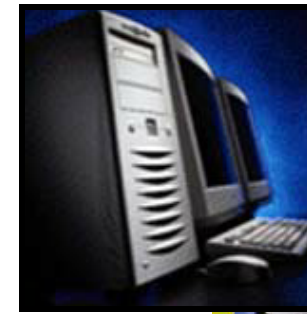
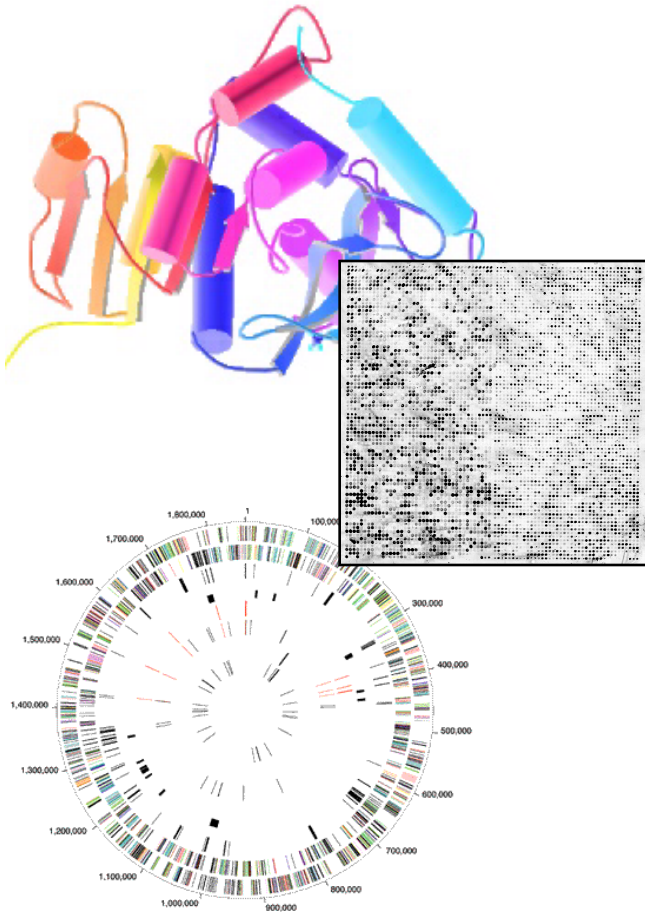


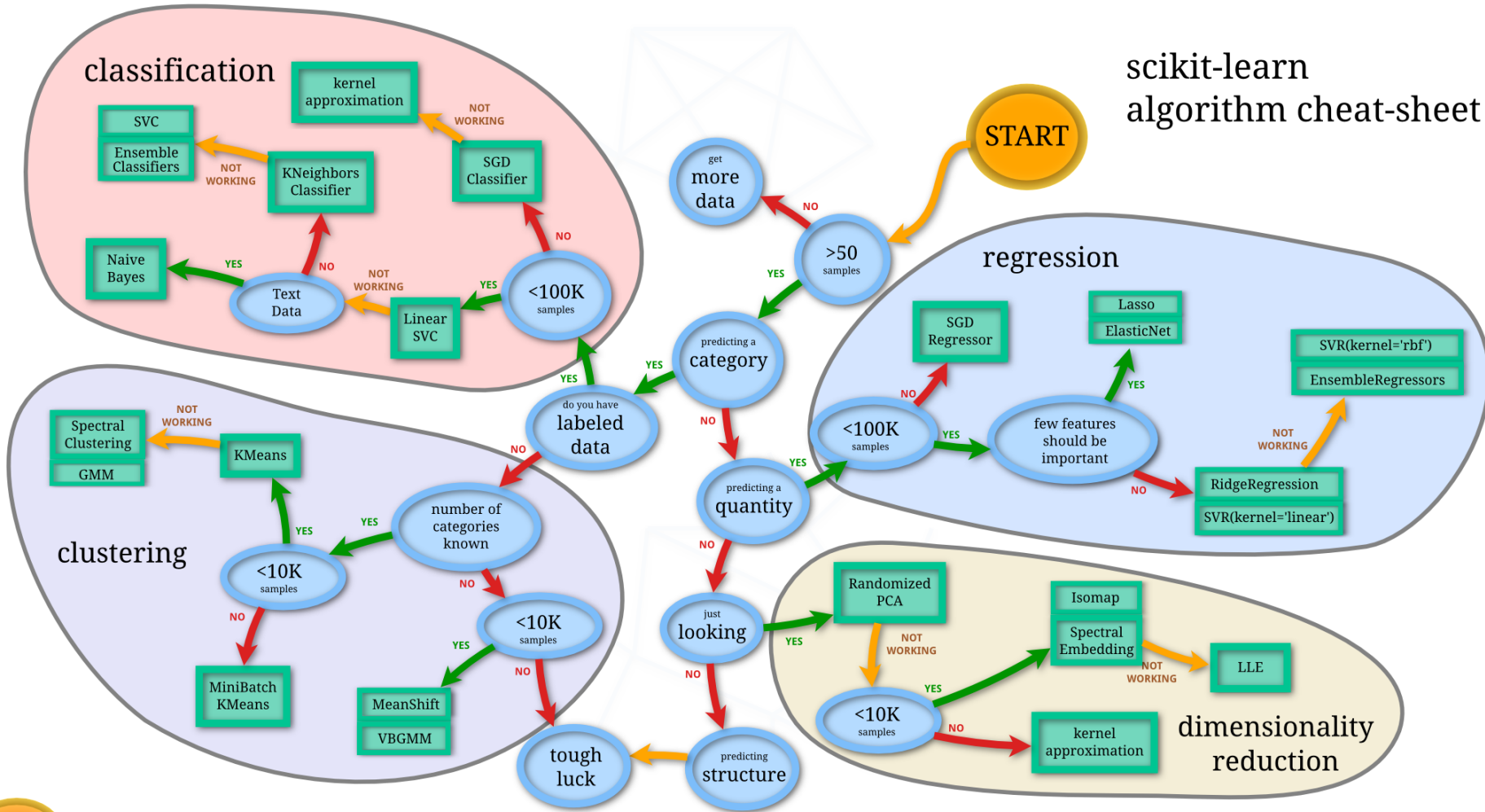
Bioinformatics: Unsupervised Datamining



Mark Gerstein, Yale University
gersteinlab.org/courses/452
(last edit in spring '15)

The World of Machine Learning

scikit-learn
algorithm cheat-sheet



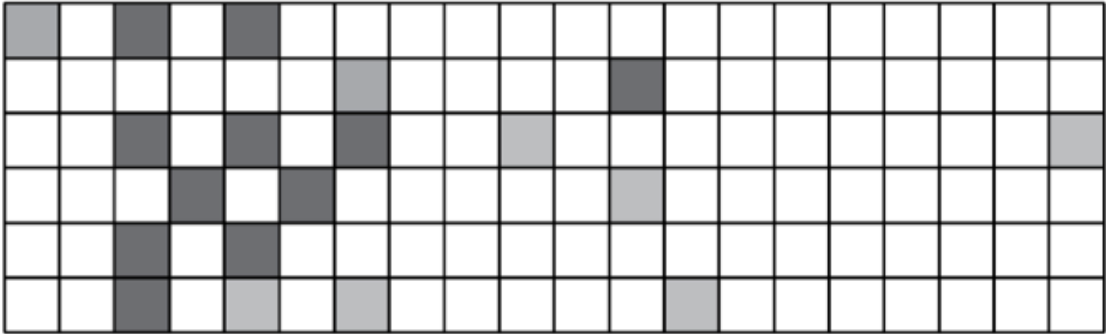
Supervised vs Unsupervised Mining

Structure of Genomic Features Matrix

1

Sites along the genome

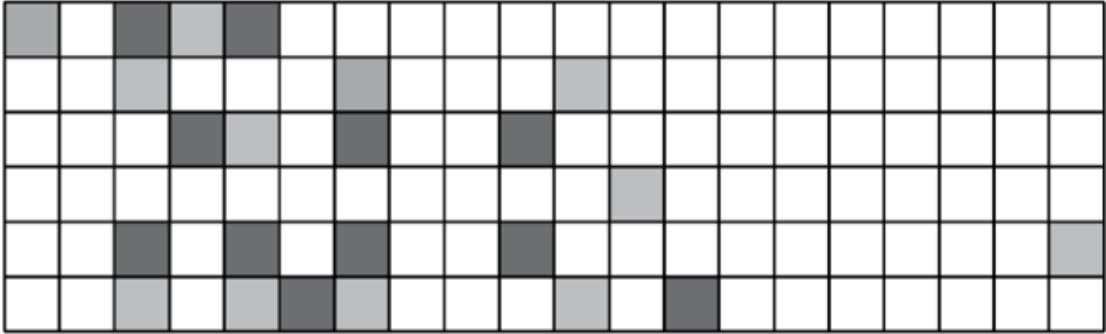
Factors
and
Chromatin
Modifications
(different
tissues)



...

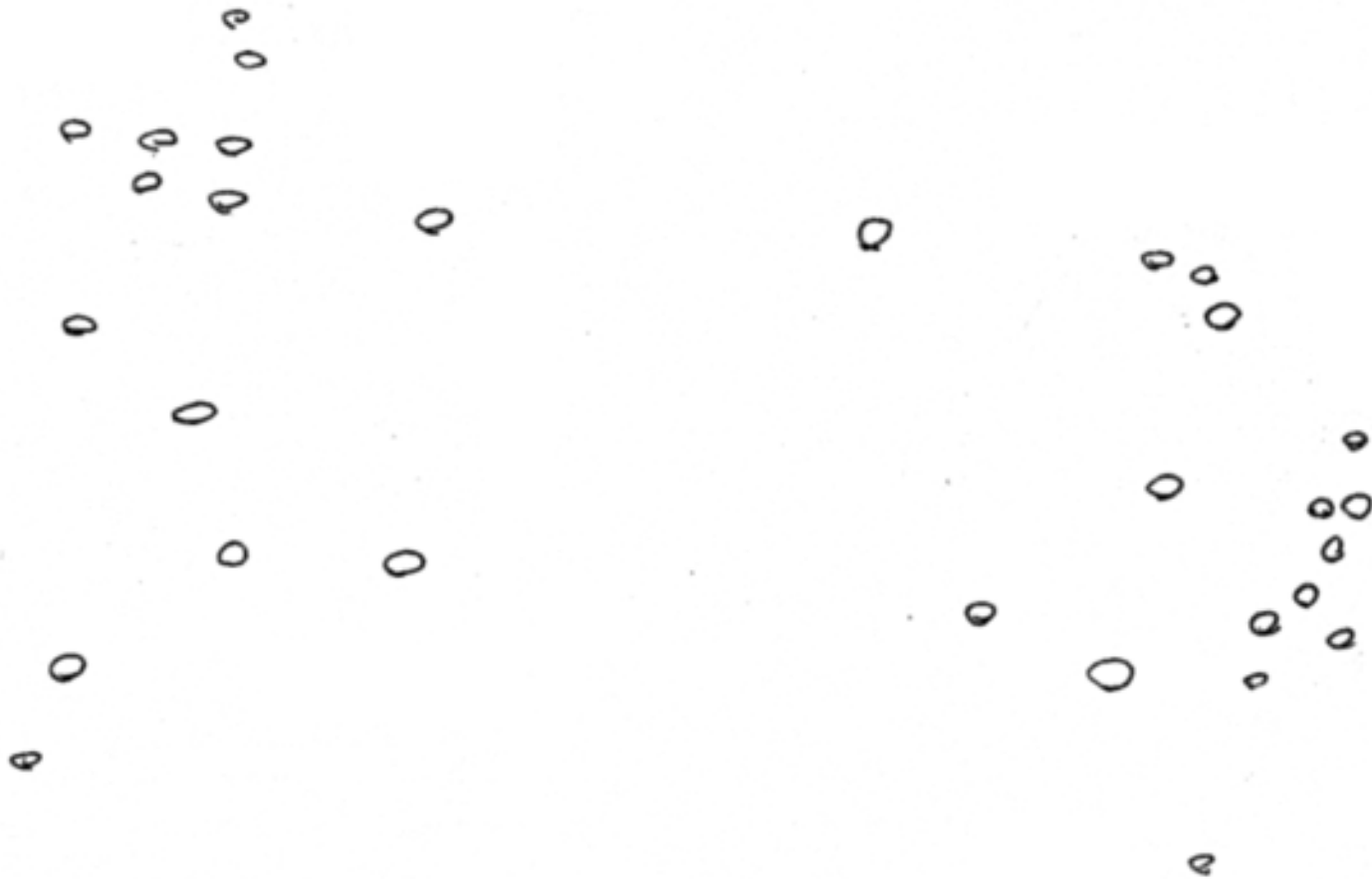
⋮ ⋮

RNA
(different
tissues)

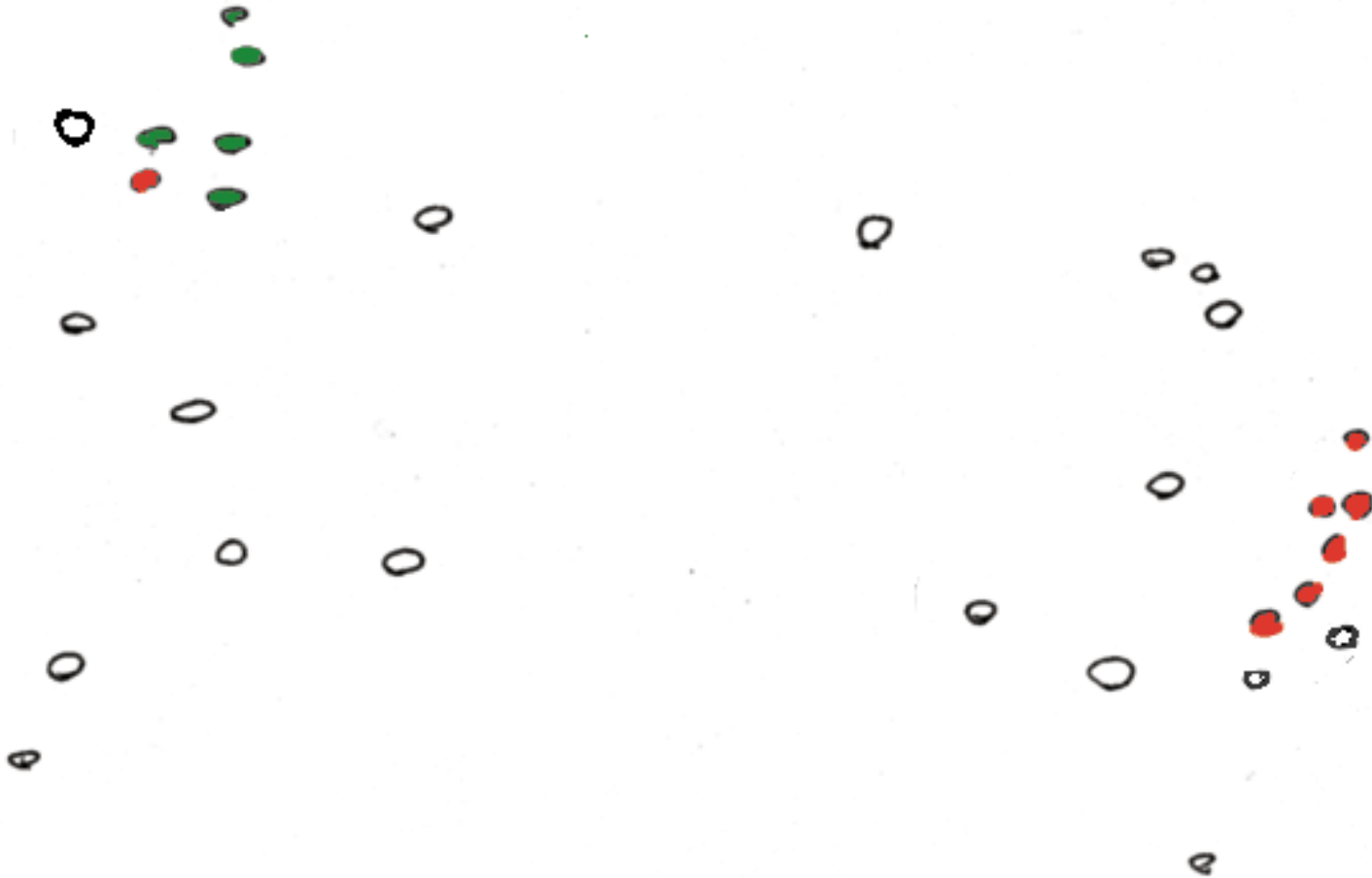


...

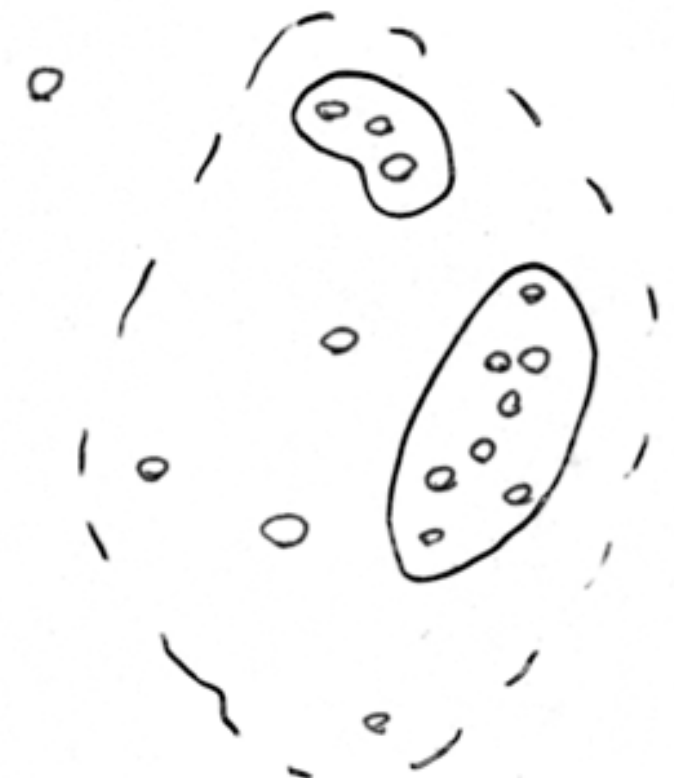
Represent predictors in abstract high dimensional space



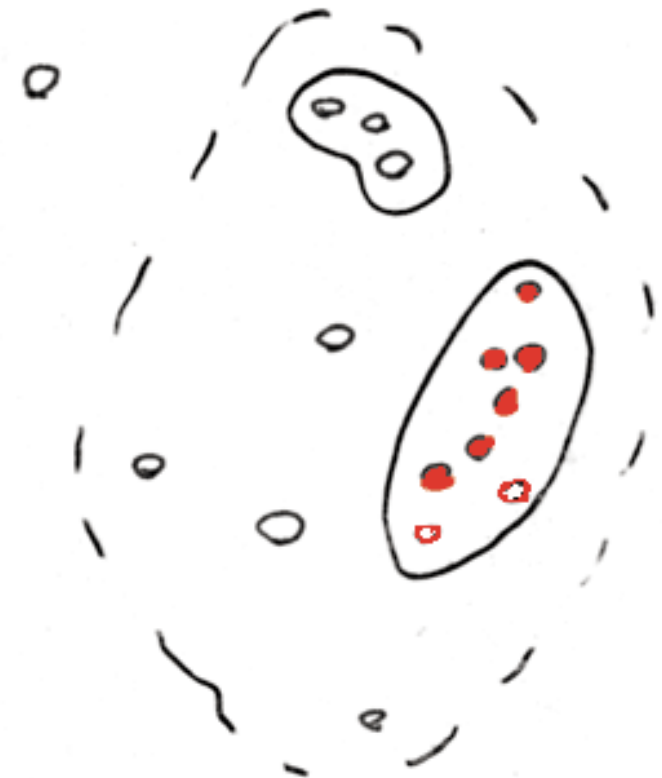
“Label” Certain Points



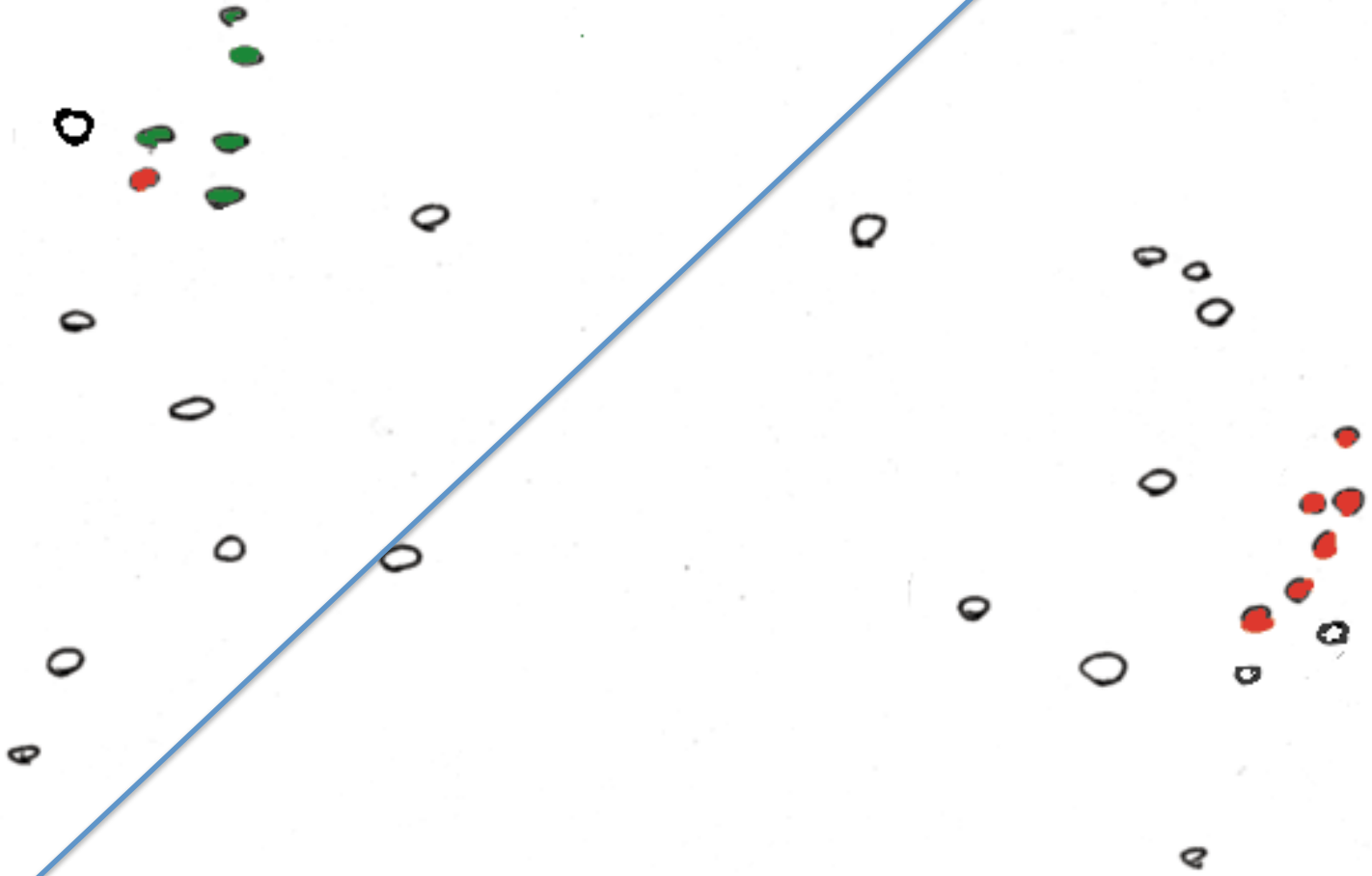
“Cluster” predictors (Unsupervised)



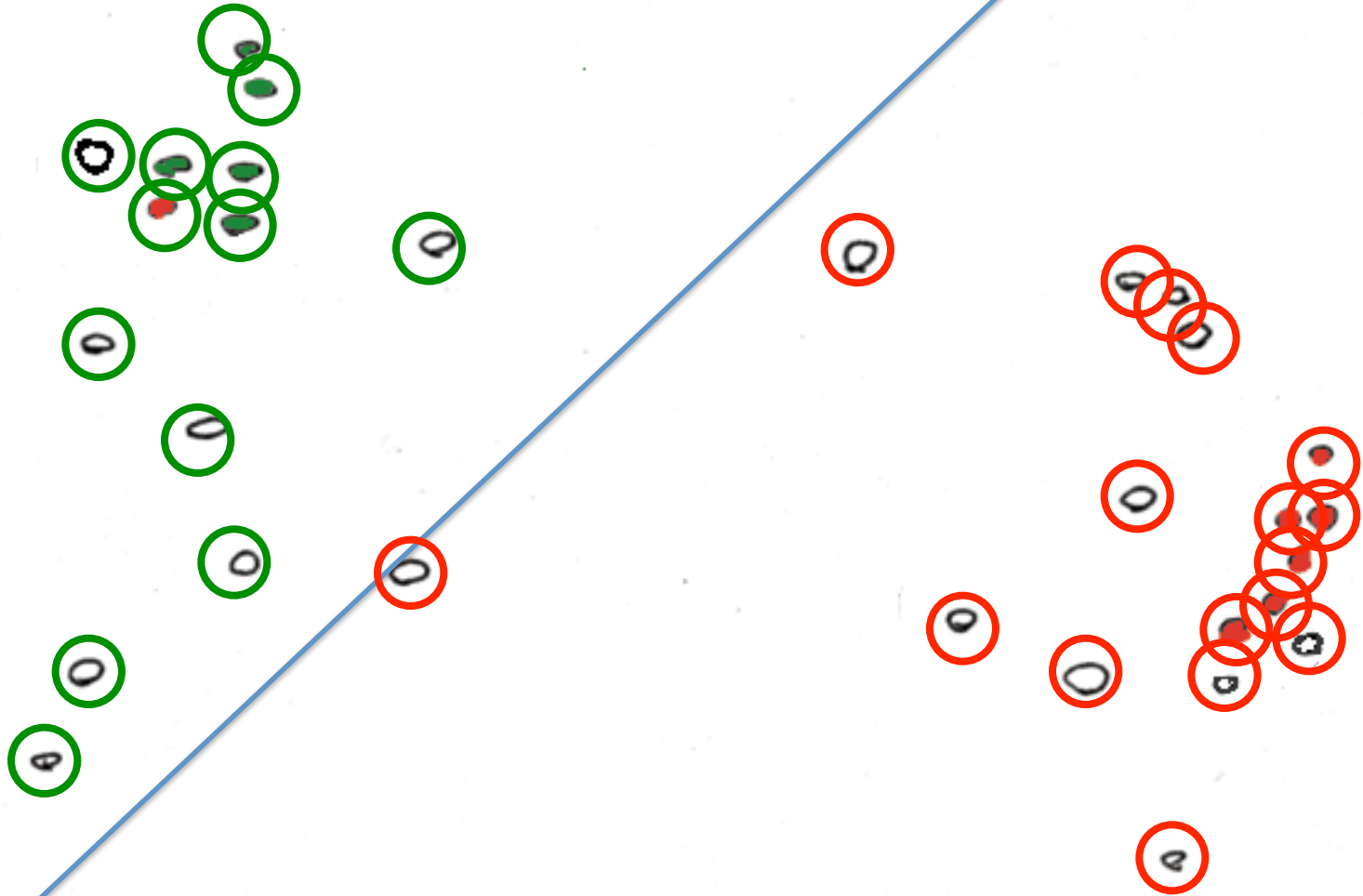
Use Clusters to predict Response (Unsupervised, guilt-by-association)



Develop Separator Based on Labeled Points (Supervised)



Predict based on Separator (Supervised)



Unsupervised Mining

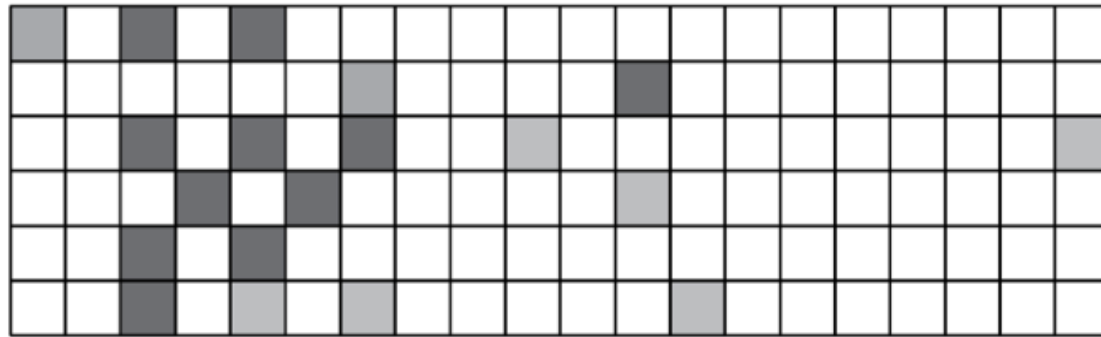
- Simple overlaps & enriched regions
- Clustering rows & columns (networks)
- PCA
- SVD (theory + appl.)
- Weighted Gene Co-Expression Network
- Biplot
- CCA

Genomic Features Matrix: Deserts & Forests

1

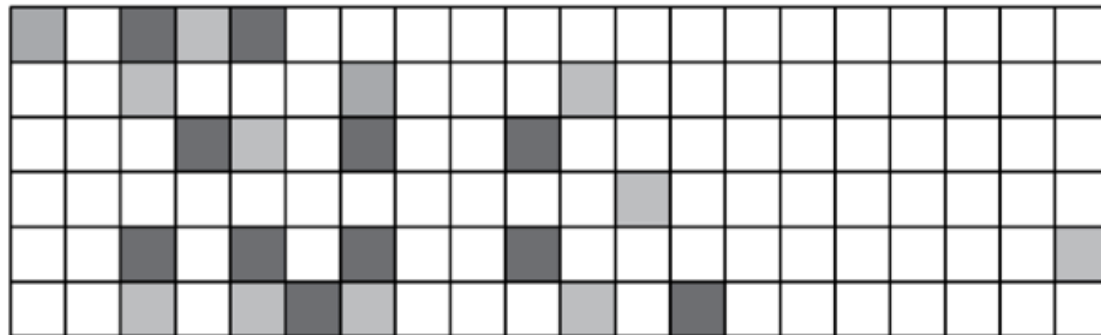
Sites along the genome

Factors
and
Chromatin
Modifications
(different
tissues)



⋮

RNA
(different
tissues)



⋮



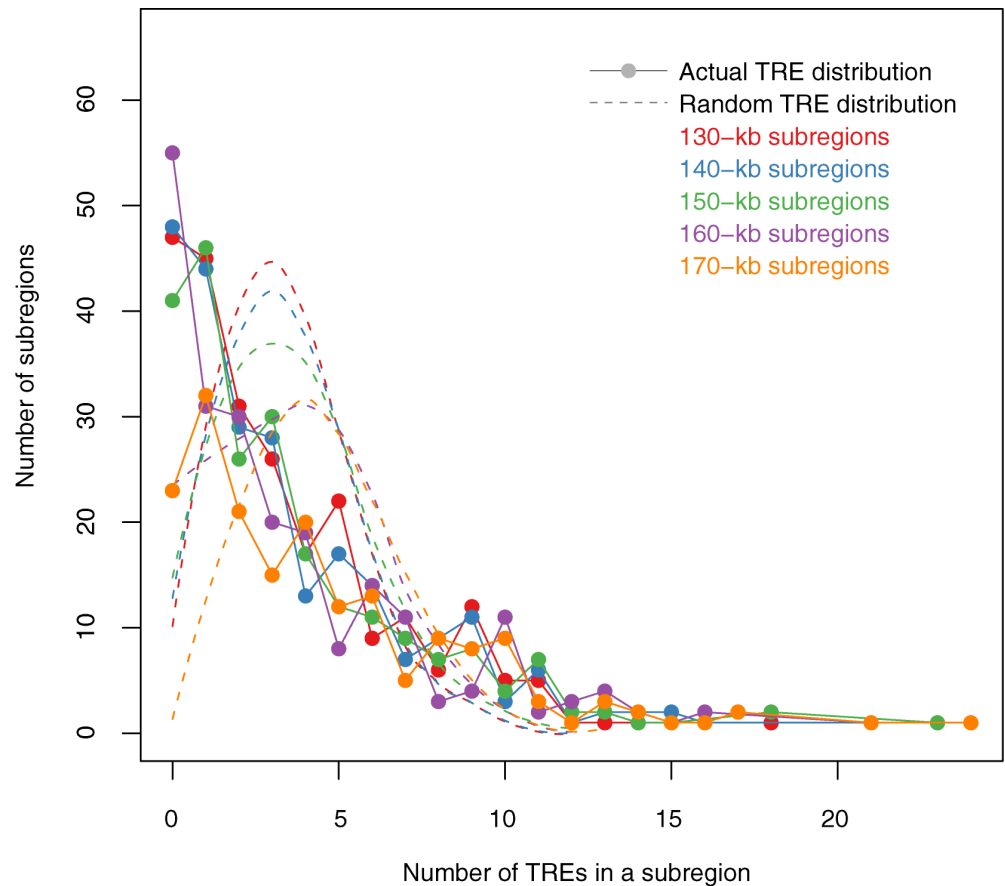
Forest



Desert

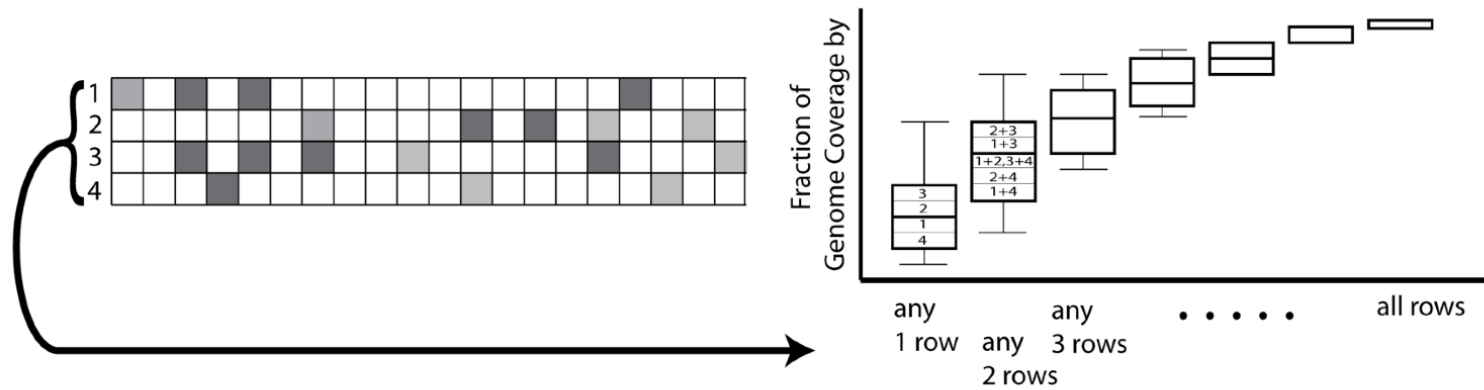
Non-random distribution of TREs

- TREs are not evenly distributed throughout the encode regions ($P < 2.2 \times 10^{-16}$).
- The actual TRE distribution is power-law.
- The null distribution is 'Poissonesque.'
- Many genomic subregions with extreme numbers of TREs.

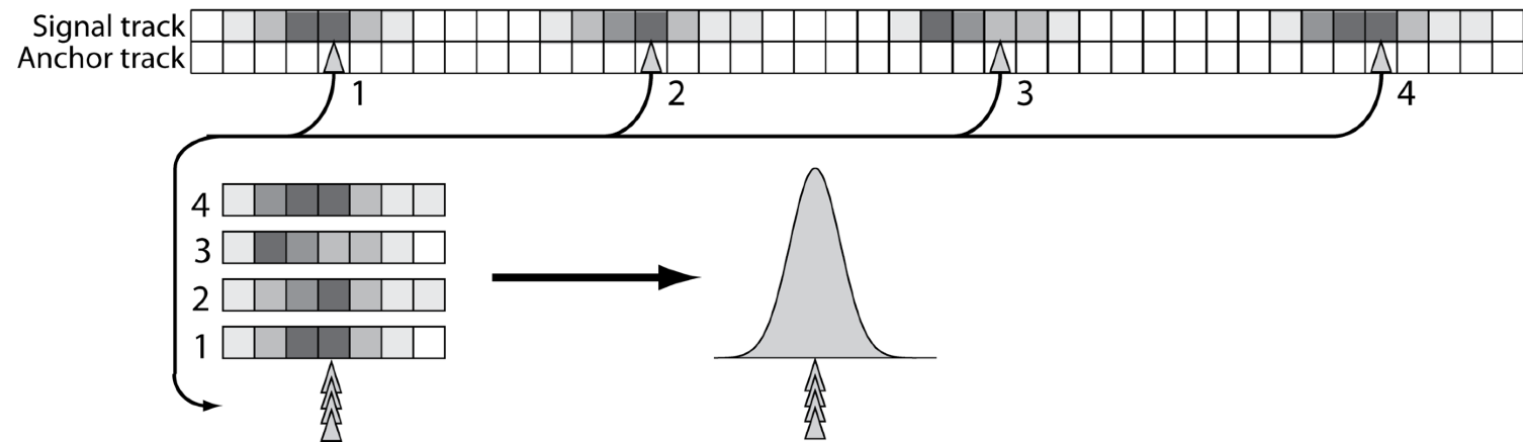


Aggregation & Saturation

B Saturation Analysis



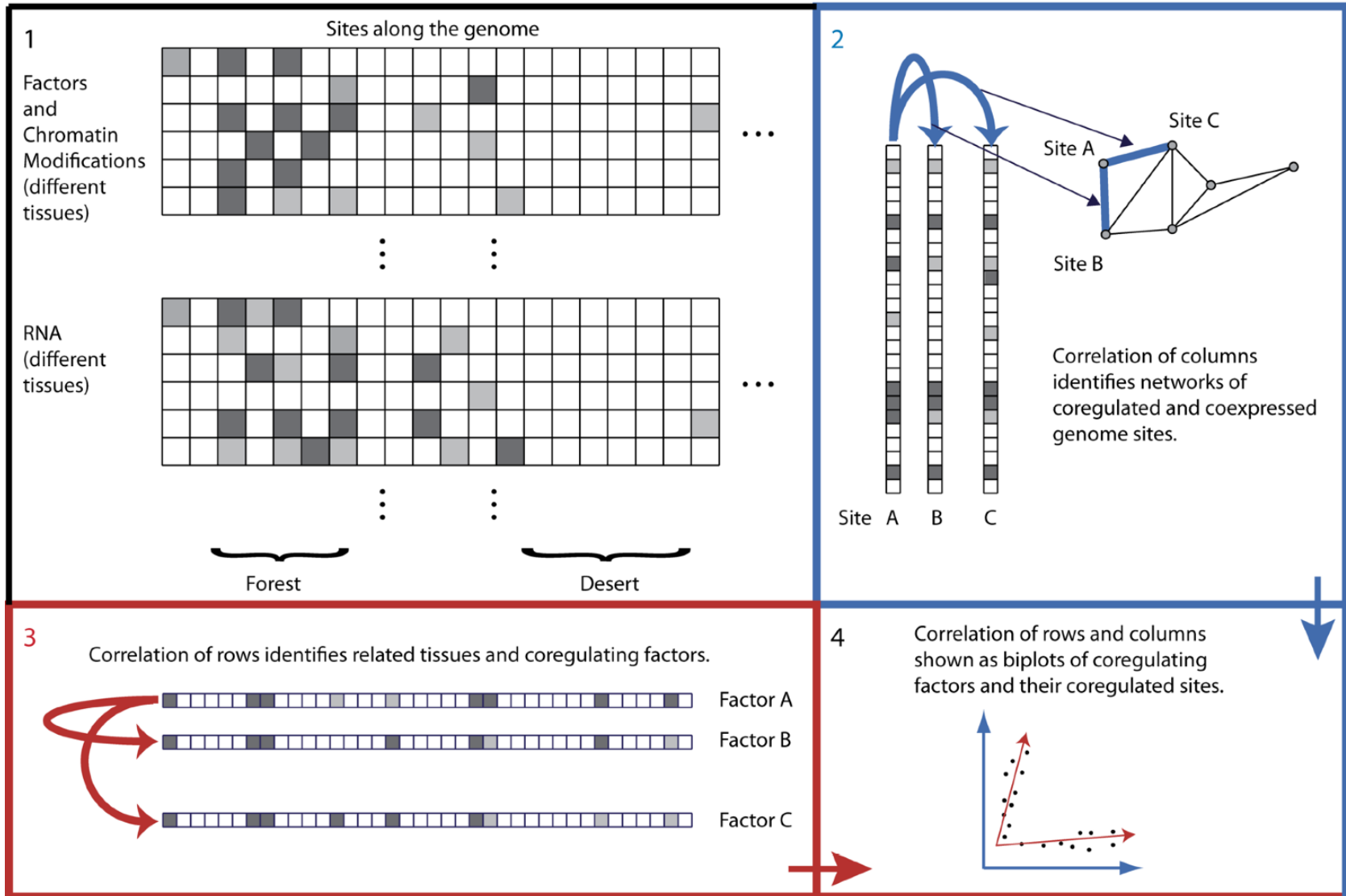
C Aggregation Analysis



Unsupervised Mining

Clustering Columns & Rows of the
Data Matrix

Correlating Rows & Columns

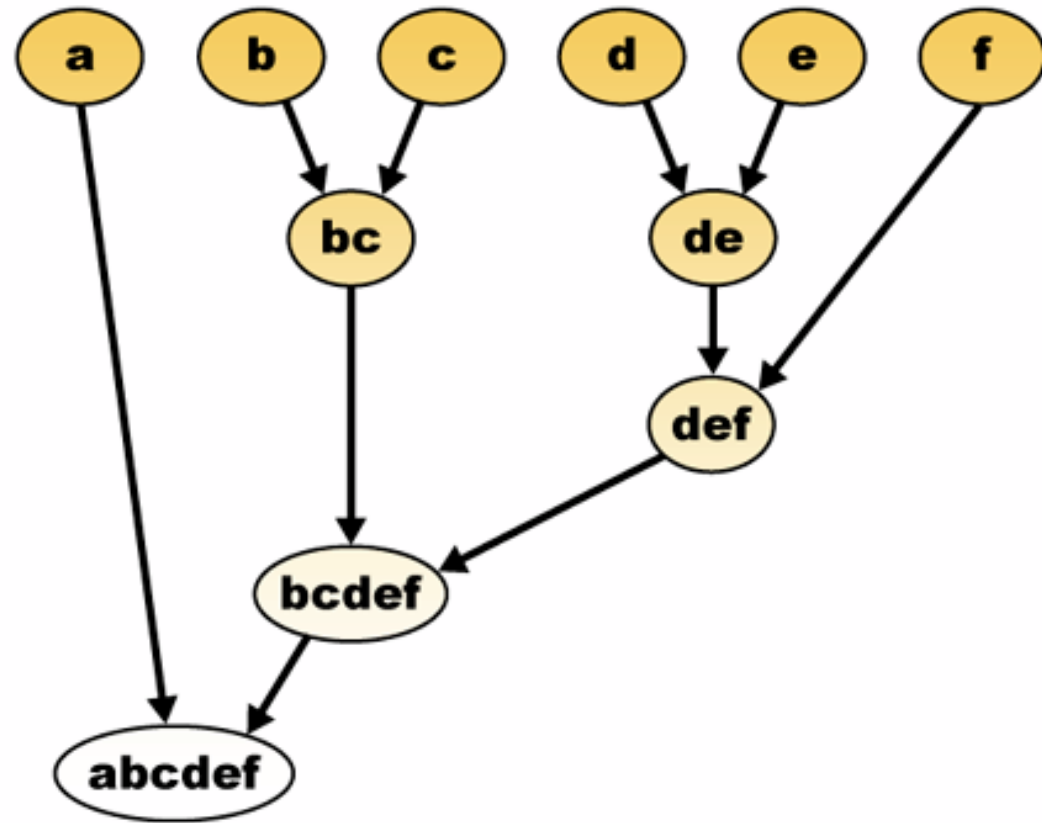


Spectral Methods Outline & Papers

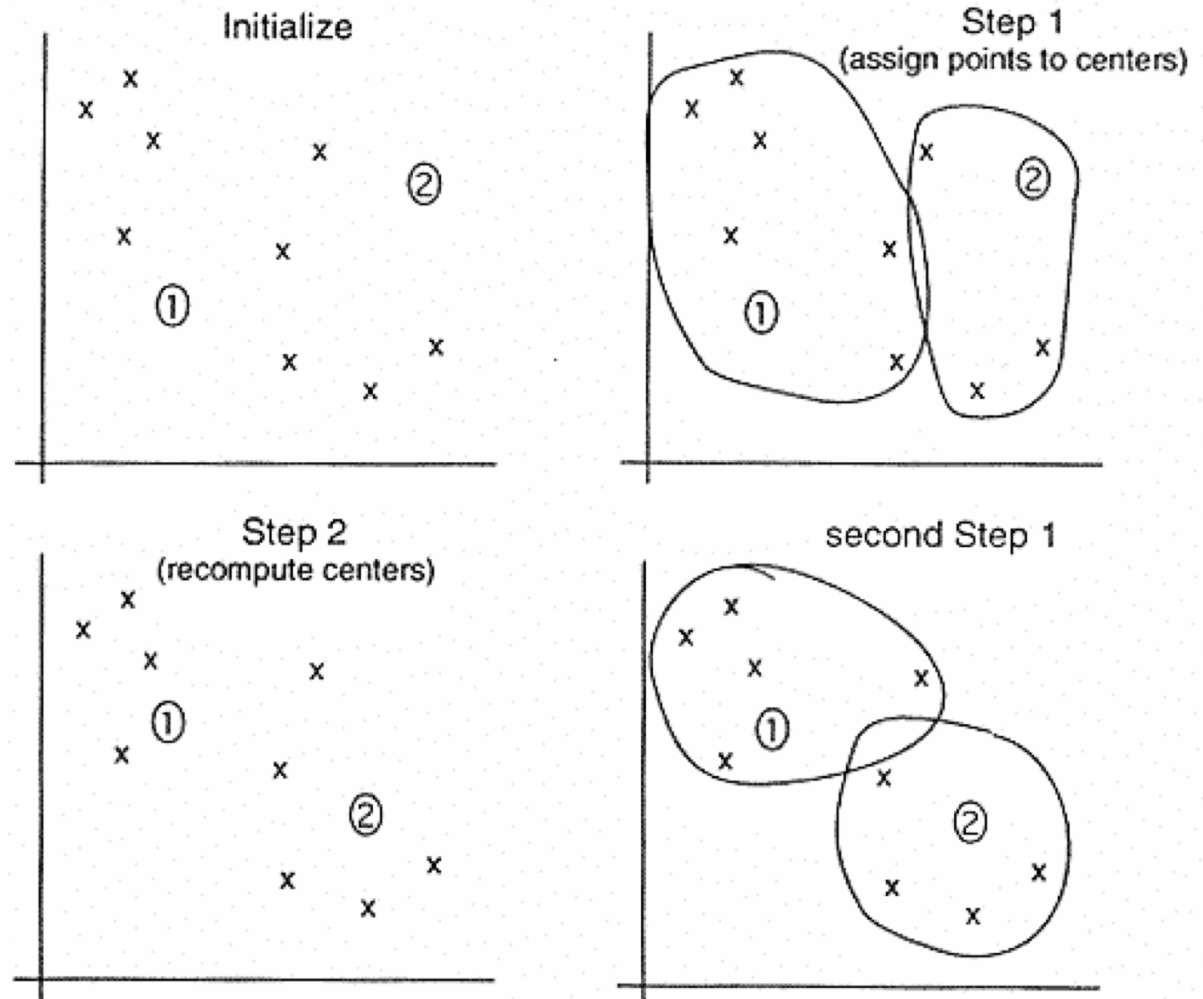
- Simple background on PCA (emphasizing lingo)
- More abstract run through on SVD
- Application to
 - O Alter et al. (2000). "Singular value decomposition for genome-wide expression data processing and modeling." PNAS 97: 10101
 - Langfelder P, Horvath S (2007) Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology 2007, 1:54
 - Z Zhang et al. (2007) "Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions." Genome Res 17: 787
 - TA Gianoulis et al. (2009) "Quantifying environmental adaptation of metabolic pathways in metagenomics." PNAS 106: 1374.

Agglomerative Clustering

- Bottom up
v top down
(K-means, know
how many
centers)
- Single or multi-
link
 - threshold for
connection?

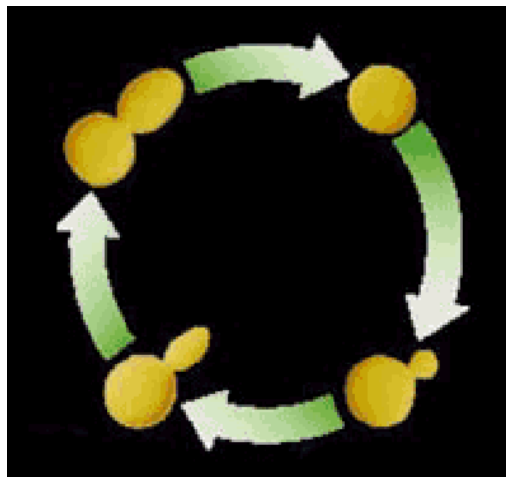


K-means

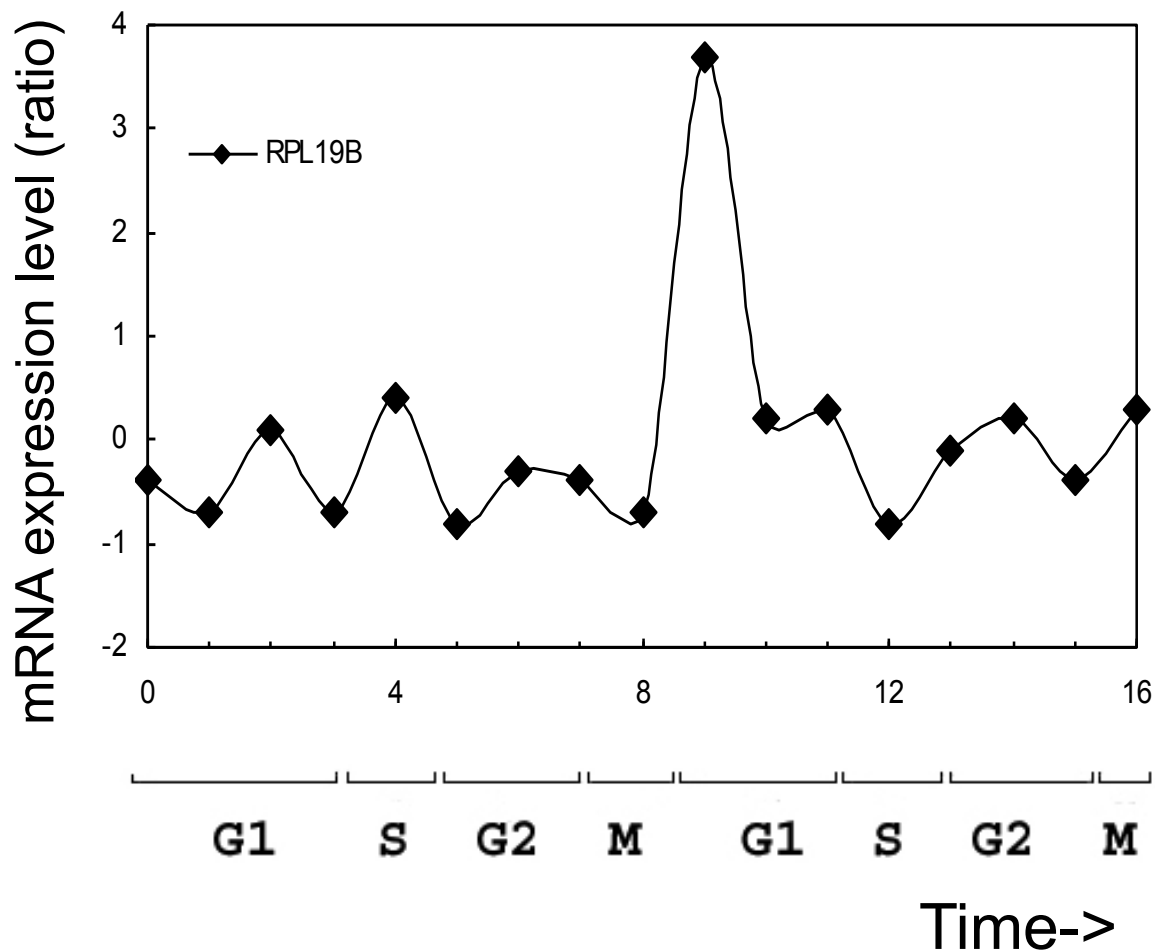


- 1) Pick ten (i.e. k ?) random points as putative cluster centers.
- 2) Group the points to be clustered by the center to which they are closest.
- 3) Then take the mean of each group and repeat, with the means now at the cluster center.
- 4) Stop when the centers stop moving.

Clustering the yeast cell cycle to uncover interacting proteins

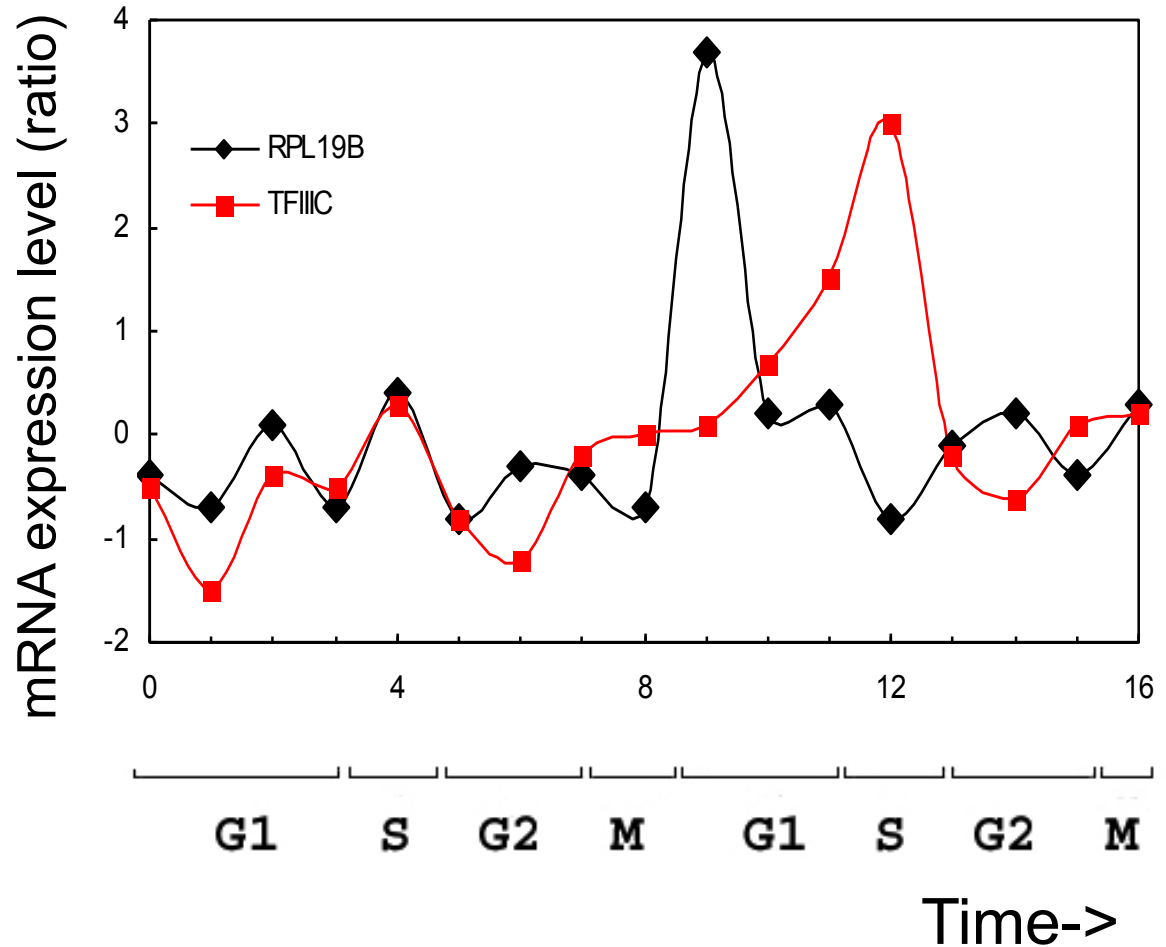
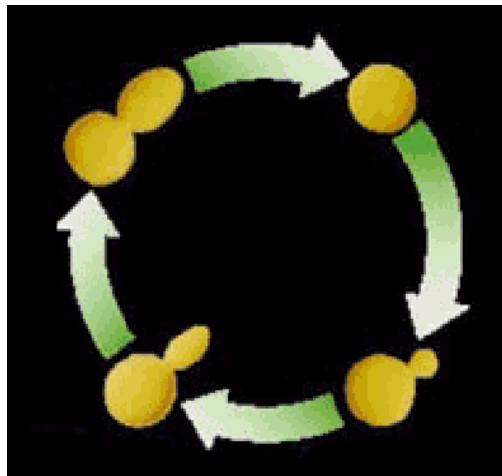


[Brown, Davis]



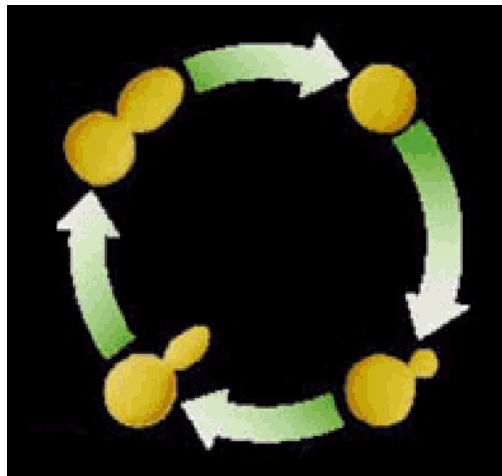
Microarray timecourse of
1 ribosomal protein

Clustering the yeast cell cycle to uncover interacting proteins

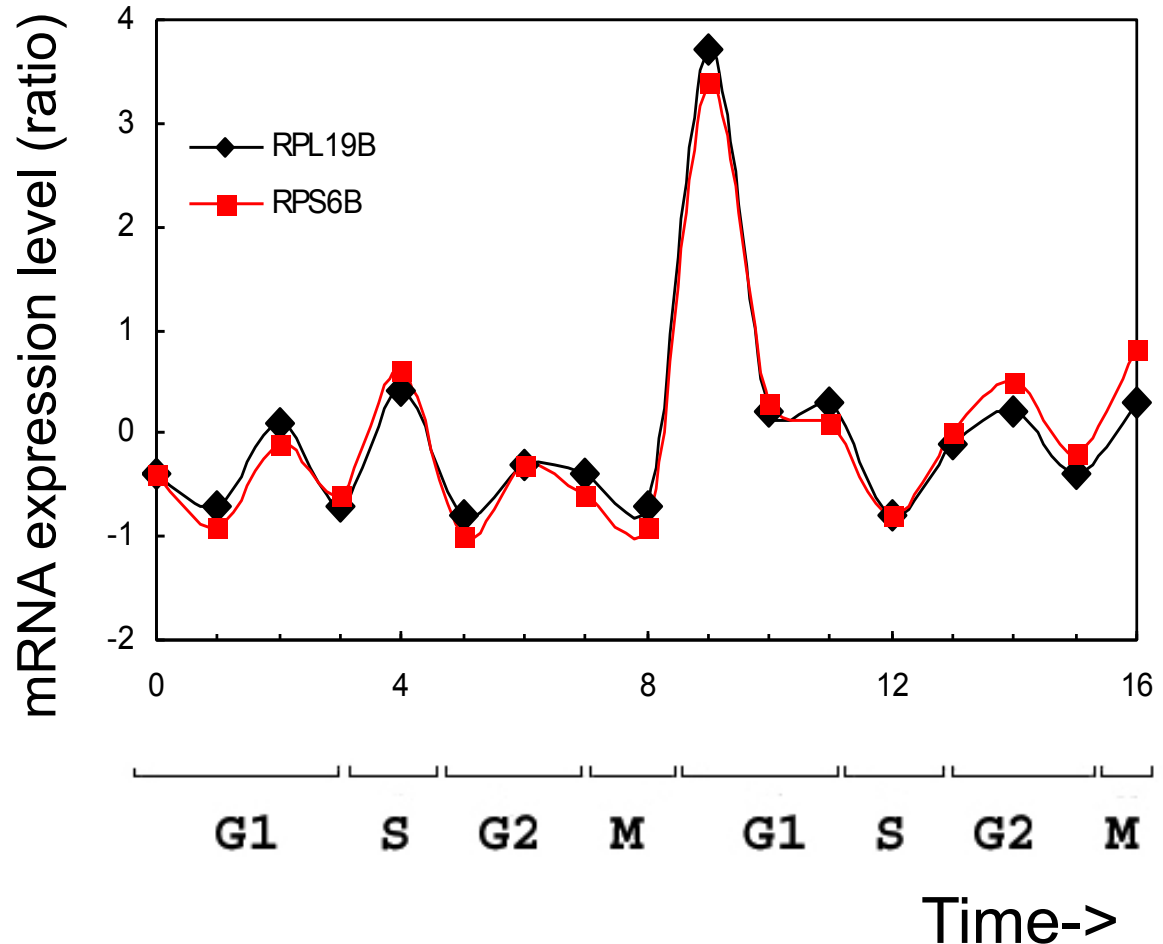


Random relationship from ~18M

Clustering the yeast cell cycle to uncover interacting proteins

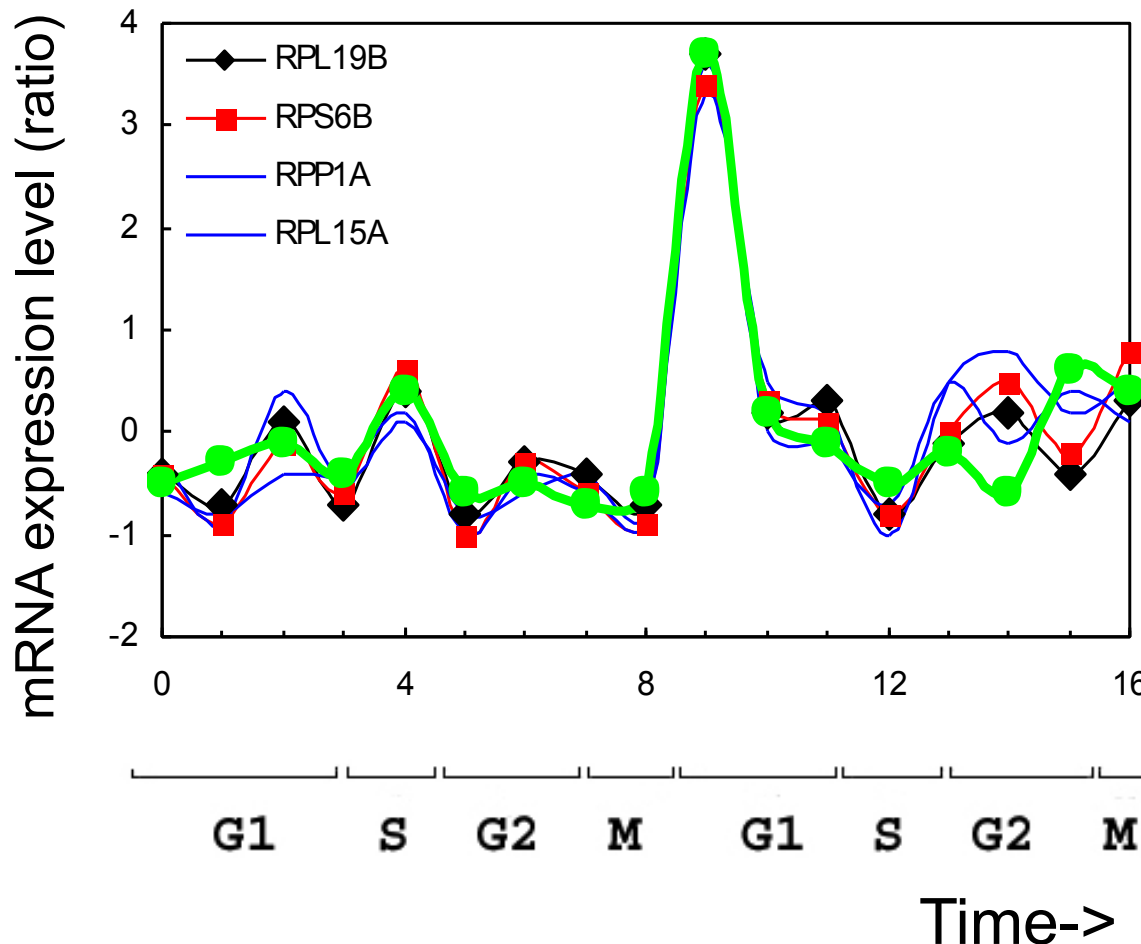
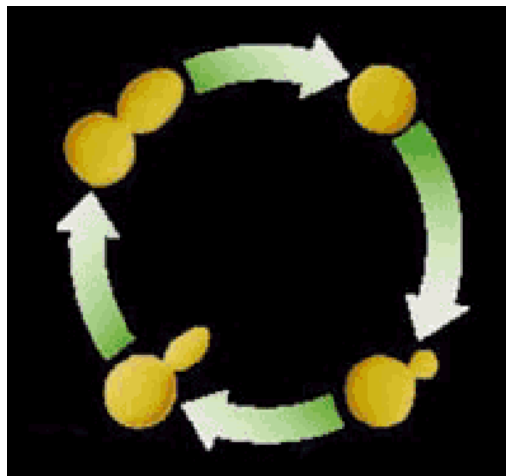


[Botstein; Church, Vidal]



Close relationship from 18M
(2 Interacting Ribosomal Proteins)

Clustering
the
yeast cell
cycle to
uncover
interacting
proteins

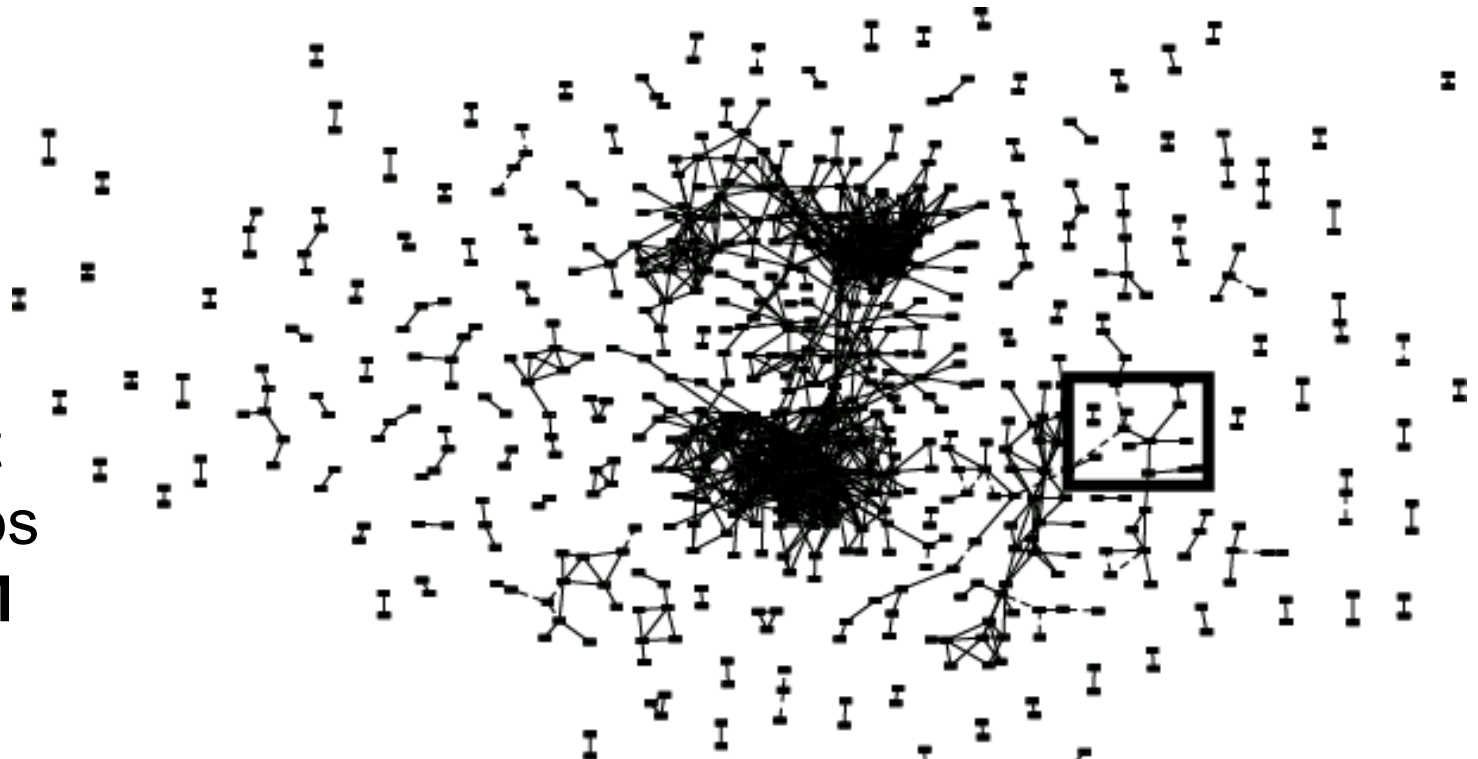


Predict Functional Interaction of
Unknown Member of Cluster



Global Network of Relationships

~470K
significant
relationships
from **~18M**
possible



Network = Adjacency Matrix

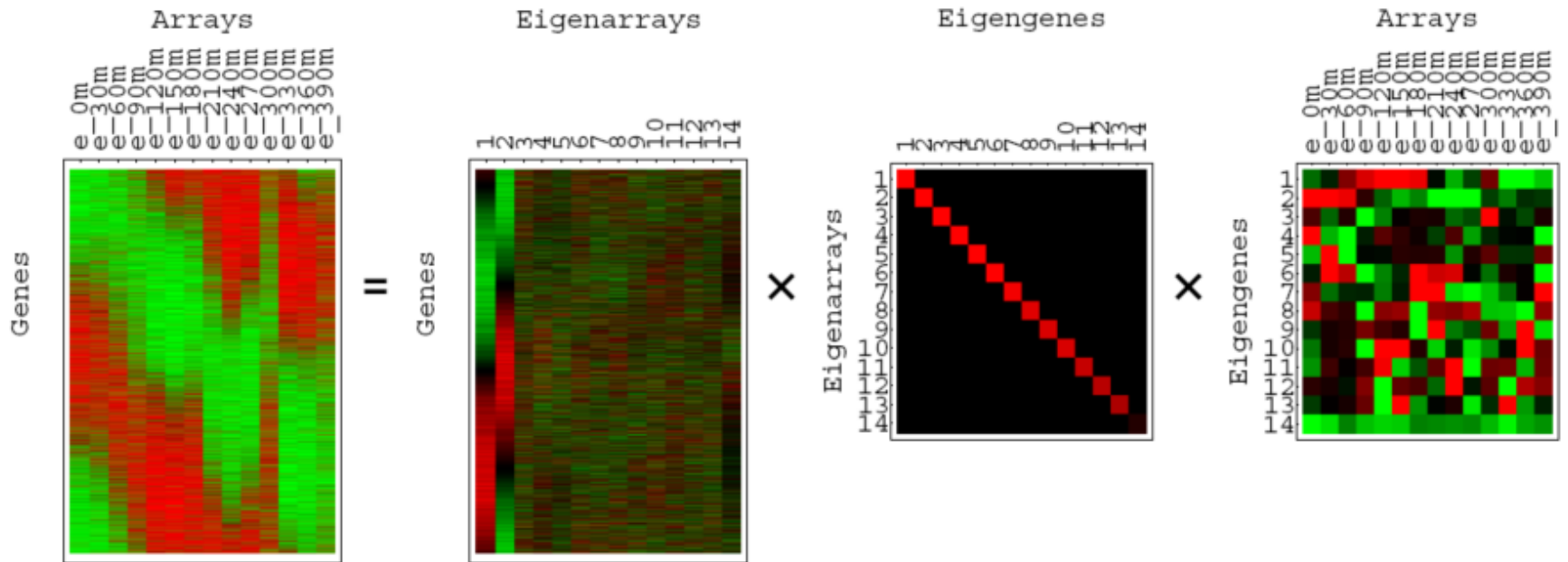
- Adjacency matrix $A=[a_{ij}]$ encodes whether/how a pair of nodes is connected.
- For unweighted networks: entries are 1 (connected) or 0 (disconnected)
- For weighted networks: adjacency matrix reports connection strength between gene pairs

Unsupervised Mining

SVD

Puts together slides prepared by
Brandon Xia with images from
Alter et al. papers

SVD for microarray data (Alter et al, PNAS 2000)



U

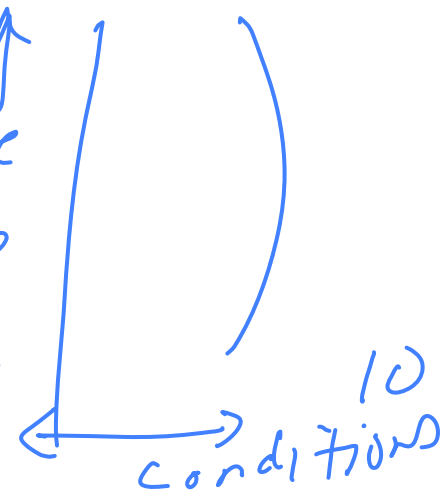
S

V^T

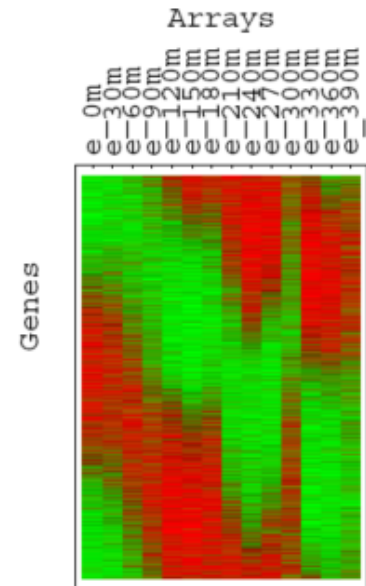
$$A = USVT$$

1000

people
genes



- A is any rectangular matrix ($m \geq n$)
- Row space: vector subspace generated by the row vectors of A
- Column space: vector subspace generated by the column vectors of A
 - The dimension of the row & column space is the rank of the matrix A: $r (\leq n)$
- A is a linear transformation that maps vector x in row space into vector Ax in column space

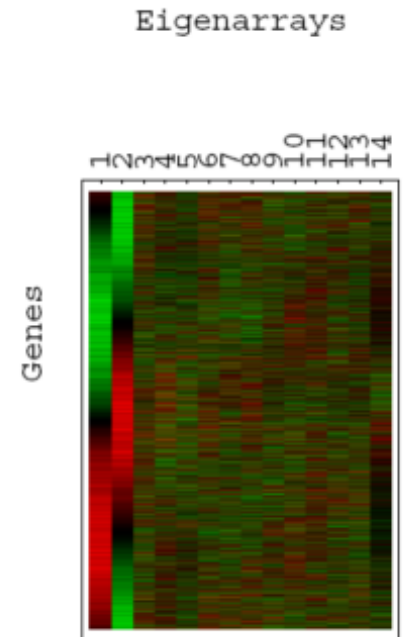


$$A = USV^T$$

- U is an “orthogonal” matrix ($m \geq n$)
- Column vectors of U form an orthonormal basis for the **column space** of A: $U^T U = I$

$$U = \begin{pmatrix} | & | & \dots & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_n \\ | & | & \dots & | \end{pmatrix}$$

- $\mathbf{u}_1, \dots, \mathbf{u}_n$ in U are eigenvectors of AA^T
 - $AA^T = USV^T V S U^T = US^2 U^T$
 - “Left singular vectors”

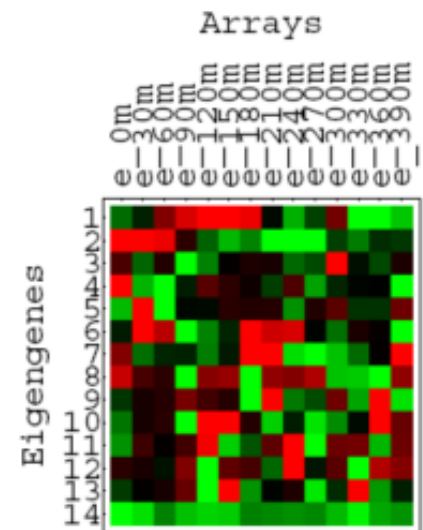


$$A = USV^T$$

- V is an orthogonal matrix (n by n)
- Column vectors of V form an orthonormal basis for the **row space** of A : $V^T V = V V^T = I$

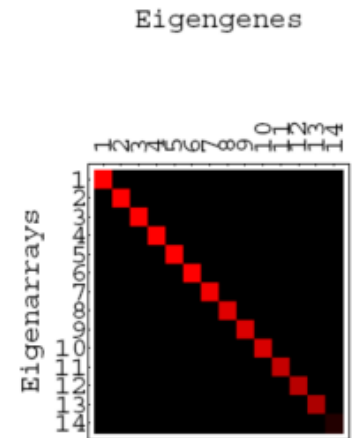
$$V = \begin{pmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \text{L} & \mathbf{v}_n \\ | & | & & | \end{pmatrix}$$

- $\mathbf{v}_1, \dots, \mathbf{v}_n$ in V are eigenvectors of $A^T A$
 - $A^T A = V S U^T U S V^T = V S^2 V^T$
 - “Right singular vectors”



$$A = USV^T$$

- S is a diagonal matrix (n by n) of non-negative singular values
- Typically sorted from largest to smallest
- Singular values are the non-negative square root of corresponding eigenvalues of $A^T A$ and AA^T



$$AV = US$$



- Means each $A\mathbf{v}_i = s_i\mathbf{u}_i$
- Remember A is a linear map from row space to column space
- Here, A maps an orthonormal basis $\{\mathbf{v}_i\}$ in row space into an orthonormal basis $\{\mathbf{u}_i\}$ in column space
- Each component of \mathbf{u}_i is the projection of a row of the data matrix A onto the vector \mathbf{v}_i

SVD of A (m by n): recap

- $A = USV^T =$ (big-"orthogonal")(diagonal)(sq-orthogonal)
- $\mathbf{u}_1, \dots, \mathbf{u}_m$ in U are eigenvectors of AA^T
- $\mathbf{v}_1, \dots, \mathbf{v}_n$ in V are eigenvectors of $A^T A$
- s_1, \dots, s_n in S are nonnegative singular values of A
- $AV = US$ means each $A\mathbf{v}_i = s_i\mathbf{u}_i$
- “Every A is diagonalized by 2 orthogonal matrices”

SVD as sum of rank-1 matrices

- $A = USV^T$
- $A = s_1 \mathbf{u}_1 \mathbf{v}_1^T + s_2 \mathbf{u}_2 \mathbf{v}_2^T + \dots + s_n \mathbf{u}_n \mathbf{v}_n^T$
- $s_1 \geq s_2 \geq \dots \geq s_n \geq 0$

an outer product
(uv^T) giving a
matrix rather than
the scalar of the
inner product

- What is the rank- r matrix \hat{A} that best approximates A ?

– Minimize
$$\sum_{i=1}^m \sum_{j=1}^n (\hat{A}_{ij} - A_{ij})^2$$

LSQ approx. If $r=1$,
this amounts to a
line fit.

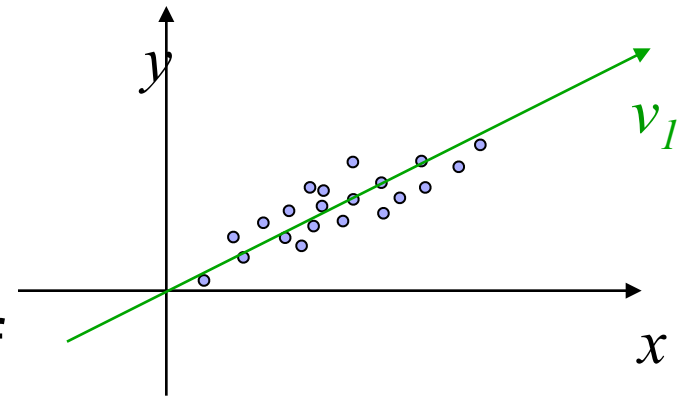
- $\hat{A} = s_1 \mathbf{u}_1 \mathbf{v}_1^T + s_2 \mathbf{u}_2 \mathbf{v}_2^T + \dots + s_r \mathbf{u}_r \mathbf{v}_r^T$
- Very useful for matrix approximation

Examples of (almost) rank-1 matrices

- Steady states with fluctuations $\begin{pmatrix} 101 & 103 & 102 \\ 302 & 300 & 301 \\ 203 & 204 & 203 \\ 401 & 402 & 404 \end{pmatrix}$
- Array artifacts? $\begin{pmatrix} 101 & 303 & 202 \\ 102 & 300 & 201 \\ 103 & 304 & 203 \\ 101 & 302 & 204 \end{pmatrix}$
- Signals? $\begin{pmatrix} 1 & 2 & -1 \\ 2 & 4 & -2 \\ -1 & -2 & 1 \\ 0 & 0 & 0 \end{pmatrix}$

Geometry of SVD in row space

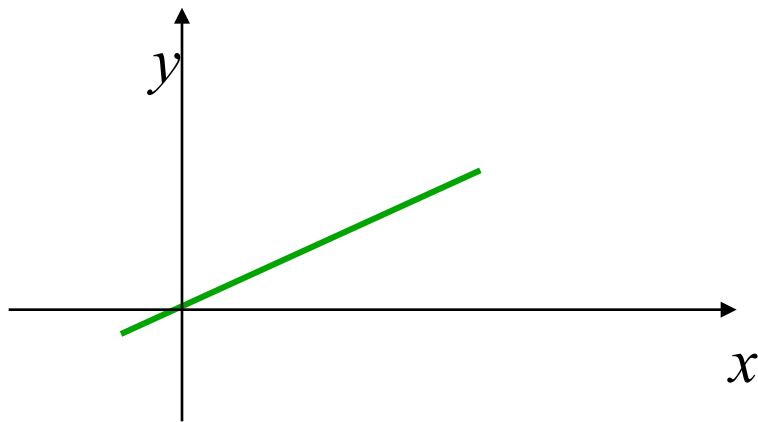
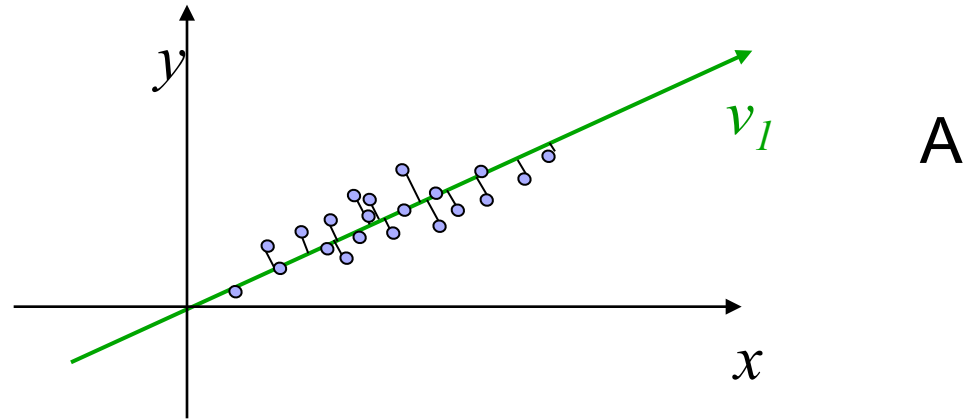
- A as a collection of m row vectors (points) in the row space of A
- $s_1 \mathbf{u}_1 \mathbf{v}_1^T$ is the best rank-1 matrix approximation for A
- Geometrically: \mathbf{v}_1 is the direction of the best approximating rank-1 subspace that goes through origin
- $s_1 \mathbf{u}_1$ gives coordinates for row vectors in rank-1 subspace
- \mathbf{v}_1 Gives coordinates for row space basis vectors in rank-1 subspace



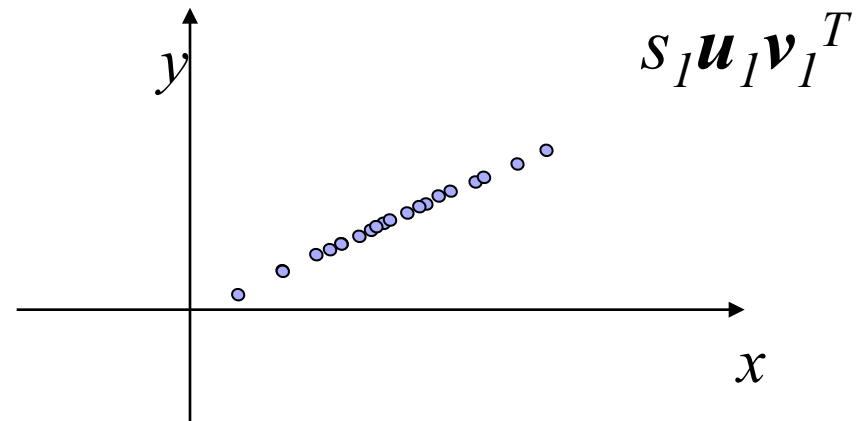
$$A \mathbf{v}_i = s_i \mathbf{u}_i$$

$$I \mathbf{v}_i = \mathbf{v}_i$$

Geometry of SVD in row space



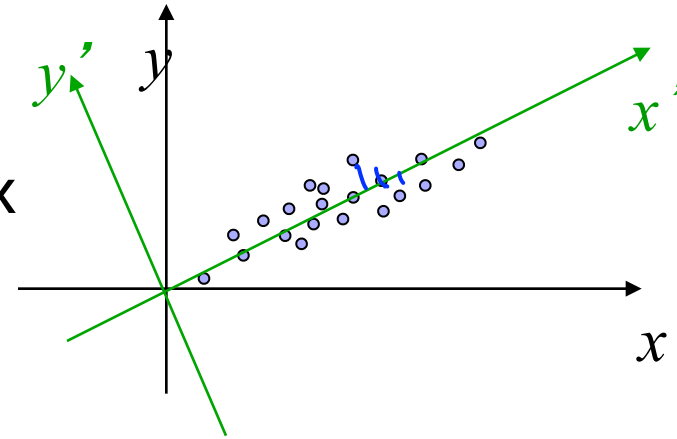
This line segment that goes through origin approximates the original data set



The projected data set approximates the original data set

Geometry of SVD in row space

- A as a collection of m row vectors (points) in the row space of A
- $s_1 \mathbf{u}_1 \mathbf{v}_1^T + s_2 \mathbf{u}_2 \mathbf{v}_2^T$ is the best rank-2 matrix approximation for A
- Geometrically: \mathbf{v}_1 and \mathbf{v}_2 are the directions of the best approximating rank-2 subspace that goes through origin
- $s_1 \mathbf{u}_1$ and $s_2 \mathbf{u}_2$ gives coordinates for row vectors in rank-2 subspace
- \mathbf{v}_1 and \mathbf{v}_2 gives coordinates for row space basis vectors in rank-2 subspace



$$A \mathbf{v}_i = s_i \mathbf{u}_i$$

$$I \mathbf{v}_i = \mathbf{v}_i$$

What about geometry of SVD in column space?

- $A = USV^T$
- $A^T = VSU^T$
- The column space of A becomes the row space of A^T
- The same as before, except that U and V are switched

Geometry of SVD in row and column spaces

- Row space
 - $s_i \mathbf{u}_i$ gives coordinates for row vectors along unit vector \mathbf{v}_i
 - \mathbf{v}_i gives coordinates for row space basis vectors along unit vector \mathbf{v}_i
- Column space
 - $s_i \mathbf{v}_i$ gives coordinates for column vectors along unit vector \mathbf{u}_i
 - \mathbf{u}_i gives coordinates for column space basis vectors along unit vector \mathbf{u}_i
- Along the directions \mathbf{v}_i and \mathbf{u}_i , these two spaces look pretty much the same!
 - Up to scale factors s_i
 - Switch row/column vectors and row/column space basis vectors
 - **Biplot....**

$$A \mathbf{v}_i = s_i \mathbf{u}_i$$

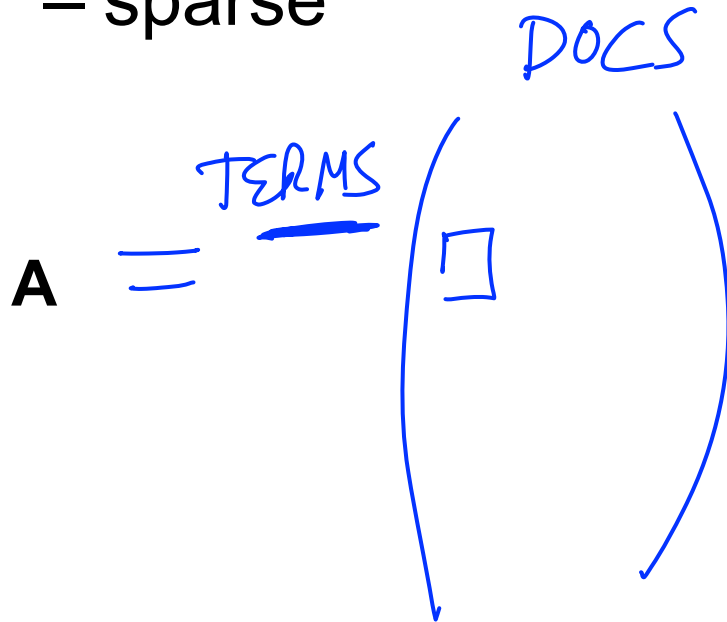
$$I \mathbf{v}_i = \mathbf{v}_i$$

$$A^T \mathbf{u}_i = s_i \mathbf{v}_i$$

$$I \mathbf{u}_i = \mathbf{u}_i$$

Additional Points

- Time Complexity (Cubic)
- Application to text mining
 - Latent semantic indexing
 - sparse



Potential problems of SVD/PCA

If the dataset...

- Lacks Independence
 - **NO PROBLEM**
- Lacks Normality
 - Normality desirable but not essential
- Lacks Precision
 - Precision desirable but not essential
- Lacks Linearity
 - **Problem:** Use other non-linear (kernel) methods
- Many Zeroes in Data Matrix (Sparse)
 - **Problem:** Use Correspondence Analysis

Conclusion

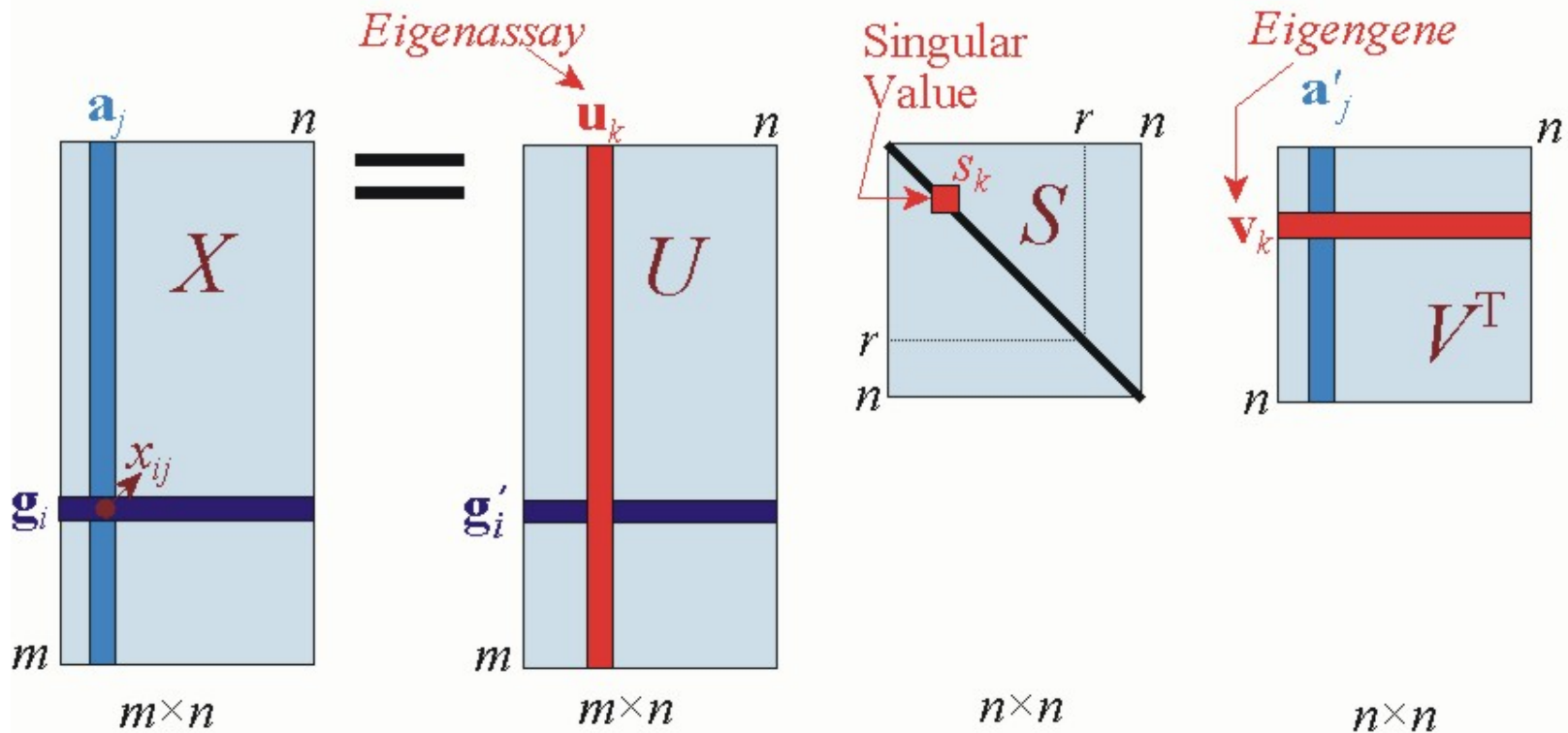
- SVD is the “absolute high point of linear algebra”
- SVD is difficult to compute; but once we have it, we have many things
- SVD finds the best approximating subspace, using **linear transformation**
- Simple SVD cannot handle translation, non-linear transformation, separation of labeled data, etc.
- Good for exploratory analysis; but once we know what we look for, use appropriate tools and model the structure of data explicitly!

Unsupervised Mining

Intuition on interpretation of SVD
in terms of genes and conditions

SVD for microarray data (Alter et al, PNAS 2000)

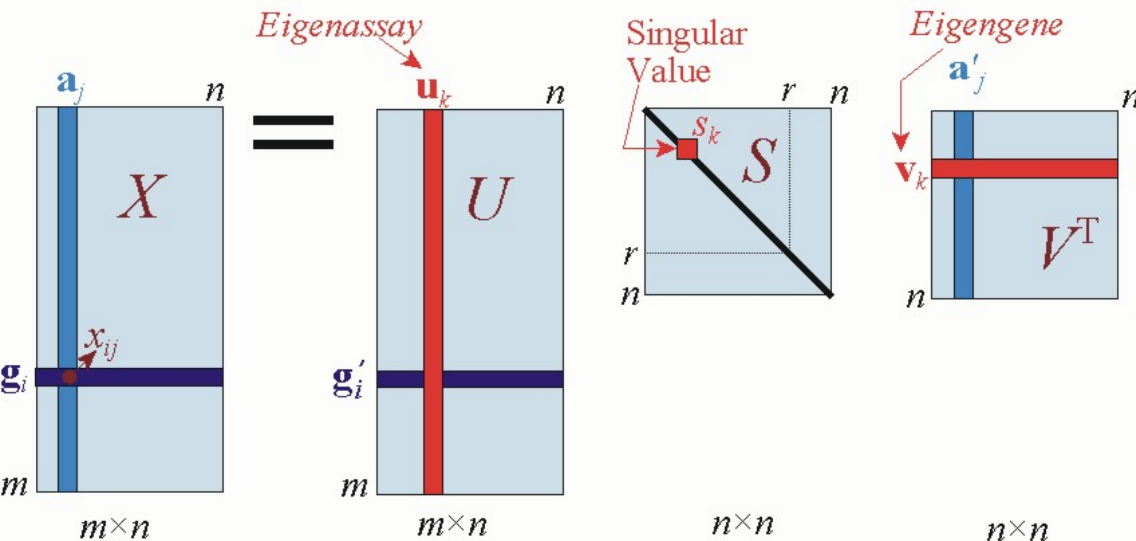
$$X = USV^T$$



Notation

- $m=1000$ genes
 - row-vectors
 - 10 eigengene (v_i) of dimension 10 conditions
- $n=10$ conditions (assays)
 - column vectors
 - 10 eigenconditions (u_i) of dimension 1000 genes

$$X = USV^T$$



Close up on Eigengenes

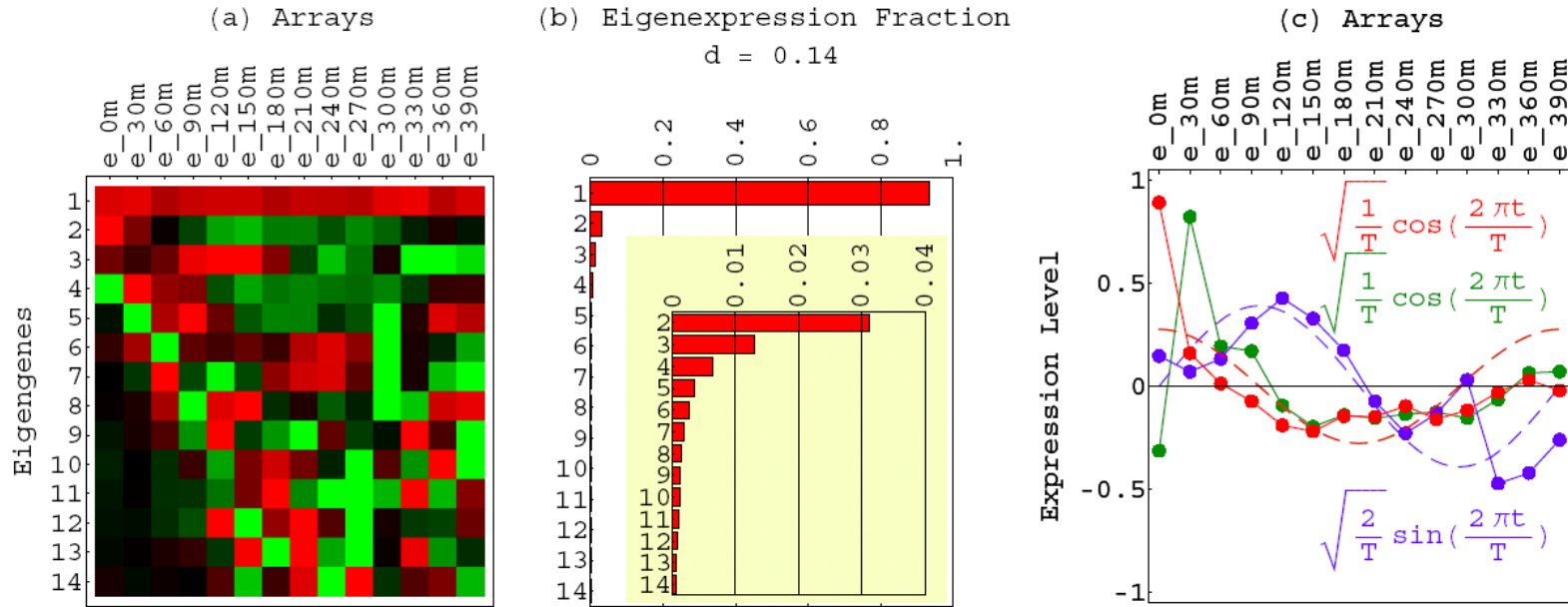
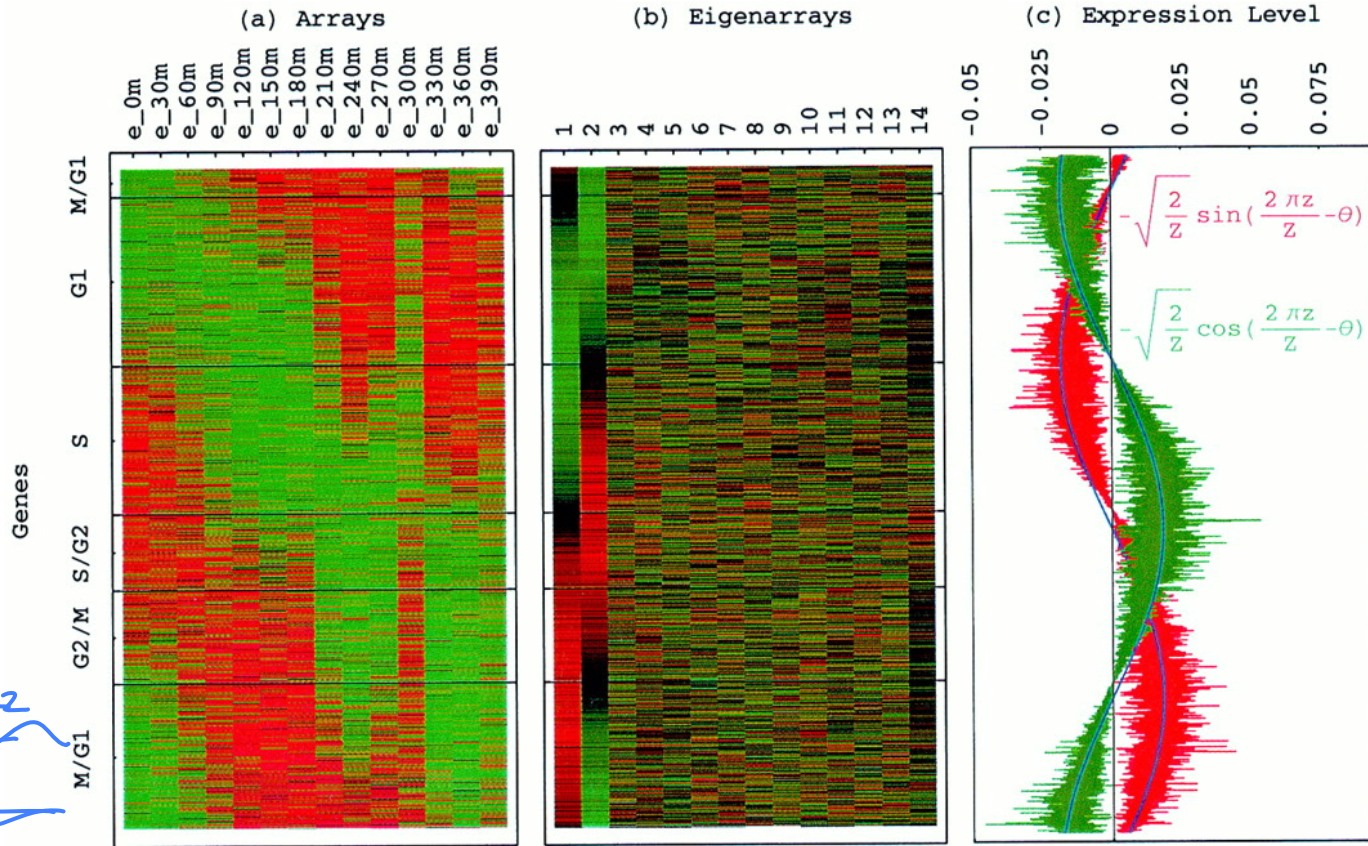


Fig. 8. Elutriation eigengenes. (a) Raster display of \hat{v}^T , the expression of 14 eigengenes in 14 arrays, with overexpression (red), no change in expression (black), and underexpression (green) around the steady state, which can be associated with the first eigengene, $|\gamma_1\rangle$. (b) Bar chart of the fraction of eigenexpression p_l of each eigengene $|\gamma_l\rangle$, showing more than 90% of the overall relative expression in $|\gamma_1\rangle$, about 3%, 1.5%, and 0.5% in $|\gamma_2\rangle$, $|\gamma_3\rangle$, and $|\gamma_4\rangle$, respectively, and a low entropy $d = 0.14 \ll 1$. (c) Line-jointed graphs of the expression levels of $|\gamma_2\rangle$ (red), $|\gamma_3\rangle$ (blue), and $|\gamma_4\rangle$ (green) in the 14 arrays, and dashed graphs of normalized cosine (blue) and sine (red) of period T .

Genes sorted by correlation with top 2 eigengenes



Alter, Orly et al. (2000) Proc. Natl. Acad. Sci. USA 97, 10101-10106

Fig. 3. Genes sorted by relative correlation with $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ of normalized elutriation. (a) Normalized elutriation expression of the sorted 5,981 genes in the 14 arrays, showing traveling wave of expression. (b) Eigenarrays expression; the expression of $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$, the eigenarrays corresponding to $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$, displays the sorting. (c) Expression levels of $|\alpha_1\rangle_N$ (red) and $|\alpha_2\rangle_N$ (green) fit normalized sine and cosine functions of period $Z \equiv N - 1 = 5,980$ and phase $\theta \approx 2\pi/13$ (blue), respectively.

Normalized elutriation expression in the subspace associated with the cell cycle

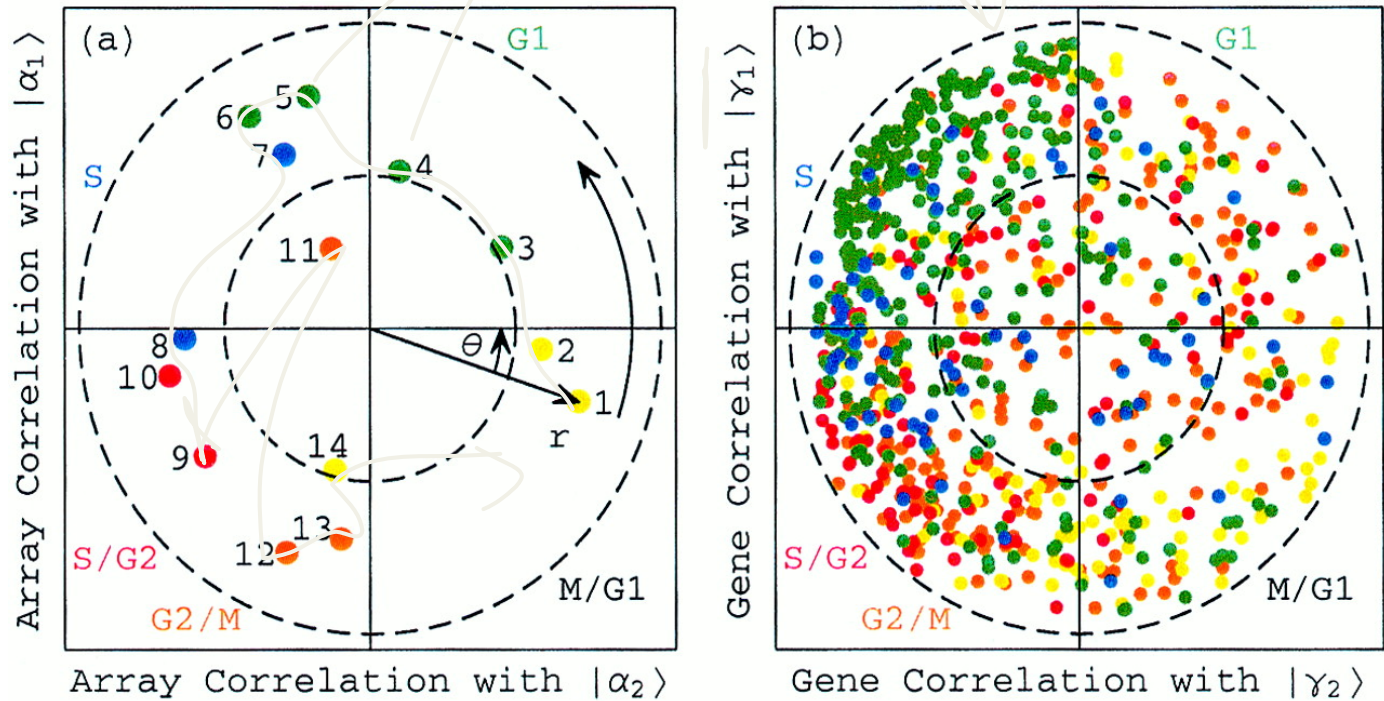
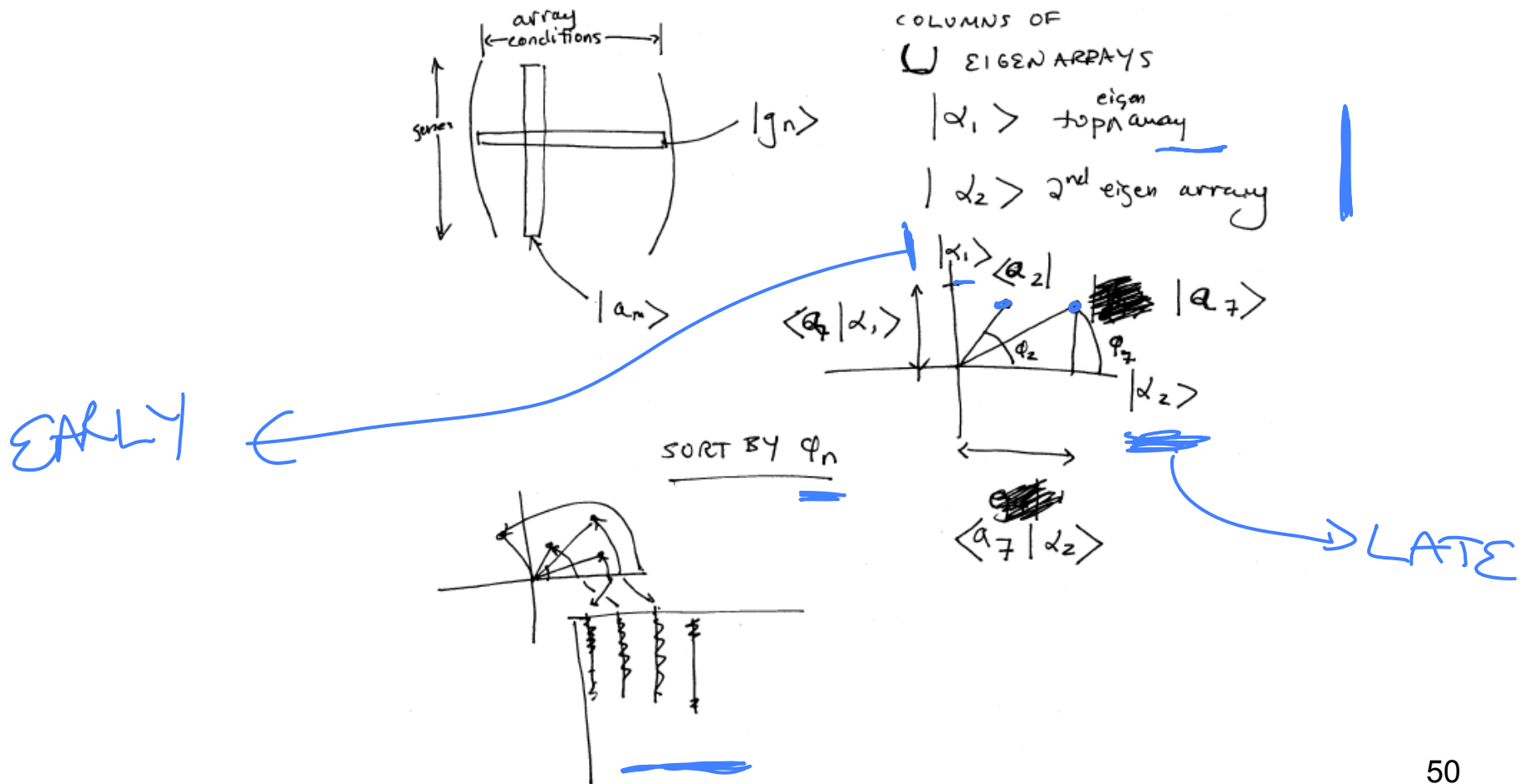


Fig. 2. Normalized elutriation expression in the subspace associated with the cell cycle. (a) Array correlation with $|\alpha_1\rangle$ along the y -axis vs. that with $|\alpha_2\rangle$ along the x -axis, color-coded according to the classification of the arrays into the five cell cycle stages, M/G1 (yellow), G1 (green), S (blue), S/G2 (red), and G2/M (orange). The dashed unit and half-unit circles outline 100% and 25% of overall normalized array expression in the $|\alpha_1\rangle$ and $|\alpha_2\rangle$ subspace. (b) Correlation of each gene with $|\gamma_1\rangle$ vs. that with $|\gamma_2\rangle$, for 784 cell cycle regulated genes, color-coded according to the classification by Spellman et al. (3).

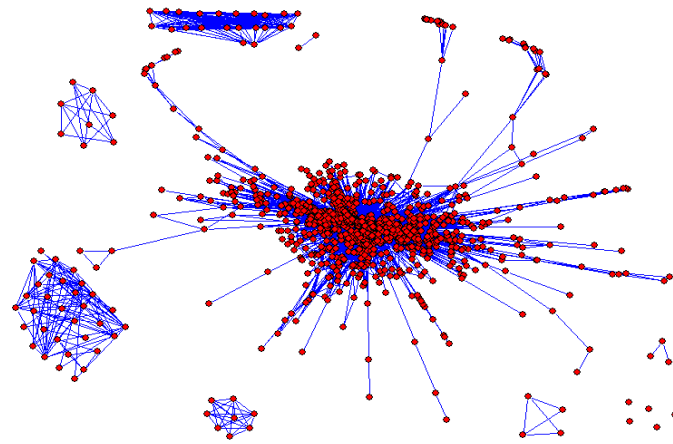
Alter, Orly et al. (2000) Proc. Natl. Acad. Sci. USA 97, 10101-10106

Plotting Experiments in Low Dimension Subspace



Unsupervised Mining

Weighted Gene Co-Expression
Network



Weighted Gene Co-Expression Network Analysis

Bin Zhang and Steve Horvath (2005)

"A General Framework for Weighted Gene Co-Expression Network Analysis",

Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1, Art. 17.

Central concept in network methodology:

Network Modules

- Modules: groups of densely interconnected genes (not the same as closely related genes)
 - a class of over-represented patterns
- Empirical fact: gene co-expression networks exhibit modular structure

Module Detection

- Numerous methods exist
- Many methods define a suitable gene-gene *dissimilarity measure and use clustering.*
- In our case: dissimilarity based on **topological overlap**
- Clustering method: Average linkage hierarchical clustering
 - branches of the dendrogram are modules

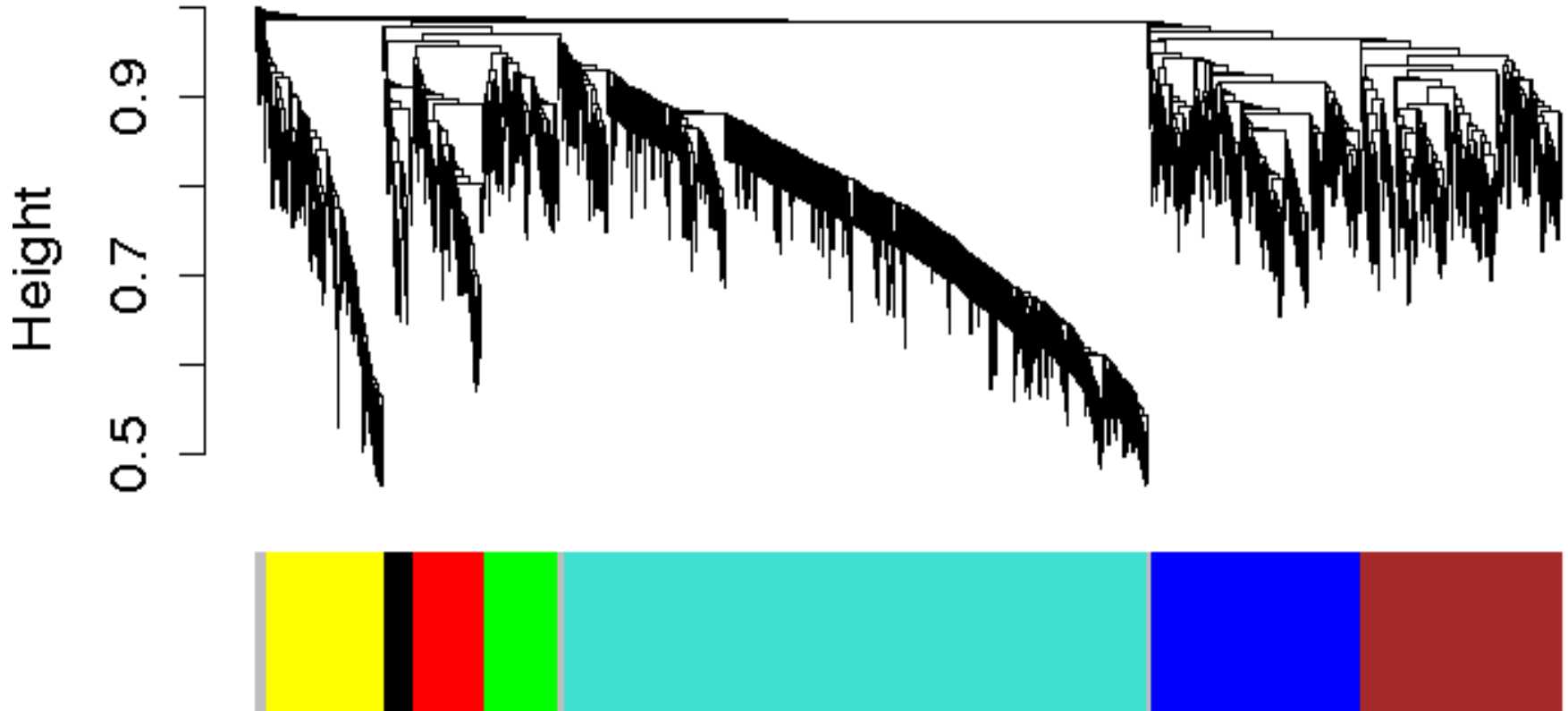
Topological overlap measure, TOM

- Pairwise measure by Ravasz *et al*, 2002
- $TOM[i,j]$ measures the overlap of the set of nearest neighbors of nodes i,j
- Closely related to *twinness*
- ***Easily generalized to weighted networks***

Example of module detection via hierarchical clustering

- Expression data from human brains, 18 samples.

Dendrogram and module colors

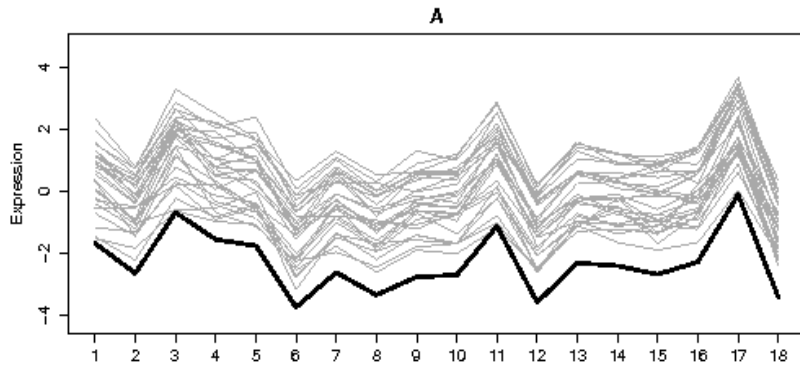


Module eigengenes

- Often: Would like to treat modules as single units
 - Biologically motivated data reduction
- Construct a representative
- Our choice: **module eigengene** = 1st principal component of the module expression matrix
- Intuitively: a kind of average expression profile
- Genes of each module must be highly correlated for a representative to really represent

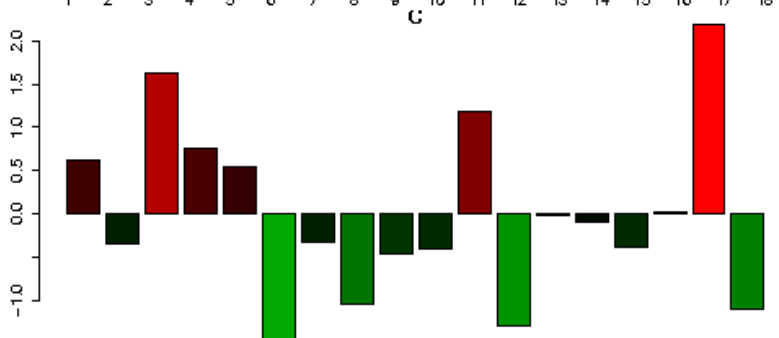
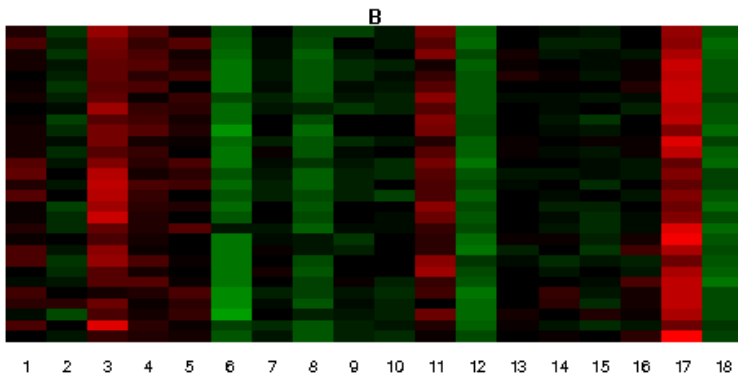
Langfelder P, Horvath S (2007) Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology 2007, 1:54

Example



Human brain expression data, 18 samples

Module consisting of 50 genes



Langfelder P, Horvath S (2007) Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology 2007, 1:54

Module eigengenes are very useful!

- Summarize each module in one synthetic expression profile
- Suitable representation in situations where modules are considered the basic building blocks of a system
 - Allow to relate modules to external information (phenotypes, genotypes such as SNP, clinical traits) via simple measures (correlation, mutual information etc)
 - Can quantify co-expression relationships of various modules by standard measures

Langfelder P, Horvath S (2007) Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology 2007, 1:54

Unsupervised Mining

Biplot

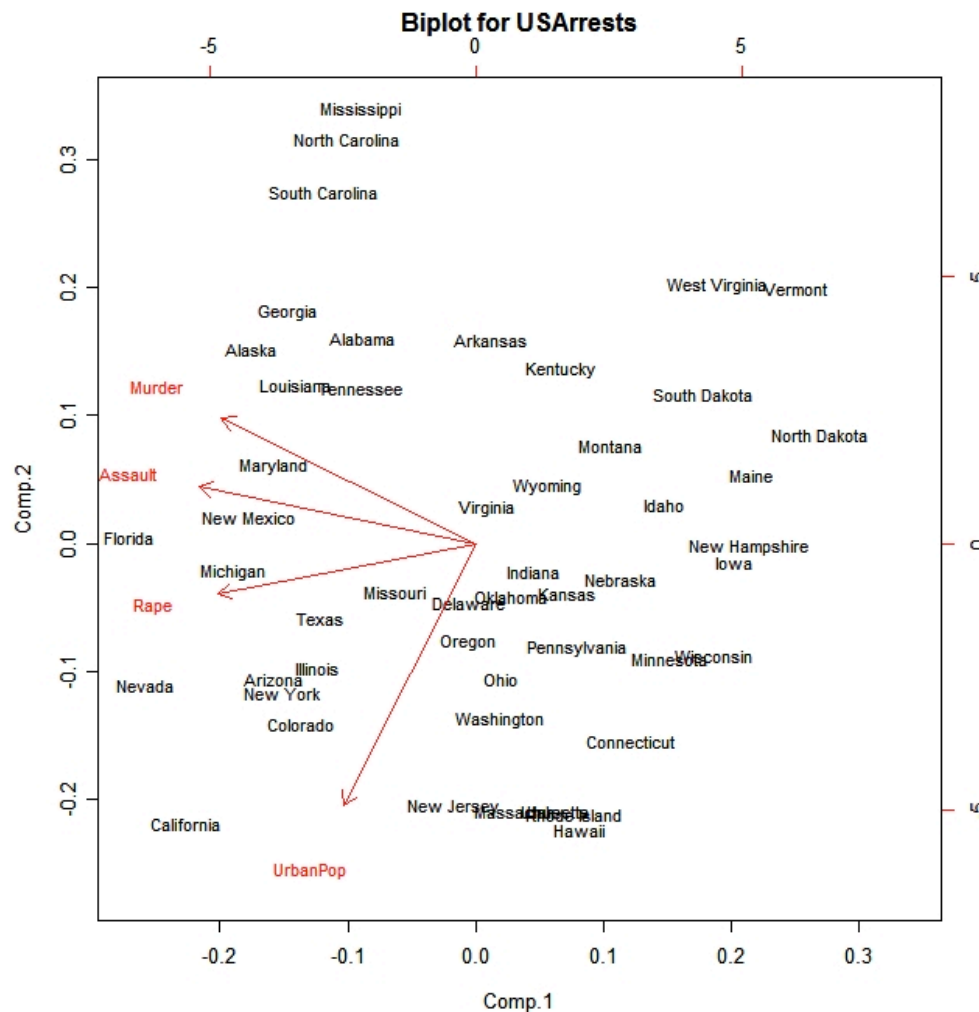
Biplot

- A biplot is a two-dimensional representation of a data matrix showing a point for each of the n observation vectors (rows of the data matrix) along with a point for each of the p variables (columns of the data matrix).
 - The prefix 'bi' refers to the two kinds of points; not to the dimensionality of the plot. The method presented here could, in fact, be generalized to a three-dimensional (or higher-order) biplot. Biplots were introduced by Gabriel (1971) and have been discussed at length by Gower and Hand (1996). We applied the biplot procedure to the following toy data matrix to illustrate how a biplot can be generated and interpreted. See the figure on the next page.
- Here we have three variables (transcription factors) and ten observations (genomic bins). We can obtain a two-dimensional plot of the observations by plotting the first two principal components of the TF-TF correlation matrix R_1 .
 - We can then add a representation of the three variables to the plot of principal components to obtain a biplot. This shows each of the genomic bins as points and the axes as linear combination of the factors.
- The great advantage of a biplot is that its components can be interpreted very easily. First, correlations among the variables are related to the angles between the lines, or more specifically, to the cosines of these angles. An acute angle between two lines (representing two TFs) indicates a positive correlation between the two corresponding variables, while obtuse angles indicate negative correlation.
 - Angle of 0 or 180 degrees indicates perfect positive or negative correlation, respectively. A pair of orthogonal lines represents a correlation of zero. The distances between the points (representing genomic bins) correspond to the similarities between the observation profiles. Two observations that are relatively similar across all the variables will fall relatively close to each other within the two-dimensional space used for the biplot. The value or score for any observation on any variable is related to the perpendicular projection from the point to the line.
- Refs
 - Gabriel, K. R. (1971), "The Biplot Graphical Display of Matrices with Application to Principal Component Analysis," *Biometrika*, 58, 453–467.
 - Gower, J. C., and Hand, D. J. (1996), *Biplots*, London: Chapman & Hall.

Introduction

- A biplot is a low-dimensional (usually 2D) representation of a data matrix **A**.

- A point for each of the m observation vectors (rows of **A**)
- A line (or arrow) for each of the n variables (columns of **A**)



PCA

TFs: a, b, c...

Genomic

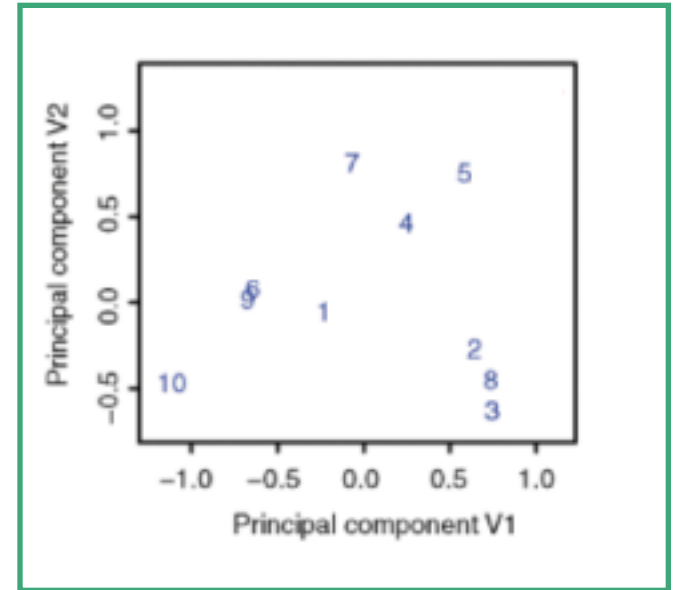
Sites: 1,2,3...

A

	a	b	c
1	21	16	28
2	14	18	25
3	14	17	22
4	14	19	33
5	17	23	28
6	20	14	34
7	22	21	30
8	15	18	22
9	18	13	36
10	24	10	32

	a	b	c
a	1.00	-0.44	0.48
b	-0.44	1.00	-0.40
c	0.48	-0.40	1.00

$A^T A$ (TF-TF corr.)

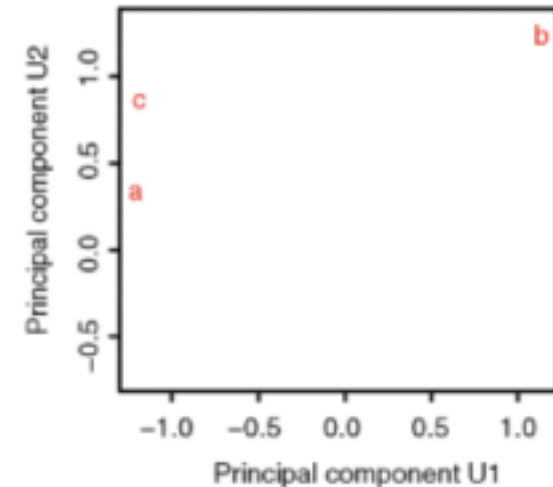


A^T

	1	2	3	4	5	6	7	8	9	10
a	21	14	14	14	17	20	22	15	18	24
b	16	18	17	19	23	14	21	18	13	10
c	28	25	22	33	28	34	30	22	36	32

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.70	0.69	0.77	0.54	0.99	0.95	0.65	0.98	0.97
2	0.70	1.00	1.00	0.99	0.98	0.79	0.89	1.00	0.84	0.50
3	0.69	1.00	1.00	0.99	0.98	0.78	0.89	1.00	0.83	0.49
4	0.77	0.99	0.99	1.00	0.95	0.85	0.94	0.98	0.89	0.59
5	0.54	0.98	0.98	0.95	1.00	0.64	0.78	0.99	0.71	0.31
6	0.99	0.79	0.78	0.85	0.64	1.00	0.98	0.74	1.00	0.93
7	0.95	0.89	0.89	0.94	0.78	0.98	1.00	0.86	0.99	0.84
8	0.65	1.00	1.00	0.98	0.99	0.74	0.86	1.00	0.80	0.43
9	0.98	0.84	0.83	0.89	0.71	1.00	0.99	0.80	1.00	0.89
10	0.97	0.50	0.49	0.59	0.31	0.93	0.84	0.43	0.89	1.00

AA^T (site-site correlation)



TFs: a, b, c...

Genomic

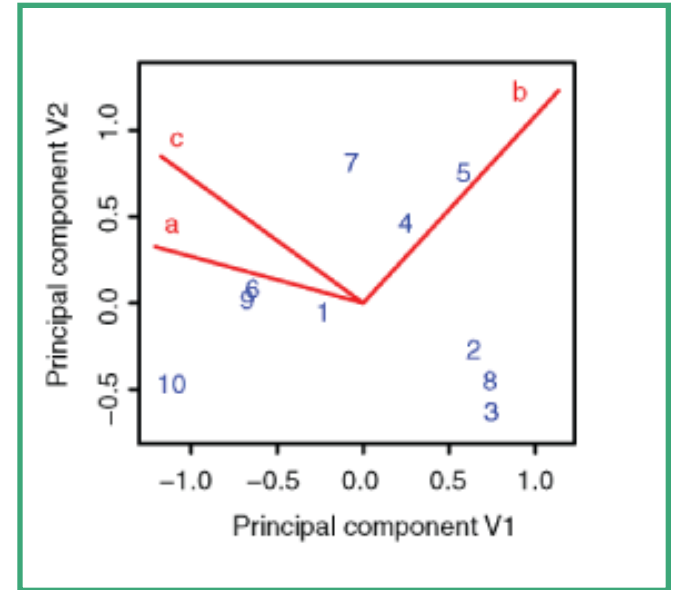
Sites: 1,2,3...

$$A=USV^T$$

	a	b	c
1	21	16	28
2	14	18	25
3	14	17	22
4	14	19	33
5	17	23	28
6	20	14	34
7	22	21	30
8	15	18	22
9	18	13	36
10	24	10	32

	a	b	c
a	1.00	-0.44	0.48
b	-0.44	1.00	-0.40
c	0.48	-0.40	1.00

$A^T A$ (TF-TF corr.)

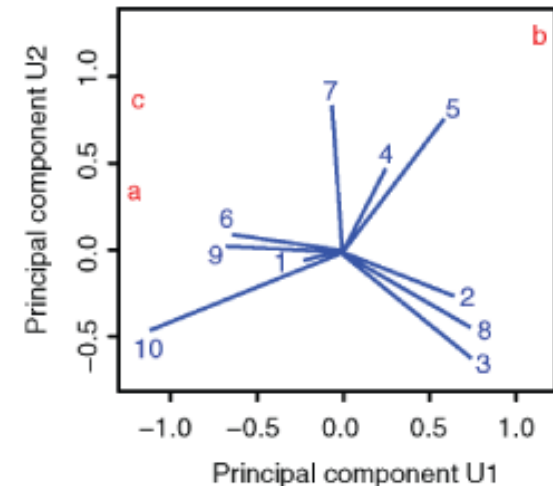


A^T

	1	2	3	4	5	6	7	8	9	10
a	21	14	14	14	17	20	22	15	18	24
b	16	18	17	19	23	14	21	18	13	10
c	28	25	22	33	28	34	30	22	36	32

	1	2	3	4	5	6	7	8	9	10
1	1.00	0.70	0.69	0.77	0.54	0.99	0.95	0.65	0.98	0.97
2	0.70	1.00	1.00	0.99	0.98	0.79	0.89	1.00	0.84	0.50
3	0.69	1.00	1.00	0.99	0.98	0.78	0.89	1.00	0.83	0.49
4	0.77	0.99	0.99	1.00	0.95	0.85	0.94	0.98	0.89	0.59
5	0.54	0.98	0.98	0.95	1.00	0.64	0.78	0.99	0.71	0.31
6	0.99	0.79	0.78	0.85	0.64	1.00	0.98	0.74	1.00	0.93
7	0.95	0.89	0.89	0.94	0.78	0.98	1.00	0.86	0.99	0.84
8	0.65	1.00	1.00	0.98	0.99	0.74	0.86	1.00	0.80	0.43
9	0.98	0.84	0.83	0.89	0.71	1.00	0.99	0.80	1.00	0.89
10	0.97	0.50	0.49	0.59	0.31	0.93	0.84	0.43	0.89	1.00

AA^T (site-site correlation)



Biplot Ex

Genomic bin	TF		
	a	b	c
1	21	16	28
2	14	18	25
3	14	17	22
4	14	19	33
5	17	23	28
6	20	14	34
7	22	21	30
8	15	18	22
9	18	13	36
10	24	10	32

Data matrix

Variable (column)
standardization

A

Transpose

Genomic bin

Transcription factor	Genomic bin									
	1	2	3	4	5	6	7	8	9	10
a	21	14	14	14	17	20	22	15	18	24
b	16	18	17	19	23	14	21	18	13	10
c	28	25	22	33	28	34	30	22	36	32

Data matrix (transposed)

Variable (column)
standardization

Genomic bin

TF	Genomic bin									
	1	2	3	4	5	6	7	8	9	10
a	-0.11	-0.90	-0.91	-0.81	-1.03	-0.26	-0.47	-0.95	-0.36	0.18
b	-0.94	-0.18	-0.16	-0.30	0.06	-0.84	-0.68	-0.09	-0.77	-1.08
c	1.05	1.08	1.07	1.12	0.97	1.10	1.15	1.04	1.13	0.90

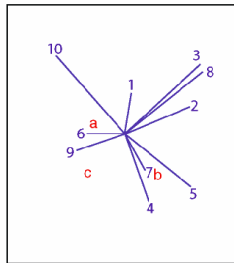
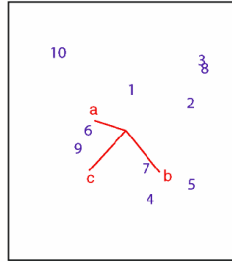
Standardized data matrix (transposed)

Genomic bin	TF		
	a	b	c
1	0.84	-0.23	-0.20
2	-1.06	0.29	-0.82
3	-1.06	0.03	-1.43
4	-1.06	0.55	0.82
5	-0.24	1.59	-0.20
6	0.57	-0.75	1.02
7	1.11	1.07	0.20
8	-0.78	0.29	-1.43
9	0.03	-1.01	1.43
10	1.65	-1.80	0.61

Standardized data matrix

Correlating factors

3D scatterplot



10D scatterplot*

Correlating bins

Genomic bin	Genomic bin									
	1	2	3	4	5	6	7	8	9	10
1	1.00	0.70	0.69	0.77	0.54	0.99	0.95	0.65	0.98	0.97
2	0.70	1.00	1.00	0.99	0.98	0.79	0.89	1.00	0.84	0.50
3	0.69	1.00	1.00	0.99	0.98	0.78	0.89	1.00	0.83	0.49
4	0.77	0.99	0.99	1.00	0.95	0.85	0.94	0.98	0.89	0.59
5	0.54	0.98	0.98	0.95	1.00	0.64	0.78	0.99	0.71	0.31
6	0.99	0.79	0.78	0.85	0.64	1.00	0.98	0.74	1.00	0.93
7	0.95	0.89	0.89	0.94	0.78	0.98	1.00	0.86	0.99	0.84
8	0.65	1.00	1.00	0.98	0.99	0.74	0.86	1.00	0.80	0.43
9	0.98	0.84	0.83	0.89	0.71	1.00	0.99	0.80	1.00	0.89
10	0.97	0.50	0.49	0.59	0.31	0.93	0.84	0.43	0.89	1.00

Correlation matrix R_2

TF	TF		
	a	b	c
a	1.00	-0.44	0.48
b	-0.44	1.00	-0.40
c	0.48	-0.40	1.00

Correlation matrix R_1

A' A

PCA

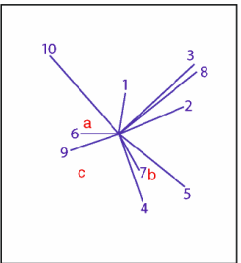
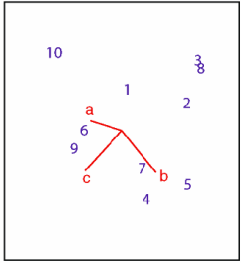
PCA'

Biplot Ex #2

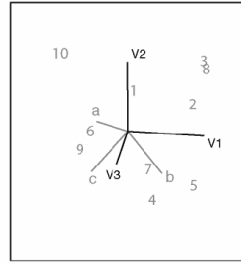
TF

	a	b	c
a	1.00	-0.44	0.48
b	-0.44	1.00	-0.40
c	0.48	-0.40	1.00

Correlation matrix R_1

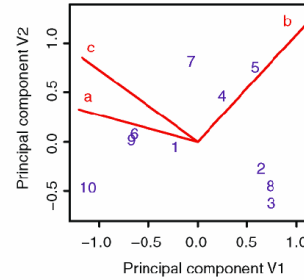


PCA *



$$A^T A = V S^2 V^T$$

Projection *

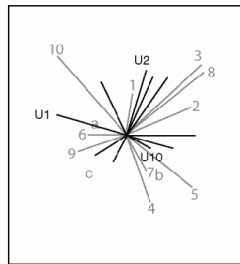


The same rank-2 approximation
of the original data matrix

$$A v_j = u_j s_j \text{ \& \ } A^T u_j = v_j s_j$$

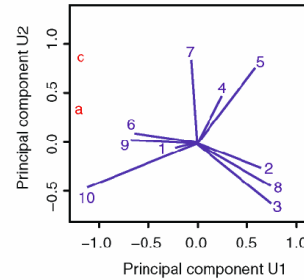
$$A = (U S^r) (V S^{1-r})^T$$

PCA *



$$A A^T = U S^2 U^T$$

Projection *



Genomic bin

	2	3	4	5	6	7	8	9	10
2	1.00	0.70	0.69	0.77	0.54	0.99	0.95	0.65	0.98
3	0.70	1.00	0.99	0.98	0.79	0.89	1.00	0.84	0.50
4	0.69	0.99	1.00	0.98	0.78	0.89	1.00	0.83	0.49
5	0.77	0.98	0.98	1.00	0.64	0.78	0.99	0.71	0.31
6	0.54	0.79	0.64	0.64	1.00	0.98	0.74	1.00	0.93
7	0.99	0.89	0.89	0.99	0.98	1.00	0.86	0.99	0.84
8	0.65	1.00	1.00	0.99	0.74	0.86	1.00	0.80	0.43
9	0.98	0.84	0.83	0.83	0.71	1.00	0.99	1.00	0.89
10	0.49	0.50	0.49	0.59	0.31	0.93	0.43	0.89	1.00

Correlation matrix R_2

*

10D scatterplots are used here for illustrative purpose only.

PCA: the correlation matrix is eigen-decomposed; then the principal components are added to the original space.

Projection: the points and axes in the original space are projected onto the plane defined by the top two principal components.

Biplot Ex #3

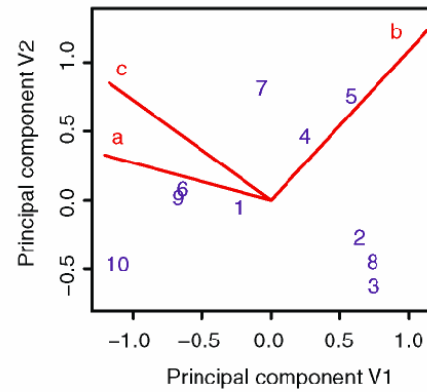
$$A \mathbf{v}_i = s_i \mathbf{u}_i$$

$$A^T \mathbf{u}_i = s_i \mathbf{v}_i$$

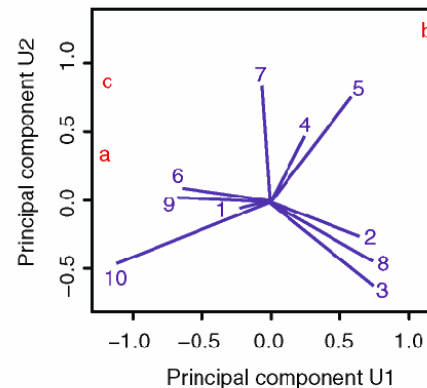
Assuming $s=1$,

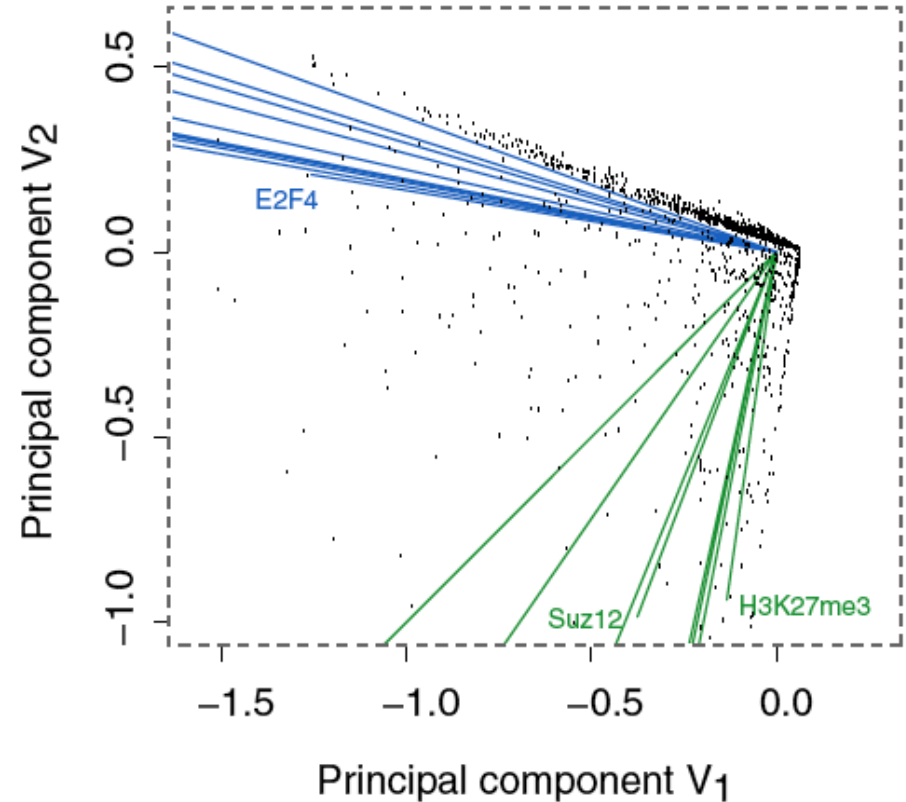
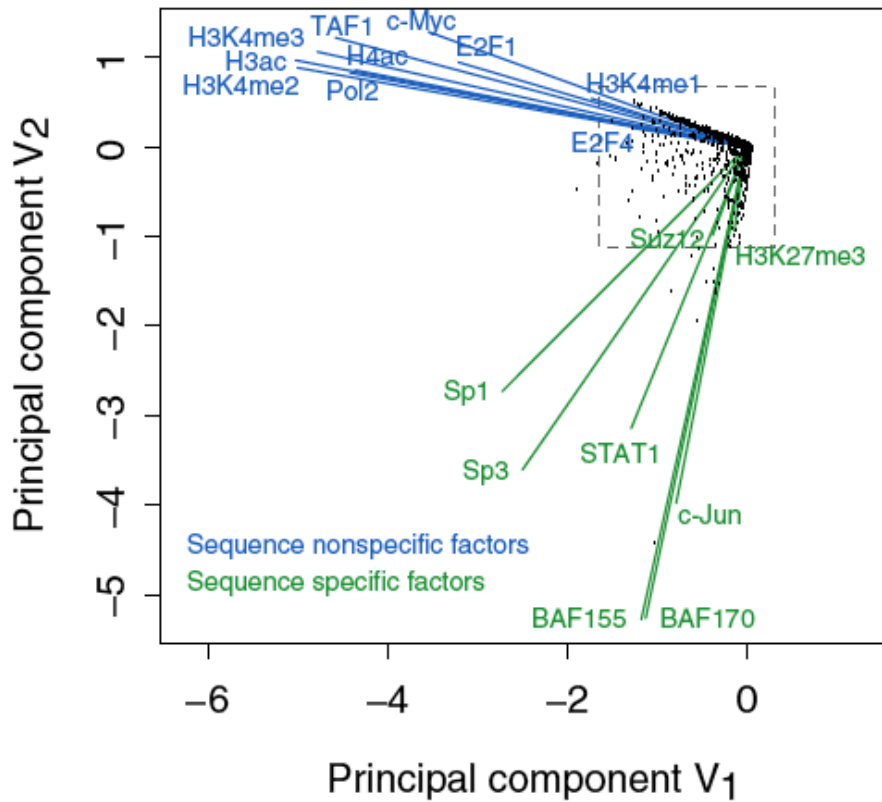
$$A \mathbf{v}_i = \mathbf{u}_i$$

$$A^T \mathbf{u}_i = \mathbf{v}_i$$



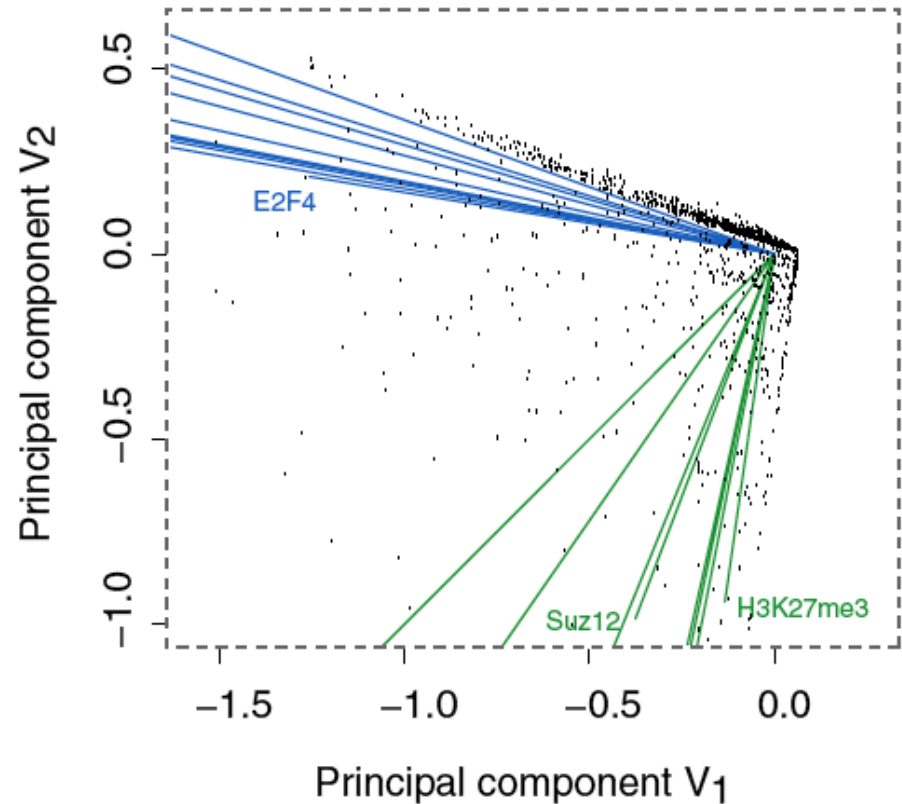
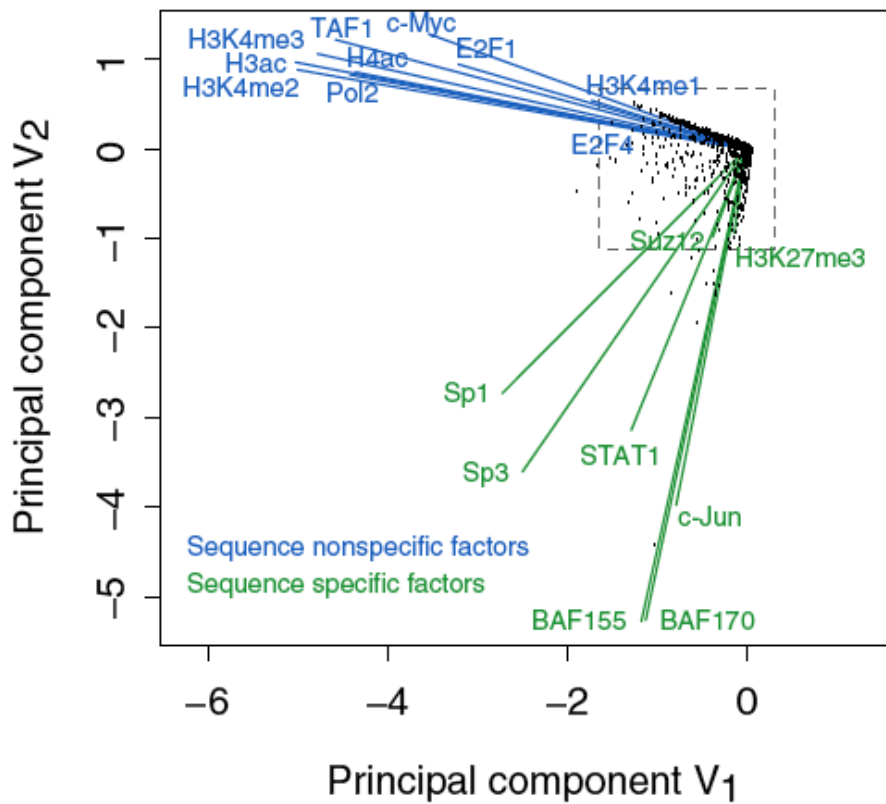
The same rank-2 approximation
of the original data matrix





Results of Biplot

- Pilot ENCODE (1% genome): 5996 10 kb genomic bins (adding all hits) + 105 TF experiments → biplot
- Angle between TF vectors shows relation b/w factors
- Closeness of points gives clustering of "sites"
- Projection of site onto vector gives degree to which site is assoc. with a particular factor



Results of Biplot

- Biplot groups TFs into sequence-specific and sequence-nonspecific clusters.
 - c-Myc may behave more like a sequence-nonspecific TF.
 - H3K27me3 functions in a transcriptional regulatory process in a rather sequence-specific manner.
- Genomic Bins are associated with different TFs and in this fashion each bin is "annotated" by closest TF cluster

Unsupervised Mining

CCA

Sorcerer II Global Ocean Survey

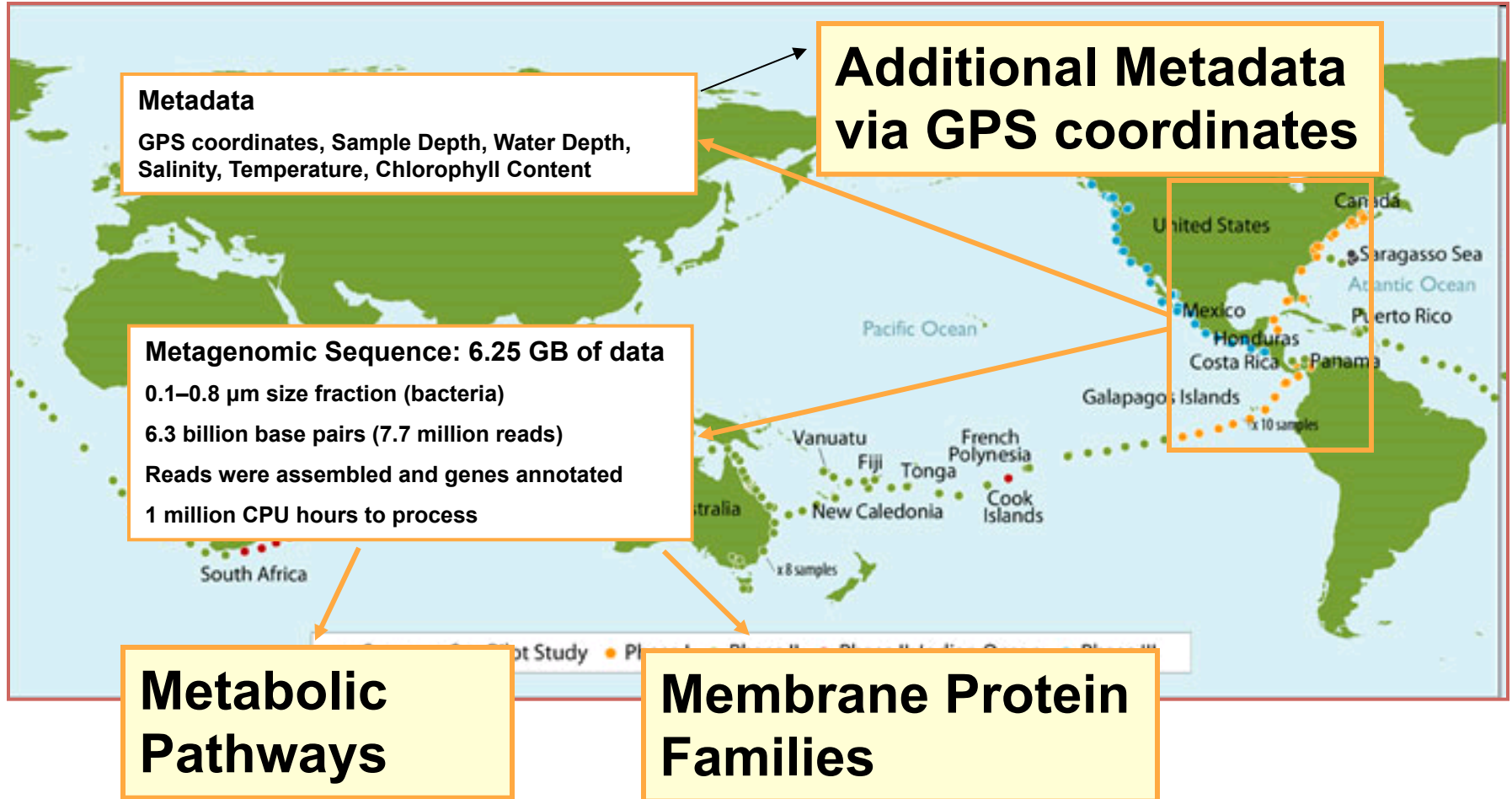


Sorcerer II journey August 2003- January 2006

Sample approximately every 200 miles



Sorcerer II Global Ocean Survey



READS → PROTEIN FAMILIES → PATHWAYS

CCGTGAGCACGATGCGC-----
 ATGCTCATGCT-----
 ATCGTGACGCGATGC-----
 CCGTGAGCACGATGCGC-----
 ATGCTCATGCT-----
 ATCGTGACGCGATGC-----
 ATGCTCATGCT-----
 GCGATCGATCGATCGTAGC-----
 TGCTGCTAGCATGCT-----
 GCGATCGATCGATCGTAGC-----
 TGCTGCTAGCATGCT-----
 CCGTGAGCACGATGCGC-----
 GATCGTAGCATGCTT-----
 CCGTGAGCACGATGCGC-----
 GCGATCGATCGATCGTAGC-----



$$P_1 = f_1 + f_2 + f_3$$

$$P_2 = f_4 + f_5 + f_6$$

Mapping Raw Metagenomic Reads to a Matrix of Families or Pathways for each Site

PATHWAYS



SITES

$$P_{1,1} = 2 + 1 + 3$$

$$P_{2,1} = 2 + 4 + 3$$

$$P_{1,2} = 5 + 2 + 6$$

$$P_{2,1} = 5 + 7 + 6$$



	Fam 1	Fam 2	Fam 157
Site 1	.01	.02			
Site 2	0	.01			
...					
Site 29					

Families Matrix

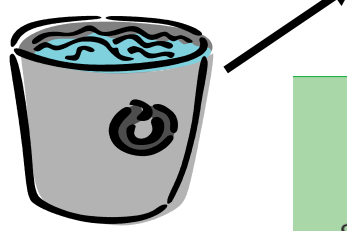
counts Fam 2 / #total protein counts at site 2

Pathway Sequences (Community Function)

Metabolic Pathways

Sites

	P1	P2	P3		
B1	3800	1400	1000		
B2	2200	100	400		
	----	----	----		



Environmental Features

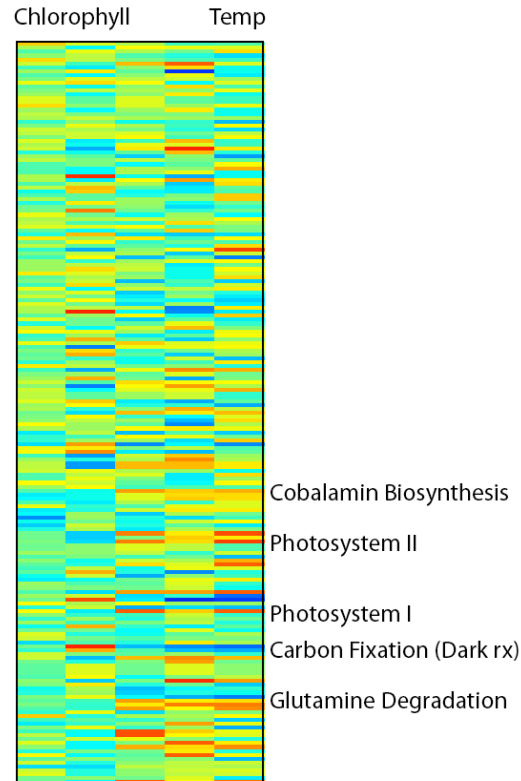
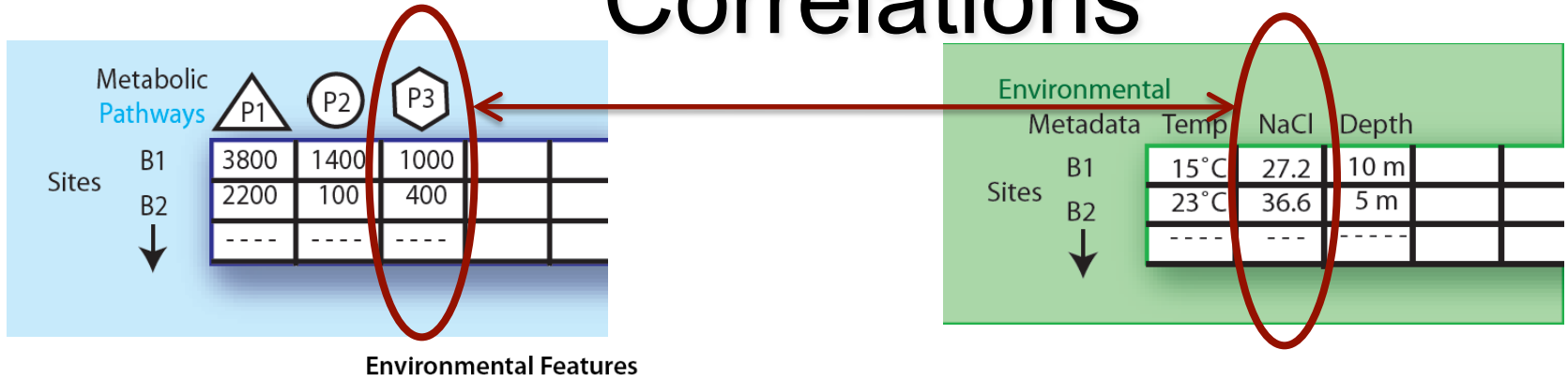
Environmental Metadata

Sites

	Temp	NaCl	Depth		
B1	15°C	27.2	10 m		
B2	23°C	36.6	5 m		
	----	---	-----		

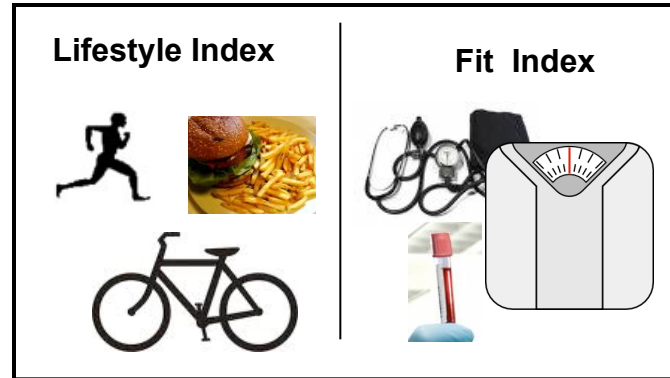
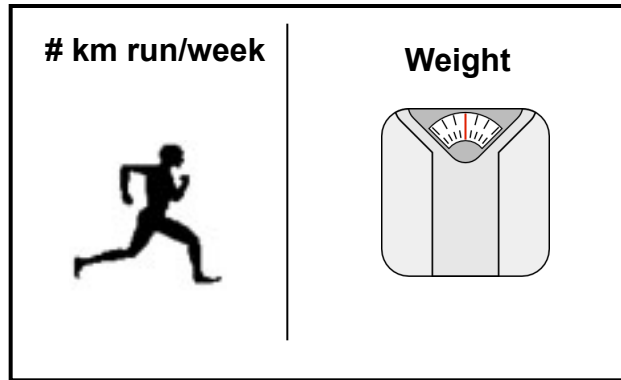
Expressing data as matrices indexed by site, env. var., and pathway usage

Simple Relationships: Pairwise Correlations



[Gianoulis et al., PNAS (in press, 2009)]

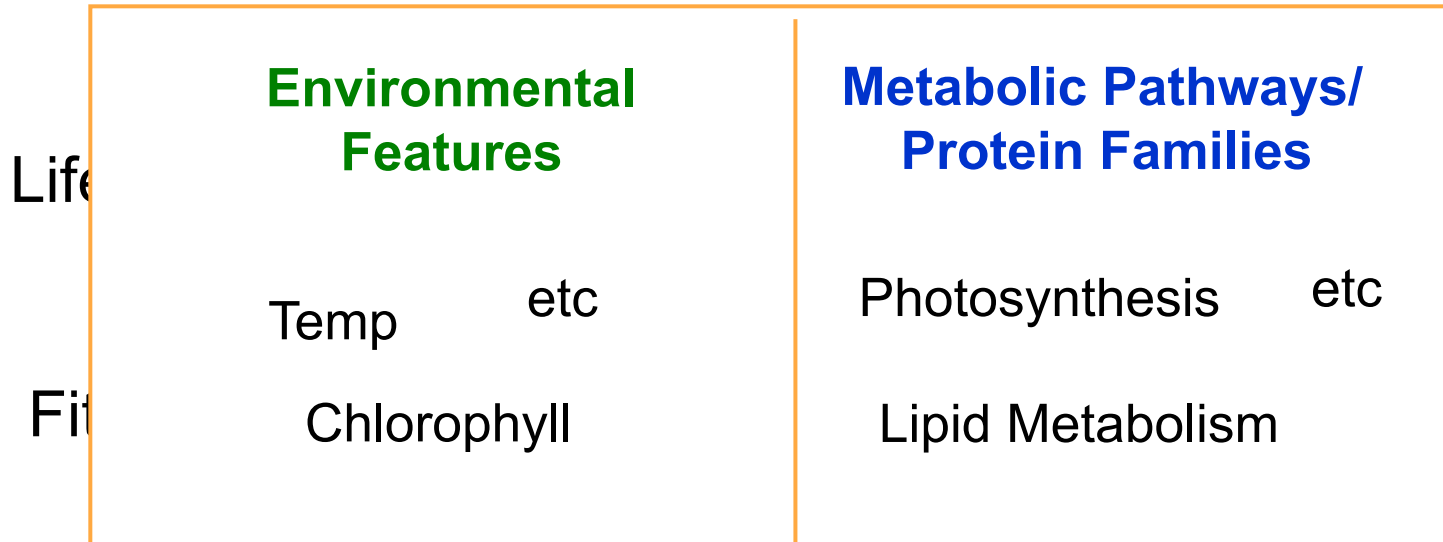
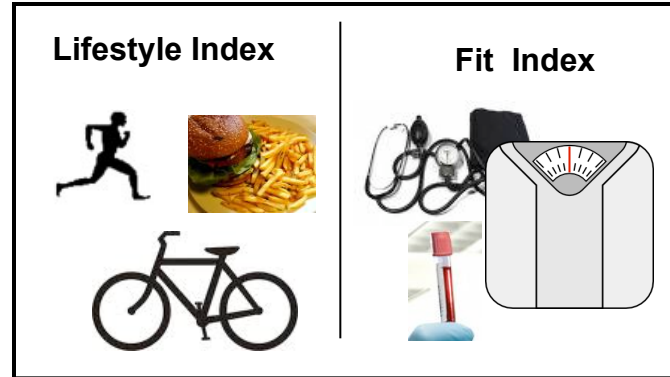
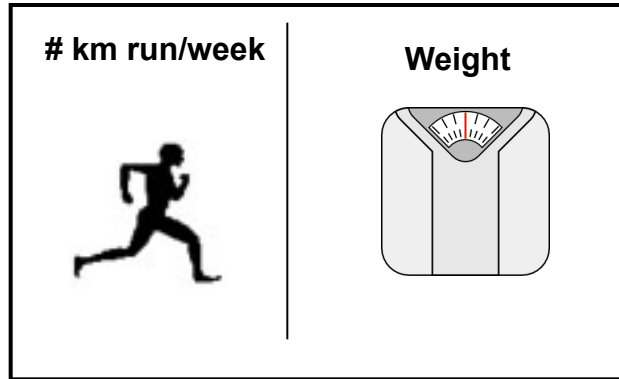
Canonical Correlation Analysis: Simultaneous weighting



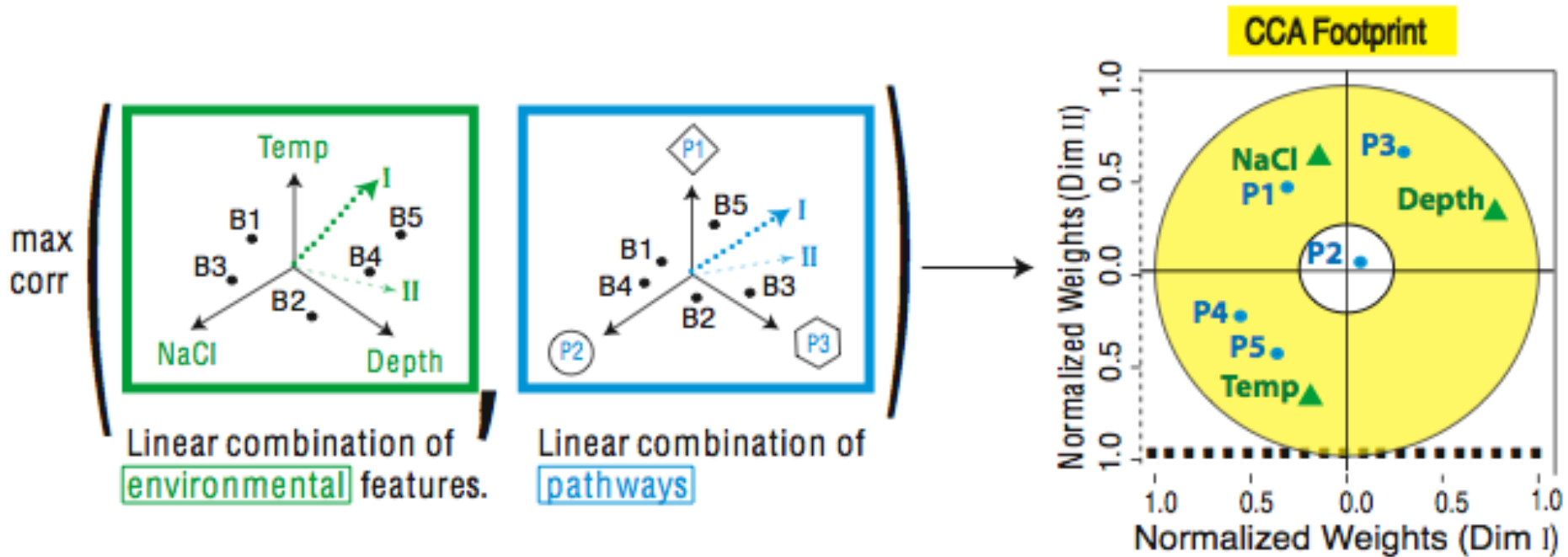
$$\text{Lifestyle Index} = a \text{  + b \text{  + c \text{ $$

$$\text{Fit Index} = a \text{  + b \text{  + c \text{ $$

Canonical Correlation Analysis: Simultaneous weighting



CCA: Finding Variables with Large Projections in "Correlation Circle"



The goal of this technique is to interpret cross-variance matrices
 We do this by defining a change of basis.

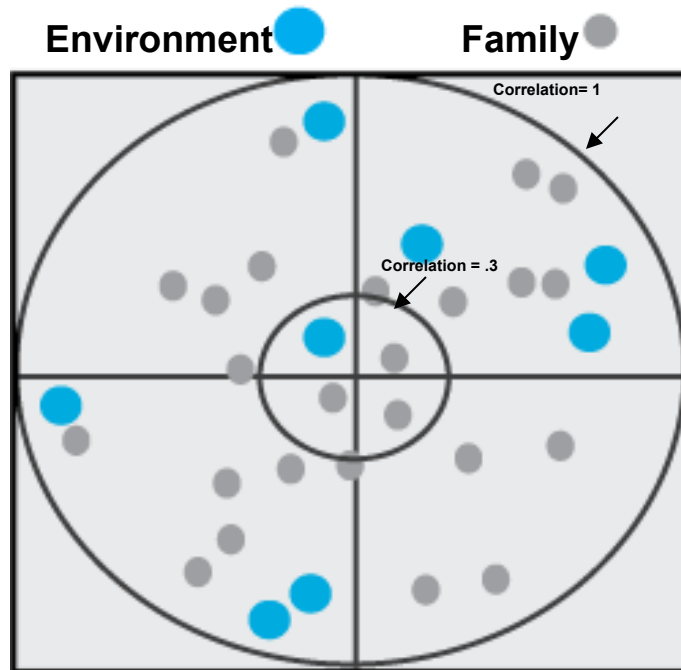
CCA results

We are defining a change of basis of the cross co-variance matrix

We want the correlations between the projections of the variables, X and Y, onto the basis vectors to be mutually maximized.

Eigenvalues \rightarrow squared canonical correlations

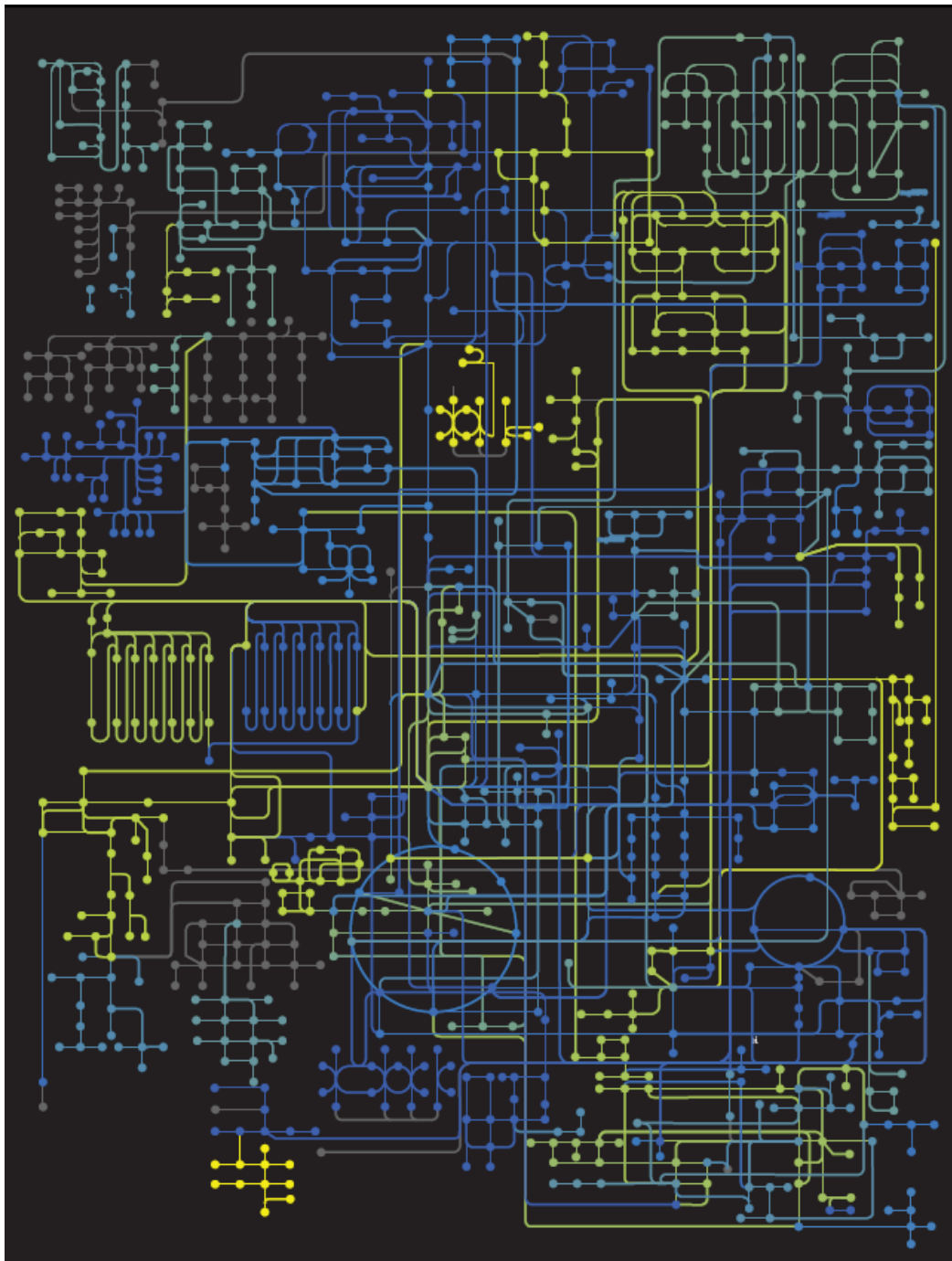
Eigenvectors \rightarrow normalized canonical correlation *basis vectors*



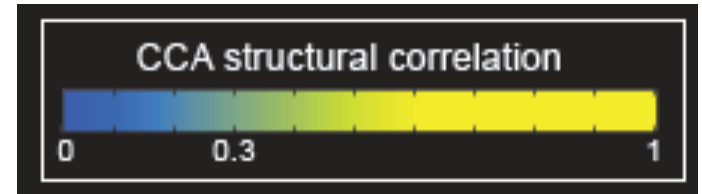
This plot shows the correlations in the first and second dimensions

Correlation Circle: The closer the point is to the outer circle, the higher the correlation

Variables projected in the same direction are correlated

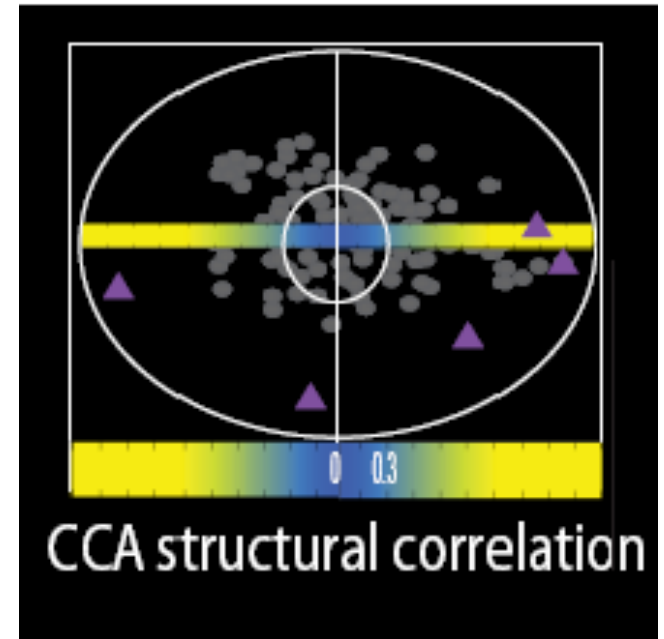


Strength of Pathway co-variation with environment

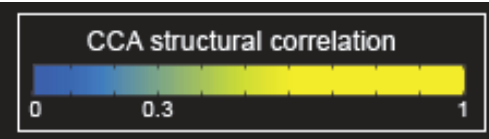


Environmentally invariant

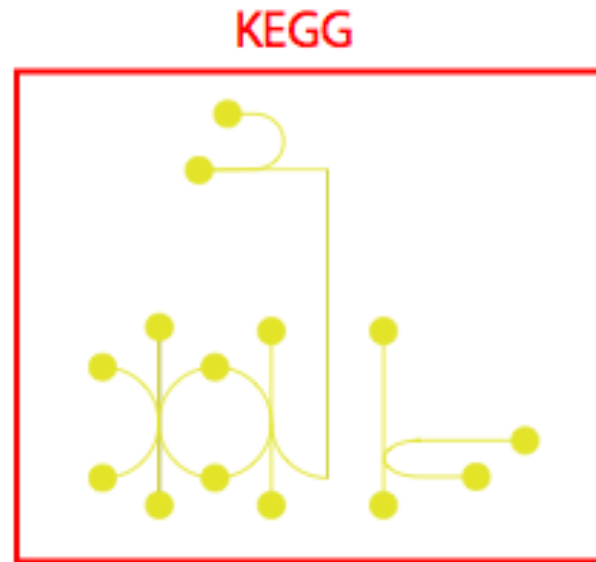
Environmentally variant



Conclusion #1: energy conversion strategy, temp and depth



Photosynthesis



Oxidative Phosphorylation

