rocha@lanl.gov

Normalization of DNA Microarray data

Luis M. Rocha
Complex Systems Modeling
CCS3 - Modeling, Algorithms, and Informatics
Los Alamos National Laboratory, MS B256
Los Alamos, NM 87545
rocha@lanl.gov or rocha@santafe.edu

rocha@lanl.gov

Normalization of DNA Microarray Data By Self-consistency and Local Regression

Thomas Kepler, Lynn Crosby, and Kevin Morgan

Little Attention is paid to a Systematic Study of Normalization. Yet it is essential to allow effective comparison of 2 or more arrays from different experimental conditions. *Implicit Assumption*: Linear response between true expression level and output intensity.

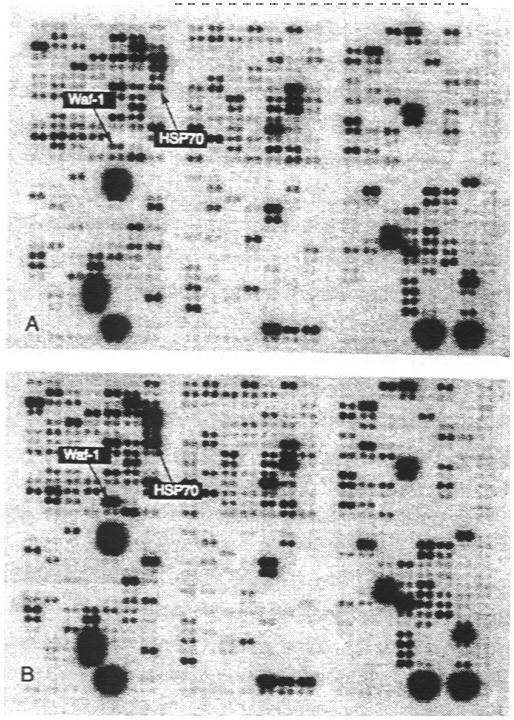
Intensity at Array Spot

$$I = vr + error$$

Unknown normalization factor

Abundance of corresponding mRNA

Needs to be determined to compare abundances



Luis Rocha
2001

http://www.c3.lanl.gov/~rocha/bioinformatics

rocha@lanl.gov

Real Response Functions

$$I = \nu r + error$$

ν is typically calculated with whole-array methods, using the median or the mean of the spot intensities or by inclusion of control mRNA. But the response function of a variety of hybridization schemes is not sufficiently linear nor consistent across assays. There may be a background constant or the intensity might saturate at large abundance.

$$I = \nu_0 + \nu_1 r + error$$

$$I = \frac{\nu_1 r}{1 + \nu_2} + error$$

- Thus simple ratio normalizations are inadequate
- Even “housekeeping” control genes cannot change the situation
  - Quantitative stability is not a priori assured nor demostrated empirically
  - Non- linearity of the response is not addressed
- 2-color probes on the same microarray do not solve the issue because of variation in relative activity and incorporation of 2 fluorescent dyes.

Luis Rocha
2001

rocha@lanl.gov

Self-consistency Error Model

Assumption: Majority of genes in any given comparison will be expressed at constant relative levels; only a minority of genes has their expression levels affected. Thus, pairs or groups of assays are normalized relative to each other by maximizing the consistency of relative expression among them.

Experimental Design: 2 treatment groups and 2 or more replicative arrays per group. Generalization is straightforward. Comparisons without replicative arrays are possible.

Logarithmic transformation converts a multiplicative normalization constant to an additive one

$$I = vr + error \quad (1)$$

Measured intensity of the *k*th spot in the *j*th replicative assay of the *i*th treatment group

$$Y_{ijk} = \alpha_{ij} + \xi_k + \delta_{ik} + \sigma_0 \varepsilon_{ijk} \quad (2)$$

$Y_{ijk} = \log I_{ijk}$

$\alpha_{ij} = \log v_{ij}$

Normalization Constants

$\xi_k + \delta_{ik} = \log r_{ik}$

Mean log abundance and treatment (specific) effects

Error standard deviation and residuals (assumed zero mean and unit variance)

Luis Rocha
2001

rocha@lanl.gov

The Complete Error Model

But a constant normalization ν is not realistic, a more flexible model is:

Measured intensity of the *k*th spot in the *j*th replicative assay of the *i*th treatment group

$$Y_{ijk} = \beta_{ij}(\xi_k) + \delta_{ik} + \sigma(\xi_k)\varepsilon_{ijk} \quad (3)$$

Normalization Function, depends on mean log abundance

Treatment (specific) effects

Error function also depends on mean log abundance

Constraints: $\sum_k \xi_k = 0$ $\sum_i n_i \delta_{ik} = 0$ n_i is the number of replicates in the ith treatment group

Kepler et al estimate parameters according to (2) then fit them as approximations to (3). The estimators for ζ, α, and δ are *x*, *a*, and *d*:

$$x_k = \overline{Y_{..k}} - \overline{Y_{...}}$$

$$a_{ij} + \overline{d_{i.}} = \overline{Y_{ij.}}$$

$$d_{ik} - \overline{d_{i.}} = \overline{Y_{i.k}} - \overline{Y_{..k}} - \overline{Y_{i..}} + \overline{Y_{...}}$$

$$\overline{Y_{ij.}} = 1/n \sum_{k=1}^{n} Y_{ijk}$$

n is the number of spots per array

Luis Rocha
2001

rocha@lanl.gov

Self-Consistency Error Model

But an additional condition is required to determine the estimator for the treatment specific effects or bias $\overline{d}_{i\cdot}$

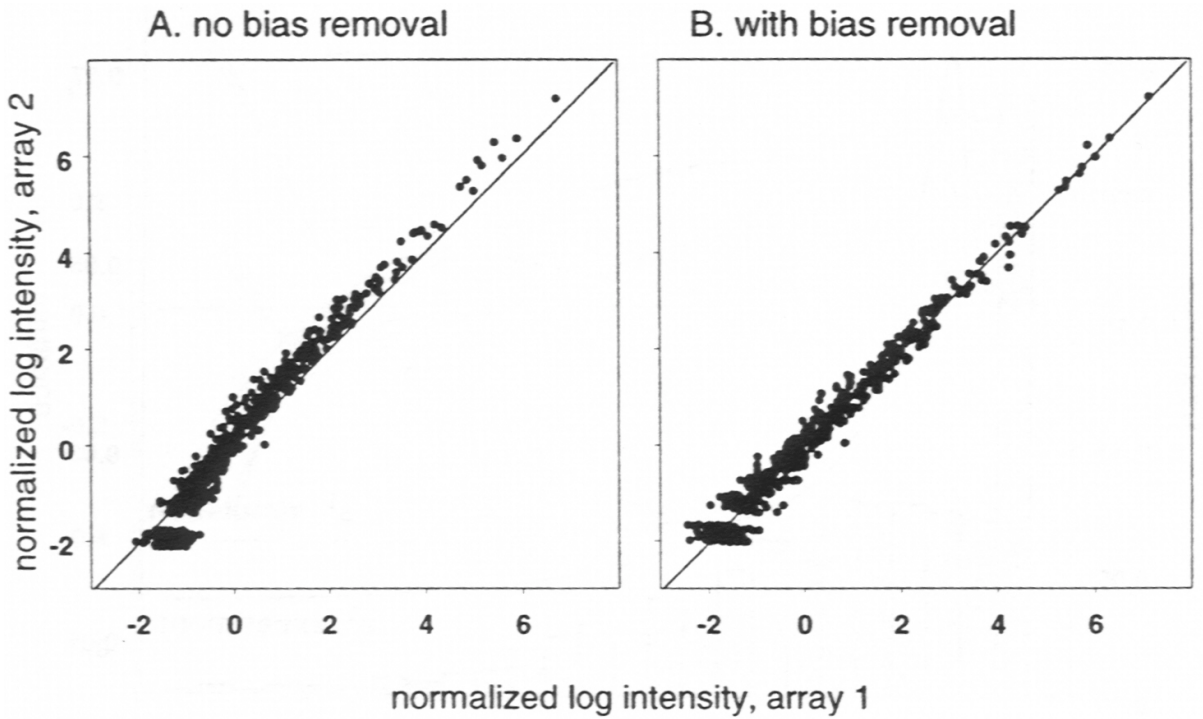
Assumption: Expression of majority of genes does not change appreciably from treatment to treatment. May not always be reasonable, but at least for living cells, maintenance must maintain a relevant background of expression will stay at stable levels.

Self-consistency: Identification of a background pattern of activity, a transcriptional "core". But which genes belong to the core depends on the normalization, and the optimal normalization depends on which genes are identified with the core. So an iterative process is used to eliminate from the core those genes with largest estimated treatment effects (bias) – until no change is observed from one iteration to the next.

rocha@lanl.gov

Example of Bias

Normalized Log Intensity in 2 replicate arrays



A. Data normalized by subtracting the mean over all spots
B. by estimating the normalization function and then subtrating inferred bias

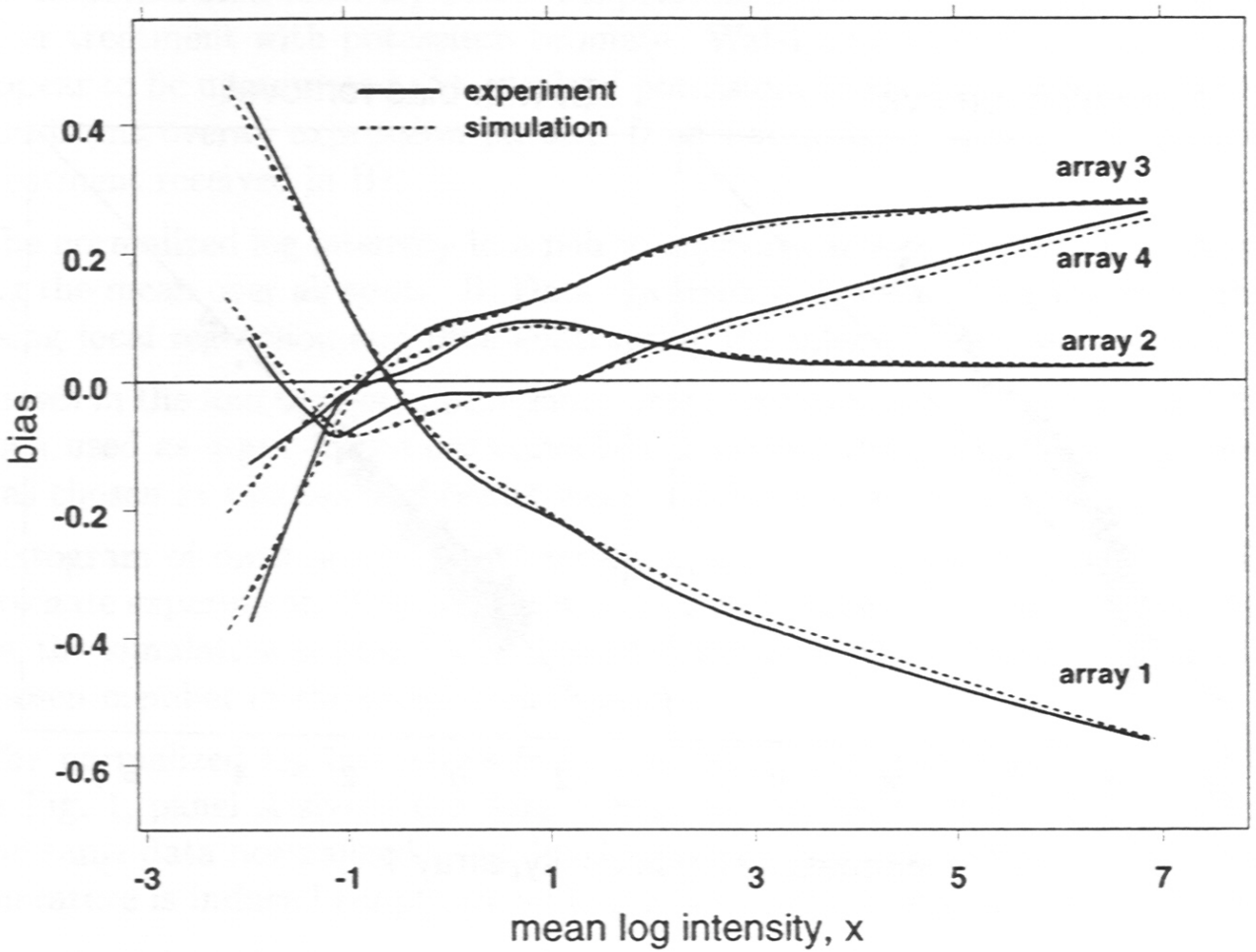
Luis Rocha
2001

http://www.c3.lanl.gov/~rocha/bioinformatics

rocha@lanl.gov

Bias from the Normalization Function

Potassium bromate experiment



Normalization bias varies with intensity

Luis Rocha
2001

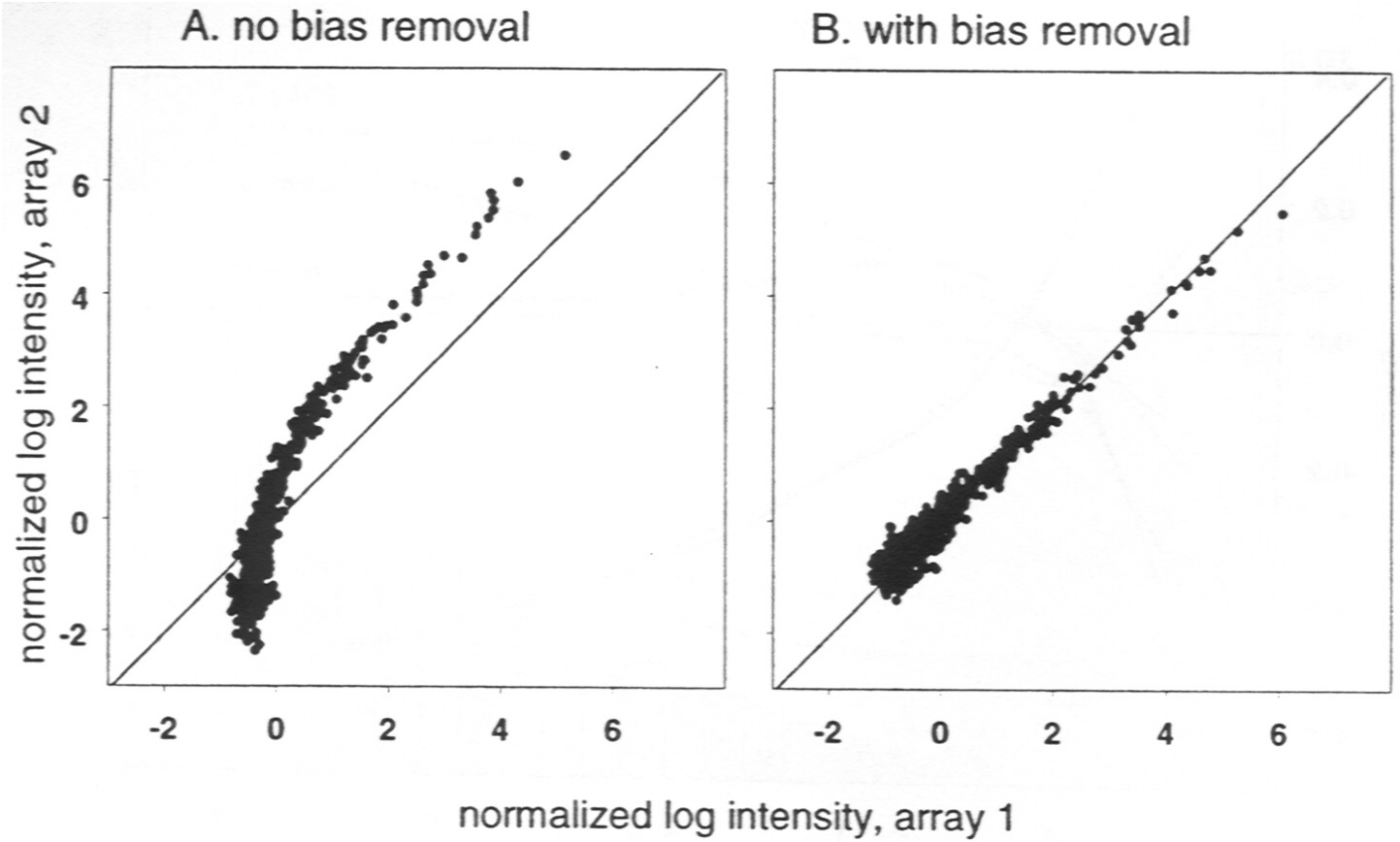
http://www.c3.lanl.gov/~rocha/bioinformatics

Los Alamos National Laboratory

rocha@lanl.gov

Normalized log Intensities

Simulated Data



Donwload from santafe.edu -> Working Papers -> 00-09-055

http://www.c3.lanl.gov/~rocha/bioinformatics

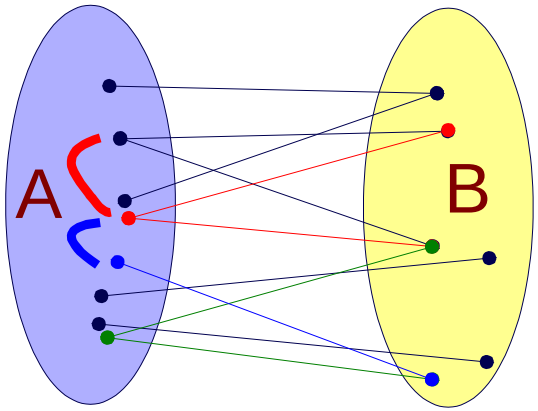
Luis Rocha
2001

Los Alamos National Laboratory

rocha@lanl.gov

Singular Value Decomposition

Higher-Order "Clustering" Also Known as: Principal Components Analysis.



- Given a relation (a matrix) between 2 sets of distinct objects. SVD is used to discover the implicit higher-order structure in the relation
  - Keyterms by Documents, Genes by Arrays
  - Higher-order means indirect relationships: Those associations between the two types of objects which are not evident by individual associations.

- In Language and IR most words have many meanings (polysemy) and there are several possible words to express the same concept (synonymy)
  - SVD is used to identify the several meanings of words and "cluster" the words that express the same concept.
- For gene expression data, we expect to find genes which participate in several networks (gene functional polysemy) and different genes to participate in the same networks (gene functional synonymy)
  - Clustering usually demands strict inclusion (except for Fuzzy)

rocha@lanl.gov

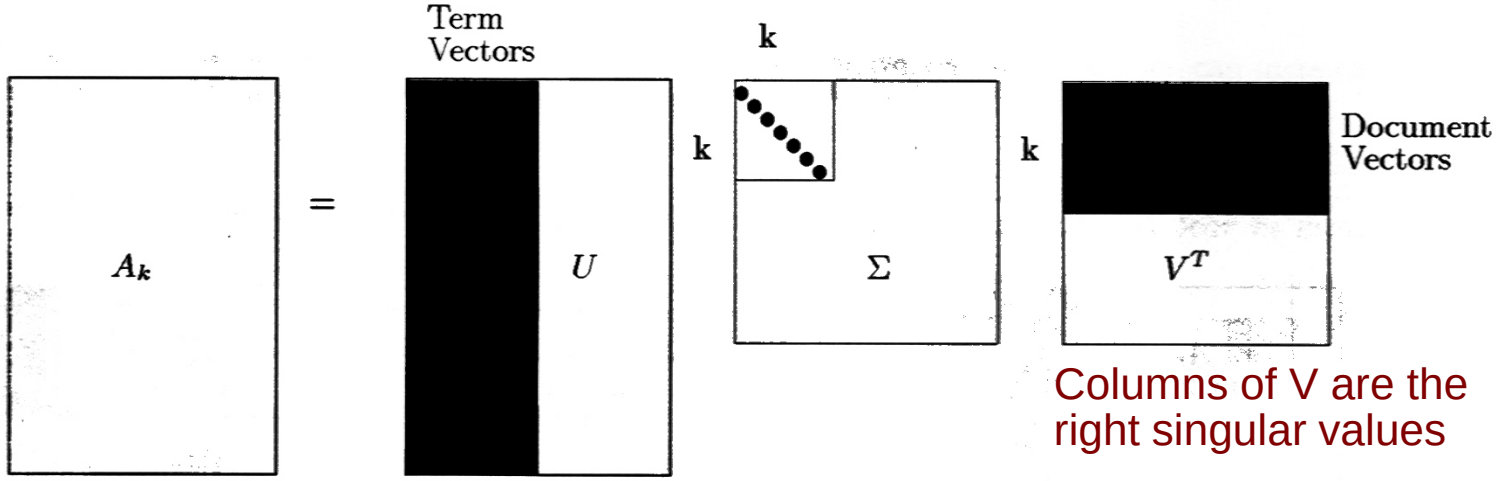
Singular Value Decomposition

Decomposition into orthogonal vectors of linear combinations of elements

Given an $m \times n$ matrix A, $m \geq n$ and rank(A)=r, the SVD of A is:

$$A = U \Sigma V^T$$

U is $m \times n$ and V $n \times m$. They are orthogonal $U^T U = V^T V = I_n$. Σ is all 0 except for the $\Sigma_{i,i} = \sigma_i$ for i=1,...r, which are the nonnegative square roots of the n eigenvalues of AA^T.



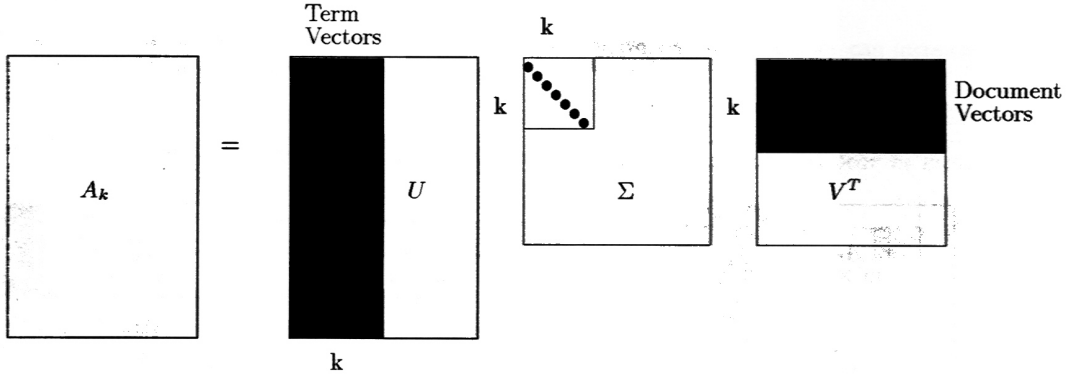
Luis Rocha
2001

rocha@lanl.gov

Singular Value Decomposition

Facts

Theorem: A_k, constructed from the k largest singular triplets of A, is the rank-k matrix that best approximates A.



- In IR, rank approximations of the strongest components are used to reduce the dimensionality of data, while removing the noise or variability of word usage.
  - Captures the important cases of synonymy and polysemy
  - Example: Keywords car, automobile, driver, and elephant.
  - Example: Search for “Demographic shifts in the U.S. with economic impact”, retrieve “The nation grew to 249.6 million people in the 1980's as more Americans left the industrial and agricultural heartlands for the South and West” – No lexical match.[Schultze, 1995]
- Neural Networks and other classifiers perform much better on the decomposed, lower dimensionality data

rocha@lanl.gov

Singular Value Decomposition

For Gene Expression Data

- Cluster analysis provides little insight into inter-relationships among groups of co-regulated genes
- Component (”spectral”) analysis yields a description of superposed behavior of gene expression networks, rather than a partition.
- Holter et al [2000] compares the superposed components to the characteristic vibration modes of a violin string which entirely specify the tone produced

rocha@lanl.gov

Holter et al SVD Analysis

PNAS, Vol. 97, no. 15, pp. 8409-8414:
www.pnas.org/cgi/doi/10.1073/pnas.150242097

Compared SVD analysis of published yeast cdc15 cell-cycle [Spellman et al 1998] and sporulation [Chu et al, 1998] data sets, as well as the data set from serum-treated human fibroblasts [Iyer et al, 1999]. Iterative normalization to guarantee zero mean row/column by subtracting mean values of raw data.

Table 1. Singular values extracted from gene expression and random data sets

cdc15, 12 points		cdc15, 15 points		spo, selected		spo, full		fibr	
Experiment	Random*	Experiment	Random*	Experiment	Random*	Experiment	Random*	Experiment	Random*
15.81	8.65	14.47	7.66	15.20	9.61	49.54	32.29	14.10	7.40
13.10	8.56	12.37	7.58	10.53	9.17	37.40	32.22	12.49	7.06
8.68	8.17	10.45	7.44	7.18	9.01	29.88	32.01	5.65	6.94
7.34	8.04	6.80	7.33	5.67	8.83	23.43	31.93	5.47	6.84
5.45	7.97	6.71	7.20	5.43	8.73	22.36	31.67	5.12	6.78
5.00	7.82	4.52	7.09	4.67	8.06	17.97	31.47	4.65	6.67
4.51	7.57	4.36	6.97					4.01	6.52
4.26	7.53	4.15	6.93					3.19	6.37
3.66	7.41	3.89	6.76					3.03	6.32
3.33	7.33	3.39	6.64					2.67	6.12
3.08	7.14	3.05	6.49					2.31	5.85
		2.89	6.47					2.17	5.68
		2.75	6.38						
		2.57	6.28						

*Random data sets contained the same number of rows and columns as the corresponding gene expression data sets. The data were generated randomly from a uniform distribution between 0 and 1 and then were polished.

rocha@lanl.gov

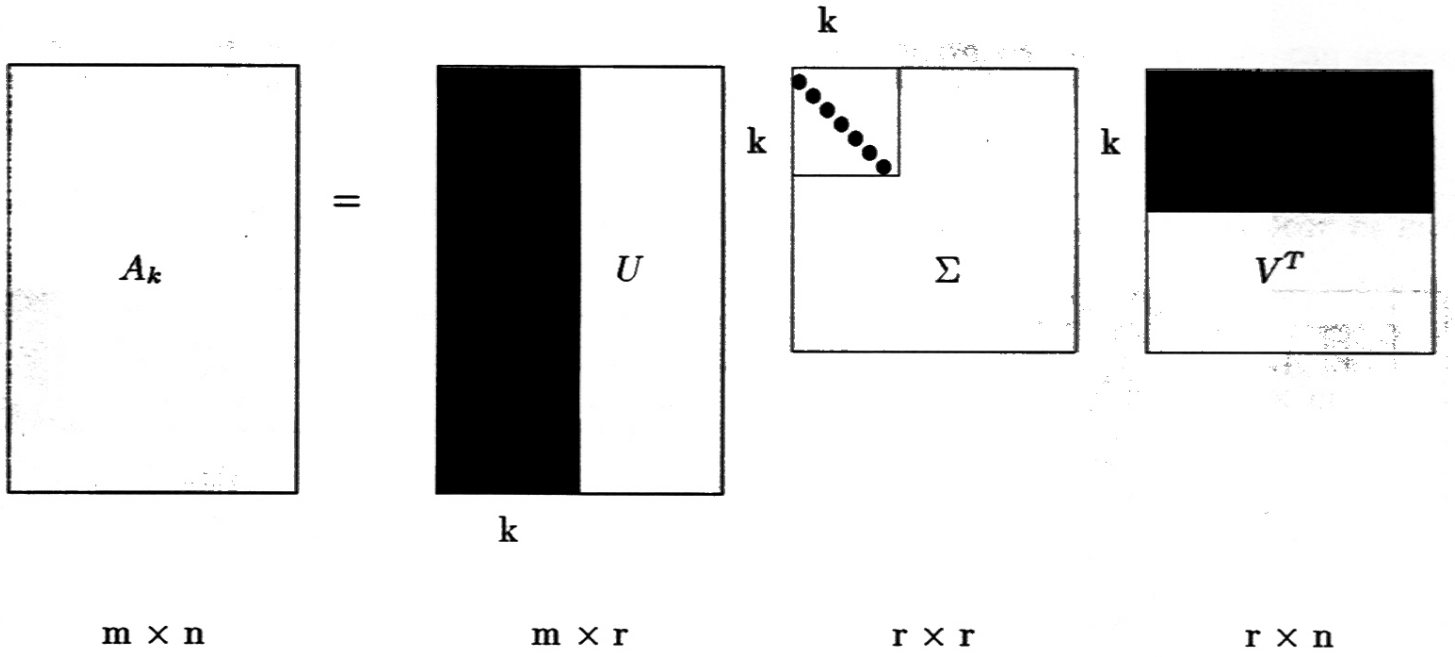
SVD of Temporal Data

Holter et al Experiments

- Most data sets contain spread out singular values with significantly greater first 2-3 values.
  - The SVD of randomly generated data does not show significant component structure structure
  - In contrast, for purely periodic data (e.g. all genes with same sinosoidal period but dephased) there would be only 2 components (a sine and cosine with same period)
- The essential behavior is captured by the first few components
  - They claim that first component represents smaller scale fluctuations and experimental noise (a steady state). Shouldn't this be filtered out by normalization in the first place?

rocha@lanl.gov

SVD for Gene Expression



Columns are time steps and rows are genes

Columns of U are eigenarrays (rows are genes)

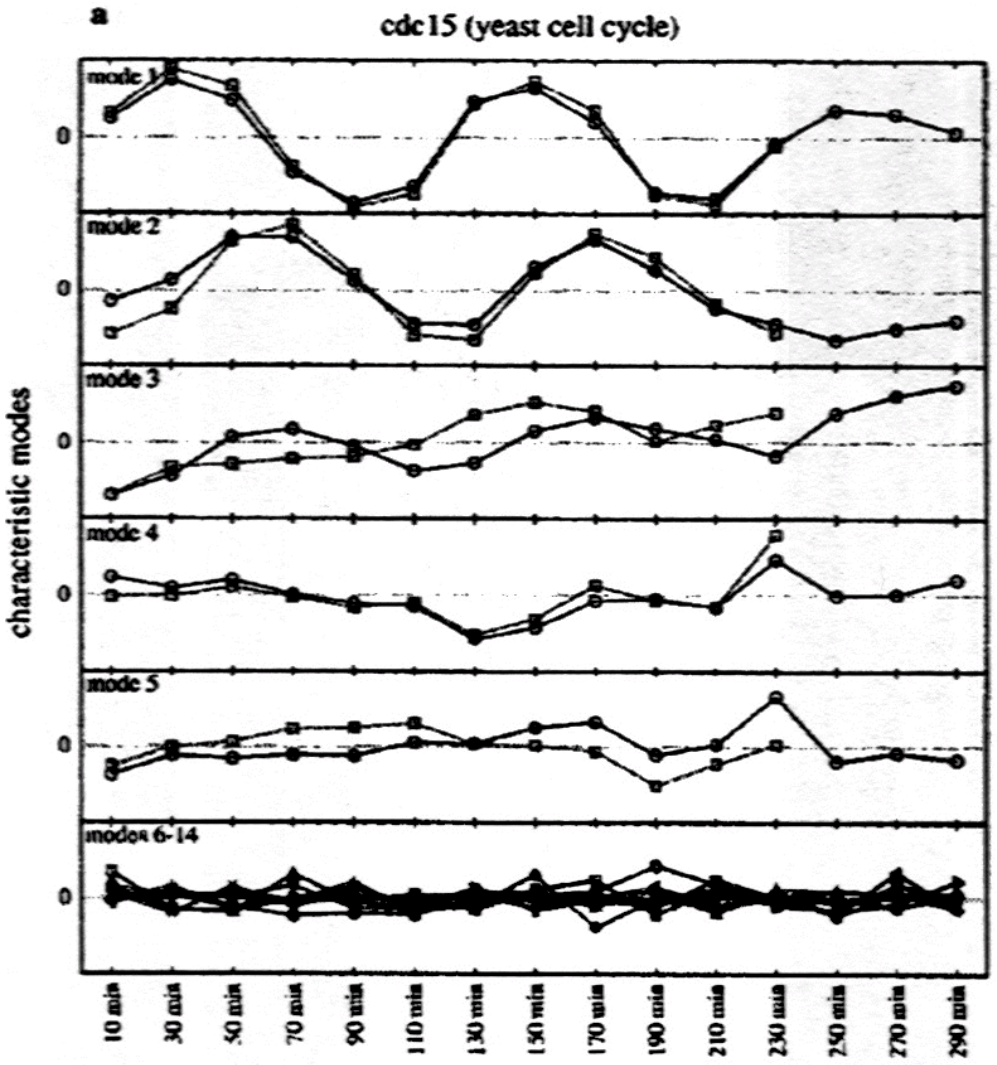
Rows of V^T are eigengenes (colums are time steps)

Luis Rocha
2001

http://www.c3.lanl.gov/~rocha/bioinformatics

rocha@lanl.gov

Holter et al SVD Analysys



- 800 genes by 15 (12) time measurements
- 2 dominant modes
  - Approximately sinusoidal and out of phase
  - Less synchronized as cell enters 3rd cycle
  - If only 12 points are used, third SV loses relevance, but 2 first components remain largely unchanged

Eigengene: rows of V^T (each column is a time instance)

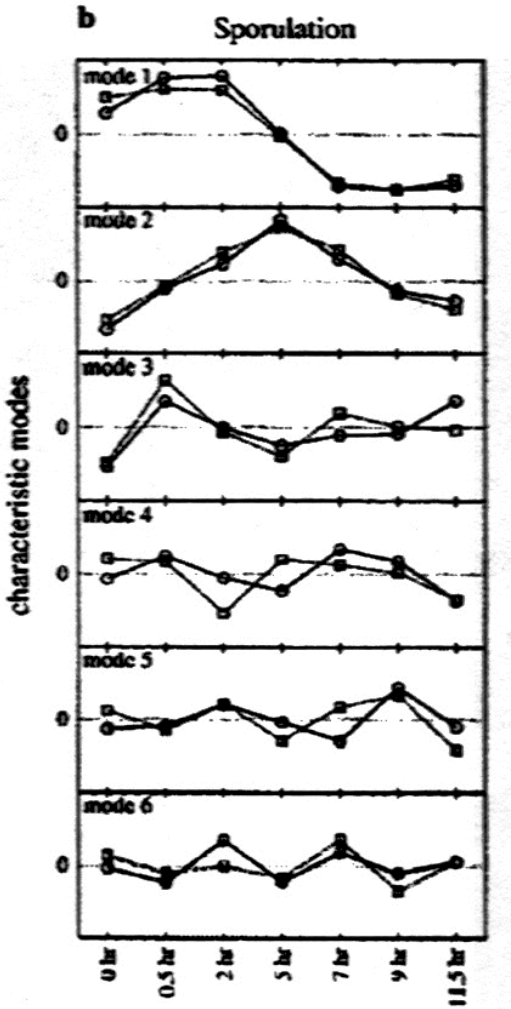
http://www.c3.lanl.gov/~rocha/bioinformatics

Luis Rocha
2001

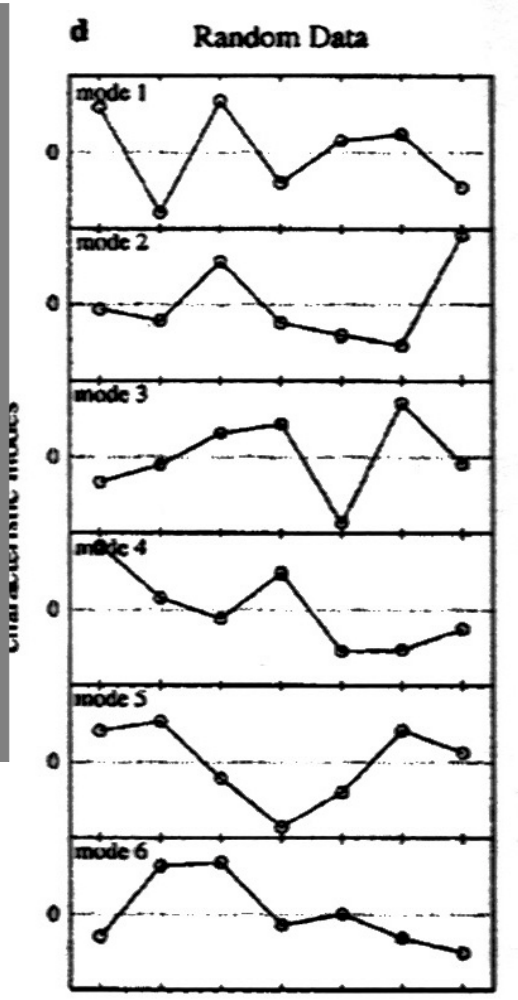
Los Alamos
National Laboratory

rocha@lanl.gov

Sporulation data set



- Sporulation specific genes (chosen by Chu et al) constrasted with whole genome data (6000)
  - The first modes demonstrate that the chosen sporulation specific genes are responsible for the essential behavior
- Random data with same dimensions
  - All modes are important



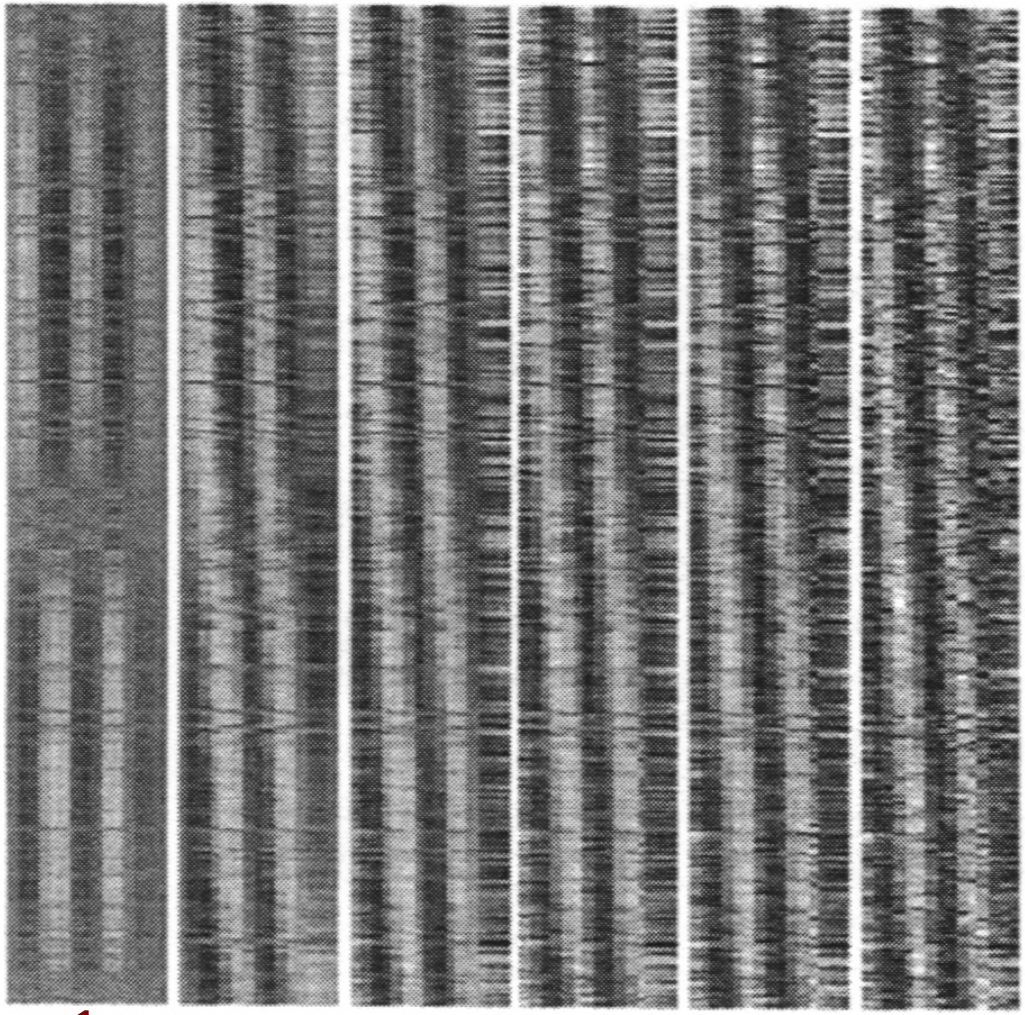
Luis Rocha
2001

http://www.c3.lanl.gov/~rocha/bioinformatics

Los Alamos
National Laboratory

rocha@lanl.gov

cdc15 Reconstruction with k-highest modes



Rows are genes
Columns are time points

It implies an undelying simplicity in genetic response

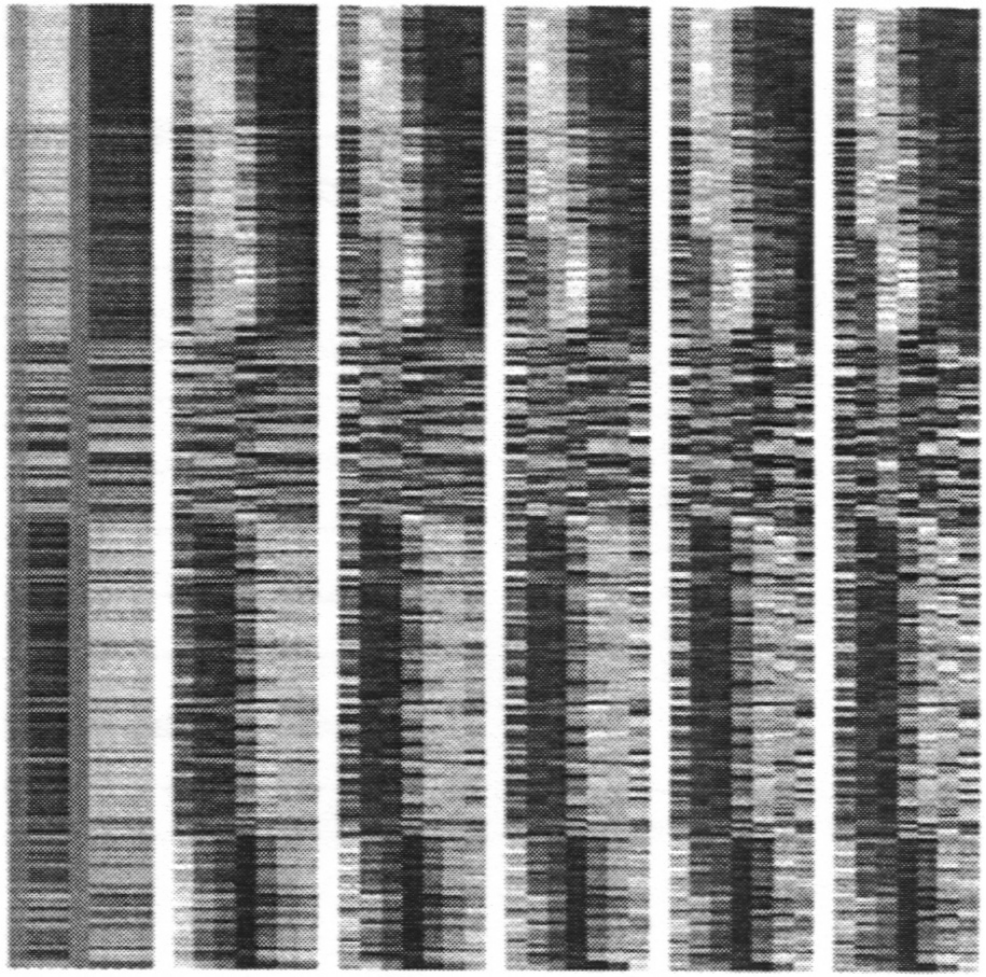
Luis Rocha
2001

http://www.c3.lanl.gov/~rocha/bioinformatics

Los Alamos
National Laboratory

rocha@lanl.gov

Sporulation Reconstruction



http://www.c3.lanl.gov/~rocha/bioinformatics

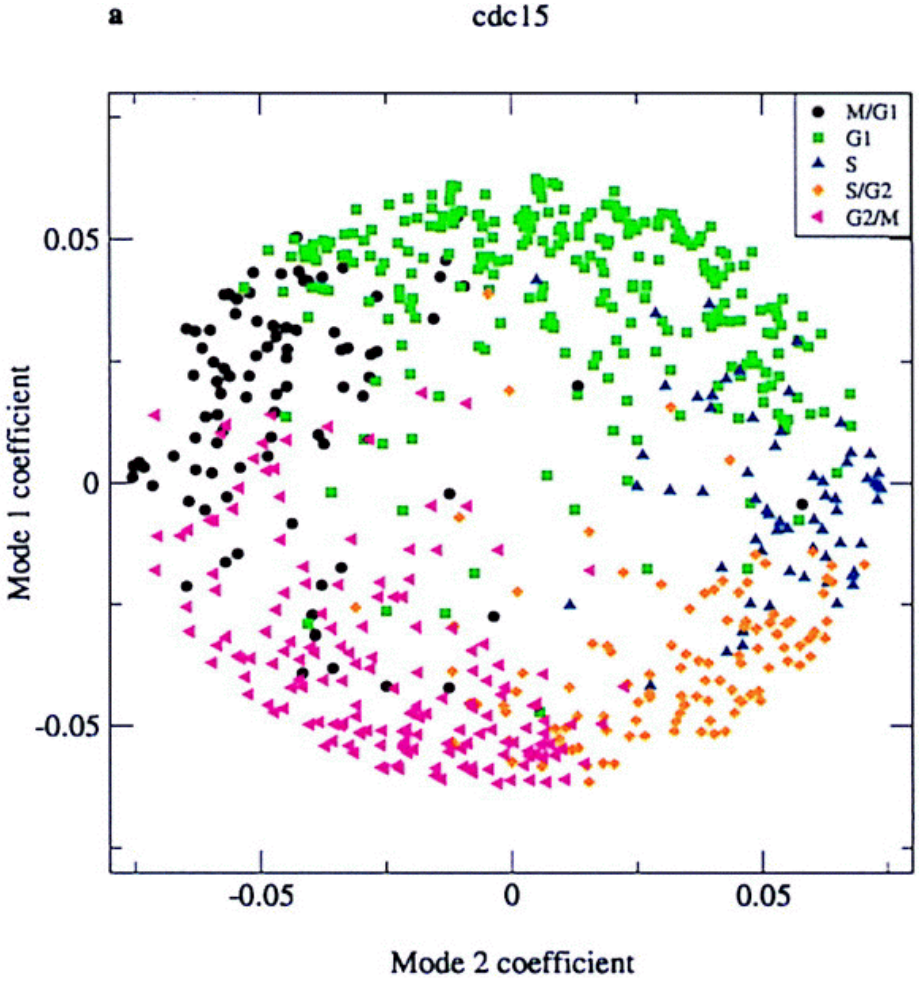
Luis Rocha
2001

Los Alamos
National Laboratory

rocha@lanl.gov

Eigenarray Coefficient Plot

Plot of the coefficients of the first 2 modes for all genes



Eigenarray: columns of *U* (each row is a gene)

Each element of the eigenarrays (coefficients) is a measure of its contribution to expression profile of a gene

Luis Rocha
2001

http://www.c3.lanl.gov/~rocha/bioinformatics

Los Alamos National Laboratory

rocha@lanl.gov

Eigenarray Coefficient Plot

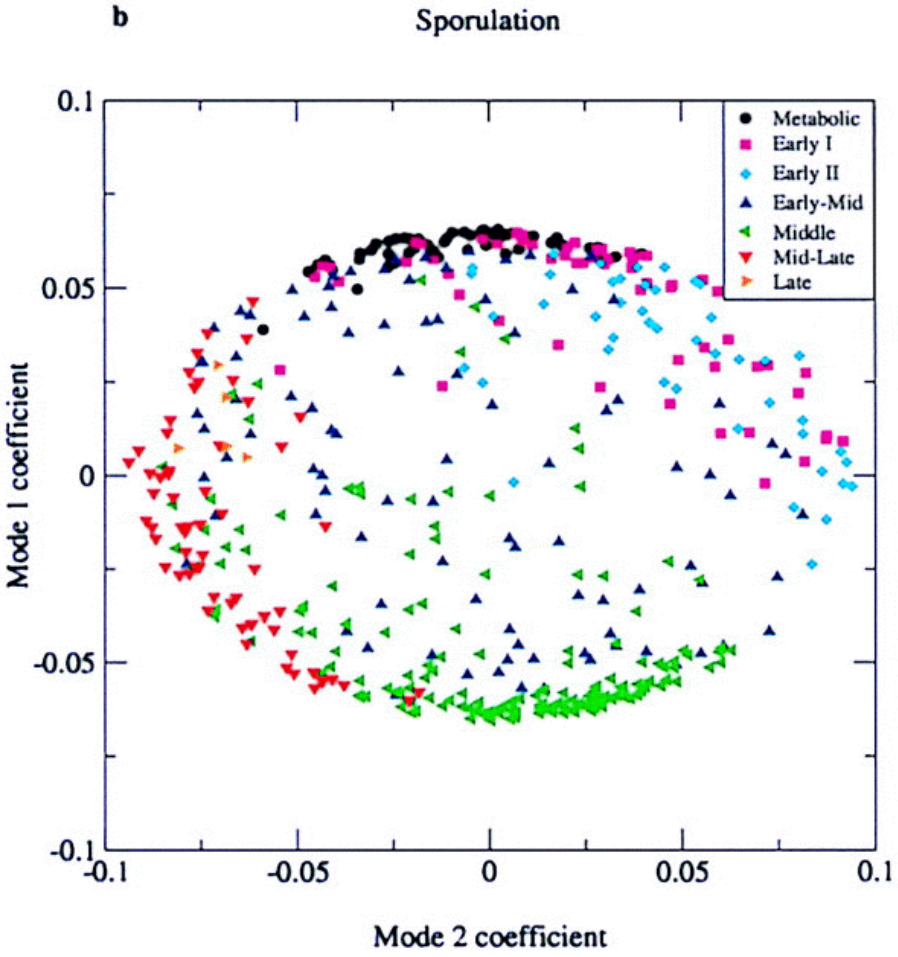
Conclusions

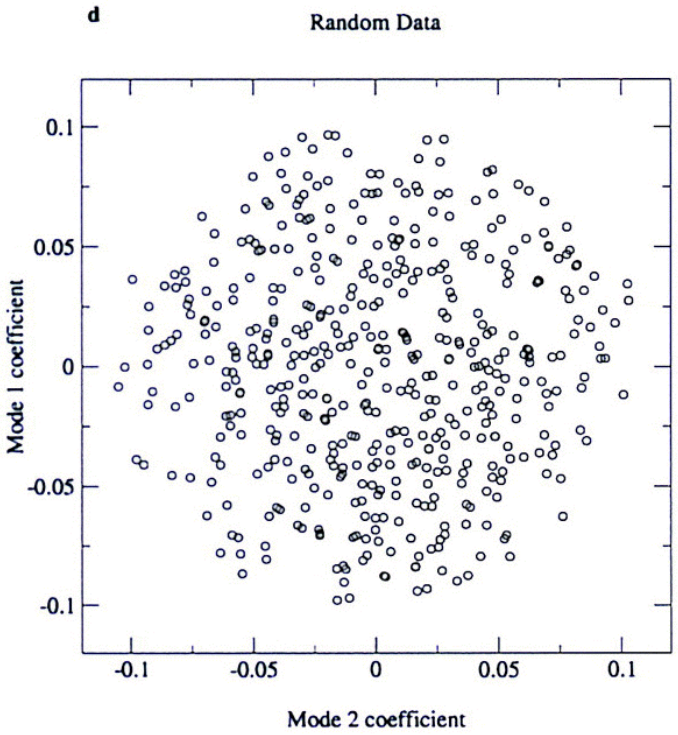
- Plot
  - More of a circle (instead of elipse) implies equal contributions from each component
  - Populated evenly implies that the coefficients vary continously
- Clusters of genes by other methods cluster in these plots, but the temporal progression in the cell cycle and in the course of sporulation is more evident in the SVD analysis
- Holter et al conclude that genes are not activated in discrete groups or blocks, as historically implied by the division of the cell cycle into phases or the sporulation response into tempotal groups. There is a continuity in expression change

rocha@lanl.gov

Coefficient Plot

Sporulation and Random





Luis Rocha
2001

http://www.c3.lanl.gov/~rocha/bioinformatics

Los Alamos
National Laboratory

rocha@lanl.gov

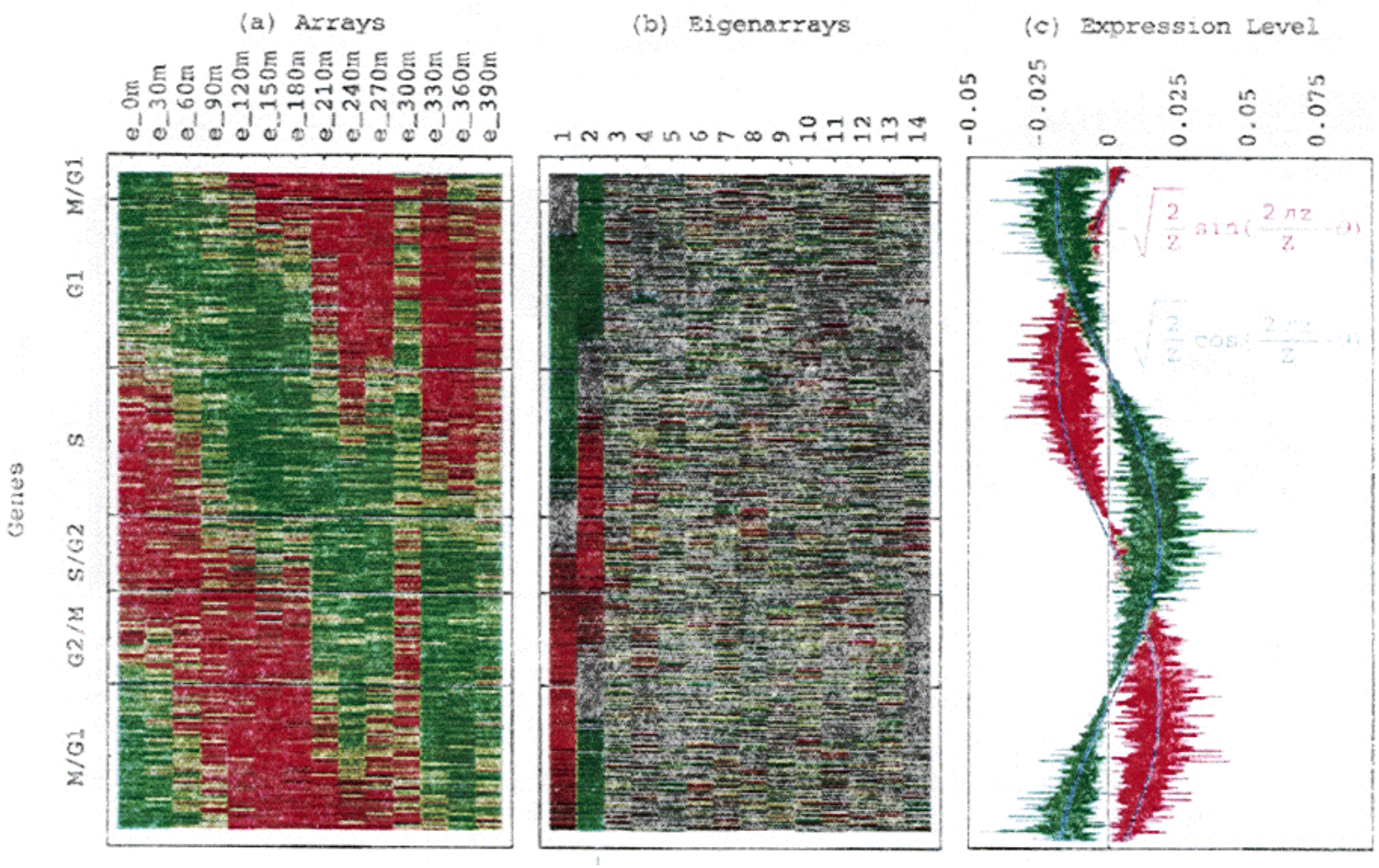
SVD and Data Processing

Alter et al [2000]. PNAS, Vol. 97, No. 18, pp. 10101-10106

- Sorting GE data according to the coefficients of genes and arrays in eigengenes and eigenarrays gives a global picture of expression dynamics
  - Genes and arrays are classified into groups of similar regulation and function (polysemy) or similar cellular state and biological phenotype respectively (synonymy)
  - ***Eigengene*** (vector of array coefficients): regulatory program or process from its expression pattern across all arrays, when this pattern is biologically interpretable. Effect on a group of genes of the change in a regulator.
  - ***Eigenarray*** (vector of gene coefficients): the cellular state which corresponds to an eigengene.
  - Wall, clusters eigenarray coefficients. Better than traditional clustering since genes affected by the same regulator are clustered together irrespective of up or down regulation
- SVD allows us to filter out the effects of particular eigengenes/eigenarrays
  - Selective Normalization of data

rocha@lanl.gov

Some results from Alter et al



First 2 eigengenes

Luis Rocha
2001

http://www.c3.lanl.gov/~rocha/bioinformatics

Los Alamos
National Laboratory

rocha@lanl.gov

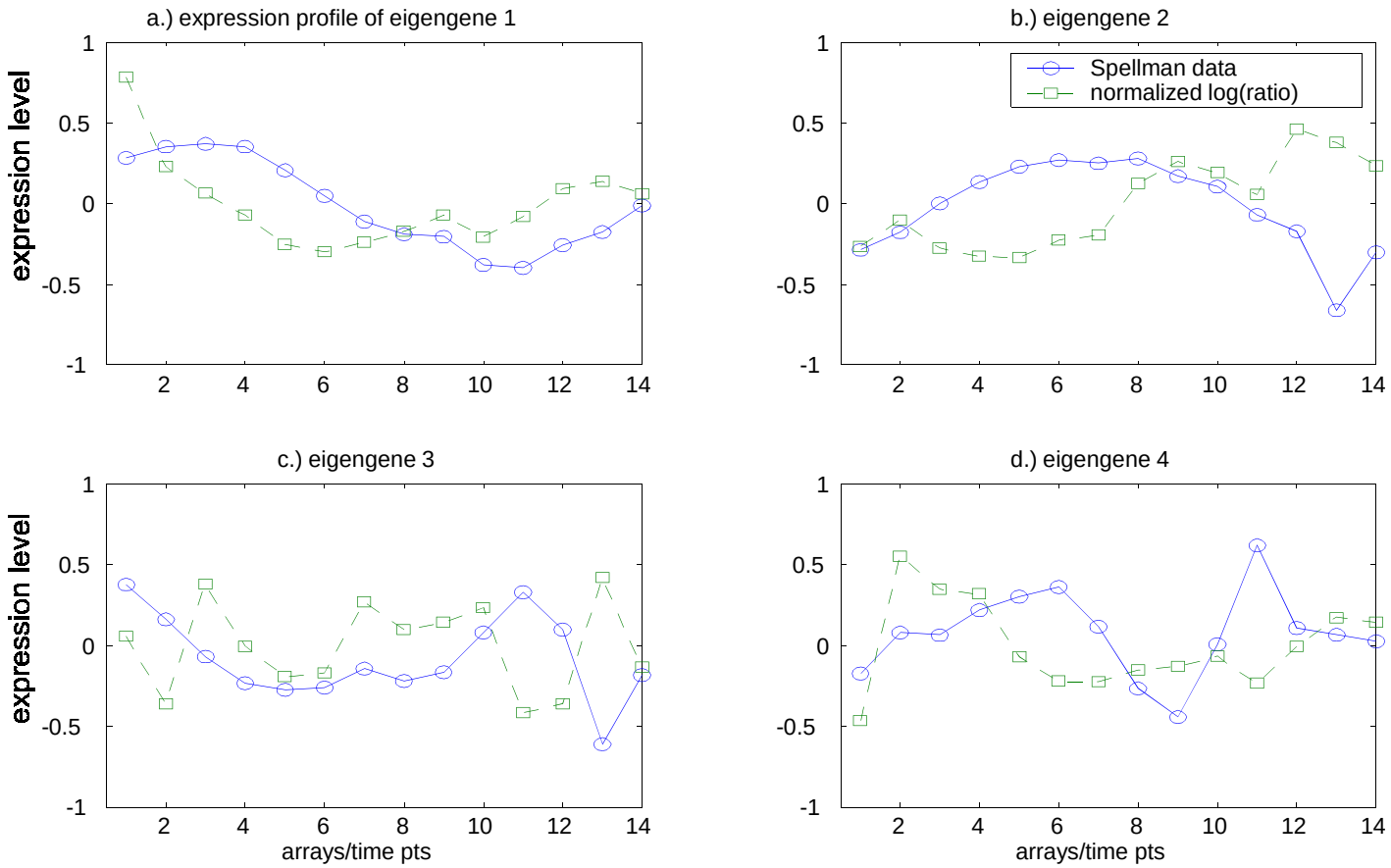
SVD and Normalization

Preliminary Results (Andreas Rechtesteiner)

- The 1st component that Alter found is responsible for 90% of expression, but our replication of his treatment was 56%
  - Alter interprets this first component as expression steady state, but his normalization should have taken care of this,as the intensity mean is subtracted from all genes...
-

rocha@lanl.gov

Different Results from Different Normalization



Luis Rocha
2001

http://www.c3.lanl.gov/~rocha/bioinformatics

Los Alamos National Laboratory

rocha@lanl.gov

References

- Normalization
  - Kepler, T., L. Crosby and K. T. Morgan [2000]."Normalization and analysis of DNA microarray data by self-consistency and local regression." *Nucleic Acids Research*. Submitted. Santa Fe Institute preprint: 00-09-055.
- SVD and Latent Semantic Analysis in IR
  - Berry, M.W., S.T. Dumais, and G.W. O'Brien [1995]."Using linear algebra for intelligent information retrieval." *SIAM Review*. Vol. 37, no. 4, pp. 573-595.
  - Kannan, R. and S. Vempala [1999]."Real-time clustering and ranking of documents on the web." Unpublished Manuscript.
  - Landauer, T.K., P.W. Foltz, and D. Laham [1998]."Introduction to Latent Semantic Analysis." *Discourse Processes*. Vol. 25, pp. 259-284.
- SVD for Gene Expression Analysis
  - Alter, O., P.O. Brown and D. Botstein [2000]."Singular value decomposition for genome-wide expression data processing and modeling." *PNAS*. Vol. 97, no. 18, pp. 10101-06.
  - Hastie, T. et al [2000]."'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns." *Genome Biology*. Vol. 1, no. 2, pp. 3.1-3.21.
  - Holter, N.S. et al [2000]."Fundamental patterns underlying gene expression profiles:Simplicity from complexity." *PNAS*. Vol. 97, no. 15, pp. 8409-14.
  - Raychaudhuri, S., J.M. Stuart and R.B. Altman [2000]."Principal components analysis to summarize microarray experiments: Application to sporulation data." ????? http://cmgm.stanford.edu.
  - Wall, M., P.A. Dyck, and T. Brettin [2001]."SVDMAN -- Singular value decomposition analysis of microarray data." *Bioinformatics*. In Press.

Luis Rocha
2001