Shrinkage-Based Similarity Metric for Cluster Analysis of Microarray Data *

Vera Cherepinsky[1], Jiawu Feng[1], Marc Rejali[1], *and* Bud Mishra[1,2†]

[1] Courant Institute, New York University, 251 Mercer Street, New York, NY 10012; and
[2] Cold Spring Harbor Lab, 1 Bungtown Road, Cold Spring Harbor, NY 11724.

June 13, 2003

ABSTRACT

The current standard correlation coefficient used in the analysis of microarray data was introduced in [1]. Its formulation is rather arbitrary. We give a mathematically rigorous correlation coefficient of two data vectors based on James-Stein Shrinkage estimators. We use the assumptions described in [1], also utilizing the fact that the data can be treated as transformed into normal distributions. While [1] uses zero as an estimator for the expression vector mean μ, we start with the assumption that for each gene, μ is itself a zero-mean normal random variable (with *a priori* distribution $\mathcal{N}(0, \tau^2)$), and use Bayesian analysis to obtain *a posteriori* distribution of μ in terms of the data. The shrunk estimator for μ differs from the mean of the data vectors and ultimately leads to a statistically robust estimator for correlation coefficients.

To evaluate the effectiveness of shrinkage, we conducted *in silico* experiments and also compared similarity metrics on a biological example using the data set from [1]. For the latter, we classified genes involved in the regulation of yeast cell cycle functions by computing clusters based on various definitions of correlation coefficients and contrasting them against clusters based on the activators known in the literature.

The estimated "false-positives" and "false-negatives" from this study indicate that using the shrinkage metric improves the accuracy of the analysis.

[1] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998), *Proc. Natl. Acad. Sci. USA* 95, 14863–14868.

1 Background

Recent advances in technology, such as microarray-based gene expression analysis, have allowed us to "look inside the cells" by quantifying their transcriptional states. While the most interesting insight can be obtained from transcriptome abundance data within a single cell under different experimental conditions, in the absence of technology to provide one with such a detailed picture, we have to make do with mRNA collected from a small, frequently unsynchronized, population of cells. Furthermore, these mRNAs will only give a partial picture, supported only by those genes that we are already familiar with and possibly missing many crucial undiscovered genes.

Of course, without the proteomic data, transcriptomes tell less than half the story. Nonetheless, it goes without saying that microarrays have already revolutionized our understanding of biology even though they only provide occasional, noisy, unreliable, partial, and occluded snapshots of the transcriptional states of cells.

In an attempt to attain functional understanding of the cell, we try to understand the underlying structure of its transcriptional state-space. Partitioning genes into closely related groups has thus become the key mathematical first step in practically all statistical analyses of microarray data.

Traditionally, algorithms for cluster analysis of genome-wide expression data from DNA microarray hybridization

*This research was conducted under the support of NSF's Qubic program, DARPA, HHMI biomedical support research grant, the US Department of Energy, the US Air Force, National Institutes of Health, the National Cancer Institute, New York State Office of Science, Technology & Academic Research, the NSF Graduate Research Fellowship, and the NYU McCracken Fellowship.

†To whom correspondence should be addressed. E-mail: mishra@nyu.edu

are based upon statistical properties of gene expressions and result in organizing genes according to similarity in pattern of gene expression. If two genes belong to a cluster then one may infer a common regulatory mechanism for the two genes or interpret this information as an indication of the status of cellular processes. Furthermore, coexpression of genes of known function with novel genes may lead to a discovery process for characterizing unknown or poorly characterized genes. In general, since false-negatives (where two coexpressed genes are assigned to distinct clusters) may cause the discovery process to ignore useful information for certain novel genes, and false-positives (where two independent genes are assigned to the same cluster) may result in noise in the information provided to the subsequent algorithms used in analyzing regulatory patterns, it is important that the statistical algorithms for clustering be reasonably robust. Unfortunately, as the microarray experiments that can be carried out in an academic laboratory for a reasonable cost are small in number and suffer from experimental noise, often a statistician must resort to unconventional algorithms to deal with small-sample data.

A popular and one of the earliest clustering algorithms reported in the literature was introduced in [1]. In this paper, the gene-expression data were collected on spotted DNA microarrays [2] and were based upon gene expression in the budding yeast *Saccharomyces cerevisiae* during the diauxic shift [3], the mitotic cell division cycle [4], sporulation [5], and temperature and reducing shocks. Each entry in a gene expression vector represents a ratio of the amount of transcribed mRNA under a particular condition with respect to its value under normal conditions. All ratio values are log-transformed to treat inductions and repressions of identical magnitude as numerically equal but opposite in sign. It is assumed that the raw ratio values follow log-normal distributions, and hence, the log-transformed data follow normal distributions. While our mathematical derivations will rely on this assumption for the sake of simplicity, we note that our approach can be generalized in a straightforward manner to deal with other situations where this assumption is violated.

The gene similarity metric employed in [1] was a form of correlation coefficient. Let G_i be 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, the classical similarity score based upon Pearson correlation coefficient is:

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

where

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

and G_{offset} is the estimated mean of the observations, i.e.,

$$G_{offset} = \bar{G} = \frac{1}{N}\sum_{i=1}^{N} G_i.$$

Note that Φ_G is simply the (rescaled) estimated standard deviation of the observations. In the analysis presented in [1], "values of G_{offset} which are not the average over observations on G *were* used when there was an assumed unchanged or reference state represented by the value of G_{offset}, against which changes were to be analyzed; in all of the examples presented there, G_{offset} was set to 0, corresponding to a fluorescence ratio of 1.0." To distinguish this modified correlation coefficient from the classical Pearson correlation coefficient, we shall refer to it as Eisen correlation coefficient. Our main innovation is in suggesting a different value for G_{offset}, namely $G_{offset} = \gamma\bar{G}$, where γ is allowed to take a value between 0.0 and 1.0. Note that when $\gamma = 1.0$, we have the classical Pearson correlation coefficient and when $\gamma = 0.0$, we have replaced it by Eisen correlation coefficient. For a non-unit value of γ, the estimator for $G_{offset} = \gamma\bar{G}$ can be thought of as the unbiased estimator $\bar{G}$ being shrunk towards the believed value for $G_{offset} = 0.0$. We address the following questions: What is the optimal value for the shrinkage parameter γ from a Bayesian point of view? (See [6] for some alternate approaches.) How do the gene expression data cluster as the correlation coefficient is modified with this optimal shrinkage parameter?

In order to achieve a consistent comparison, we leave the rest of the algorithms undisturbed. Namely, once the similarity measure has been assumed, we cluster the genes using the same hierarchical clustering algorithm as the one used by Eisen *et al.* Their hierarchical clustering algorithm is based on the centroid-linkage method (referred to as "average-linkage method" of Sokal and Michener [7] in [1]) and is discussed further in section 3. The modified algorithm has been implemented by the authors within the "NYUMAD" microarray database system and can be freely downloaded from: `http://bioinformatics.cat.nyu.edu/nyumad/clustering/`. The clusters created in this manner were used to compare the effects of choosing differing similarity measures.

2 Model

Recall that equations (1) and (2) define a correlation coefficient $S(X,Y)$ and the corresponding estimated standard deviation Φ, respectively, and let

$$G_{offset} = \gamma \bar{G} \quad \text{for} \quad G \in \{X, Y\}.$$

A family of such correlation coefficients can be parametrized by $0 \leq \gamma \leq 1$.

- *Pearson Correlation Coefficient* uses

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

- *Eisen et al.* (in [1]) use

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

- 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).

Next, we derive the proposed similarity metric. In our setup, the microarray data is given in the form of the levels of M genes expressed under N experimental conditions. The data can be viewed as

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

where $M \gg N$ and $\{X_{ij}\}_{i=1}^{N}$ is the data vector for gene j.

We begin by rewriting S in our notation:

$$\begin{aligned} & S(X_j, X_k) \qquad (3) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_{ij} - (X_j)_{offset}}{\Phi_j}\right)\left(\frac{X_{ik} - (X_k)_{offset}}{\Phi_k}\right), \\ & {\Phi_j}^2 = \frac{1}{N}\sum_{i}\Big(X_{ij} - (X_j)_{offset}\Big)^2 \end{aligned}$$

In the most general setting, we can make the following assumptions on the data distribution: let all values X_{ij} for gene j have a Normal distribution with mean θ_j and standard deviation β_j (variance ${\beta_j}^2$); i.e.,

$$X_{ij} \sim \mathcal{N}(\theta_j, {\beta_j}^2) \quad \text{for} \quad i = 1, \ldots, N$$

with j fixed ($1 \leq j \leq M$), where θ_j is an unknown parameter (taking different values for different j). To estimate θ_j, it is convenient to assume that θ_j is itself a random variable taking values close to zero:

$$\theta_j \sim \mathcal{N}(0, \tau^2).$$

The assumed distribution aids us in obtaining the estimate of θ_j given in (6).

For convenience, let us also assume that the data are range-normalized, so that ${\beta_j}^2 = \beta^2$ for every j. If this assumption does not hold on the given data set, it is easily corrected by scaling each gene vector appropriately. Following common practice, we adjusted the range to scale to an interval of unit length, i.e., its maximum and minimum values differ by 1. Thus,

$$X_{ij} \sim \mathcal{N}(\theta_j, \beta^2) \quad \text{and} \quad \theta_j \sim \mathcal{N}(0, \tau^2).$$

Replacing $(X_j)_{offset}$ in (3) by the exact value of the mean θ_j yields a *Clairvoyant* correlation coefficient of X_j and X_k. In reality, since θ_j is itself a random variable, it must be estimated from the data. Therefore, to get an explicit formula for $S(X_j, X_k)$ we must derive estimators $\widehat{\theta_j}$ for all j.

In Pearson correlation coefficient, θ_j is estimated by the vector mean $\overline{X}_{\cdot j}$; Eisen correlation coefficient corresponds to replacing θ_j by 0 for every j, which is equivalent to assuming $\theta_j \sim \mathcal{N}(0,0)$ (i.e., $\tau^2 = 0$.) We propose to find an estimate of θ_j (call it $\widehat{\theta_j}$) that takes into account both the prior assumption and the data.

First, let us obtain the posterior distribution of θ_j from the prior $\mathcal{N}(0, \tau^2)$ and the data. This derivation can be done either from the Bayesian considerations, or via the James-Stein Shrinkage estimators (see [8], or [9] for a more recent review). Here, we discuss the former method.

Assume initially that $N = 1$, i.e., we have one data point for each gene, and denote the variance by σ^2 for the moment:

$$X_j \sim \mathcal{N}(\theta_j, \sigma^2) \quad \text{and} \quad \theta_j \sim \mathcal{N}(0, \tau^2).$$

From these assumptions, we get (see [10] for full details)

$$\begin{aligned} \mathbf{E}(\theta_j | X_j) &= \frac{\tau^2}{\sigma^2 + \tau^2}\, X_j \\ &= \left(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right) X_j, \qquad (4) \\ Var(\theta_j | X_j) &= \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}. \end{aligned}$$

Now, if $N > 1$ is arbitrary, X_j becomes a vector $X_{\cdot j}$. In [10] we show (by using likelihood functions) that the vector of values $\{X_{ij}\}_{i=1}^{N}$, with $X_{ij} \sim \mathcal{N}(\theta_j, \beta^2)$, can be treated as a single data point $Y_j = \overline{X}_{\cdot j} = \sum_{i=1}^{N} X_{ij}/N$ from the distribution $\mathcal{N}(\theta_j, \beta^2/N)$.

Thus, following the same derivation with $\sigma^2 = \beta^2/N$, we have a Bayesian estimator for θ_j given by $\mathbf{E}(\theta_j|X_{\cdot j})$:

$$\widehat{\theta_j} = \left(1 - \frac{\beta^2/N}{\beta^2/N + \tau^2}\right) Y_j. \tag{5}$$

Unfortunately, (5) cannot be used in (3) directly, because τ^2 and β^2 are unknown, so must be estimated from the data. The details of the estimation are presented in [10].

The resulting explicit estimate for θ_j is

$$\begin{aligned} \widehat{\theta_j} & \\ &= \left(1 - W \cdot \frac{\widehat{\beta^2}}{N}\right) Y_j \\ &= \underbrace{\left(1 - \frac{M-2}{MN(N-1)} \cdot \frac{\sum_{k=1}^{M}\sum_{i=1}^{N}(X_{ik} - Y_k)^2}{\sum_{k=1}^{M} {Y_k}^2}\right)}_{\gamma} Y_j \\ &= \gamma \overline{X}_{\cdot j}, \end{aligned} \tag{6}$$

where $W = \frac{M-2}{\sum_{j=1}^{M} {Y_j}^2}$ is an estimator for $1/(\beta^2/N + \tau^2)$.

Finally, we substitute $\widehat{\theta_j}$ from equation (6) into the correlation coefficient in (3) wherever $(X_j)_{offset}$ appears to obtain an explicit formula for $S(X_{\cdot j}, X_{\cdot k})$.

3 Algorithm & Implementation

The implementation of hierarchical clustering proceeds in a greedy manner, always choosing the most similar pair of elements (starting with genes at the bottom-most level) and combining them to create a new element. The "expression vector" for the new element is simply the weighted average of the expression vectors of the two elements that were combined. This structure of repeated pair-wise combinations is conveniently represented in a binary tree, whose leaves are the set of genes and internal nodes are the elements constructed from the two children nodes. The algorithm is described below in pseudocode.

3.1 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^*.

As each internal node can be labeled by a value representing the similarity between its two children nodes, one can create a set of clusters by simply breaking the tree into subtrees by eliminating all the internal nodes with labels below a certain predetermined threshold value.

The implementation of generalized hierarchical clustering with options to choose different similarity measures has been incorporated into NYUMAD (NYU MicroArray Database), an integrated system to maintain and analyze biological abundance data along with associated experimental conditions and protocols. To enable widespread utility, NYUMAD supports the MAGE-ML standard (web site at `http://www.mged.org/Workgroups/MAGE/mage-ml.html`) for the exchange of gene expression data, defined by the Microarray Gene Expression Data (MGED) Group. More detailed information about NYUMAD can be found at `http://bioinformatics.cat.nyu.edu/nyumad/`.

4 Results

4.1 Mathematical Simulation

To compare the performance of these algorithms, we started with a relatively simple *in silico* experiment. In such an experiment, one can create two genes X and Y and simulate N (about 100) experiments as follows:

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

where α_i, chosen from a uniform distribution over a range $[L, H]$ ($\mathcal{U}(L,H)$), is a "bias term" introducing a correlation (or none if all α's are zero) between X and Y. $\theta_X \sim \mathcal{N}(0, \tau^2)$ and $\theta_Y \sim \mathcal{N}(0, \tau^2)$ are the means of X and Y, respectively.

Similarly, σ_X and σ_Y are the standard deviations for X and Y, respectively.

Note that, with this model

$$S(X,Y) = \frac{1}{N}\sum_{i=1}^{N} \frac{(X_i - \theta_X)}{\sigma_X}\frac{(Y_i - \theta_Y)}{\sigma_Y}$$

if the exact values of the mean and variance are used.

The model was implemented in Mathematica [11]; the following parameters were used in the simulation: $N = 100$, $\tau \in \{0.1, 10.0\}$ (representing very low or high variability among the genes), $\sigma_X = \sigma_Y = 10.0$, and $\alpha = 0$ representing no correlation between the genes or $\alpha \sim \mathcal{U}(0,1)$ representing some correlation between the genes. Once the parameters were fixed for a particular *in silico* experiment, the gene-expression vectors for X and Y were generated many thousand times, and for each pair of vectors $S_c(X,Y)$, $S_p(X,Y)$, $S_e(X,Y)$, and $S_s(X,Y)$ were estimated by four different algorithms and further examined to see how the estimators of S varied over these trials. These four different algorithms estimated S according to equations (1), (2) as follows: *Clairvoyant* estimated S_c using the true values of θ_X, θ_Y, σ_X, and σ_Y; *Pearson* estimated S_p using the unbiased estimators $\bar{X}$ and $\bar{Y}$ of θ_X and θ_Y (for X_{offset} and Y_{offset}), respectively; *Eisen* estimated S_e using the value 0.0 as the estimator of both θ_X and θ_Y; and *Shrinkage* estimated S_s using the shrunk biased estimators $\widehat{\theta_X}$ and $\widehat{\theta_Y}$ of θ_X and θ_Y, respectively. In the latter three, the standard deviation was estimated as in (2). The histograms corresponding to these *in silico* experiments can be found in Figure 1. Our observations are summarized in Table 1.

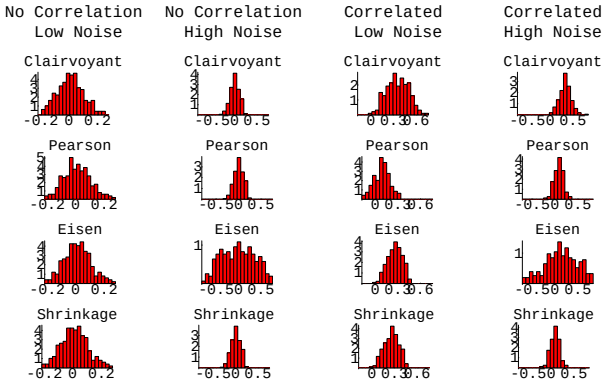

Figure 1: Histograms representing the performance of four different estimators of correlation between genes.

Table 1: Summary of observations from mathematical simulation of gene expression models of correlated and uncorrelated genes. The distributions of S as estimated by S_c (Clairvoyant), S_p (Pearson), S_e (Eisen), and S_s (Shrinkage), are characterized by the means μ and standard deviations δ. When there is no correlation ($\alpha = 0$) and low noise ($\tau = 0.1$), all methods do equally well. When there is no correlation but the noise is high ($\tau = 10$), all methods except Eisen do equally well; Eisen has too many false-positives. When the genes are correlated ($\alpha \sim \mathcal{U}(0,1)$) and the noise is low, all methods except Pearson do equally well; Pearson has too many false-negatives. Finally, when the genes are correlated and the noise is high, all methods do equally poorly, introducing false-negatives; Eisen may also have false-positives.

Params			Distributions			
α	τ		S_c	S_p	S_e	S_s
0	0.1	μ	-0.000297	-0.000269	-0.000254	-0.000254
		δ	0.0996	0.0999	0.0994	0.0994
0	10	μ	-0.000971	-0.000939	-0.00119	-0.000939
		δ	0.0994	0.100	0.354	0.100
$\mathcal{U}(0,1)$	0.1	μ	0.331	0.0755	0.248	0.245
		δ	0.132	0.0992	0.0915	0.0915
$\mathcal{U}(0,1)$	10	μ	0.333	0.0762	0.117	0.0762
		δ	0.133	0.100	0.368	0.0999

In summary, one can conclude that for the same clustering algorithm, Pearson tends to introduce more false-negatives and Eisen tends to introduce more false-positives than Shrinkage. Shrinkage, on the other hand, reduces these errors by combining the good properties of both algorithms.

4.2 Biological Example

We then proceeded to test the algorithms on a biological example. We chose a biologically well-characterized system, and analyzed the clusters of genes involved in the yeast cell cycle. These clusters were computed using the hierarchical clustering algorithm with the underlying similarity measure chosen from the following three: Pearson, Eisen, or Shrinkage. As a reference, the computed clusters were compared to the ones implied by the common cell-cycle functions and regulatory systems inferred from the roles of various transcriptional activators (see Figure 2).

Note that our experimental analysis is based on the assumption that the groupings suggested by the ChIP (Chromatin ImmunoPrecipitation) analysis are, in fact, correct and thus, provide a direct approach to compare various cor-

relation coefficients. It is quite likely that the ChIP-based groupings themselves contain many false relations (both positives and negatives) and corrupt our inference in some unknown manner. Nonetheless, we observe that the trends of reduced false positives and negatives in shrinkage analysis with these biological data are consistent with the analysis based on mathematical simulation and hence, reassuring.

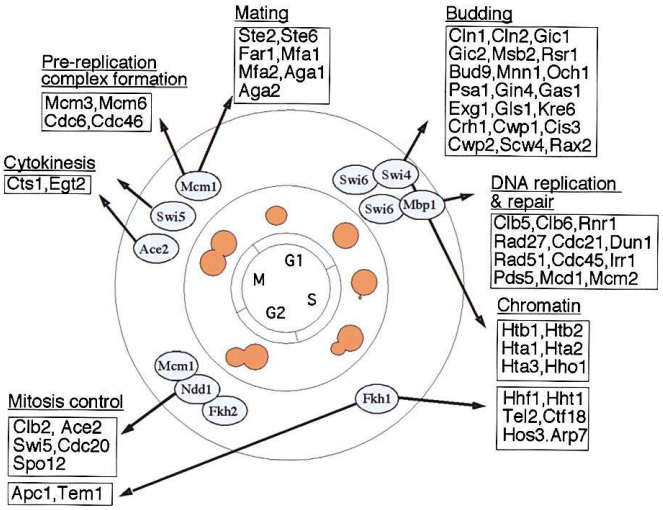

Figure 2: Regulation of cell cycle functions by the activators (Figure 5 in [12]).

In the work of Simon *et al.* ([12]), genome-wide location analysis was used to determine how the yeast cell cycle gene expression program is regulated by each of the nine known cell cycle transcriptional activators: Ace2, Fkh1, Fkh2, Mbp1, Mcm1, Ndd1, Swi4, Swi5, and Swi6. It was also found that cell cycle transcriptional activators which function during one stage of the cell cycle regulate transcriptional activators that function during the next stage. This serial regulation of transcriptional activators together with various functional properties suggests a simple way of partitioning some selected cell cycle genes into nine clusters, each one characterized by a group of transcriptional activators working together and their functions (see Table 2): for instance, Group 1 is characterized by the activators Swi4 and Swi6 and the function of budding; Group 2 is characterized by the activators Swi6 and Mbp1 and the function involving DNA replication and repair at the juncture of G1 and S phases, etc.

Upon closer examination of the data, we observed that the data in its raw "pre-normalized" form is inconsistent with the assumptions used in deriving γ:

1. The gene vectors are not range-normalized, so ${\beta_j}^2 \neq \beta^2$ for every j, and
2. The N experiments are not necessarily independent.

Table 2: Genes in our data set, grouped by transcriptional activators and cell-cycle functions.

	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

Range-normalization and subsampling of experiments were used prior to clustering in an attempt to alleviate these shortcomings. The clusters on the processed data set, thresholded at the cut-off value of 0.60, are listed in Tables 3, 4, and 5. The choice of the threshold parameter is discussed further in section 5.

Our initial hypothesis can be summarized as follows: *Genes expressed during the same cell cycle stage, and regulated by the same transcriptional activators should be in the same cluster.* We compared the performance of the similarity metrics based on the degree to which each of them deviated from this hypothesis. Below we list some of the observed deviations from the hypothesis.

Possible False-Positives:

- Bud9 (Group 1: Budding), Egt2 (Group 7: Cytokinesis), and Cdc6 (Group 8: Pre-replication complex formation) are placed in the same cluster by all three metrics: (E68, S49, and P51).
- Mcm2 (Group 2: DNA replication and repair) and Mcm3 (Group 8) are placed in the same cluster by all three metrics: (E68, S15, and P15),
- For more examples, see [10].

Possible False-Negatives:

- Group 1: Budding (Table 2) is split into five clusters by the Eisen metric:
  {Cln1, Och1} ∈ E58, {Cln2, Msb2, Rsr1, Bud9, Mnn1, Exg1} ∈ E68, Gic1 ∈ E29, Gic2 ∈ E64, and {Kre6, Cwp1} ∈ E33;
  into four clusters by the Shrinkage metric:
  {Cln1, Bud9, Och1} ∈ S49, {Cln2, Gic2, Msb2, Rsr1, Mnn1, Exg1} ∈ S6, Gic1 ∈ S32, and {Kre6, Cwp1} ∈ S65;
  and into eight clusters by the Pearson metric:
  {Cln1, Och1} ∈ P1, {Cln2, Rsr1, Mnn1} ∈ P15, Gic1 ∈ P29, Gic2 ∈ P2, {Msb2, Exg1} ∈ P3, Bud9 ∈ P51, Kre6 ∈ P11, and Cwp1 ∈ P62.

We introduced a new notation to represent the resulting cluster sets, as well as a scoring function to aid in their comparison.

Each cluster set can be written as follows:

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

where x denotes the group number (as described in Table 2), 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 in Table 2. A value of "$*$" for z_j denotes that cluster j contains additional genes, although none of them are cell cycle genes. The cluster set can then be scored according to the following measure:

$$\mathrm{FP}(\gamma) = \frac{1}{2}\sum_{x}\sum_{j=1}^{n_x} y_j \cdot z_j \tag{7}$$

$$\mathrm{FN}(\gamma) = \sum_{x}\sum_{1\le j<k\le n_x} y_j \cdot y_k \tag{8}$$

$$\mathrm{Error\_score}(\gamma) = \mathrm{FP}(\gamma) + \mathrm{FN}(\gamma) \tag{9}$$

Table 2 contains those genes from Figure 2 that were present in our data set. The following tables contain these genes grouped into clusters by a hierarchical clustering algorithm using the three metrics (Eisen in Table 3, Shrinkage in Table 4, and Pearson in Table 5) thresholded at a correlation coefficient value of 0.60. Genes that have not been grouped with any others at a similarity of 0.60 or higher are absent from the tables; in the subsequent analysis they are treated as *singleton* clusters.

The subsampled data yielded the estimate $\gamma \simeq 0.66$. In our set notation, the resulting Shrinkage clusters with the corresponding error score computed as in (9) can be written as follows:

$$\begin{aligned}
&\gamma = 0.66(S) \Longrightarrow \\
\{1 &\to \{\{6,6\},\{3,2\},\{2,5\},\{1,*\}\}, \\
2 &\to \{\{6,6\},\{2,5\},\{1,1\}\}, \\
3 &\to \{\{5,2\},\{1,*\}\}, \\
4 &\to \{\{2,5\},\{1,3\},\{1,6\}\}, \\
5 &\to \{\{1,*\}\}, \\
6 &\to \{\{3,1\},\{1,6\}\}, \\
7 &\to \{\{1,*\},\{1,4\}\}, \\
8 &\to \{\{1,*\},\{1,1\},\{1,4\},\{1,6\}\}, \\
9 &\to \{\{1,*\},\{1,*\}\} \\
\}& \\
&\mathrm{Error_score}(0.66) = 76 + 88 = 164
\end{aligned}$$

The error scores for the Eisen ($\gamma = 0.0$) and Pearson ($\gamma = 1.0$) cluster sets, computed according to (9), are

$$\mathrm{Error\_score}(0.0) = 370 + 79 = 449$$
$$\mathrm{Error\_score}(1.0) = 69 + 107 = 176$$

From the data shown in the tables, as well as by comparing the error scores, one can conclude that for the same

Table 3: RN Subsampled Data, $\gamma = 0.0$ (Eisen)

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

Table 4: RN Subsampled Data, $\gamma = 0.66$ (Shrinkage)

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

Table 5: RN Subsampled Data, $\gamma = 1.0$ (Pearson)

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

clustering algorithm and threshold value, Pearson tends to introduce more false-negatives and Eisen tends to introduce more false-positives than Shrinkage, as Shrinkage reduces these errors by combining the good properties of both algorithms. This observation is consistent with our mathematical analysis and the simulation presented in section 4.1.

We have also conducted a more extensive computational analysis of Eisen's data, but omitted it from this paper due to space limitations. This analysis appears in a full technical report available for download from `http://www.cs.nyu.edu/cs/faculty/mishra/` ([10]).

5 Discussion

Microarray-based genomic analysis and other similar high-throughput methods have begun to occupy an increasingly important role in biology, as they have helped to create a visual image of the state-space trajectories at the core of the cellular processes. This analysis will address directly to the observational nature of the "new" biology. As a result, we need to develop our ability to "see," accurately and reproducibly, the information in the massive amount of quantitative measurements produced by these approaches—or be able to ascertain when what we "see" is unreliable and forms a poor basis for proposing novel hypotheses. Our investigation demonstrates the fragility of many of these analysis algorithms when used in the context of a small number of experiments. In particular, we see that a small perturbation of, or a small error in, the estimation of a parameter (the shrinkage parameter) has a significant effect on the overall conclusion. The errors in the estimators manifest themselves by missing certain biological relations between two genes (false-negatives) or by proposing phantom relations between two otherwise unrelated genes (false-positives).

A global picture of these interactions can be seen in Figure 3, the Receiver Operator Characteristic (ROC) figure, with each curve parametrized by the cut-off threshold in the

range of $[-1, 1]$. An ROC curve ([13]) for a given metric plots sensitivity against (1−specificity), where

Sensitivity = fraction of positives detected by a metric

$$= \frac{\mathrm{TP}(\gamma)}{\mathrm{TP}(\gamma) + \mathrm{FN}(\gamma)}, \tag{10}$$

Specificity = fraction of negatives detected by a metric

$$= \frac{\mathrm{TN}(\gamma)}{\mathrm{TN}(\gamma) + \mathrm{FP}(\gamma)}, \tag{11}$$

and $\mathrm{TP}(\gamma)$, $\mathrm{FN}(\gamma)$, $\mathrm{FP}(\gamma)$, and $\mathrm{TN}(\gamma)$ denote the number of True Positives, False Negatives, False Positives, and True Negatives, respectively, arising from a metric associated with a given γ. (Recall that γ is 0.0 for Eisen, 1.0 for Pearson, and is computed according to (6) for Shrinkage, which yields 0.66 on this data set.) For each pair of genes, $\{j, k\}$, we define these events using our hypothesis (see section 4.2) as a measure of truth:

TP: $\{j, k\}$ are in same group (see Table 2) and $\{j, k\}$ are placed in same cluster;

FP: $\{j, k\}$ are in different groups, but $\{j, k\}$ are placed in same cluster;

TN: $\{j, k\}$ are in different groups and $\{j, k\}$ are placed in different clusters; and

FN: $\{j, k\}$ are in same group, but $\{j, k\}$ are placed in different clusters.

$\mathrm{FP}(\gamma)$ and $\mathrm{FN}(\gamma)$ were already defined in equations (7) and (8), respectively, and we define

$$\mathrm{TP}(\gamma) = \sum_{x} \sum_{j=1}^{n_x} \binom{y_j}{2} \tag{12}$$

and

$$\mathrm{TN}(\gamma) = \mathrm{Total} - (\mathrm{TP}(\gamma) + \mathrm{FN}(\gamma) + \mathrm{FP}(\gamma)) \tag{13}$$

where Total= $\binom{44}{2} = 946$ is the total # of gene pairs $\{j, k\}$ in Table 2.

The ROC figure suggests the best threshold to use for each metric, and can also be used to select the best metric to use for a particular sensitivity.

The dependence of the error scores on the threshold can be more clearly seen from Figure 4. It shows that the conclusions we draw in section 4.2 hold for a wide range of threshold values, and hence a threshold value of 0.60 is a reasonable representative value.

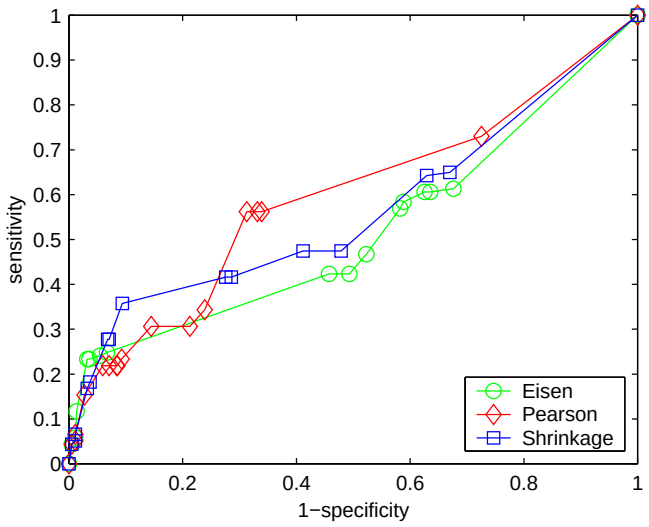

Figure 3: Receiver Operator Characteristic curves. Each curve is parametrized by the cut-off value $\theta \in \{1.0, 0.95, \ldots, -1.0\}$

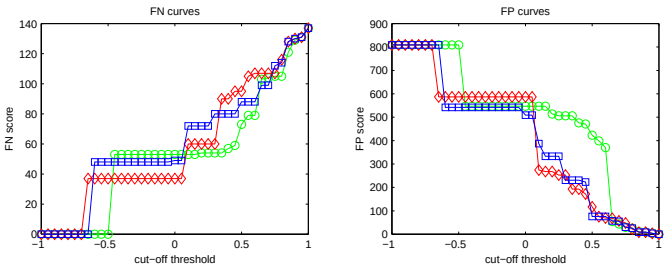

Figure 4: FN and FP curves, plotted as functions of θ.

As a result, in order to study the clustering algorithms and their effectiveness, one may ask the following questions. If one must err, is it better to err on the side of more false-positives or more false-negatives? What are the relative costs of these two kinds of errors? In general, since false-negatives may cause the inference process to ignore useful information for certain novel genes, and since false-positives may result in noise in the information provided to the algorithms used in analyzing regulatory patterns, intelligent answers to our questions depend crucially on how the cluster information is used in the subsequent discovery processes.

References

[1] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998), *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868.

[2] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., and Davis, R.W. (1996), *Proc. Natl. Acad. Sci. USA* **93**, 10614–10619.

[3] DeRisi, J.L., Iyer, V.R., and Brown, P.O. (1997), *Science* **278**, 680–686.

[4] Spellman, P.T., Sherlock, G., Zhang, M., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998), *Mol. Biol. Cell* **9**, 3273–3297.

[5] Chu, S., DeRisi, J.L., Eisen, M.B., Mulholland, J., Botstein, D., Brown, P.O., and Herskowitz, I. (1998), *Science* **282**, 699–705.

[6] MacKay, D.J.C. (1992), *Neural Computation* **4**, 415–447.

[7] Sokal, R.R. and Michener, C.D. (1958), *Univ. Kans. Sci. Bull.* **38**, 1409–1438.

[8] James, W. and Stein, C. (1961), in *Proc. 4th Berkeley Symp. Math. Statist.*, (ed. Neyman, J.), Vol. **1**, 361–379, Univ. of California Press.

[9] Hoffman, K. (2000), *Statistical Papers*, **41(2)**, 127–158.

[10] Cherepinsky, V., Feng, J., Rejali, M., and Mishra, B. (2003) (unpublished, PDF available for download from `http://www.cs.nyu.edu/cs/faculty/mishra/`).

[11] Wolfram, S. (1999), *The Mathematica Book*, Cambridge Univ. Pr. (4th edition).

[12] Simon, I., Barnett, J., Hannett, N., Harbison, C.T., Rinaldi, N.J., Volkert, T.L., Wyrick, J.J., Zeitlinger, J., Gifford, D.K., Jaakkola, T.S., and Young, R.A. (2001), *Cell* **106**, 697–708.

[13] Egan, J.P. (1975), *Signal Detection Theory and ROC analysis*, Academic Press, New York.