Shrinkage-Based Similarity Metric for Cluster Analysis of Microarray Data

Vera Cherepinsky[1]
(with Jiawu Feng[1], Marc Rejali[1], and Bud Mishra[1,2])
[1] Courant Institute of Mathematical Sciences, NYU
[2] Watson School of Biological Sciences, CSHL

Yale University
April 1, 2003

Transcriptional State of a Cell

Transcriptional State of a Cell

◊ Transcriptional state of a cell can be characterized by detecting and quantitating gene expression levels:

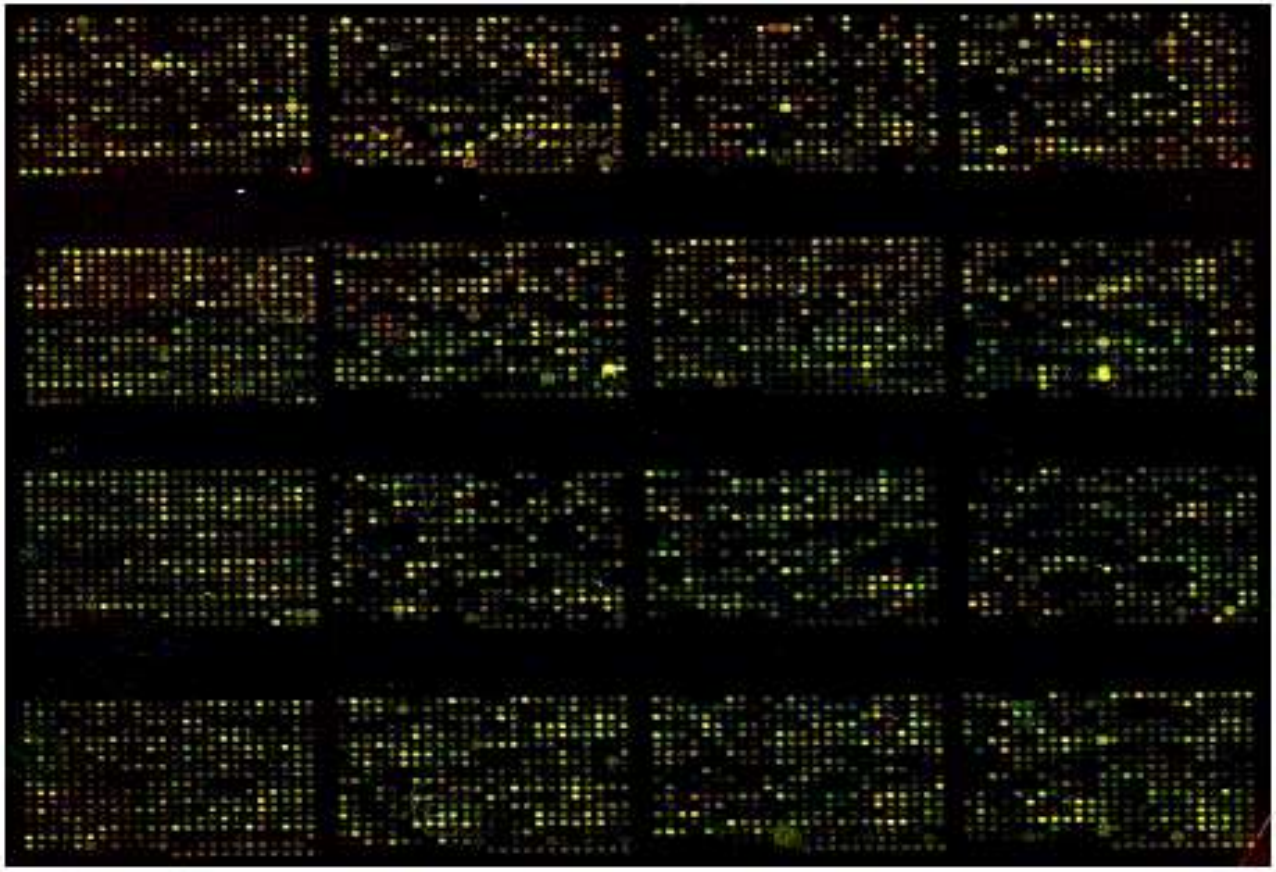
- Northern blots
- S1 nuclease protection
- differential display
- sequencing of cDNA libraries
- serial analysis of gene expression (cDNA)
- Array based technologies:
  - ◇ spotted arrays
  - ◇ oligonucleotide arrays

Gene Expression Data

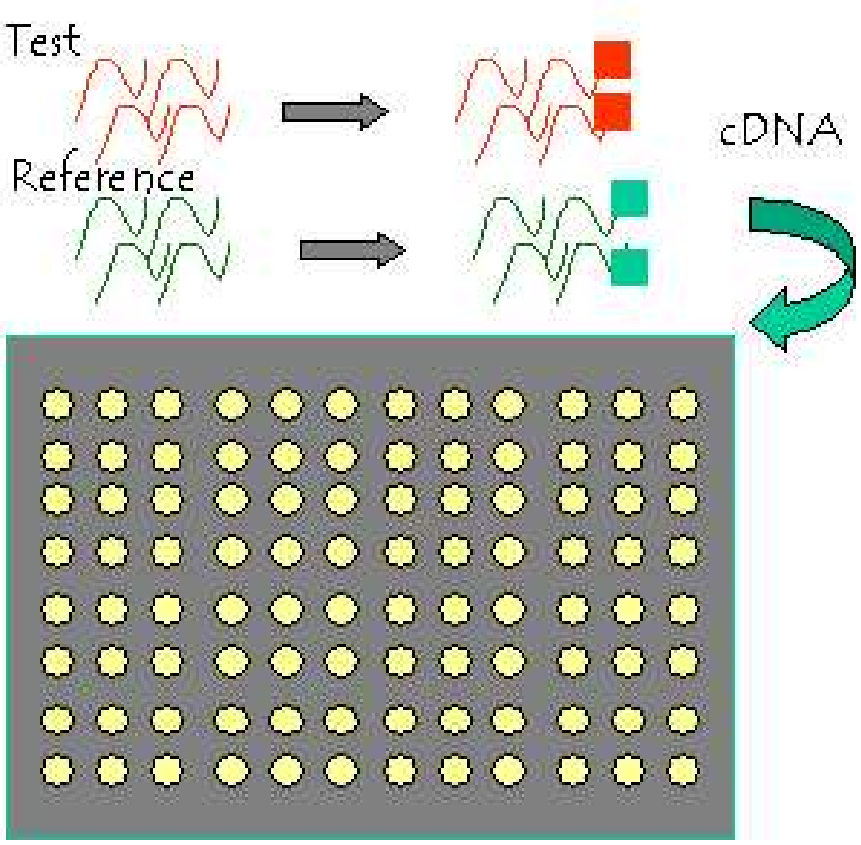
◊ Microarrays enable one to simultaneously measure the activity of up to 30,000 ($\sim 10^4$–10^5) genes.

◊ In particular, the amount of mRNA for each gene in a given sample (or a pair of samples) can be measured.

Spotted Arrays



Spotted Arrays



Two samples (reference and test) of mRNA are converted to cDNA, labeled with fluorochrome dyes and allowed to hybridize to the array.

Cluster Analysis

Cluster Analysis

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14863–14868, December 1998
Genetics

Cluster analysis and display of genome-wide expression patterns

MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡

*Department of Genetics and †Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, 300 Pasteur Avenue, Stanford, CA 94305

Contributed by David Botstein, October 13, 1998

In the above, Eisen *et al.* claim to use "standard statistical algorithms to arrange genes according to similarity in pattern of gene expression."

Distances & Correlations

Let G_i equal the (log-transformed) primary data for gene G in condition i. For any two genes X and Y observed over a series of N conditions, we can compute a similarity score as follows:

$$S(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - X_{offset}}{\Phi_X}\right)\left(\frac{Y_i - Y_{offset}}{\Phi_Y}\right), \quad (1)$$

where

$$\Phi_G = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(G_i - G_{offset}\right)^2}$$

Let $G_{offset} = \gamma\bar{G}$.

Metric Comparison

◊ *Pearson Correlation Coefficient:*

$$G_{offset} = \bar{G} = \frac{1}{N}\sum_{j=1}^{N} G_i, \quad \text{or} \quad \gamma = 1$$

◊ *Eisen:* (prone to False Positives)

$$G_{offset} = 0 \quad \text{for every gene } G, \quad \text{or} \quad \gamma = 0$$

◊ We propose using the general form of equation (1) to derive a similarity metric which is dictated by the data and reduces the occurrence of false-positives (relative to the Eisen metric) and false-negatives (relative to the Pearson correlation coefficient).

Shrinkage Metric: Result

$$S(X_j, X_k) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_{ij} - (X_j)_{offset}}{\Phi_j}\right)\left(\frac{X_{ik} - (X_k)_{offset}}{\Phi_k}\right),$$

where

$$\begin{aligned}
(X_j)_{offset} &= \widehat{\theta_j} \\
&= \left(1 - \frac{\widehat{1}}{\frac{\beta^2}{N} + \tau^2}\frac{\widehat{\beta^2}}{N}\right) Y_j \\
&= \underbrace{\left(1 - \left(\frac{M-2}{\sum_{k=1}^{M} {Y_k}^2}\right) \cdot \frac{1}{N} \cdot \frac{1}{M(N-1)} \sum_{k=1}^{M}\sum_{i=1}^{N} (X_{ik} - Y_k)^2\right)}_{\gamma} Y_j \qquad (2) \\
&= \gamma \overline{X}_{\cdot j}
\end{aligned}$$

and $Y_j = \overline{X}_{\cdot j}$.

Simulation

Simulation

◊ Simulation Model:

- Random Variables: X_i and Y_i:

$$\begin{aligned} X_i &= \theta_X + \sigma_X(\alpha_i(X,Y) + \mathcal{N}(0,1)) \\ Y_i &= \theta_Y + \sigma_Y(\alpha_i(X,Y) + \mathcal{N}(0,1)) \end{aligned}$$

- $\theta_X \sim \mathcal{N}(0,\tau^2)$; $\theta_Y \sim \mathcal{N}(0,\tau^2)$ are the means,
- $\alpha_i \sim$ Uniform(L,H) - Bias term (or $\alpha_i = 0$ for no bias).

◊ $S(X,Y) = \frac{1}{N}\sum_{i=1}^{N} \frac{(X_i-\theta_X)}{\sigma_X}\frac{(Y_i-\theta_Y)}{\sigma_Y} = \frac{1}{N}\left[\left(\sum_i \alpha_i^2\right) + \chi^2_N + 2\mathcal{N}(0,1)\sum_i \alpha_i\right]$

- N = Number of Experiments = 100.

Key Parameters

- N = Number of Experiments = 100
- $\tau \in \{0.1, 10.0\}$ $\leftarrow$ Very low or very high variability among the genes
- $\sigma_X = \sigma_Y = 10.0$
- $\alpha = 0 (\sim \mathcal{U}(0,0))$ $\leftarrow$ no correlation or
  $\alpha \sim \mathcal{U}(0,1)$ $\leftarrow$ some correlation between the genes.

Key Methods

(*Clairvoyant Metric Parameters*)

$$S[X_{1\_}, X_{2\_}] := \frac{1}{NExpt}\left(\frac{X_1-\theta_1}{\sigma_1} \cdot \frac{X_2-\theta_2}{\sigma_2}\right);$$

(*Pearson Metric Parameters*)

$$\mu_1 = \text{Mean}[X_1];\ \mu_2 = \text{Mean}[X_2];$$

$$\beta_1 = \sqrt{\frac{(X_1-\mu_1).(X_1-\mu_1)}{NExpt-1}};$$

$$\beta_2 = \sqrt{\frac{(X_2-\mu_2).(X_2-\mu_2)}{NExpt-1}};$$

$$S_p[X_{1\_}, X_{2\_}] := \frac{1}{NExpt-1}\left(\frac{X_1-\mu_1}{\beta_1} \cdot \frac{X_2-\mu_2}{\beta_2}\right);$$

(*Eisen Metric Parameters*)

$$me_1 = me_2 = 0;$$

$$be_1 = \sqrt{\frac{(X_1-me_1).(X_1-me_1)}{NExpt-1}};$$

$$be_2 = \sqrt{\frac{(X_2-me_2).(X_2-me_2)}{NExpt-1}};$$

$$S_e[X_{1\_}, X_{2\_}] := \frac{1}{NExpt-1}\left(\frac{X_1-me_1}{be_1} \cdot \frac{X_2-me_2}{be_2}\right);$$

(*Shrinkage Metric Parameters*)

$$ms_1 = \left(1 - \frac{{\sigma_1}^2}{{\sigma_1}^2+\tau^2 NExpt}\right)\mu_1;$$

$$ms_2 = \left(1 - \frac{{\sigma_2}^2}{{\sigma_2}^2+\tau^2 NExpt}\right)\mu_2;$$

$$bs_1 = \sqrt{\frac{(X_1-ms_1).(X_1-ms_1)}{NExpt-1}};$$

$$bs_2 = \sqrt{\frac{(X_2-ms_2).(X_2-ms_2)}{NExpt-1}};$$

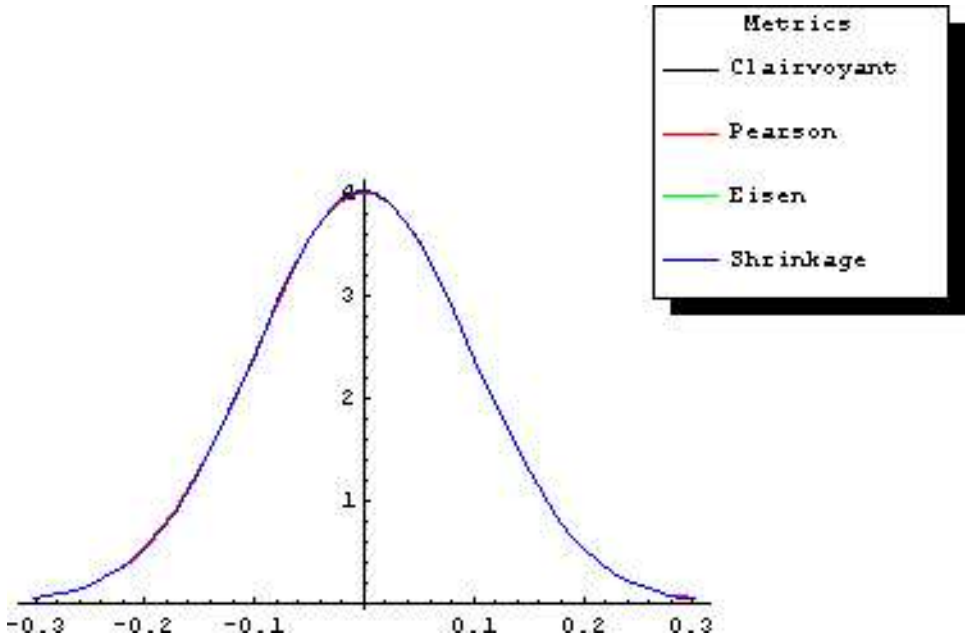
$$S_s[X_{1\_}, X_{2\_}] := \frac{1}{NExpt-1}\left(\frac{X_1-ms_1}{bs_1} \cdot \frac{X_2-ms_2}{bs_2}\right);$$

Experiment 1

◊ 1a. When X and Y are not correlated and the noise in the input is low, Pearson does as well as Eisen or Shrinkage:

- $\tau = 0.1$;
- $\alpha = 0$;
- $NExpt = 100$;
- $\sigma_X = \sigma_Y = 10$;

	True	Clairvoyant	Pearson	Eisen	Bayes
Mean=	0.	-0.000297241	-0.000268684	-0.000253954	-0.000254259
S.D.=	0.	0.0995578	0.0999037	0.0994011	0.0994009



Uncorrelated Genes

◊ If two genes are uncorrelated,

◊ and their "base-level values do not vary much"

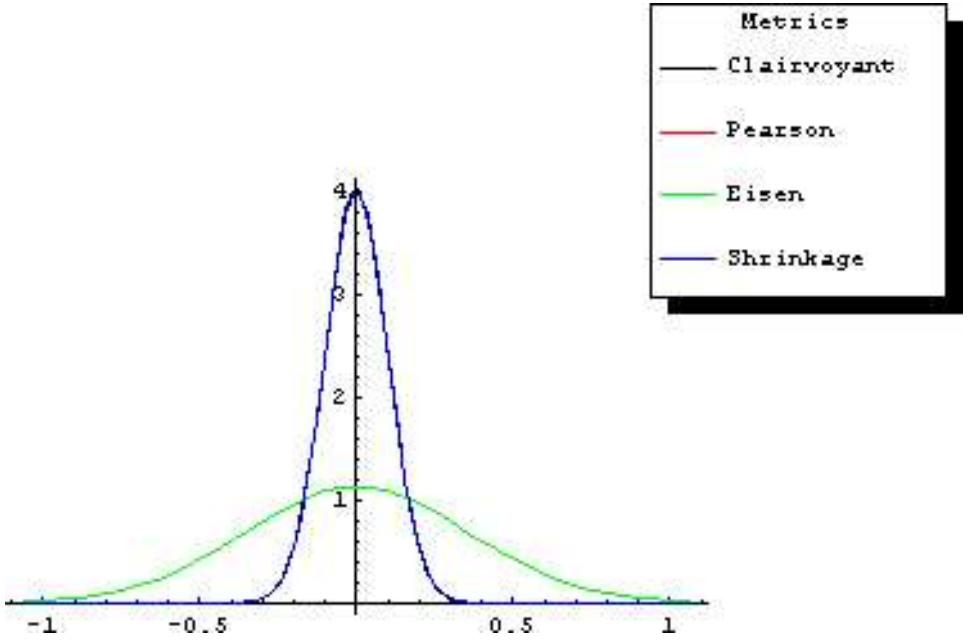
  – All the methods do equally well

◊ True Negatives

Experiment 2

◊ 1b. When X and Y are not correlated but the noise in the input is high, Eisen does much more poorly:

	True	Clairvoyant	Pearson	Eisen	Bayes
Mean=	0.	-0.000971174	-0.000939357	-0.00119089	-0.000939366
S.D.=	0.	0.0993954	0.100216	0.353616	0.100207



- $\tau = 10$;
- $\alpha = 0$;
- $NExpt = 100$;
- $\sigma_X = \sigma_Y = 10$;

Uncorrelated Genes

◊ If two genes are uncorrelated,

◊ and their "base-level values vary quite a bit"

  – All the methods except Eisen *et al.* do equally well

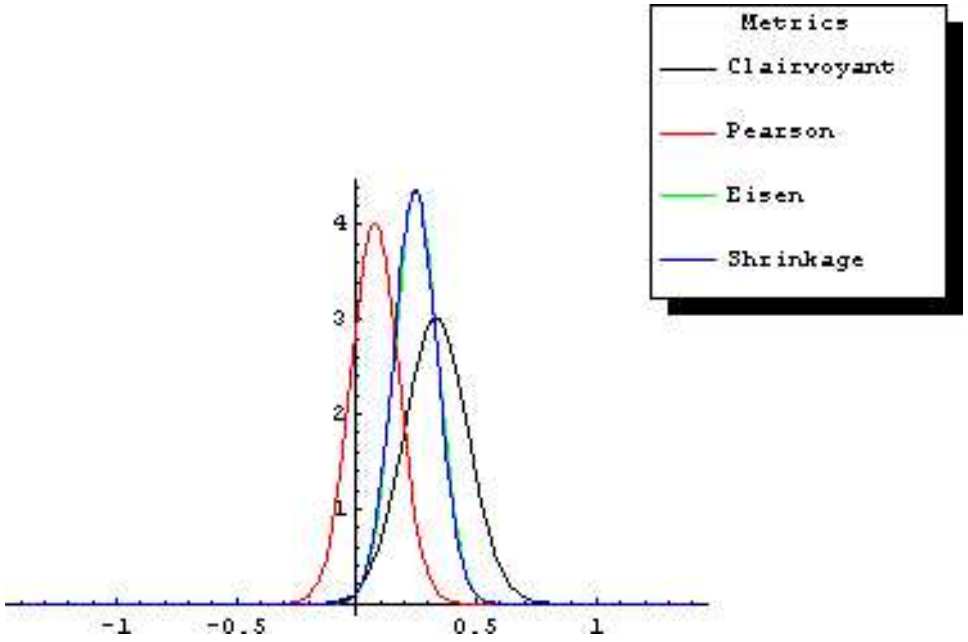
◊ False-Positives for Eisen

Experiment 3

◊ 2a. When X and Y are correlated and the noise in the input is low, Pearson does worse than Eisen or Shrinkage:

- $\tau = 0.1$;
- $\alpha \sim \mathcal{U}(0, 1)$;
- $NExpt = 100$;
- $\sigma_X = \sigma_Y = 10$;

	True	Clairvoyant	Pearson	Eisen	Bayes
Mean=	0.333112	0.331247	0.0754506	0.247842	0.24513
S.D.=	0.0299233	0.132424	0.0992482	0.0915289	0.0915238



Correlated Genes

◊ If two genes are correlated,

◊ and their "base-level values do not vary much"

  – All the methods except Pearson's do equally well

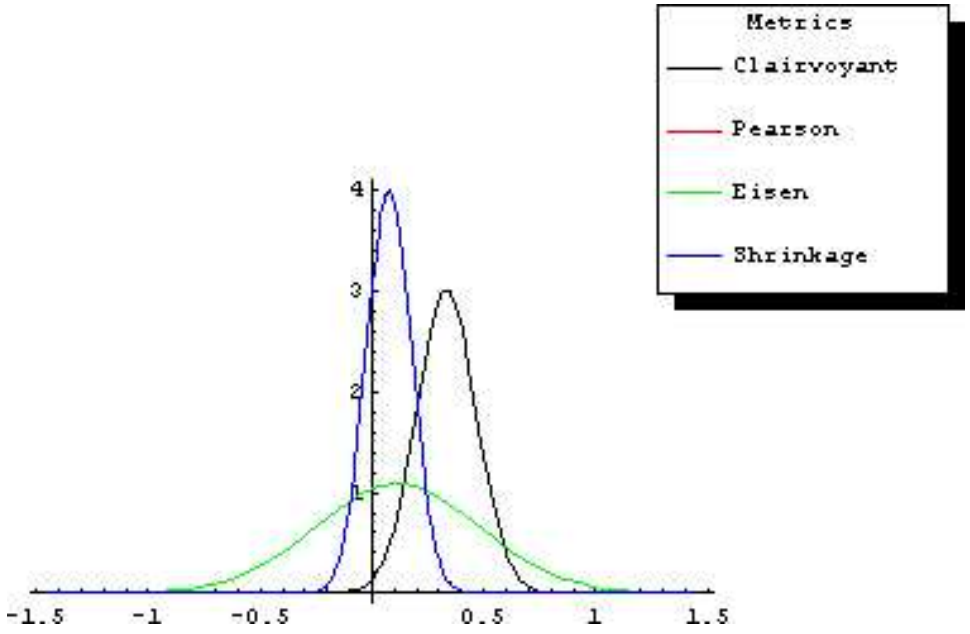
◊ False-Negatives for Pearson

Experiment 4

◊ 2b. When X and Y are correlated and the noise in the input is high, all algorithms fail, i.e., introduce errors:

- $\tau = 10$;
- $\alpha \sim \mathcal{U}(0, 1)$;
- $NExpt = 100$;
- $\sigma_X = \sigma_Y = 10$;

	True	Clairvoyant	Pearson	Eisen	Bayes
Mean=	0.333233	0.332629	0.0761588	0.116725	0.0761723
S.D.=	0.0298209	0.132613	0.0999562	0.367705	0.099945



Correlated Genes

◊ If two genes are correlated,

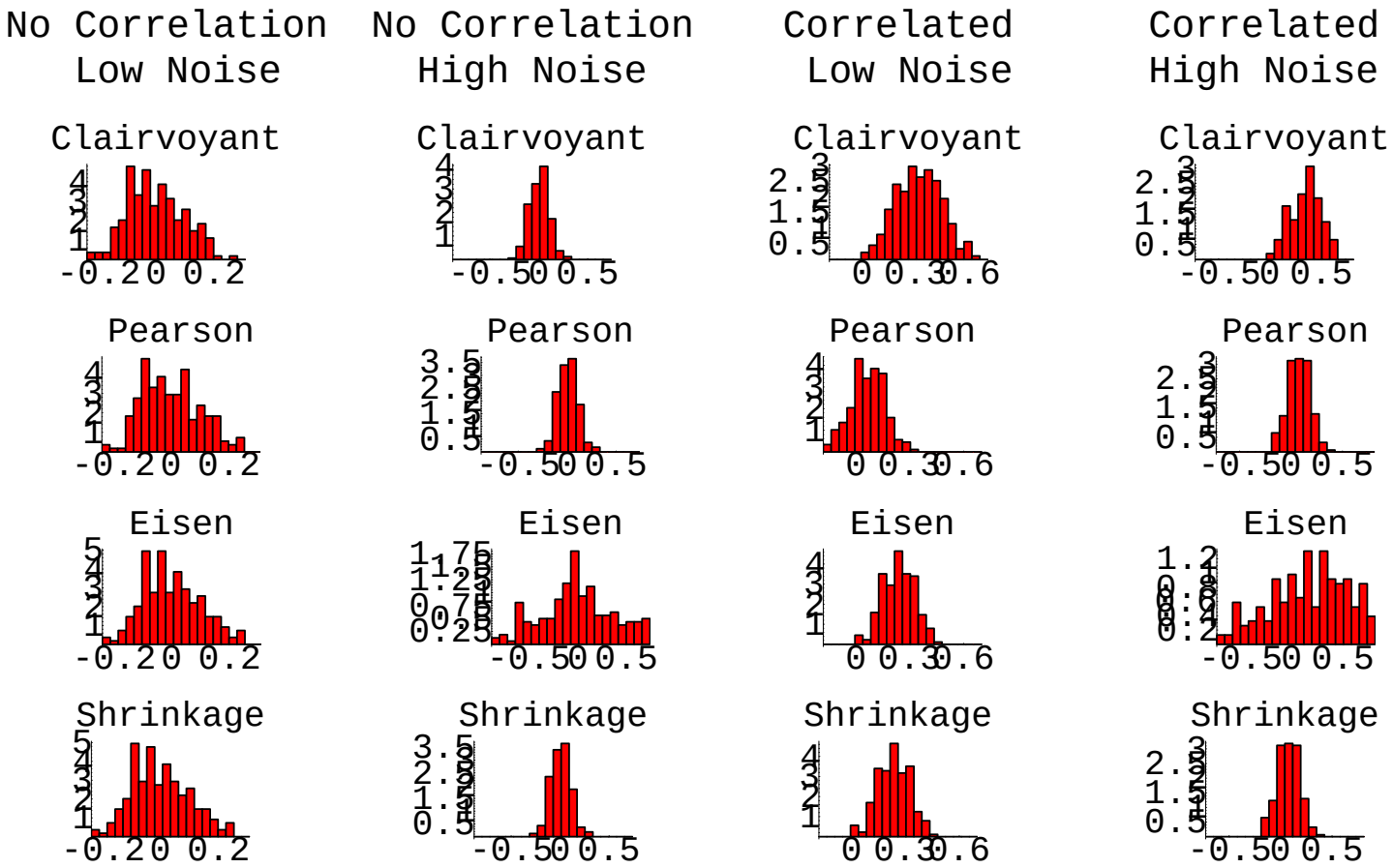
◊ and their "base-level values vary quite a bit"

  – All the methods do equally poorly

◊ False-Negatives

◊ (Eisen may also have some False-Positives.)

Histogram Comparison

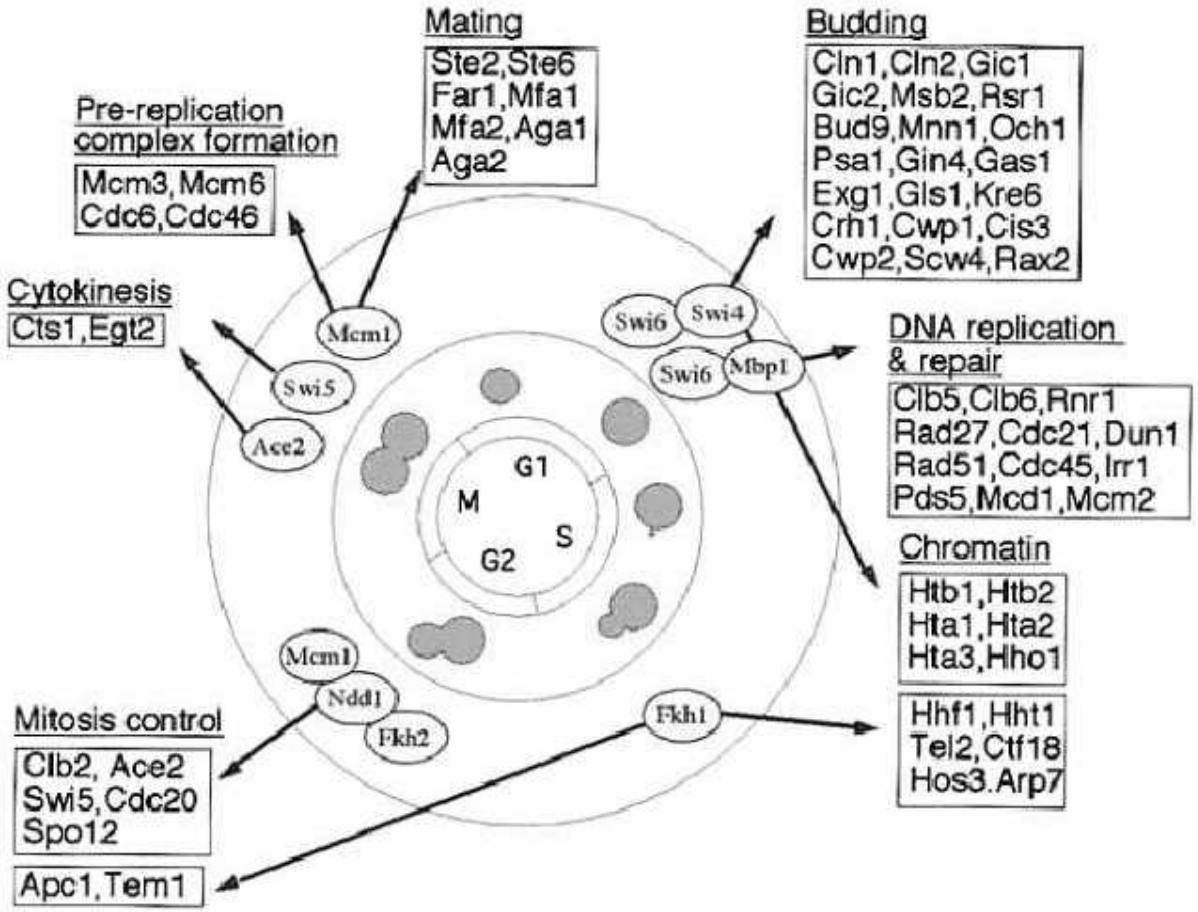


Summary

	Uncorrelated/ Small Variance	Uncorrelated/ Large Variance	Correlated/ Small Variance	Correlated/ Large Variance
Pearson	OK	OK	False Negatives	False
Eisen	OK	False Positives	OK	False
Shrinkage	OK	OK	OK	False

Yeast Cell Cycle

Yeast Cell Cycle



Clusters based on Transcriptional Activators

Reduced table of targets of cell cycle activators, based on the availability of genes in our data set.

	Activators	Genes	Functions
1	Swi4, Swi6	Cln1, Cln2, Gic1, Gic2, Msb2, Rsr1, Bud9, Mnn1, Och1, Exg1, Kre6, Cwp1	Budding
2	Swi6, Mbp1	Clb5, Clb6, Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45, Mcm2	DNA replication and repair
3	Swi4, Swi6	Htb1, Htb2, Hta1, Hta2, Hta3, Hho1	Chromatin
4	Fkh1	Hhf1, Hht1, Tel2, Arp7	Chromatin
5	Fkh1	Tem1	Mitosis Control
6	Ndd1, Fkh2, Mcm1	Clb2, Ace2, Swi5, Cdc20	Mitosis Control
7	Ace2, Swi5	Cts1, Egt2	Cytokinesis
8	Mcm1	Mcm3, Mcm6, Cdc6, Cdc46	Pre-replication complex formation
9	Mcm1	Ste2, Far1	Mating

Clustering Method used for Yeast Data

Hierarchical clustering pseudocode

Given $\{\{X_{ij}\}_{i=1}^{N}\}_{j=1}^{M}$:

Switch:

Pearson: $\gamma = 1$;

Eisen: $\gamma = 0$;

Shrinkage: {

Compute $W = (M-2) \Big/ \sum_{j=1}^{M} \overline{X}_{\cdot j}{}^2$

Compute $\widehat{\beta^2} = \sum_{j=1}^{M} \sum_{i=1}^{N} \left(X_{ij} - \overline{X}_{\cdot j}\right)^2 \Big/ (M(N-1))$

$\gamma = 1 - W \cdot \widehat{\beta^2}/N$

}

While (# clusters > 1) do

- ◇ Compute similarity table:
  $S(G_j, G_k) = \frac{\sum_i \left(G_{ij} - (G_j)_{offset}\right)\left(G_{ik} - (G_k)_{offset}\right)}{\sqrt{\sum_i \left(G_{ij} - (G_j)_{offset}\right)^2 \cdot \sum_i \left(G_{ik} - (G_k)_{offset}\right)^2}}$,
  where $(G_\ell)_{offset} = \gamma \overline{G_\ell}$.
- ◇ Find (j^*, k^*) :
  $S(G_{j^*}, G_{k^*}) \geq S(G_j, G_k)$
  $\forall$ clusters j, k
- ◇ Create new cluster $N_{j^*k^*}$
  = weighted average of
  G_{j^*} and G_{k^*}.
- ◇ Take out clusters j^* and k^*.

Clusters based on Eisen *et al.*

RN Subsampled Data, Eisen clusters ($\gamma = 0.0$)		
E58	Swi4/Swi6	Cln1, Och1
E68	Swi4/Swi6	Cln2, Msb2, Rsr1, Bud9, Mnn1, Exg1
	Swi6/Mbp1	Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45, Mcm2
	Swi4/Swi6	Htb1, Htb2, Hta1, Hta2, Hho1
	Fkh1	Hhf1, Hht1, Arp7
	Fkh1	Tem1
	Ndd1/Fkh2/Mcm1	Clb2, Ace2, Swi5
	Ace2/Swi5	Egt2
	Mcm1	Mcm3, Mcm6, Cdc6
E29	Swi4/Swi6	Gic1
E64	Swi4/Swi6	Gic2
E33	Swi4/Swi6	Kre6, Cwp1
	Swi6/Mbp1	Clb5, Clb6
	Swi4/Swi6	Hta3
	Ndd1/Fkh2/Mcm1	Cdc20
	Mcm1	Cdc46
E73	Fkh1	Tel2
E23	Ace2/Swi5	Cts1
E43	Mcm1	Ste2
E66	Mcm1	Far1

Clusters based on Pearson

RN Subsampled Data, Pearson clusters ($\gamma = 1.0$)		
P1	Swi4/Swi6	Cln1, Och1
P15	Swi4/Swi6 Swi6/Mbp1 Mcm1	Cln2, Rsr1, Mnn1 Cdc21, Dun1, Rad51, Cdc45, Mcm2 Mcm3
P29	Swi4/Swi6	Gic1
P2	Swi4/Swi6	Gic2
P3	Swi4/Swi6 Swi6/Mbp1	Msb2, Exg1 Rnr1
P51	Swi4/Swi6 Ndd1/Fkh2/Mcm1 Ace2/Swi5 Mcm1	Bud9 Clb2, Ace2, Swi5 Egt2 Cdc6
P11	Swi4/Swi6	Kre6
P62	Swi4/Swi6 Swi6/Mbp1 Swi4/Swi6 Ndd1/Fkh2/Mcm1 Mcm1	Cwp1 Clb5, Clb6 Hta3 Cdc20 Cdc46
P49	Swi6/Mbp1 Swi4/Swi6 Fkh1	Rad27 Htb1, Htb2, Hta1, Hta2, Hho1 Hhf1, Hht1
P10	Fkh1 Mcm1	Tel2 Mcm6
P23	Fkh1	Arp7
P50	Fkh1	Tem1
P69	Ace2/Swi5	Cts1
P42	Mcm1	Ste2
P13	Mcm1	Far1

Clusters based on Shrinkage

RN Subsampled Data, Shrinkage clusters (here, $\gamma = 0.66$)		
S49	Swi4/Swi6 Ace2/Swi5 Mcm1	Cln1, Bud9, Och1 Egt2 Cdc6
S6	Swi4/Swi6 Swi6/Mbp1	Cln2, Gic2, Msb2, Rsr1, Mnn1, Exg1 Rnr1, Rad27, Cdc21, Dun1, Rad51, Cdc45
S32	Swi4/Swi6	Gic1
S65	Swi4/Swi6 Swi6/Mbp1 Fkh1 Ndd1/Fkh2/Mcm1 Mcm1	Kre6, Cwp1 Clb5, Clb6 Tel2 Cdc20 Cdc46
S15	Swi6/Mbp1 Mcm1	Mcm2 Mcm3
S11	Swi4/Swi6 Fkh1	Htb1, Htb2, Hta1, Hta2, Hho1 Hhf1, Hht1
S60	Swi4/Swi6	Hta3
S30	Fkh1 Ndd1/Fkh2/Mcm1	Arp7 Clb2, Ace2, Swi5
S62	Fkh1	Tem1
S53	Ace2/Swi5	Cts1
S14	Mcm1	Mcm6
S35	Mcm1	Ste2
S36	Mcm1	Far1

Comparison of Results

Hypothesis: Genes expressed during the same cell cycle stage, and regulated by the same transcriptional activators should be in the same cluster.

Deviations from hypothesis:
Possible False Positives:

- Bud9(1) + Egt2(7) + Cdc6(8): in E68, P51, and S49.
- Mcm2(2) + Mcm3(8): in E68, P15, and S15.
- {Cln2, Rsr1, Mnn1}(1) + {Cdc21, Dun1, Rad51, Cdc45}(2): in E68, P15, and S6.
- {Htb1, Htb2, Hta1, Hta2, Hho1}(3) + {Hhf1, Hht1}(4): in E68, P49, and S11.
- In addition, E68 also contains Tem1(5) and {Clb5, Ace2, Swi5}(6).

Possible False Negatives: Group 1 (Budding) is split into

- 5 clusters by Eisen,
- 8 clusters by Pearson, and
- 4 clusters by Shrinkage.

Notation for Cluster comparison

◊ Each cluster set can be written as follows:

$$\left\{x \rightarrow \{\{y_1, z_1\}, \{y_2, z_2\}, \ldots, \{y_{n_x}, z_{n_x}\}\}\right\}_{x=1}^{\#\text{ of groups}}$$

- x denotes the group number,
- n_x is the number of clusters group x appears in, and
- for each cluster $j \in \{1, \ldots, n_x\}$ there are
  - ◇ y_j genes from group x and
  - ◇ z_j genes from other groups.

Eisen, Shrinkage, and Pearson clusters in Set Notation

$\gamma = 0.0 (E) \Longrightarrow$

$\{1 \rightarrow \{\{6,23\},\{2,0\},\{2,5\},\{1,0\},\{1,0\}\},$
$2 \rightarrow \{\{7,22\},\{2,5\}\},$
$3 \rightarrow \{\{5,24\},\{1,6\}\},$
$4 \rightarrow \{\{3,26\},\{1,0\}\},$
$5 \rightarrow \{\{1,28\}\},$
$6 \rightarrow \{\{3,26\},\{1,6\}\},$
$7 \rightarrow \{\{1,0\},\{1,28\}\},$
$8 \rightarrow \{\{3,26\},\{1,6\}\},$
$9 \rightarrow \{\{1,0\},\{1,0\}\}\}$

$\gamma = 0.66 (S) \Longrightarrow$

$\{1 \rightarrow \{\{6,6\},\{3,2\},\{2,5\},\{1,0\}\},$
$2 \rightarrow \{\{6,6\},\{2,5\},\{1,1\}\},$
$3 \rightarrow \{\{5,2\},\{1,0\}\},$
$4 \rightarrow \{\{2,5\},\{1,3\},\{1,6\}\},$
$5 \rightarrow \{\{1,0\}\},$
$6 \rightarrow \{\{3,1\},\{1,6\}\},$
$7 \rightarrow \{\{1,0\},\{1,4\}\},$
$8 \rightarrow \{\{1,0\},\{1,1\},\{1,4\},\{1,6\}\},$
$9 \rightarrow \{\{1,0\},\{1,0\}\}\}$

$\gamma = 1.0 (P) \Longrightarrow$

$\{1 \rightarrow \{\{3,6\},\{2,0\},\{2,1\},\{1,0\},\{1,0\},\{1,0\},\{1,5\},\{1,5\}\},$
$2 \rightarrow \{\{5,4\},\{2,4\},\{1,2\},\{1,7\}\},$
$3 \rightarrow \{\{5,3\},\{1,5\}\},$
$4 \rightarrow \{\{2,6\},\{1,0\},\{1,1\}\},$
$5 \rightarrow \{\{1,0\}\},$
$6 \rightarrow \{\{3,3\},\{1,5\}\},$
$7 \rightarrow \{\{1,0\},\{1,5\}\},$
$8 \rightarrow \{\{1,1\},\{1,5\},\{1,5\},\{1,8\}\},$
$9 \rightarrow \{\{1,0\},\{1,0\}\}\}$

Scoring Function

◊ Each cluster set can be scored according to:

$$\begin{aligned} \text{FP}(\gamma) &= \frac{1}{2}\sum_{x}\sum_{j=1}^{n_x} y_j \cdot z_j \\ \text{FN}(\gamma) &= \sum_{x}\sum_{1\le j<k\le n_x} y_j \cdot y_k \\ \text{Error\_score}(\gamma) &= \text{FP}(\gamma) + \text{FN}(\gamma) \end{aligned}$$

◊ For previously listed cluster sets:

- Error_score(0.0) = 370 + 79 = 449 (Eisen)
- Error_score(0.66) = 76 + 88 = 164 (Shrinkage)
- Error_score(1.0) = 69 + 107 = 176 (Pearson)

Choice of Cut-off Threshold

A Receiver Operator Characteristic (ROC) curve plots sensitivity against (1−specificity), with the curve parametrized by the cut-off threshold in the range of $[-1, 1]$. Here,

Sensitivity = fraction of positives detected by a metric

$$= \frac{\mathrm{TP}(\gamma)}{\mathrm{TP}(\gamma) + \mathrm{FN}(\gamma)},$$

Specificity = fraction of negatives detected by a metric

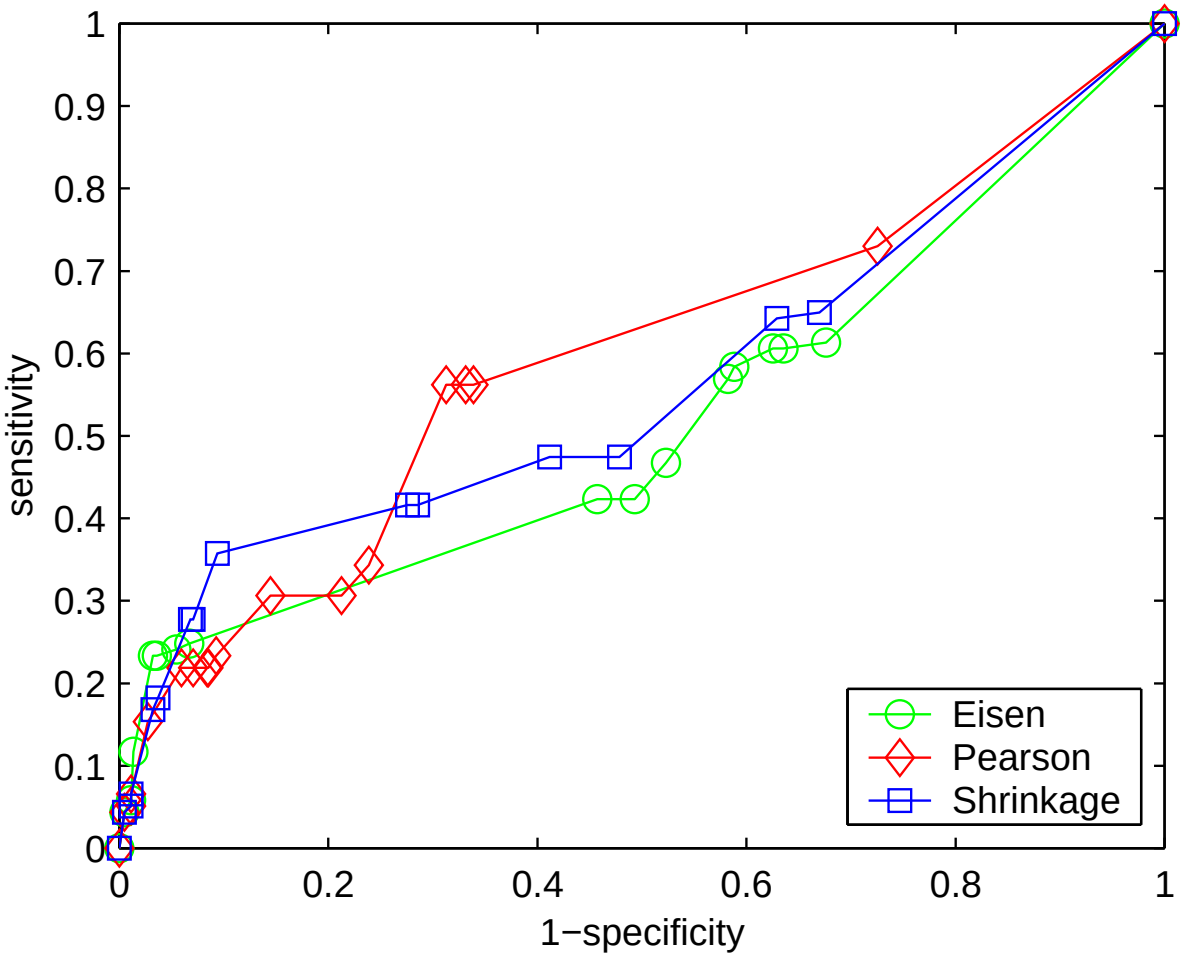
$$= \frac{\mathrm{TN}(\gamma)}{\mathrm{TN}(\gamma) + \mathrm{FP}(\gamma)},$$

ROC Definitions (cont'd)

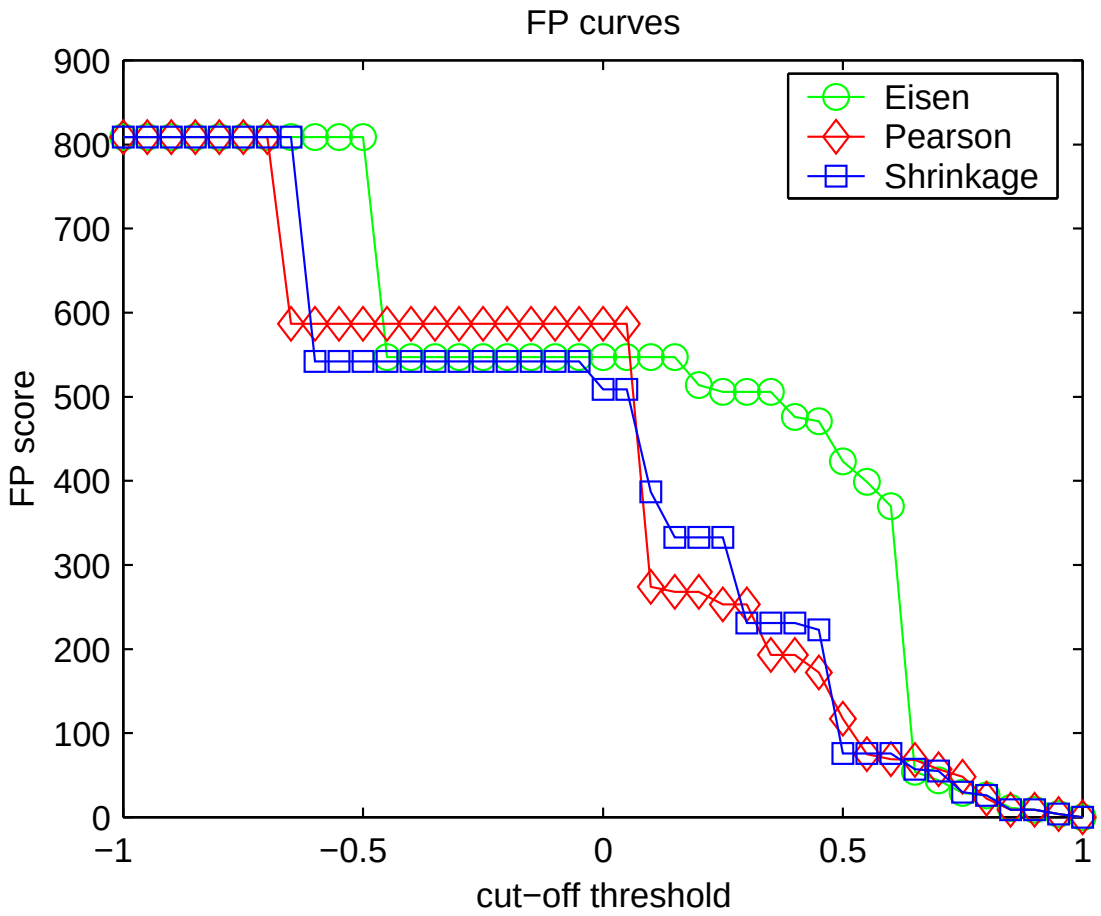
TP(γ), FN(γ), FP(γ), and TN(γ) denote the number of True Positives, False Negatives, False Positives, and True Negatives, respectively, arising from a metric associated with a given γ.

- FP(γ) and FN(γ) defined under scoring function
- $\mathrm{TP}(\gamma) = \sum_{x} \sum_{j=1}^{n_x} \binom{y_j}{2}$
- $\mathrm{TN}(\gamma) = \mathrm{Total} - (\mathrm{TP}(\gamma) + \mathrm{FN}(\gamma) + \mathrm{FP}(\gamma))$
- Total= $\binom{44}{2} = 946$ is the total # of gene pairs $\{j, k\}$ in Transcriptional Activator table.

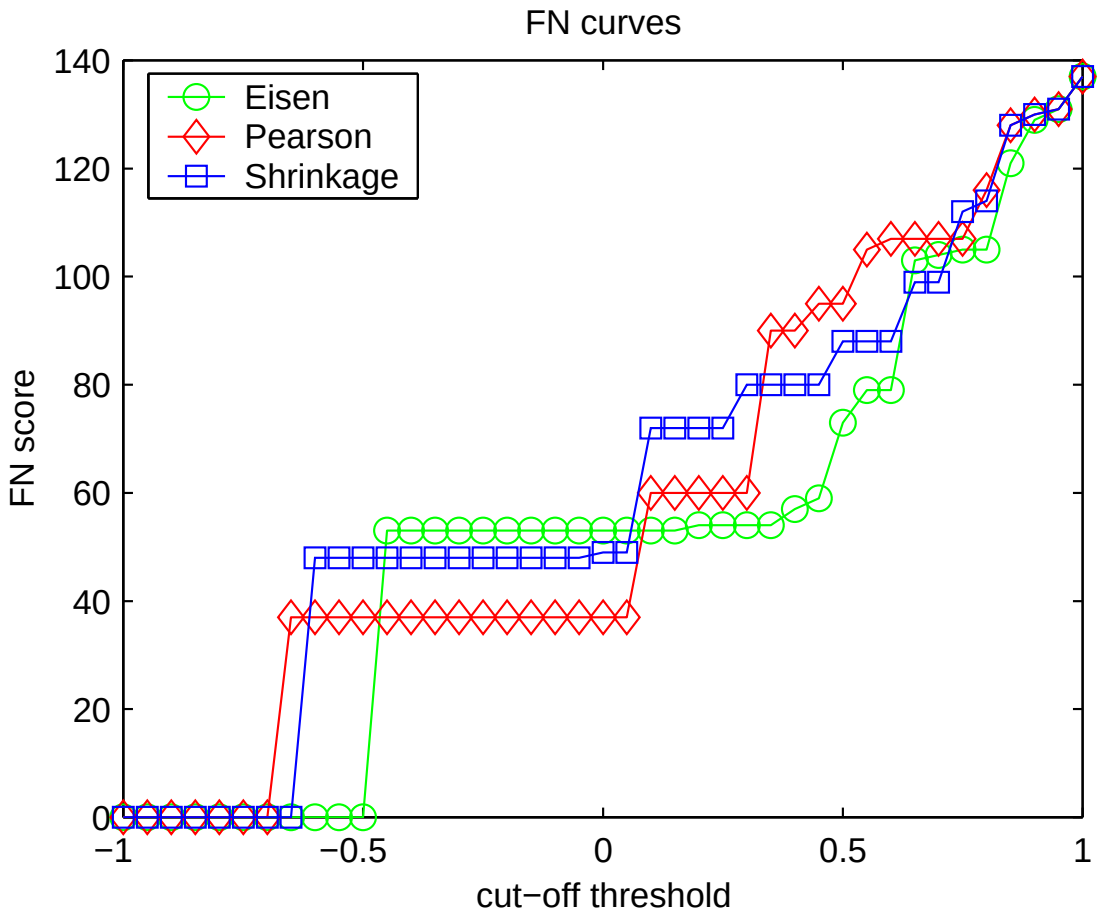
ROC Curves



FP count as a function of threshold



FN count as a function of threshold



References

◊ Full technical report:
Cherepinsky, V., Feng, J., Rejali, M., and Mishra, B. (2003)
(to be published on NCSTRL,
PDF available for download from
http://www.cs.nyu.edu/cs/faculty/mishra/)

◊ Yeast cell cycle Transcriptional Activators data:
Simon, I. *et al.* (2001), *Cell* **106**, 697–708.

◊ Stein Estimation - a recent review:
Hoffman, K. (2000), *Statistical Papers*, **41(2)**, 127–158.

Derivation

Pearson Correlation Coefficient

◊ Random Variables: X & Y

- $X = (X_1, \ldots, X_N) = \{X_i\}_{i=1}^N$; $\mu_X = \left(\sum_i X_i\right) \Big/ N$; ${\sigma_X}^2 = \left(\sum_i (X_i - \mu_X)^2\right) \Big/ N$
- $Y = (Y_1, \ldots, Y_N) = \{Y_i\}_{i=1}^N$; $\mu_Y = \left(\sum_i Y_i\right) \Big/ N$; ${\sigma_Y}^2 = \left(\sum_i (Y_i - \mu_Y)^2\right) \Big/ N$

◊ $S(X,Y) = \frac{1}{N} \sum_i (X_i - \mu_X)(Y_i - \mu_Y)/(\sigma_X \sigma_Y) =$ R.V.

◊ $S(X,Y) =$ Ratio of two χ^2 distributions, and hence an F distribution. Its variance depends on N.

- Its statistical significance can be estimated from the distributions of X & Y and hence, it is a function of N.
- For small values of N (e.g., 100), its statistical significance is poor.
- Prior beliefs about μ_X and μ_Y can improve the reliability of $S(X,Y)$. E.g., $\mu_X \approx 0$ and $\mu_Y \approx 0$.

◊ This argument suggests a Bayesian approach that accounts for prior knowledge.

Bayesian Analysis

NYU SHRINK

Bayesian Approach

◊ Given:

$$\left\{\left\{X_{ij}\right\}_{i=1}^{N}\right\}_{j=1}^{M}, \quad \text{where } M \gg N$$

are data points.

– $\left\{X_{ij}\right\}_{i=1}^{N}$ is a data vector for

– gene j $(1 \leq j \leq M)$, corresponding to

– N experimental conditions: $1 \leq i \leq N$.

Prior Belief

◊ A prior belief:

- $\{X_{ij}\}_{i=1}^{N} \sim \mathcal{N}(\theta_j, {\beta_j}^2)$,
- where $\theta_j \sim \mathcal{N}(0, \tau^2)$.

$\Longrightarrow$ Prior distribution of θ_j is given by:

$$\pi(\theta_j) = \frac{1}{\sqrt{2\pi}\tau} \exp\left(-{\theta_j}^2/2\tau^2\right)$$

◊ We wish to obtain the posterior distribution of θ_j, $\pi(\theta_j|X)$.

◊ From the posterior distribution we compute $\mathbf{E}_X(\theta_j)$.

Bayes' Theorem

- $p(\theta|y)p(y) = p(\theta, y) = p(\theta)p(y|\theta)$
- $p(\theta|y) = cp(y|\theta)p(\theta) \propto l(\theta|y)p(\theta) = f(\theta|y)$
- $$p(\theta|y) = f(\theta|y) \Big/ \left[\int_{-\infty}^{\infty} f(\theta'|y)d\theta'\right]$$
  - where $f(\theta|y) \propto p(y|\theta)p(\theta)$

Combining a Normal Prior with a Normal Likelihood

◊ Assume two random variables θ and y:

- Assume their variances are known. . . (An assumption that will have to be relaxed subsequently.)

- Suppose *a priori* θ is distributed as $\theta \sim \mathcal{N}(\theta_0, {\sigma_0}^2)$.

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\left((\theta - \theta_0)/\sigma_0\right)^2/2\right], \quad -\infty < \theta < \infty$$

- The likelihood function of θ is proportional to a normal function: $y \sim \mathcal{N}(\theta, {\sigma_1}^2)$

$$l(\theta|y) \propto \exp\left[-\left((\theta - x)/\sigma_1\right)^2/2\right],$$

where x is some function of the variable y.

Posterior Distribution

◊ The posterior distribution of θ given the data y is

$$p(\theta|y) = p(\theta)l(\theta|y) \Big/ \int_{-\infty}^{\infty} p(\theta')l(\theta'|y)d\theta'$$
$$= f(\theta|y) \Big/ \int_{-\infty}^{\infty} f(\theta'|y)d\theta' \,, \quad -\infty < \theta < \infty,$$

where

$$f(\theta|y) \propto p(\theta) \cdot l(\theta|y)$$
$$\propto (1/\sqrt{2\pi}\sigma_0) \exp\left[-\left((\theta-\theta_0)/\sigma_0\right)^2/2\right] \times \exp\left[-\left((\theta-x)/\sigma_1\right)^2/2\right]$$

◊ Simplify. . .

$$f(\theta|y) = \exp\left[-\frac{1}{2}\{((\theta-\theta_0)/\sigma_0)^2 + ((\theta-x)/\sigma_1)^2\}\right]$$

Simplification

◊ Now use the following identity:

$$A(z-a)^2 + B(z-b)^2 = (A+B)(z-c)^2 + \frac{AB}{A+B}(a-b)^2,$$

where

$$c = \frac{Aa+Bb}{A+B}$$

◊ The critical parameter c is simply the weighted average of a and b with weights A and B, respectively.

Final Result

◊ It follows that

$$[(\theta-\theta_0)/\sigma_0]^2+[(\theta-x)/\sigma_1]^2$$
$$=\left(1/{\sigma_0}^2+1/{\sigma_1}^2\right)(\theta-\theta_X)^2+\text{Terms independent of }\theta\ldots$$

◊ Thus

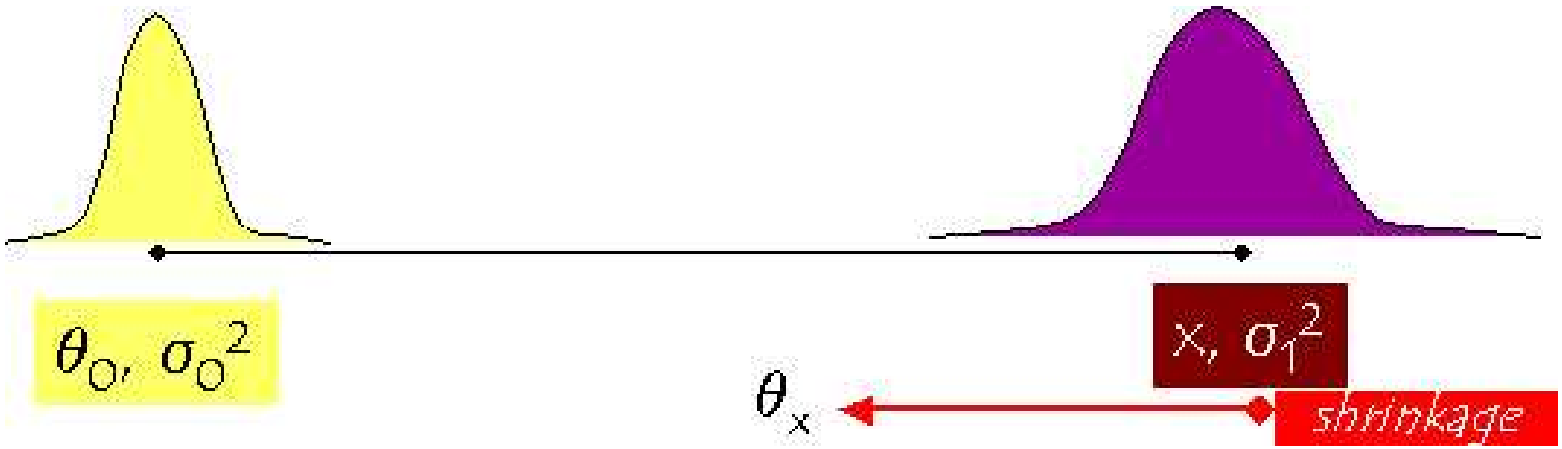
$$\begin{aligned}\theta_X &= \left(\theta_0/{\sigma_0}^2+x/{\sigma_1}^2\right)/\left(1/{\sigma_0}^2+1/{\sigma_1}^2\right)\\ &= \left({\sigma_1}^2\theta_0+{\sigma_0}^2x\right)/\left({\sigma_1}^2+{\sigma_0}^2\right)\end{aligned}$$

◊ Since ${\sigma_0}^2>0$ and ${\sigma_1}^2>0$, we have $\theta_0\le\theta_X\le x$.

– If ${\sigma_0}^2\gg{\sigma_1}^2$ (there is more uncertainty in θ_0 than in x), then $\theta_X\approx x$. . . In other words, if our observation is much better than our prior belief, then we put more weight on our observation.

– Conversely, if ${\sigma_1}^2\gg{\sigma_0}^2$, then $\theta_X\approx\theta_0$. Put more trust in our prior beliefs than the observation.

Shrinking



$$\theta_X = \left(\theta_0/{\sigma_0}^2 + x/{\sigma_1}^2\right) / \left(1/{\sigma_0}^2 + 1/{\sigma_1}^2\right) = \left({\sigma_1}^2\theta_0 + {\sigma_0}^2 x\right) / \left({\sigma_1}^2 + {\sigma_0}^2\right)$$

◊ A simpler form

$$\theta_X = \left[1 - \left\{{\sigma_1}^2/({\sigma_1}^2 + {\sigma_0}^2)\right\}\{1 - \theta_0/x\}\right] x$$

◊ The observation x is "shrunk" towards the belief θ_0.

◊ The estimator swaps "bias" for "variance".

Recall our Model

◊ Prior belief:

- $\{X_{ij}\}_{i=1}^{N} \sim \mathcal{N}(\theta_j, {\beta_j}^2)$,
- where $\theta_j \sim \mathcal{N}(0, \tau^2)$.

◊ Thus $p(\theta_j|\{X_{ij}\}_{i=1}^{N}) \sim \mathcal{N}(\theta_{jX}, {\sigma_{jX}}^2)$

$$\begin{aligned} \theta_{jX} &= \left[1 - ({\beta_j}^2/N)/({\beta_j}^2/N + \tau^2)\right] \mathbf{E}[X_{.j}] \\ {\sigma_{jX}}^2 &= {\beta_j}^2/(N + {\beta_j}^2/\tau^2) \end{aligned}$$

◊ $S(X_j, X_k) = \frac{1}{N}\sum_i (X_{ij} - \theta_{jX})(X_{ik} - \theta_{kX})/(\sigma_{jX}\sigma_{kX})$

James-Stein Estimator

◊ $\theta_{jX} = \left[1 - (\beta_j{}^2/N)/(\beta_j{}^2/N + \tau^2)\right] \mathbf{E}[X_{\cdot j}]$

– But since neither $\beta_j{}^2$ nor τ^2 are known a priori, *they have to be estimated*.

◊ Note that $\mathbf{E}[X_{\cdot j}] \sim \mathcal{N}(\theta_j, \beta_j{}^2/N)$, and hence

$$Q = \sum_{j=1}^{M} \mathbf{E}[X_{\cdot j}]^2 \sim \left(\tau^2 + \beta_j{}^2/N\right) \chi^2_M$$

– Thus, $\mathbf{E}\left[(M-2)/\sum_{j=1}^{M} \mathbf{E}[X_{\cdot j}]^2\right]$ is an unbiased estimator for $1/(\beta_j{}^2/N + \tau^2)$.

◊ Similarly: $S = \sum_{i=1}^{N} (X_{ij} - \mathbf{E}[X_{\cdot j}])^2 \sim \beta_j{}^2 \chi^2_N$.

– Thus, $\mathbf{E}\left[S/N(N+2)\right] = \beta_j{}^2/N$.

James-Stein Estimator: Final Form

$$\theta_{jX} = \left[1 - \frac{M-2}{N(N+2)} \frac{S}{Q}\right] \mathbf{E}[X_{.j}]$$

where

$$S = \sum_{i=1}^{N} \left(X_{ij} - \mathbf{E}[X_{.j}]\right)^2 \sim {\beta_j}^2 \chi^2_N$$

$$Q = \sum_{j=1}^{M} \mathbf{E}[X_{.j}]^2 \sim \left(\tau^2 + {\beta_j}^2/N\right) \chi^2_M$$