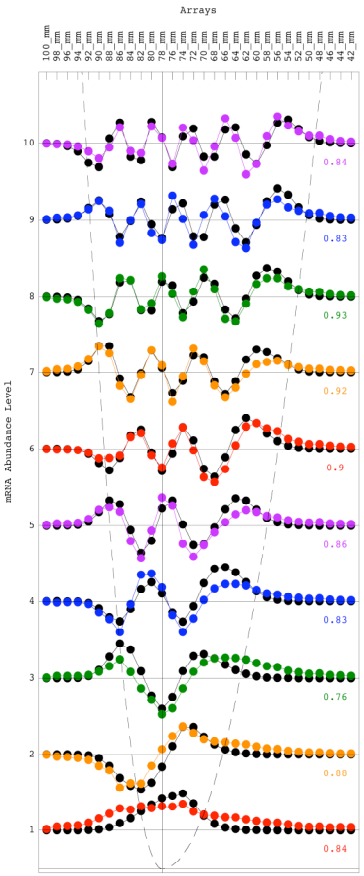
	Tempus	Locus ☉		
1582.	23 Nove. H. 16. 0	11.41 ♐	98345	158852
	26 Dece. H. 8.30	15. 4 ♑	98226	162104
	30 Dece. H. 8.10	19. 9 ♑	98252	162443
1583.	26 Janua. H. 6.15	16.33 ♒	98624	164421
1584.	21 Dece. H. 14. 0	10.16 ♑	98207	164907
1585.	24 Janua. H. 9. 0	14.53 ♒	98595	166210
	4 Febr. H. 6.40	26.10 ♒	98830	166400
	12 Mart. H. 10.30	2.16 ♈	99858	166170
1587.	25 Janua. H. 17. 0	16. 1 ♒	98611	166232
	4 Mart. H. 13.24	24. 0 ♓	99595	164737
	10 Mart. H. 11.30	29.52 ♓	99780	164382
	21 April. H. 9.30	10.48 ♉	101010	161027
1589.	8 Mart. H. 16.24	28.36 ♓	99736	161000
	13 April. H. 11.15	3.38 ♉	100810	157141
	15 April. H. 12. 5	5.36 ♉	100866	156900
	6 Maji. H. 11.20	25.49 ♉	101366	154326
1591.	13 Maji. H. 14. 0	2.10 ♊	101467	147891
	6 Junii H. 12.20	24.59 ♊	101769	144981
	10 Junii H. 11.50	28.47 ♊	101789	144526
	28 Junii H. 10.24	15.51 ♋	101770	142608
1593.	21 Julii H. 14. 0	8.26 ♌	101498	138376
	22 Aug. H. 12.20	9.11 ♍	100761	138463
	29 Aug. H. 10.20	11.54 ♍	100562	138682
	3 Octo. H. 8. 0	20.15 ♎	99500	140697
1595.	17 Sept. H. 16.45	4.18 ♎	99990	143222
	27 Octo. H. 12.20	13.59 ♏	98851	147890
	3 Nove. H. 12. 0	21. 2 ♏	98694	148773
	18 Dece. H. 8. 0	6.43 ♑	98200	154539



Kepler, *Astronomia Nova* (Voegelinus, Heidelberg, 1609), reproduced by permission of the Harry Ransom Humanities Research Center of the University of Texas, Austin, TX).

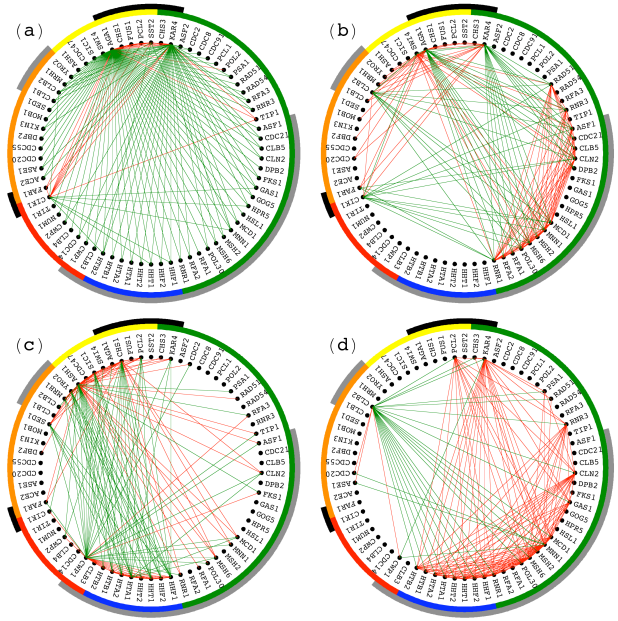
Data Patterns: Vectors, Matrices and Cuboids

SVD



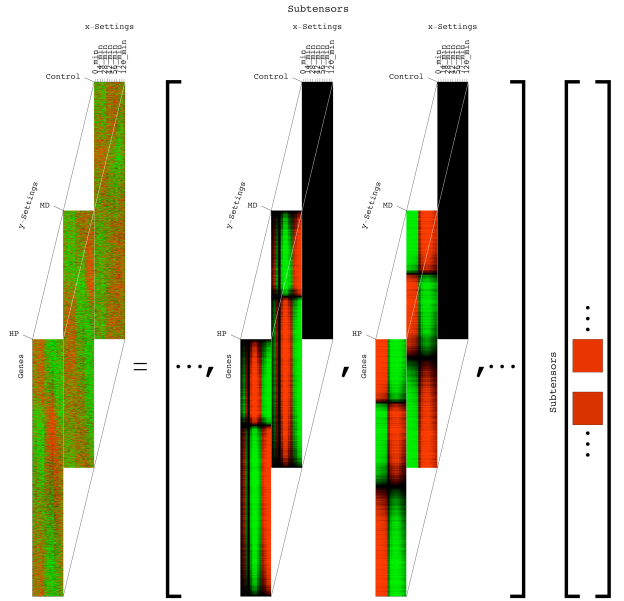
"Asymmetric Hermite functions" reveal asymmetry in gel electrophoresis thermal broadening of RNA bands.

Comparative HOEVD



"Subnetworks" common or exclusive among multiple networks elucidate pathway-dependency of gene regulation.

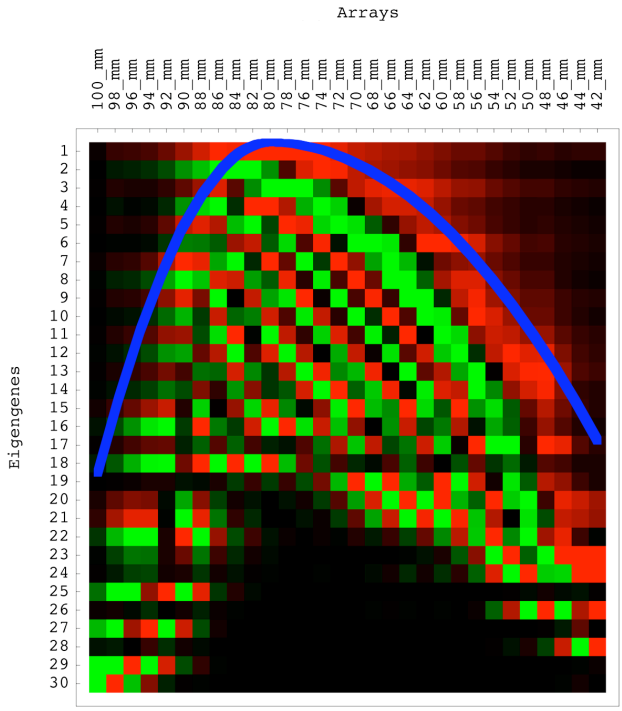
Integrative HOSVD

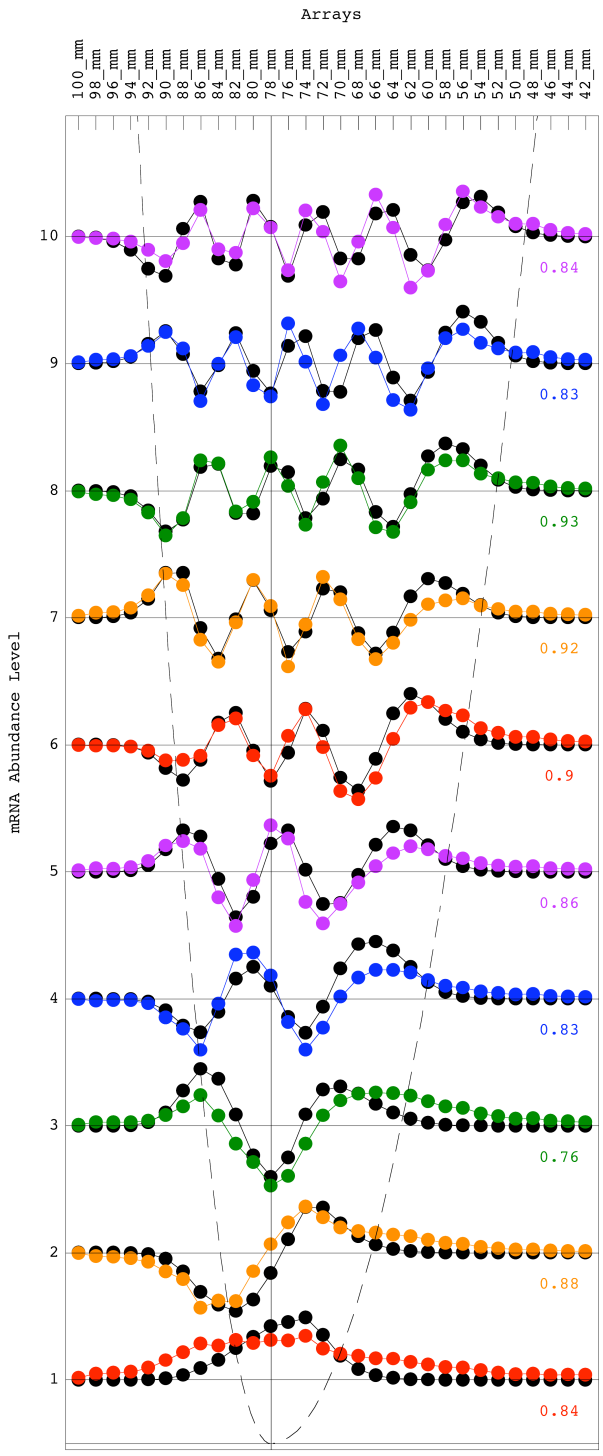


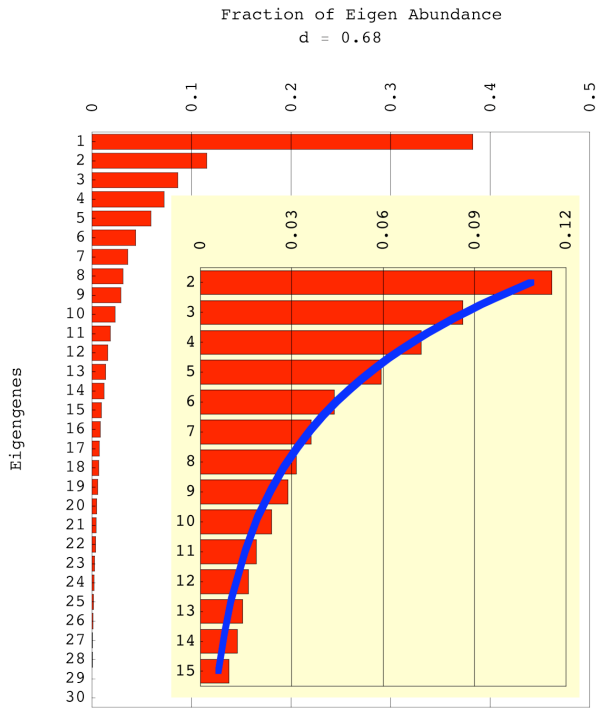
"Subtensors" of data from different studies uncover independent processes and their causal coordination.

SVD Modeling of Genome-Wide mRNA Lengths Distribution Predicts a Physical Principle

Alter & Golub, *PNAS* 103, 11828 (2006);
http://www.bme.utexas.edu/research/orly/harmonic_oscillator/.

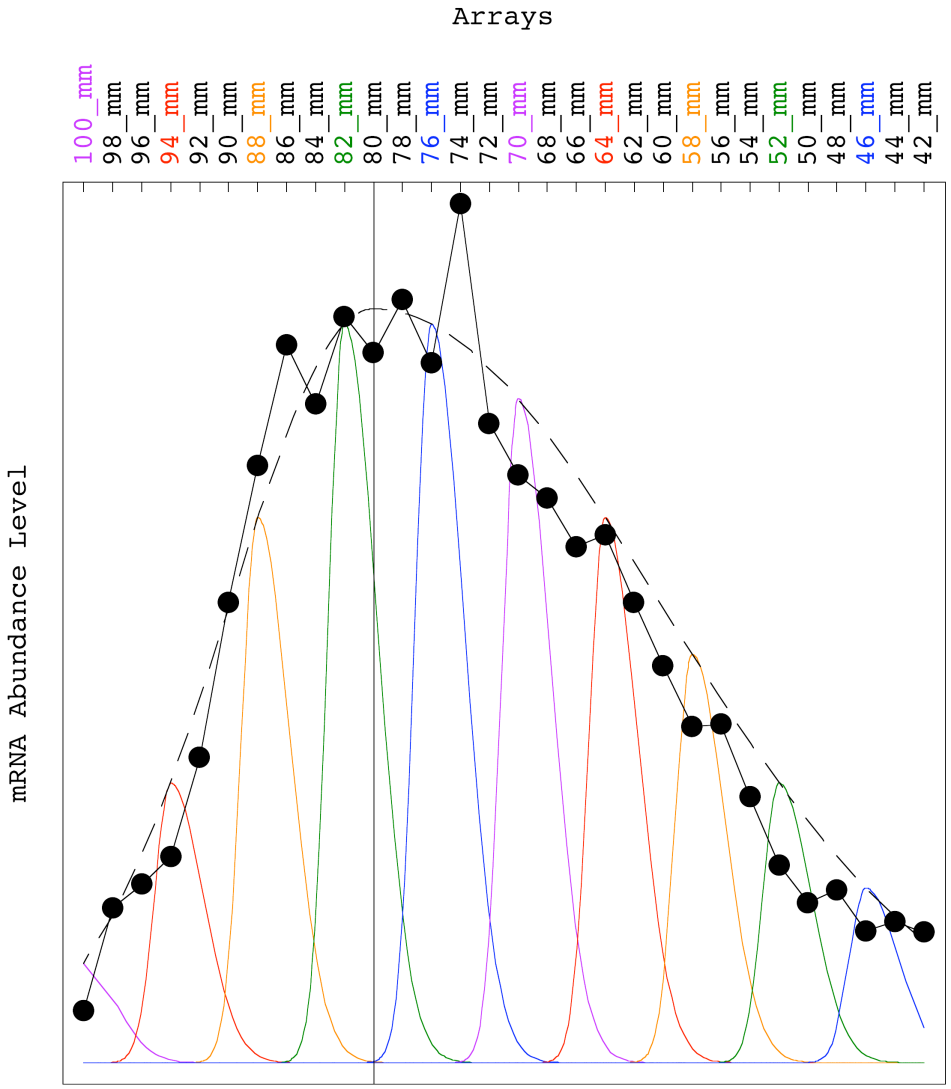






Hurowitz & Brown, *Genome Biology* 5, R2 (2003).

“Asymmetric” Generalized Coherent State Model of mRNA Lengths Distribution

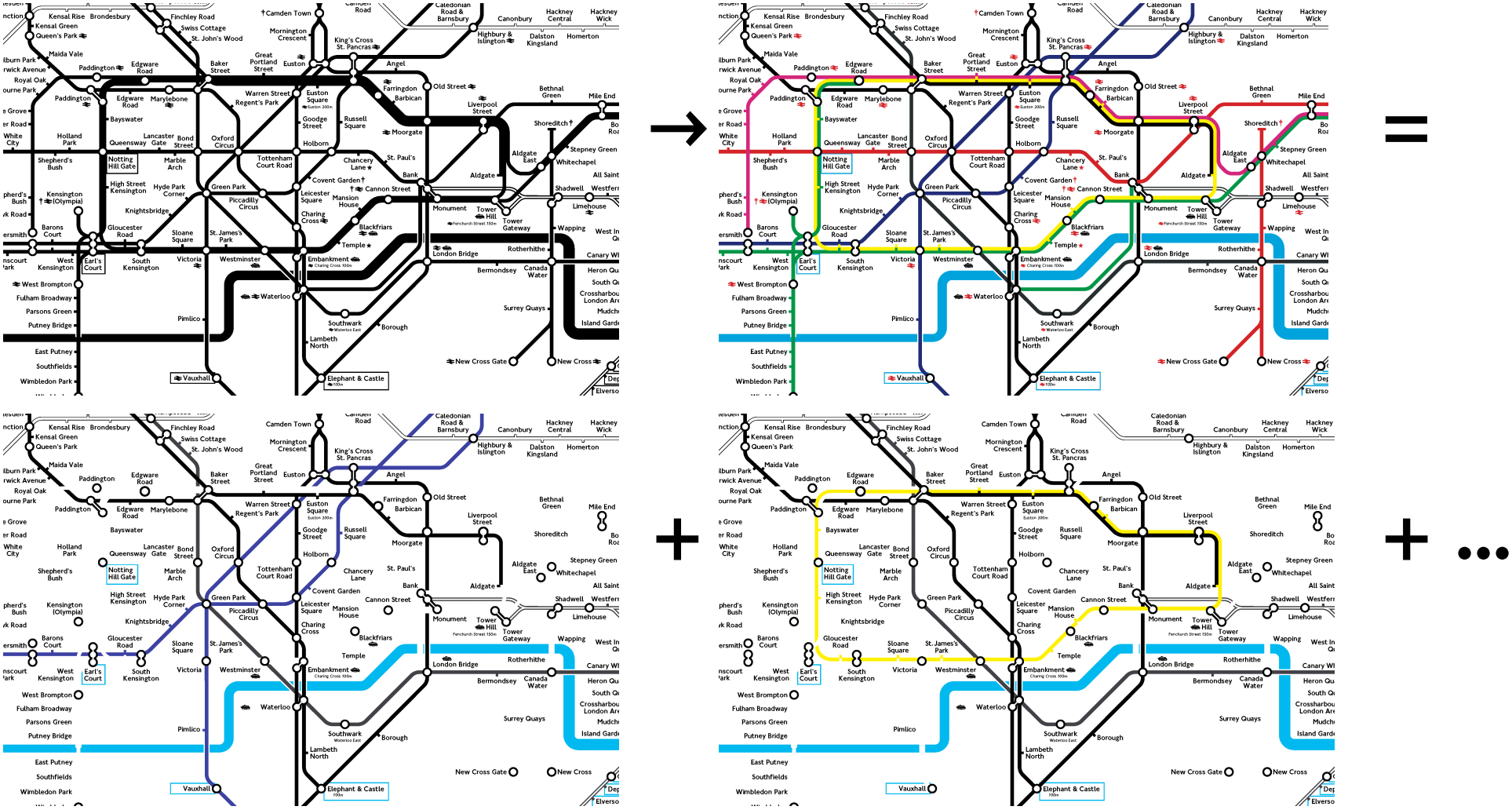


Prediction: The peak of a moving RNA band is moving toward the front of the band and away from its back. Studies of DNA band broadening in gel electrophoresis showed different broadening of a moving than a stationary band, but did not suggest asymmetry.

Hypothesis: Two competing evolutionary forces determine the distribution of mRNA gene transcripts, in the manner of the restoring force of the harmonic oscillator.

Networks are Tensors of "Subnetworks"

Alter & Golub, *PNAS* 102, 17559 (2005);
http://www.bme.utexas.edu/research/orly/network_decomposition/.



The 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.

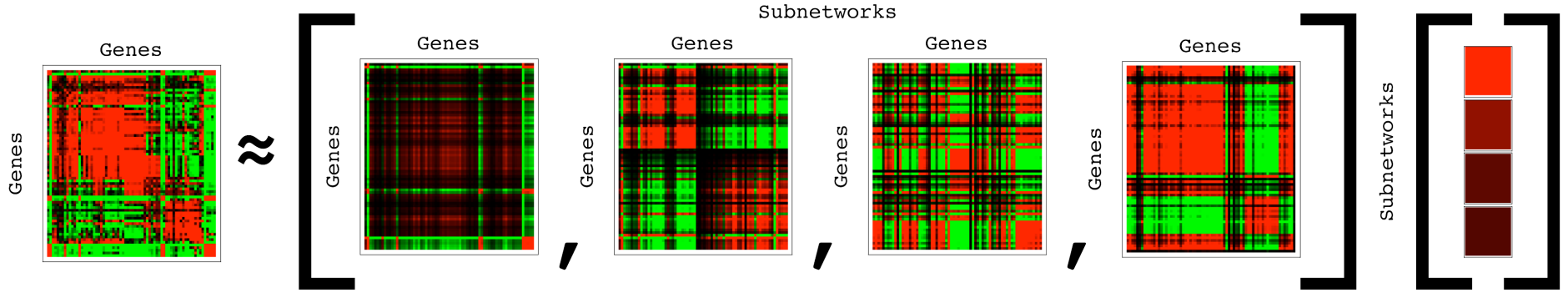
Eigenvalue Decomposition (EVD)

EVD formulates a genes × genes nondirectional network as a linear superposition of genes × genes decorrelated and decoupled rank-1 subnetworks, which can be associated with functionally independent pathways.

EVD of the network $\hat{a}_1$,

$$\hat{a}_1 = \hat{e}_1 \hat{e}_1^T = \hat{u}_1 \hat{\epsilon}_1^2 \hat{u}_1^T = \sum_{m=1}^{M_1} \epsilon_{1,m}^2 |\alpha_{1,m}\rangle\langle\alpha_{1,m}|,$$

is computed from the SVD of the data signal $\hat{e}_1 = \hat{u}_1 \hat{\epsilon}_1 \hat{v}_1^T$.



Yeast Cell Cycle mRNA Expression With Pheormone Synchronization

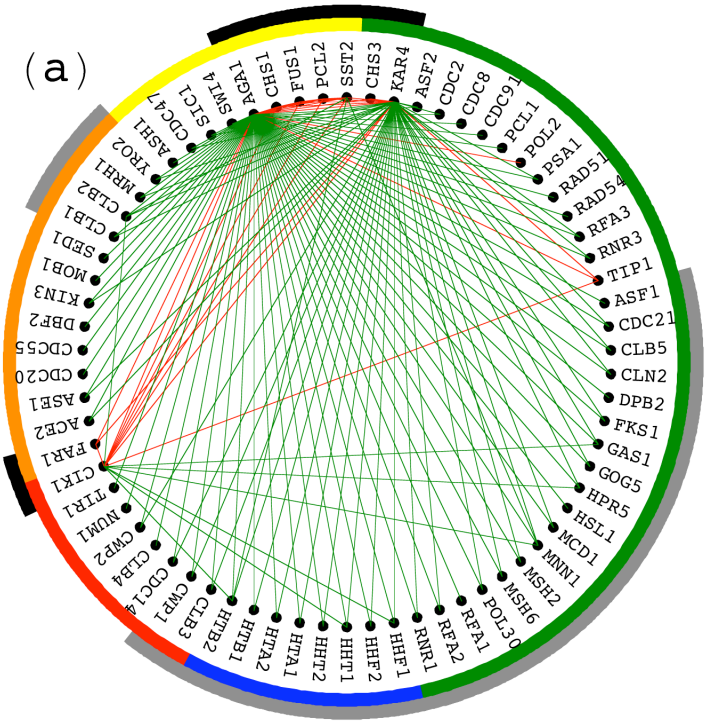
Spellman et al., *MBC* 9, 3273 (1998).

Math Variables → Biology

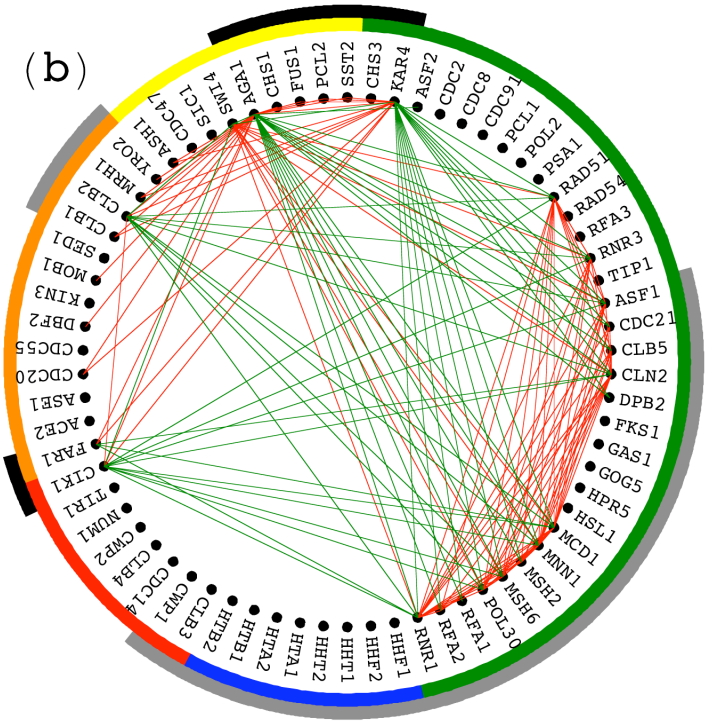
Significant EVD subnetworks → functionally independent pathways:

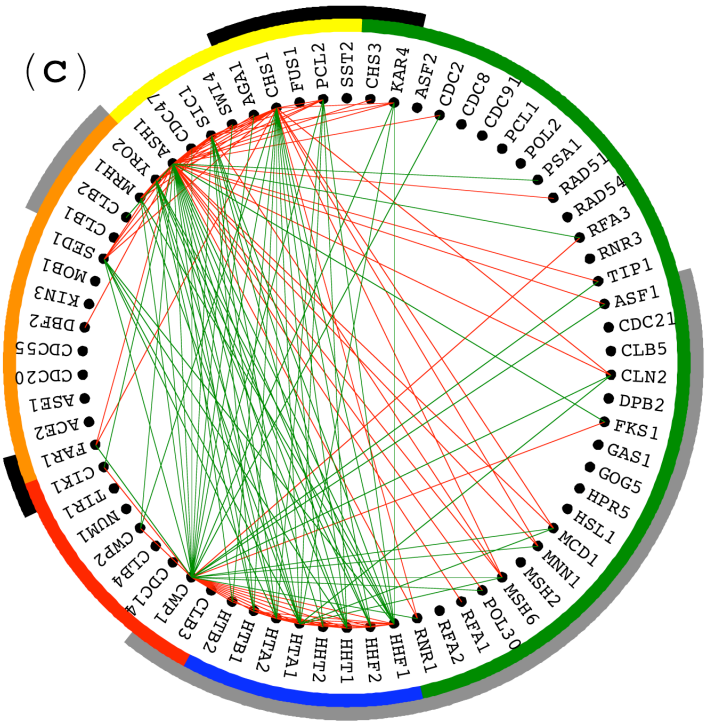
Pheromone Signaling Pathway

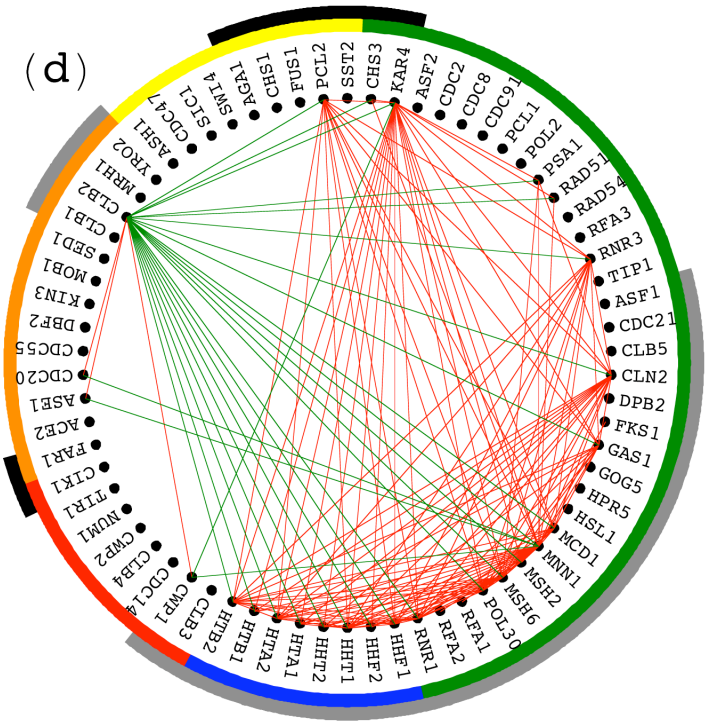
KAR4 || *CIK1*



Pheromone Arrest Exit & G_1 Entry







Cell Cycle S ↔ M

KAR4 || −*CIK1*

Cell Cycle G_1 ↔ G_2

Interpretation of the Subnetworks: Probabilistic Associations by Annotations

Classification	Subnetwork	Most likely parallel association	P value of parallel association	Most likely antiparallel association	P value of antiparallel association
Cell Cycle	1	S S	1.7×10^{-22}	M/G_1 S	5.1×10^{-7}
	2	G_1 G_1	1.3×10^{-29}	G_1 G_2/M	3.2×10^{-11}
	3	S S	2.1×10^{-30}	M/G_1 S	2.6×10^{-25}
	4	G_1 S	2.1×10^{-28}	G_1 G_2/M	5.7×10^{-24}
Pheromone Response	1	Up Up	4.0×10^{-53}	Down Up	2.2×10^{-50}
	2	Down Down	1.6×10^{-11}	Down Up	9.8×10^{-17}
	3	Down Down	6.2×10^{-6}	Down Down	1.6×10^{-11}
	4	Down Down	8.0×10^{-32}	Down Down	2.5×10^{-6}

The P value of a given association by annotation is calculated 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 $x \subseteq X = N(N-1)/2$ pairs of genes with either largest and smallest levels of correlations in the subnetwork

$$P(x; y, Y, X) = \binom{X}{x}^{-1} \sum_{z=y}^{x} \binom{Y}{z} \binom{X-Y}{x-z}.$$

where $\binom{X}{x} = X!x!^{-1}(X-x)!^{-1}$ is the binomial coefficient.

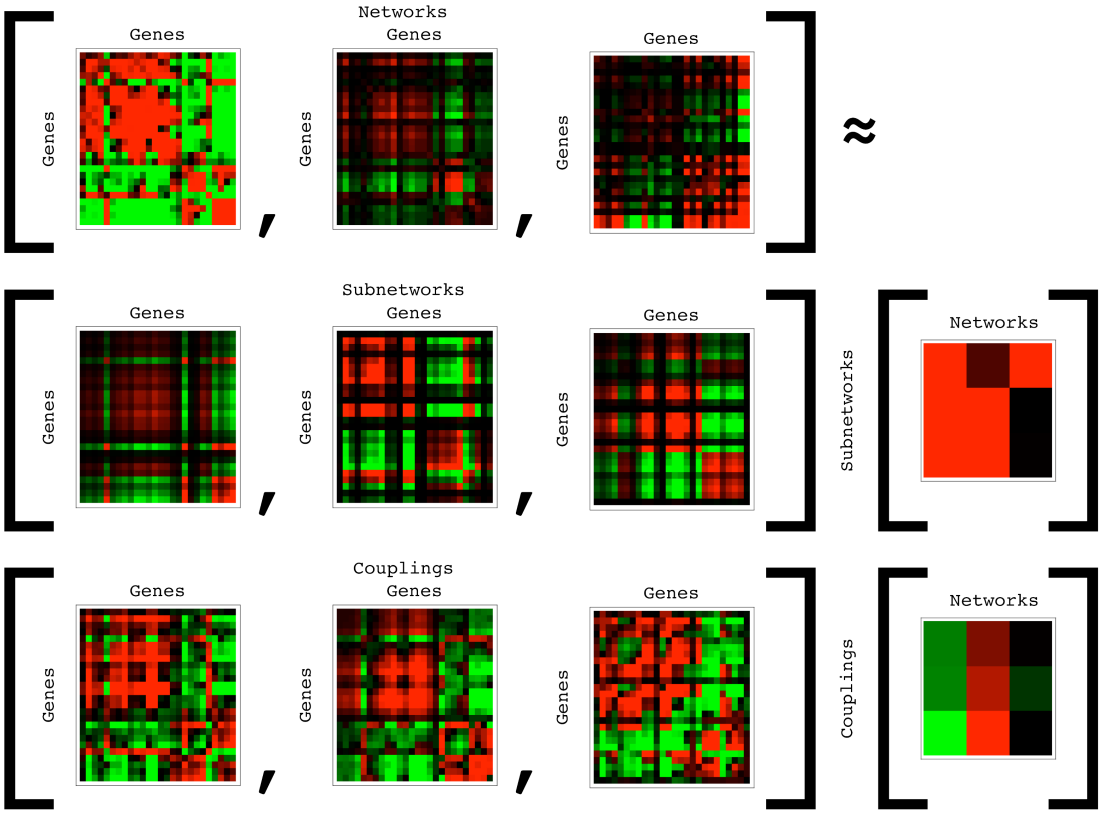
A Comparative Higher-Order EVD (HOEVD)

Formulates a series of networks as linear superpositions of decorrelated rank-1 subnetworks and the rank-2 couplings among them.

This HOEVD of the tensor of networks $\{\hat{a}_k\}$,

$$\hat{a} \equiv \sum_{k=1}^{K} \hat{a}_k = \hat{u}(\sum_{k=1}^{K} \hat{\epsilon}_k^2)\hat{u}^T = \hat{u}\hat{\epsilon}^2\hat{u}^T,$$

is computed from the SVD of the appended signals $\hat{e} \equiv (\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_K) = \hat{u}\hat{\epsilon}\hat{v}^T$.



Cell Cycle Expression

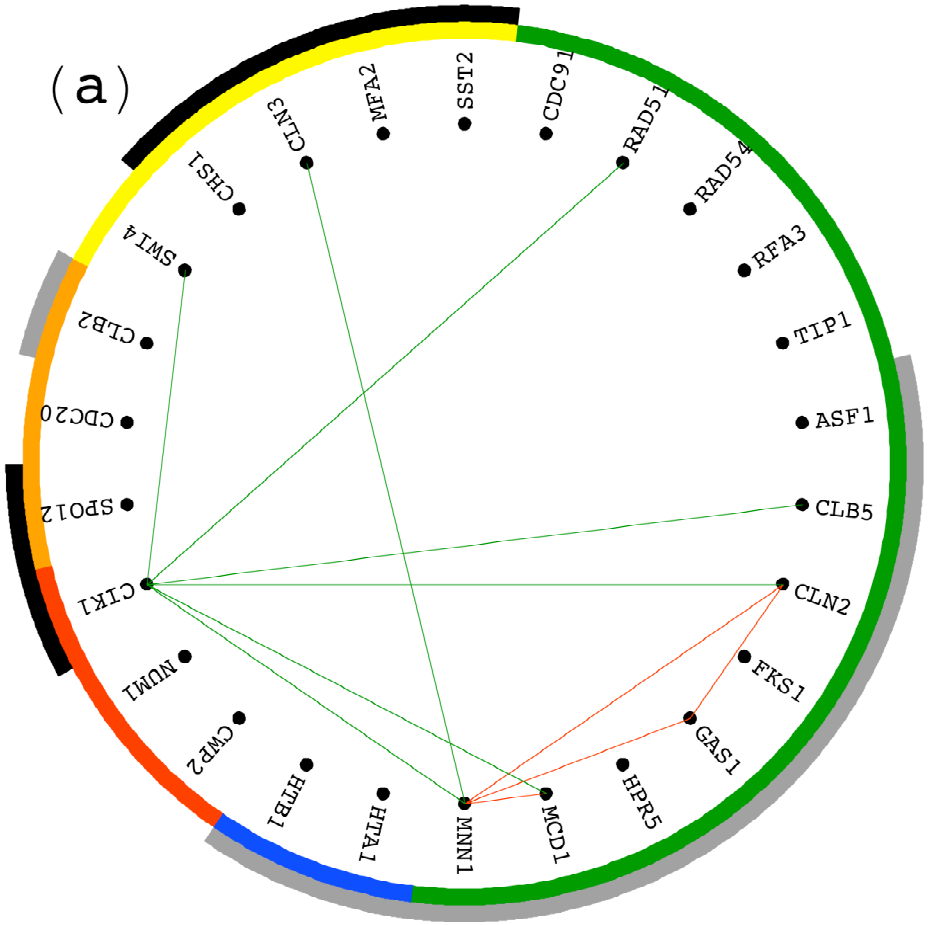
Spellman et al., *MBC* 9, 3273 (1998).

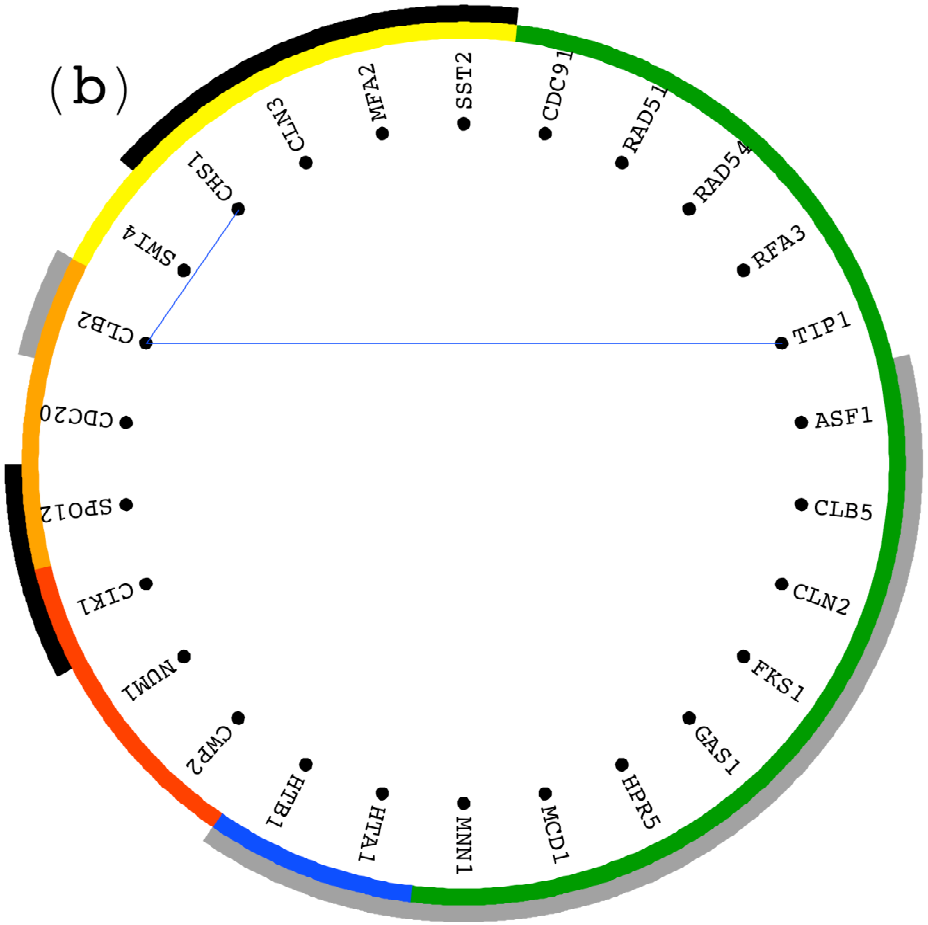
Development and Cell Cycle Transcription Factors' Binding

Lee et al., *Science* 298, 799 (2002).

Math Operations → Biology

Boolean functions of subnetworks and couplings → pathway-dependent relations among genes common or exclusive among the networks:





Known Relations:

Pheromone Response **AND** $G_1 \leftrightarrow G_2$ **AND** Transition Between These

CLN2 || −*CIK1*

Novel Relations:

Pheromone Response **AND** $G_1 \leftrightarrow G_2$ **AND NOT** Transition Between These

CLB2 || ±*TIP1*

A Higher-Order SVD

Linear transformation of the tensor data from genes × *x*-settings × *y*-settings space to reduced "eigenarrays" × "*x*-eigengenes" × "*y*-eigengenes" space.

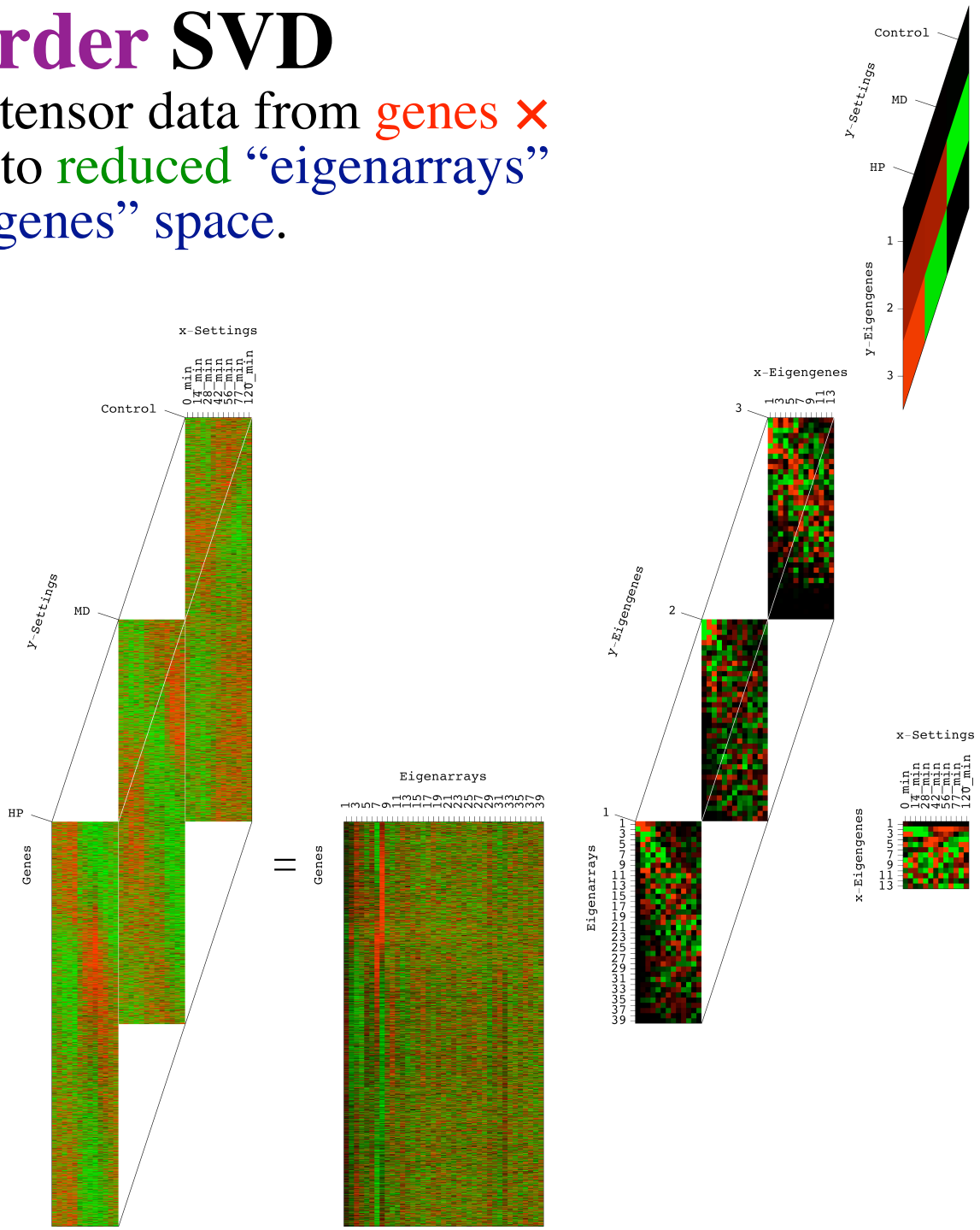
This HOSVD is computed from each SVD of the data tensor unfolded along all axes perpendicular to one given axis,

$$\mathcal{T} = \mathcal{R} \times_a U \times_b V_x \times_c V_y.$$

De Lathauwer, De Moor & Vandewalle, *SIMAX* 21, 1253 (2000); Kolda, *SIMAX* 23, 243 (2001); Zhang & Golub, *SIMAX* 23, 543 (2001).

mRNA Expression From Cell Cycle Time Courses Under Different Conditions of Oxidative Stress

Shapira, Segal & Botstein, *MBC* 15, 5659 (2004); Spellman et al., *MBC* 9, 3273 (1998).

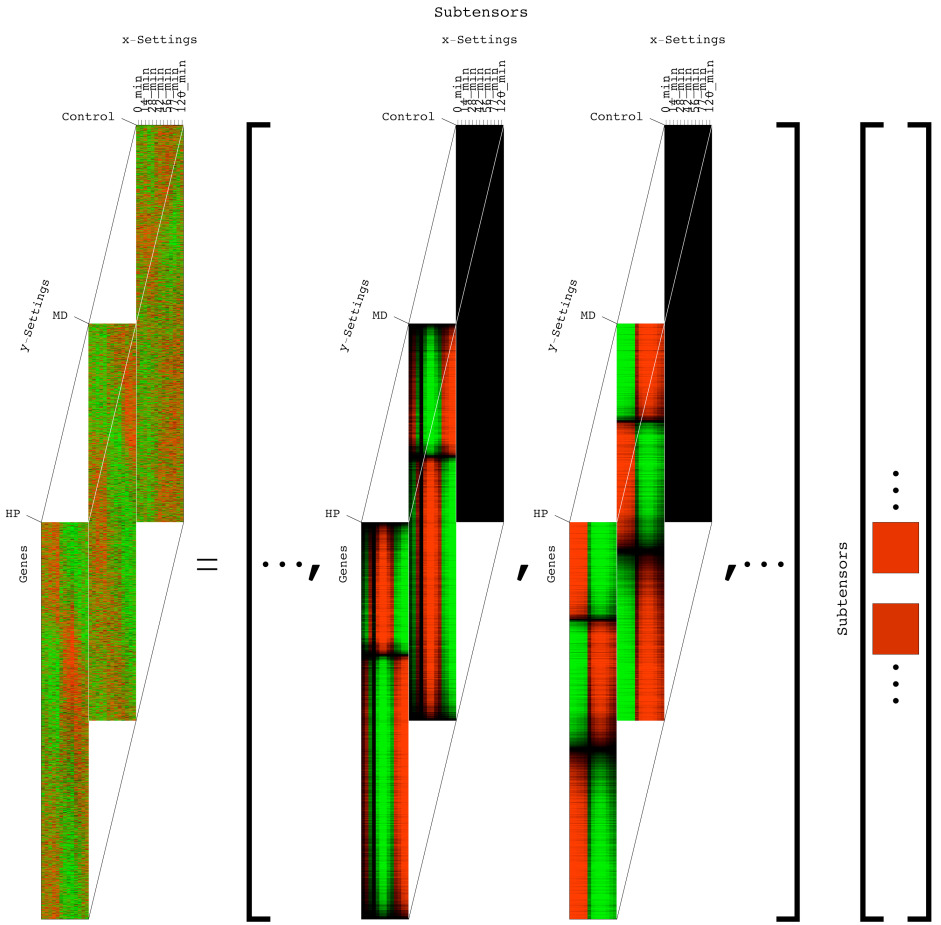


HOSVD for Integrative Data Analysis

Omberg, Golub & Alter, *PNAS* 104, 18371 (2007);
http://www.bme.utexas.edu/research/orly/HOSVD/.

The tensor data is a superposition of all rank-1 "subtensors," i.e., outer products of an eigenarray, an *x*- and a *y*-eigengene,

$$\mathcal{T} \equiv \sum_{a=1}^{LM} \sum_{b=1}^{L} \sum_{c=1}^{M} \mathcal{R}_{abc} \mathcal{S}(a,b,c).$$



The "fraction" computed from the higher-order singular values, indicates the significance of the corresponding subtensor,

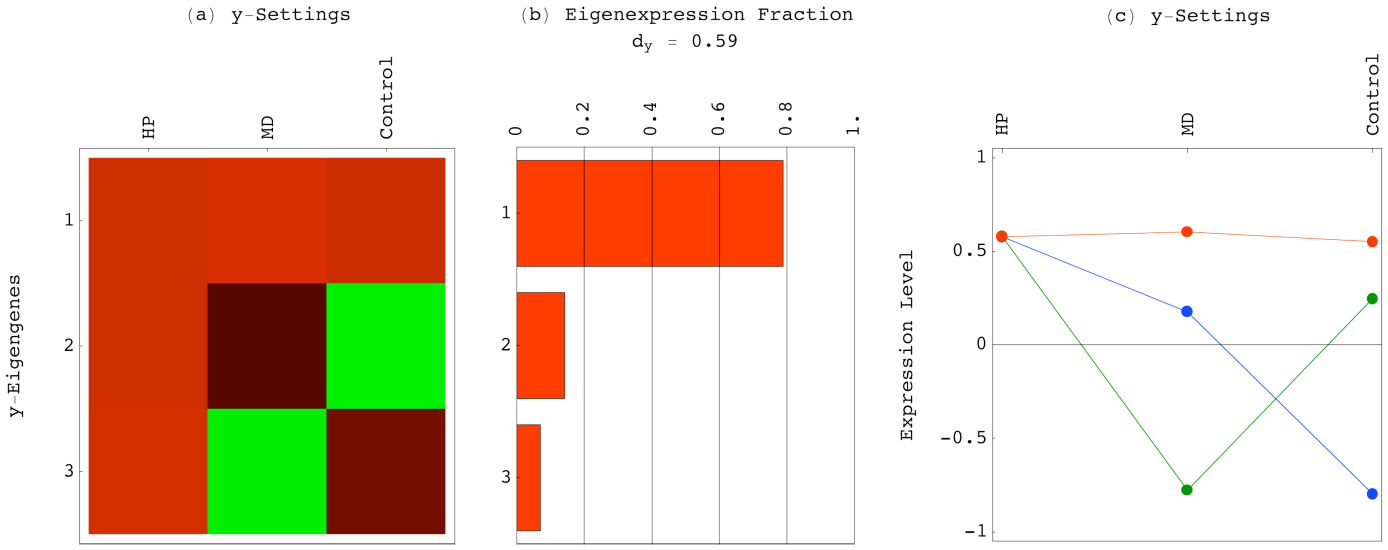
$$\mathcal{P}_{abc} = \mathcal{R}_{abc}^2 / \sum_{a=1}^{LM} \sum_{b=1}^{L} \sum_{c=1}^{M} \mathcal{R}_{abc}^2.$$

The "normalized entropy" measures the complexity of the data tensor,

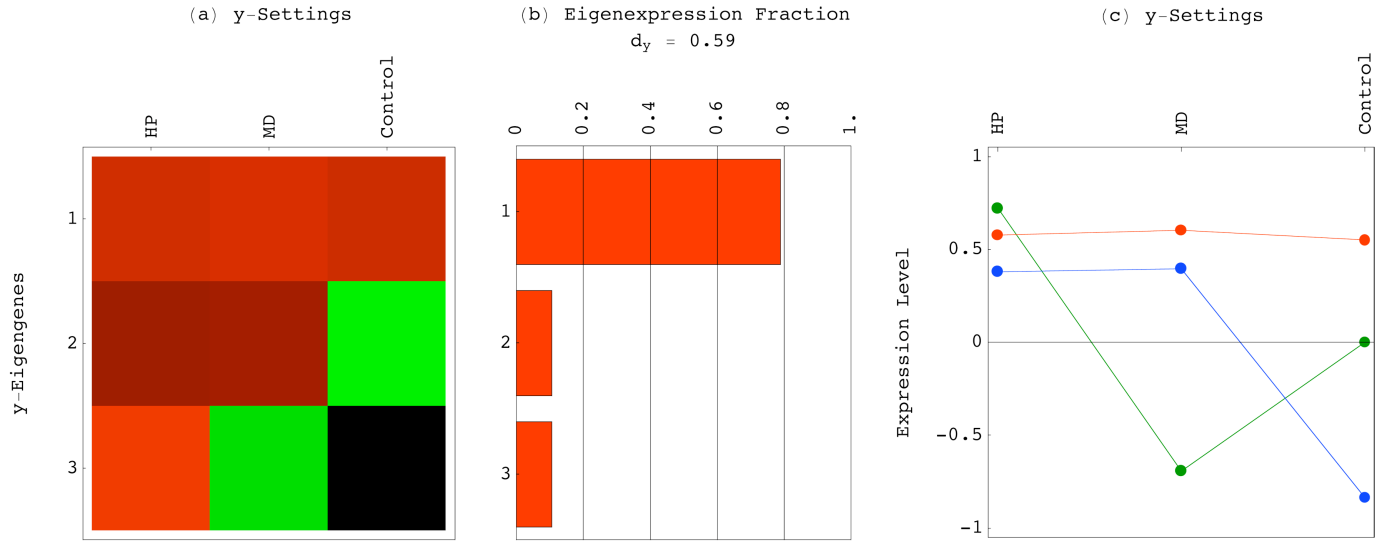
$$0 \leq d = \frac{-1}{2\log(LM)} \sum_{a=1}^{LM} \sum_{b=1}^{L} \sum_{c=1}^{M} \mathcal{P}_{abc} \log(\mathcal{P}_{abc}) \leq 1.$$

Rotation in an Approximately Degenerate Eigenvector Space

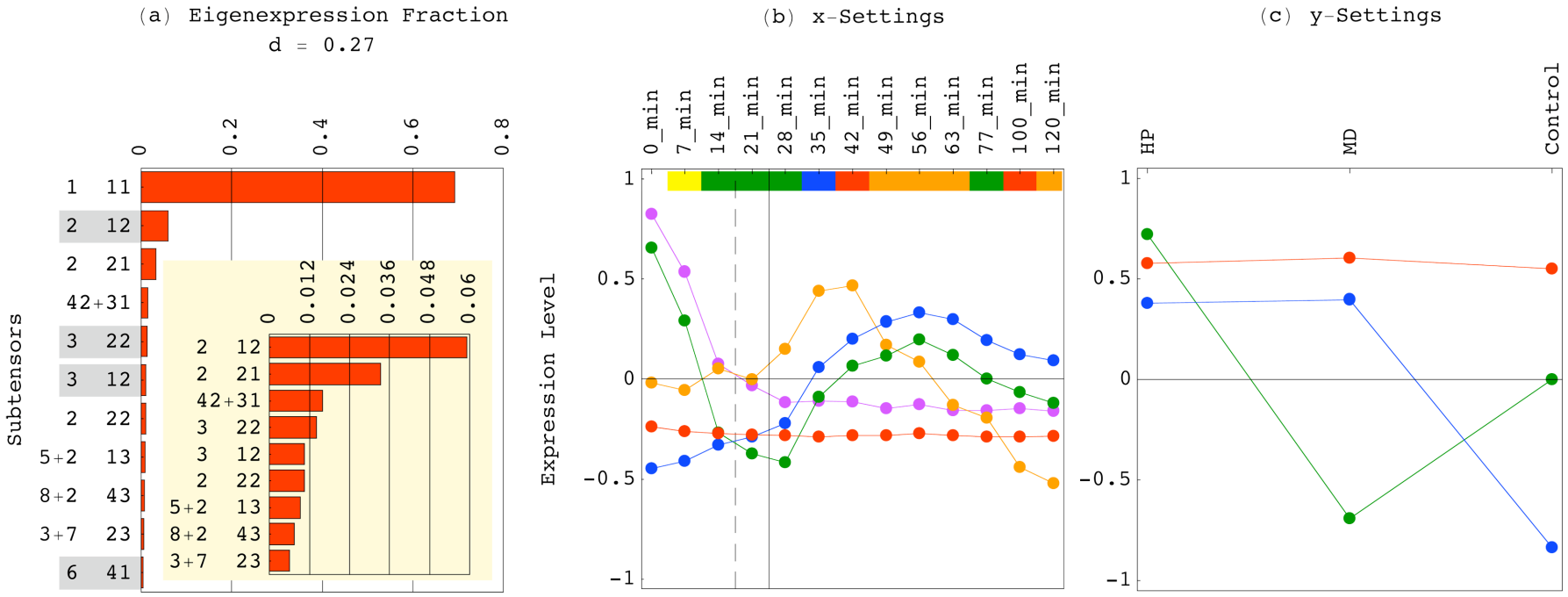
The 2nd and 3rd *y*-eigengenes are approximately degenerate, i.e., not unique:



This HOSVD is reformulated with a unique orthogonal rotation of the 2nd and 3rd *y*-eigengenes under the constraint that the 3rd *y*-eigengene in the control time course is at steady state:



Rotation in an Approximately Degenerate Subtensor Space



An "approximately degenerate subtensor space" is defined as that which is span by, e.g., the subtensors

$$\mathcal{S}(a,b,c) \text{ and } \mathcal{S}(k,b,c),$$

which satisfy

$$|\mathcal{R}_{abc}| \approx |\mathcal{R}_{kbc}|.$$

This HOSVD is reformulated with a unique single rank-1 subtensor that is composed of these two subtensors,

$$\mathcal{R}_{a+k,b,c}\mathcal{S}(a+k,b,c) =$$

$$\mathcal{R}_{abc}\mathcal{S}(a,b,c) + \mathcal{R}_{kbc}\mathcal{S}(k,b,c).$$

Math Variables → Biology

Significant subtensors → independent biological programs or experimental phenomena:

$\mathcal{S}(k,l,m)$	$\mathcal{P}_{klm}$	$\mathcal{R}_{klm}$
1,1,1	70%	>0

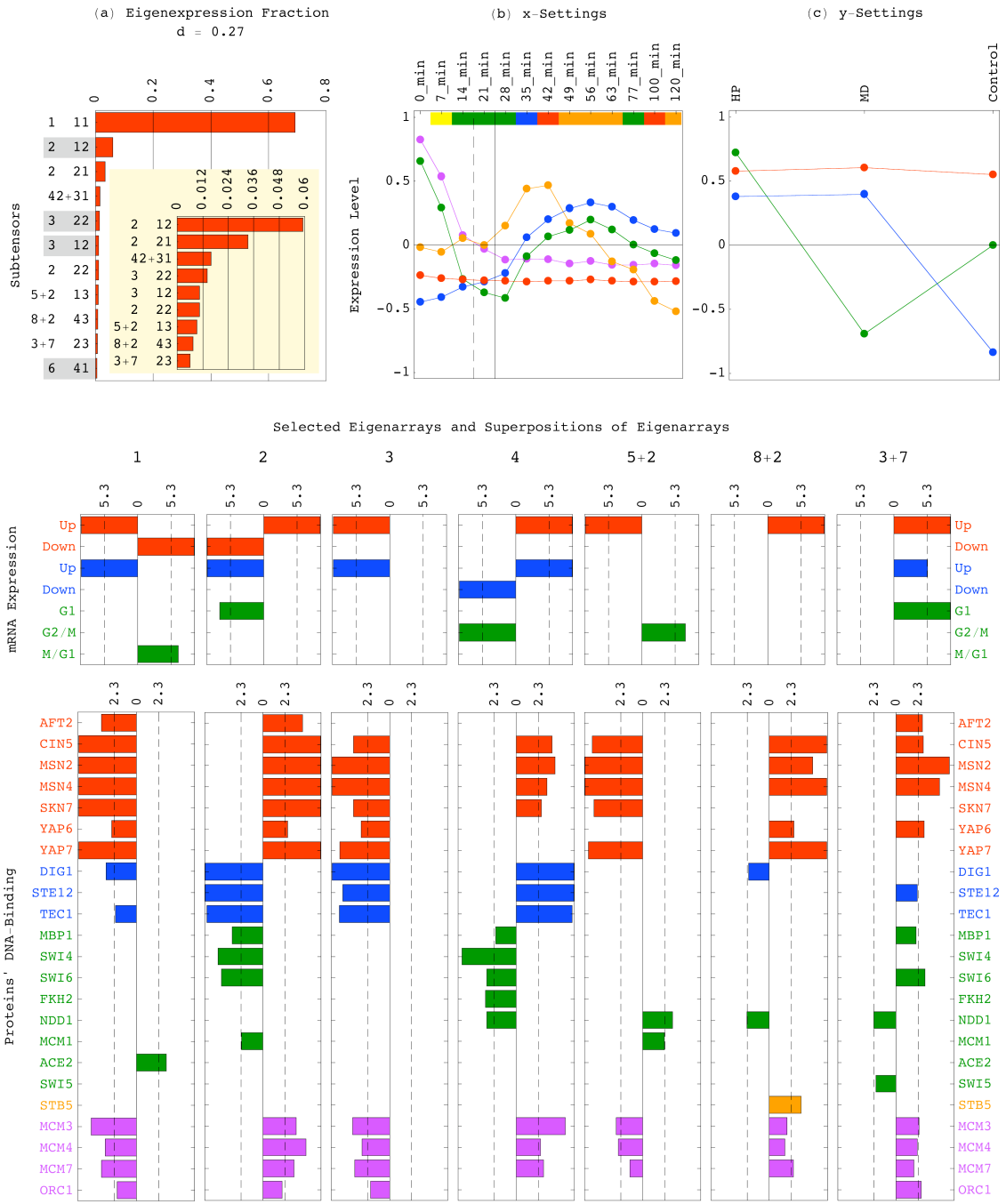
Steady State

2,1,2	6%	<0
2,2,1	3.3%	>0
2,2,2	1%	>0

Oxidative Stress in Time and Across Conditions

4,2+3,1	1.6%	>0
3,2,2	1.4%	<0
3,1,2	1%	<0

Pheromone Responses



Math Operations → Biology

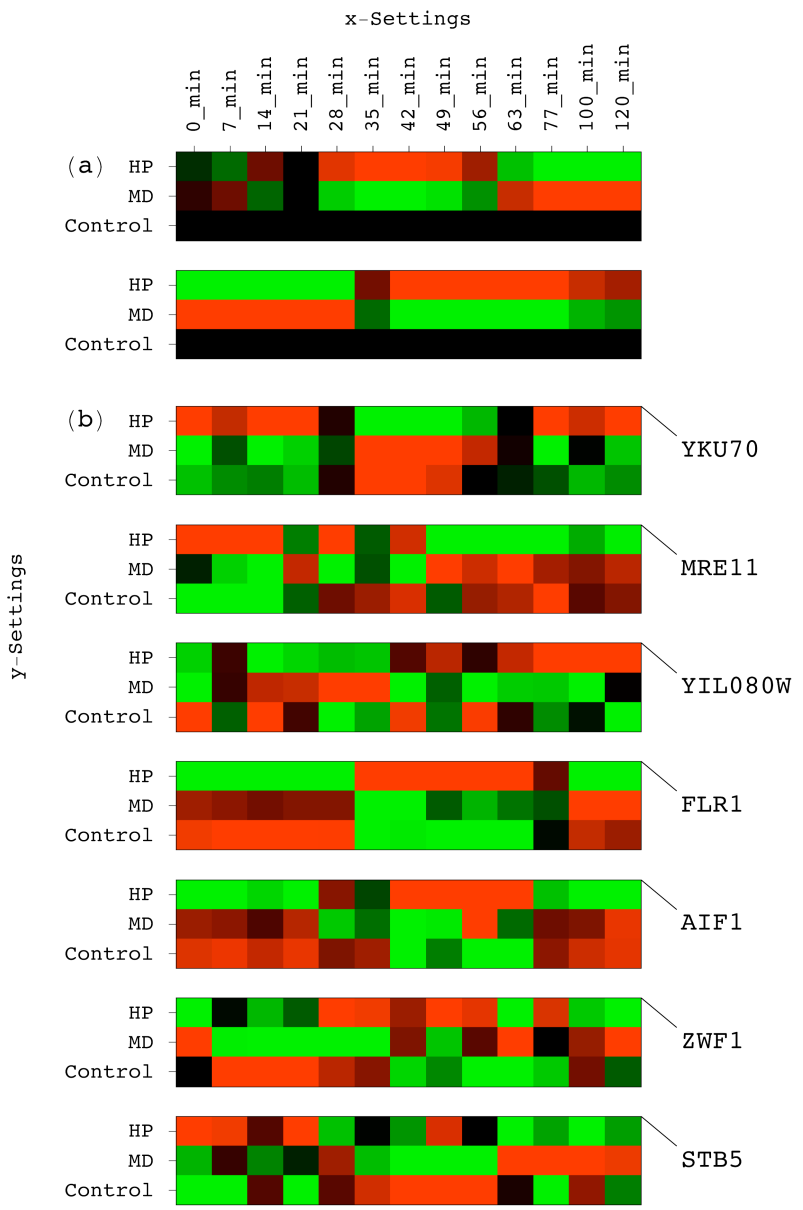
$\mathcal{S}(k,l,m)$	$\mathcal{P}_{klm}$	$\mathcal{R}_{klm}$
5+2,1,3	0.9%	>0
8+2,4,3	0.75%	>0
3+7,2,3	0.6%	>0

HP vs. MD-Induced Expression

Flattery-O'Brien & Dawes,
J. Biol. Chem. 273, 8564 (1998).

Classification identifies genes significant in terms of the information that they capture in each subtensor → global picture of time-dependence of HP vs. MD-induced expression:

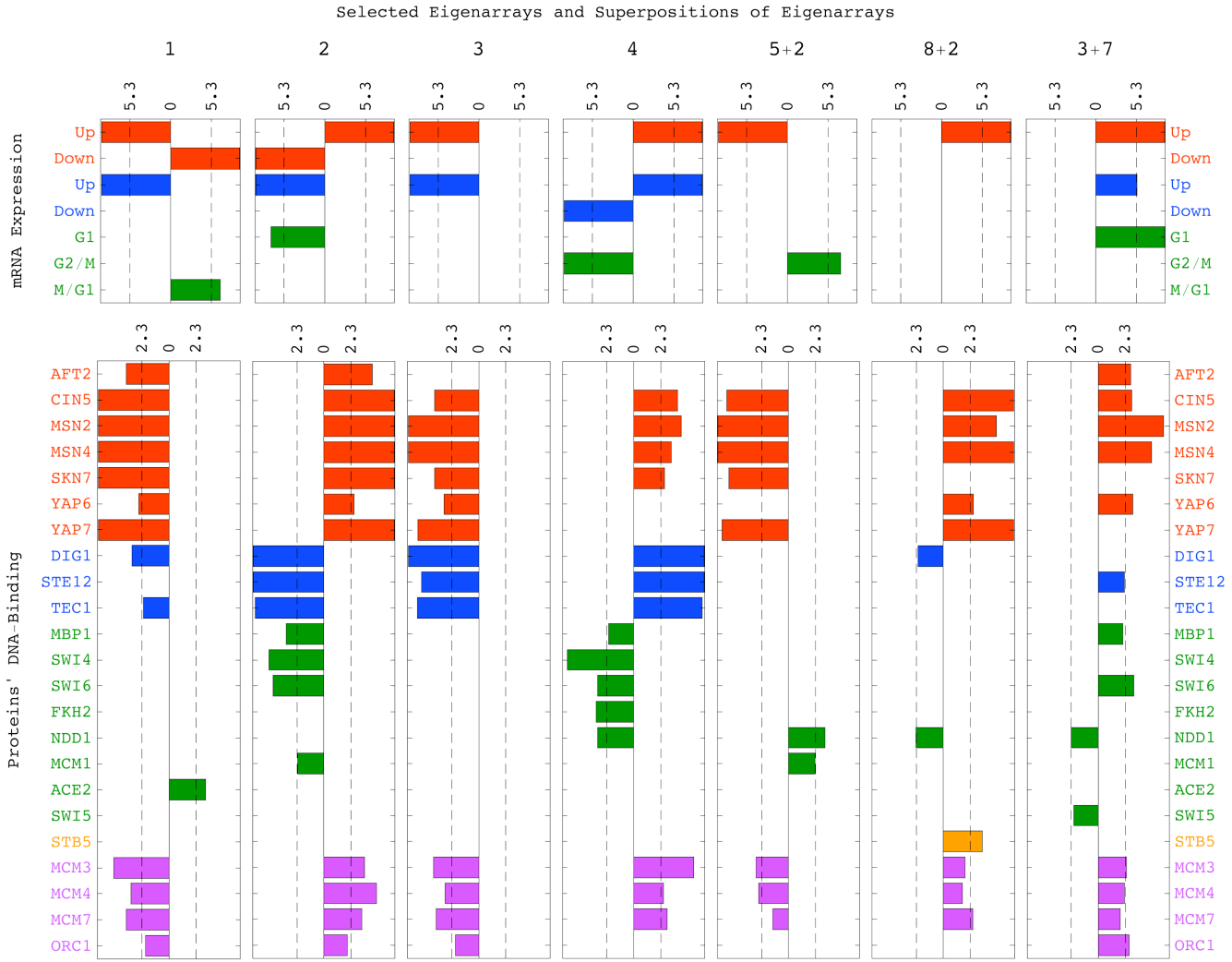
The conserved genes *YKU70*, *MRE11*, *AIF1* and *ZWF1*, and the processes of retrotransposition, apoptosis and the oxidative pentose phosphate cycle that they are involved in, play significant, yet previously unrecognized, roles in the differential effects of HP and MD on cell cycle progression.



Math Variables & Operations → Biology

HOSVD uncovers independent data patterns across each variable and the interactions among them → global picture of the causal coordination among biological processes and experimental phenomena:

DNA ↔ RNA Correlation



Overexpression of binding targets of Mcm3, Mcm4 and Mcm7 correlates with expression in response to environmental stress and overexpression of oxidative stress activators-bound genes.

→ DNA damage and apoptosis, as caused by oxidative stress and overexpression of *AIF1*, inhibit binding of origins by degradation of the pre-replicative complex protein Cdc6.

Cocker et al., *Nature* 379, 180 (1996);

Blanchard et al., *MBC* 13, 1536 (2002).

Overexpression of binding targets of replication initiation proteins correlates with reduced, or even inhibited, binding of the origins.

→ This correlation is equivalent to a recently discovered correlation between the binding of these proteins and reduced expression of adjacent genes, which might be due to a previously unknown mechanism of regulation.

Alter & Golub, *PNAS* 101, 16577 (2004).

→ This correlation is in agreement with the recent observation that reduced efficiency of activation of origins correlates with local transcription.

Donato, Chung & Tye, *PLoS Genet*. 2, E141 (2006);

Snyder, Sapolsky & Davis, *MCB* 8, 2184 (1988).

HO GSVD for Comparative Analysis of DNA Microarray Data from Different Organisms

Ponnapalli, Saunders, Golub & Alter, submitted;

Ponnapalli, Golub & Alter, *2006 Stanford/Yahoo! Workshop on Algorithms for Modern Massive Data Sets*; http://www.stanford.edu/group/mmds-program.pdf

An HO GSVD that extends to higher orders most of the mathematical properties of GSVD,

$$D_1 = U_1 \Sigma_1 X^{-1},$$

$$D_2 = U_2 \Sigma_2 X^{-1},$$

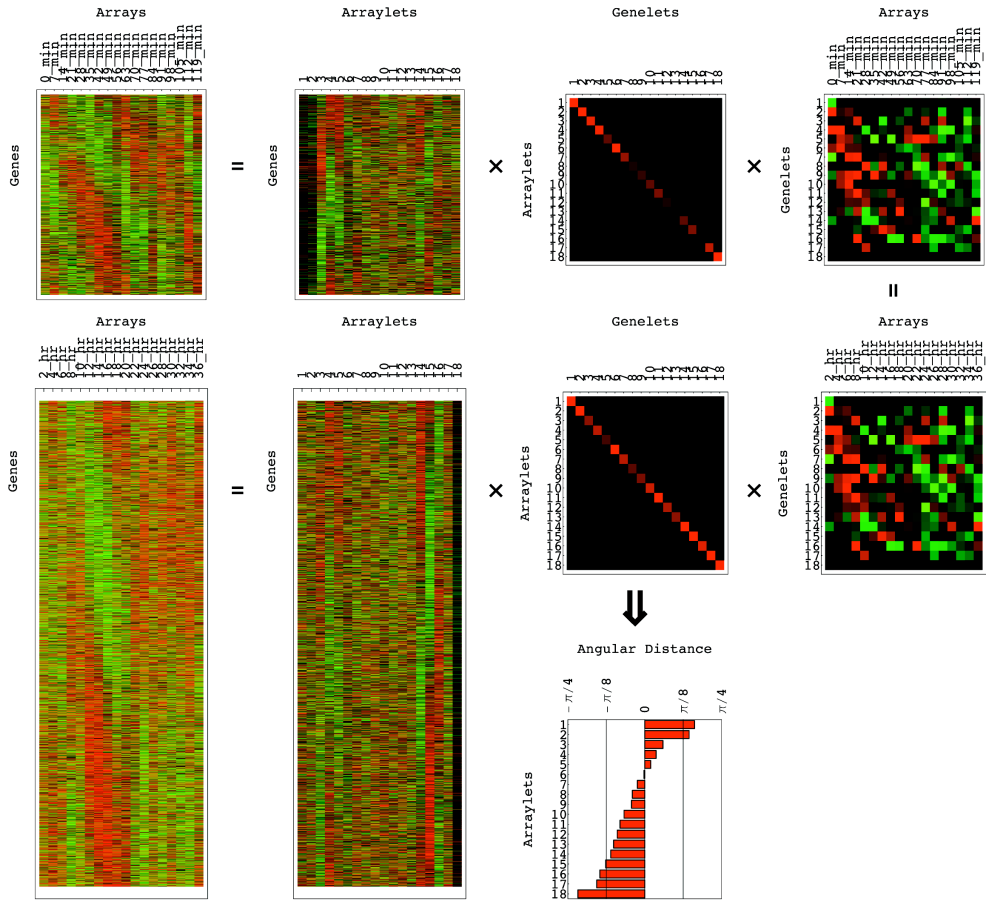
$$\vdots$$

$$D_N = U_N \Sigma_N X^{-1}.$$

GSVD is the only analysis tool to date that is not limited to comparison of orthologous or homologous genes.

Alter, Brown & Botstein, *PNAS* 100, 3351 (2003).

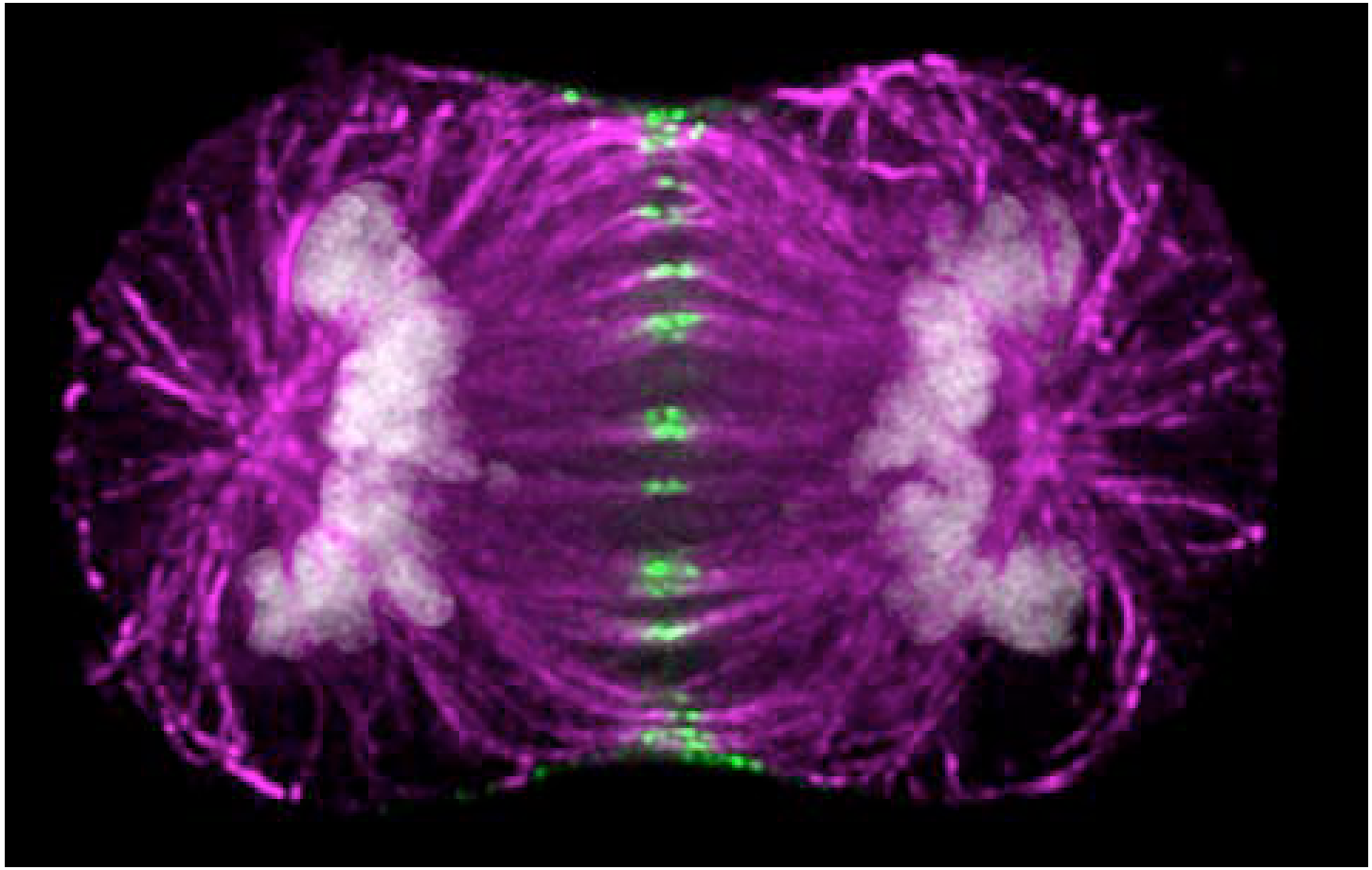
Yeast Spellman et al. *MBC* 9, 3273 (1998).



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

Matrix & Tensor Models Will Enable a Future

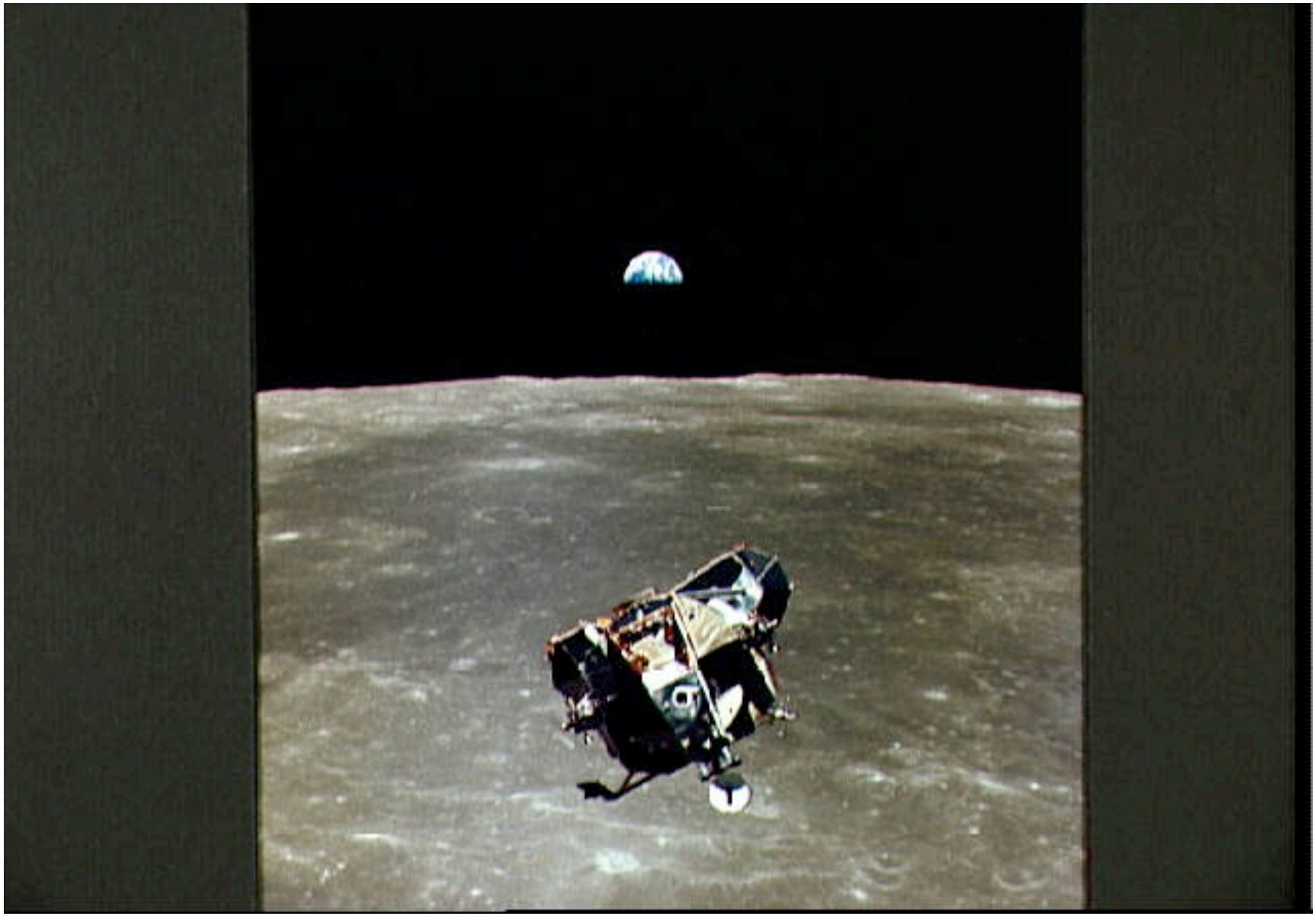
where cellular processes could be controlled in real time and in vivo.



Andrews & Swedlow, Nikon Small World (2002).

Cancer and disease could be stopped or reversed.
Damaged tissues could be engineered to regenerate.
Aging could be slowed or even halted altogether.

Today, NASA can control the trajectories of its spacecraft, because…



… their motion is understood and can be predicted mathematically.

Thanks to –

Collaborators:

John F. X. Diffley
Cancer Research UK, London

Gene H. Golub
Computer Science, Stanford

Robin R. Gutell
Integrative Biology, UT

Vishy Iyer
Molecular Genetics, UT

David Botstein
Genomics Institute, Princeton

Patrick O. Brown
Biochemistry, Stanford

Students:

Kayta Kobayashi, Pharmacy, UT

Chase Krumpelman, ECE, UT

Joel R. Meyerson, BME, UT

Chaitanya Muralidhara, CMB, UT

Larsson Omberg, Physics, UT

Sri Priya Ponnapalli, ECE, UT

Deepu Sudhakar, BME, UT

Funding:

NHGRI K01 Development Award in Genomic Research

NHGRI R01 HG004302

And, thank you!!!