# Singular value decomposition for genome-wide expression data processing and modeling

Presented by Jing Qiu

April 23, 2002

# Outline

- Biological Background
- Mathematical Framework:Singular Value Decomposition
  - SVD calculation
  - Pattern Inference
  - Data normalization
  - Data Sorting
- Biological Data Analysis I: Elutriation-synchronized cell cycle.

# Biological Background–Cell cycle regulation

- Cell cycle: the program for cell growth and division

- Four broad phases:$G_1$(and $G_0$), S, $G_2$,and M

- Cell cycle diagram
  - $G_1$(cell growth and protein for DNA synthesis)
  - $S$ (DNA replication, two daughter cell)
  - $G_2$(cell growth and protein synthesis)
  - $M$ (split apart)

- Application: cancer(uncontrolled cell growth and proliferation)

# Biological Background–Microarray Experiment

- Diagram of Central Dogma of Molecule Biology
- 100,000 genes in mammalian genome
- each cell express 15,000 of these genes
- each gene is expressed at a different level
- cell cycle regulated genes have different expression levels across the cycle period
- Microarrays:Massive parallel analysis of gene expression
- The process diagram

# Mathematical Framework:Notation

- $\widehat{e}$ denotes a matrix;
- $|v\rangle$ denotes a column vector
- $\langle u|$ denotes a row vector
- $\widehat{e}|v\rangle$, $\langle u|\widehat{e}$, $\langle u|v\rangle$ all denote inner products
- $|v\rangle\langle u|$ denotes outer product.
- $|a_m\rangle \equiv \widehat{e}|m\rangle$—the $m$th column of $\widehat{e}$
- $\langle g_n| \equiv \langle n|\widehat{e}$–the $n$th row of the matrix $\widehat{e}$.

## Mathematical Framework:Data of interest

- Data of interest: $\widehat{e}$ (dim=N $\times$ M)(See Fig 13a)

- N rows–N genes of a model organism $\langle g_n| \equiv \langle n|\widehat{e}$,—the relative expression of the $n$th gene across different arrays.

- M columns–M different samples(arrays) $|a_m\rangle \equiv \widehat{e}|m\rangle$,—the genome-wide relative expression measured by the $m$th array.

# Mathematical Framework:SVD of the data

- $\widehat{e} = \widehat{u}_{N\times N}\widehat{\varepsilon}_{N\times M}\widehat{v}^T_{M\times M} = \widehat{u}^*_{N\times M}\widehat{\varepsilon}^*_{M\times M}\widehat{v}^T_{M\times M}$

  where $\widehat{u}_{N\times N} = [\widehat{u}^*_{N\times M}, \widehat{h}_{(N-M)\times M}]$,

  $\widehat{\varepsilon}_{N\times M} = [\widehat{\varepsilon}^*_{M\times M}, 0_{(N-M)\times M}]^T$

- $\widehat{\varepsilon}^* = \begin{bmatrix} \varepsilon_1 & & \\ & \ddots & \\ & & \varepsilon_M \end{bmatrix}, \varepsilon_1 \geq \varepsilon_2 \geq ... \geq \varepsilon_L \geq 0.$

  - $\varepsilon_l$–the eigenexpression of the $l$th eigengene in the $l$th eigenarray.

- $\widehat{u}$—the N-genes $\times$ L-eigenarray basis sets

  - $|\alpha_l\rangle$,the genome-wide expression in the $l$th eigenarray.

- $\widehat{v}^T$—M-eigengenes $\times$ M-arrays basis sets
  - $\langle\gamma_l|$,the $l$th eigengene across the different arrays
- The “fraction of eigenexpression”:

$$p_l = \varepsilon_l^2 / \sum_{k=1}^{M} \varepsilon_k^2$$

- “Shannon entropy ”

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

  - measure the complexity of the dataset
  - $d = 0$– captured by a single eigengene(eigenarray)
  - $d = 1$–all eigengenes are equally expressed
- Figure 13 and Fig14

## Mathematical Framework: SVD calculation

- $\widehat{a} = \widehat{e}^T \widehat{e} = \widehat{v}\widehat{\varepsilon}^2\widehat{v}^T$

- diagonalizing $\widehat{a}$(dim=$M \times M$; $M \ll N$) to get $\widehat{v}$ and $\widehat{\varepsilon}$

- $\widehat{u} = \widehat{e}\widehat{v}\widehat{\varepsilon}^{-1}$.

## Mathematical Framewor:Pattern Inference

- An eigengene $\langle \gamma_l |$ represents a regulatory process when this pattern is bilogically interpretable.

- $\langle n | \alpha_l \rangle \equiv \frac{\langle g_n | \gamma_l \rangle}{\varepsilon_l}$–the relative amplitude of the $l$th eigengene pattern in $\langle g_n |$ relative to all other genes.

- The corresponding eigenarray $| \alpha_l \rangle$ represents the cellular state which corresponds to this process.

# Mathematical Framework:Data normalization

- Normalizing the data by filtering out the eigen-genes $\langle\gamma_l|$ which are inferred to represent noise

- $\widehat{e} \rightarrow \widehat{e} - \varepsilon_l|\alpha_l\rangle\langle r_l| = \widehat{u}\widehat{\varepsilon}'_{l-}\widehat{v}$ where

$$\widehat{\varepsilon}'_{l-} = \begin{bmatrix} \varepsilon_1 & & & & & & \\ & \ddots & & & & & \\ & & \varepsilon_{l-1} & & & & \\ & & & 0 & & & \\ & & & & \varepsilon_{l+1} & & \\ & & & & & \ddots & \\ & & & & & & \varepsilon_M \end{bmatrix}$$

# Mathematical Framework:Data sorting

- Sort data by similarity in the expression of any chosen subsets of the eigengenes

- Correlation plot: Scatter plot of $(C_{k,n}, C_{l,n})$, where

$$C_{k,n} \equiv \frac{\langle \gamma_k | g_n \rangle}{\langle g_n | g_n \rangle}$$

- $r_n = \frac{\sqrt{(|\langle \gamma_k | g_n \rangle|^2 + |\langle \gamma_l | g_n \rangle|^2)}}{\langle g_n | g_n \rangle}$: Amplitude of expression of the $n$th gene in the subspace spanned by $\langle \gamma_k |$ and $\langle \gamma_l |$ relative to its overall expression:

- $\phi_n \equiv tan^{-1} \frac{\langle \gamma_k | g_n \rangle}{\langle \gamma_l | g_n \rangle}$: The phase of the $n$th gene in the transition from the expression pattern $\langle \gamma_l |$ to $\langle \gamma_k |$ and back to $\langle \gamma_l |$.

- sort the genes according to $\phi_n$.

## Biological Data Analysis:Data

- Elutriation-synchronized cell cycle data of budding yeast

- $N = 5981$ genes, $784$ of which were classified as cell cycle regulated in Spellman et al.,1998.
  $M = 14$ arrays in interval of 30 minutes
  $L = min(M, N) = 14$.

# Biological Data Analysis:Pattern inference

- $\langle \gamma_1 |$ describes the time-invariant relative expression during the cell cycle.(Fig 14a, constantly red)

- Others show oscillation during the cell cycle.

- Low entropy:$d = .14 \ll 1$–weak perturbation of a steady state of expression(figure 14b)

- $\langle \gamma_1 |$ captures more than $90\%$ of the overall relative expression:

- $\langle \gamma_2 |$,$\langle \gamma_3 |$,$\langle \gamma_4 |$ capture $3\%, 1\%, .5\%$ of the overall relative expression, respectively.

- The time variation of $\langle \gamma_3 |$ fits a normalized sine function of period T;(fig 14c)

- The time variations of $\langle\gamma_2|$ and $\langle\gamma_4|$ fit a cosine function of period,(fig 14c)

- $\langle\gamma_2|$ show decreasing expression on transition from $t = 0$ to $30$ min

- $\langle\gamma_4|$ show increasing expression on transition from $t = 0$ to $30$ min

- Inference:
  - $\langle\gamma_1|$ represents experimental additive constants superimposed on a steady gene expression state.
  - $\langle\gamma_3|$ represents expression oscillation during a cell cycle
  - $\langle\gamma_2|$ and $\langle\gamma_4|$ represent initial transient increase and decrease in expression in response to the elutriation, respectively.

# Biological Data Anlalysis: Data Normalization

- Filter out the first eigengene to remove the steady state of expression:

$$\widehat{e} \rightarrow \widehat{e}_C \equiv \widehat{e} - \varepsilon_1 |\alpha_1\rangle\langle\gamma_1|$$

- $\widehat{e}_C \rightarrow \widehat{e}_{LV} \equiv [log(e^2_{C,nm})]_{N\times M}$ where

$$e_{C,nm} = \langle n|\widehat{e}_C|m\rangle$$

and $e^2_{C,nm}$ is the variance in the measured expression of the $n$th gene in the $m$th array.

- Figure 15a displays the eigengenes for $\widehat{e}_{LV}$

- $|\gamma_1\rangle_{LV}$ captures more than $80\%$ of the overall information in this dataset.

- It describes a weak initial transient increase superimposed on a time-invariant scale of expression variance (maybe a response to the elutriation).

- The time-invariant scale of expression variance suggests a steady scale of the data.

- It also suggests the time-invariant multiplicative constant noise may be superimposed on the data.

- Filter out $|\gamma_1\rangle_{LV}$, removing the steady scale of expression variance,

$$\widehat{e}_{LV} \rightarrow \widehat{e}_{CLV} \equiv \widehat{e}_{LV} - \varepsilon_{1,LV}|\alpha_1\rangle_{LV}\langle\gamma_1|_{LV}$$

- $\widehat{e}_{CLV} \rightarrow \widehat{e}_N \equiv [sign(e_C, nm)\sqrt{exp(e_{CLV,nm})}]_{N\times M}$

- $\widehat{e}_N$ tabulates expression pattern that are approximately centered at the steady state with variance

which are approximately normalized by the steady scale of the expression variance.

- $\langle \gamma_1 |_N$ and $\langle \gamma_2 |_N$(of similar significance), capture together more than $40\%$ of the overall normalized expression,$d = 0.88$.(Figure 1)

- The time variations of $\langle \gamma_1 |_N$ and $\langle \gamma_2 |_N$ fit normalized sine and cosine functions of period T and initial phase $\theta \approx 2\pi/13$.

- Inference: $\langle \gamma_1 |_N$ and $\langle \gamma_2 |_N$ represent cell cycle expression oscillations and assume $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ represent the corresponding cell cycle cellular states.

## Biological Data Analysis:Data sorting

- Assume $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ approximately represent all cell cycle cellular states

- All arrays(except $|a_1 1\rangle$) have at least $25\%$ of their normalized expression in this subspace.

- Suggest:$|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ are sufficient to approximate the elutriation array expression

- The array order according to $\phi_m$ is similar to the cell cycle time points. (an order of the cell cycle progress)

- Inferences:

  - $|\alpha_1\rangle_N$ is associated with the cell cycle cellular state of transition from $G_1$ to $S$.

  - -$|\alpha_1\rangle_N$ is associated with the transition from $G_2/M$ to $M/G_1$.

  - $|\alpha_2\rangle_N$ is associated with the cell cycle cellular state of transition from $M/G_1$ to $G_1$.

  - -$|\alpha_2\rangle_N$ is associated with the transition from $S$ to $S/G_2$.

- The phase of $|a_1\rangle$, $\phi_1 = -\theta \approx \frac{-2\pi}{13}$ corresponds to the 30-min delay between the start of the experiment and that of the cell cycle stage $G_1$.

- Figure 2b

- Recall $\langle\gamma_1|_N$ and $\langle\gamma_2|_N$ are inferred to approximately represent all cell cycle expression oscillation

- Expect $\gamma_n \approx 1$–cell cycle regulated
  $\gamma_n \approx 0$–not regulated by the cell cycle at all.

- 641(classified as cell cycle regulated) have more than $25\%$ of their normalized expression in this subspace.

- This sortinggives a different classification from that by Spellman et al(the poor quality of the elutriation expression data).

- With all genes sorted, the gene variations of $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ fit normalized sine and cosine functions(Figure 3).

- The sorted and normalized elutriation expression fit approximately a traveling wave of expression varying sinusoidally across both genes and arrays.

# Some points

- SVD calculation:$\widehat{\varepsilon}^{-1}$ don't necessarily exist.

- Data Sorting

  - What if more then 2 significant eigengenes?

  - $C_{k,n} \equiv \frac{\langle \gamma_k | g_n \rangle}{\langle g_n | g_n \rangle}$ or $\equiv \frac{\langle \gamma_k | g_n \rangle}{\sqrt{\langle g_n | g_n \rangle}}$?

- What smoothing method to fit $|\gamma_1\rangle_N$?

- The disagreement due to poor quality of the data(too simple)?
  Are the first two eigengenes enough to explain all?(capture only around $40\%$and Figure 2b–not good projection)