Report in fulfillment of the area exam requirement in the MIT Department of Electrical Engineering and Computer Science. June 2001.

Noise Models in Gene Array Analysis

Ron Dror

Abstract

Gene arrays measure the expression levels of thousands of genes simultaneously, providing an extremely powerful tool for biology and medicine. Unfortunately, the high level of noise in the resulting measurements often obscures the biological processes of interest. After providing an overview of gene array technology, we consider several recent expression level estimation methods which deal with this noise explicitly. We also consider the implications of estimation methods based on noise models in higher-level processing of gene array data.

1 Introduction

The past decade has witnessed an explosion of genetic sequence data, culminating with the publication in February of the draft sequence of the human genome [4, 30]. Researchers have identified tens thousands of coding regions of the genetic sequence commonly known as genes. The gene set of yeast and other simple organisms has been completely characterized, while an estimated two-thirds of human genes have been identified [4]. Despite all the excitement, sequencing the genome and identifying coding sequences is only a first step in understanding the control and function of an organism at the cellular level. The greater challenge is to understand what each gene does, when each is active, and how they interact.

Gene array technologies developed over the past five years provide a valuable tool for answering such questions. These arrays simultaneously measure cellular concentrations of thousands of messenger RNAs (mRNAs). Since transcription from DNA to mRNA is the first step in the creation of a protein from a gene, mRNA concentrations provide a measure of the activity of a gene, also known as the *expression level*.[1] Prior to the development of gene arrays, researchers could measure individual mRNA concentrations using techniques such as the Northern blot. Modern gene arrays allow simultaneous measurement of expression levels of all known genes in a cell, providing

[1] DNA codes for RNA in a process known as transcription. Some genes code for RNAs, such as transfer RNA or ribosomal RNA, which play a functional role in their own right. The majority of genes are transcribed to messenger RNAs, which code for an amino acid (protein) sequence. Proteins are constructed from mRNAs through a process known as translation. The most meaningful measure of the activity of a gene is the concentration of the corresponding protein. Because gene arrays only measure RNA concentrations, however, we refer to mRNA concentrations as expression levels in this paper.

a far more complete characterization of the "state" of the cell than previously possible. Despite its considerable expense, this technology has been rapidly adopted by both academic and commercial laboratories. It has been used to characterize the function of novel genes [16], to describe novel mechanisms for gene regulation [14], and to identify cancer cell lines [11]; in the future, it may become a commonplace diagnostic tool for human disease.

The explosion of data provided by gene array technologies has led to an urgent need for novel statistical and computational methods. A number of computational techniques are already in common use. Some estimate actual expression levels or ratios of expression levels from noisy raw measurements. Others infer relationships between genes based on expression levels or structure gene array output graphically to aid humans in mining the data visually. Additional techniques estimate statistical confidence levels for experimental results, or serve to design efficient experiments for a particular purpose. Unfortunately, the computational and statistical portions of this field have struggled to keep pace with the rapidly developing biology, such that some widely used computational techniques leave tremendous room for improvement.

Rather than attempting to survey the entire body of computational work in gene array analysis, this paper focuses on error models for gene array measurements and on estimation of expression levels from noisy data. Section 3 examines in detail the estimation methods proposed by Li and Wong [20] and by Hughes *et al.* [16]. Section 4 relates these basic estimation methods to higher-level analysis in the context of the work of Friedman *et al.* [9] and of Hughes *et al.*

2 Gene array technologies

Two different gene array technologies are currently in common use. Both rely on the propensity of mRNA or DNA to hybridize with single strands of complementary DNA immobilized on a solid substrate. The two technologies differ in the length of the immobilized DNA strands, the number of spots used to detect the expression level of each gene, and the method of manufacture.

The GeneChip®, developed, marketed and patent-protected by Affymetrix, Inc., relies on DNA oligonucleotides only 25 base pairs in length [21, 31]. Affymetrix manufactures these chips on a silicon substrate using a photolithography process similar to that standard in the manufacture of integrated circuits. The photolithography process allows precise control of the sequence of single-stranded DNA in each spot on the chip, such that each spot contains thousands of identical single-stranded oligonucleotides. To measure gene expression levels for a population of cells, a researcher

extracts mRNA from a cell culture, processes the mRNA in a number of chemical reactions which amplify it and label it with a fluorescent tag, and runs the resulting solution over the chip for 16 to 24 hours. The amount of RNA which has hybridized to a spot is quantified by using laser confocal microscopy to measure the fluorescent luminance of the spot.

An Affymetrix array uses 40 different oligonucleotide sequences at 40 adjacent spots on the array to detect each type of mRNA molecule. 20 of these spots contain "Perfect Match" (PM) sequences complementary to subsequences of the target gene's mRNA. Each of the remaining 20 "Mismatch" (MM) sequences differs by a single base pair from its corresponding PM sequence. Although the PM sequences are carefully chosen by Affymetrix to be unique to their target gene, each oligonucleotide can suffer from cross-hybridization with other similar genes. Because the corresponding MM sequence typically experiences a similar degree of cross-hybridization, the difference between pairs of corresponding PM and MM responses provides a measure of the concentration of the mRNA of interest. Affymetrix's software produces an expression level measurement for each gene by averaging the differences for the 20 probe pairs.[2]

Spotted arrays, popularized by the Brown lab at Stanford [25], rely on full-length strands of DNA complementary to the target mRNA. A solution containing DNA strands of a particular sequence is spotted onto a glass slide using a robotic arm. By spotting thousands of solutions containing different DNA strands onto the glass slide, one produces a complete microarray. The length of the spotted DNA confers much greater hybridization specificity than that of the Affymetrix oligonucleotides, such that only one spot is required per gene. The robotic spotting process, however, is less precise than photolithography, so the spots on a spotted array may differ substantially in shape and size. To deal with this problem, one labels mRNA samples from the test sample and from a control sample using two different fluorescent probes, one red and one green.[3] One mixes the two together before running the solution over the microarray. While the absolute fluorescent intensity of each spot depends heavily on the size of the spot, the ratio of red to green for a particular spot provides a fairly accurate measure of the ratio of the target mRNA concentrations in the two samples.

[2]The Affymetrix software applies a method known as *superscoring* to eliminate some probe pair responses before averaging. The software calculates the mean and standard deviation of PM − MM responses across the 20 probe set, excluding the minimum and maximum values. PM − MM values more than k standard deviations from the mean are excluded from the final average. k is user defined, with a default value of 3 [2].

[3]Standard spotted array techniques[25, 18] use labelled single-stranded DNA, which is reverse transcribed from cellular mRNA, rather than the mRNA itself.

3 Error models in expression level estimation

Gene array experiments involve a large number of error-prone steps which lead to a high level of noise in the resulting data [26]. Several potential sources of measurement noise are listed in Table 1.

This noise raises several practical questions in interpreting experimental results. For example, how should one combine multiple observations of the same transcript level into a single estimate? How should one determine ratios of transcript levels under different conditions, given one or more observations of each? How should one handle the negative observations reported by Affymetrix chips when the MM readings exceed the PM readings? How can one quantify the statistical significance of a result based on gene array data?

Current biological practice typically addresses these issues through simple heuristic measures. For example, investigators often deal with negative values by applying a technique known as "flooring," in which observations below some small, positive threshold are set to that threshold while observations above that threshold are accepted as accurate [14]. Two transcript levels are typically considered significantly different if one exceeds the other by a factor of two, after flooring to some value [14, 29].

Several authors have recently addressed the noise issue explicitly. Lee *et al.* [19] repeated a gene array experiment three times and showed that the results differed substantially, driving home the point that repetition can increase the significance of conclusions from gene array experiments. Kerr *et al.* [17] applied an ANOVA model to microarray experiments and used bootstrap methods to obtain confidence intervals for the results. Chen *et al.* [3] and Ermolaeva *et al.* [7] used a ratio distribution to determine the statistical significance of an observed change in expression levels. Hughes *et al.* [16] suggested statistics for estimation and uncertainty from multiple repetitions. Hartemink *et al.* [12] developed a maximum likelihood method for whole-chip normalization based on a noise model for Affymetrix chip data. Li and Wong [20] developed an error model for Affymetrix data at the level of individual PM and MM probes. This section focuses on the two approaches which have gained the most attention recently, those of Li and Wong and of Hughes *et al.*

3.1 A model for probe-level noise in Affymetrix chips

Li and Wong [20] proposed that one can improve estimates of expression levels for Affymetrix chips by performing statistical analysis on the individual PM and MM probe outputs, rather than using the average difference reported by the Affymetrix software. Their analysis applies specifically to

Affymetrix chips.

3.1.1 Probe response model

Li and Wong model the distribution of individual PM − MM probe differences. They consider a set of J probe pairs probes targeted at a single gene, and they assume that experimental results are available for $I > 1$ chips containing this probe set. Their model for individual probe responses takes the form

$$y_{ij} = \mathrm{PM}_{ij} - \mathrm{MM}_{ij} = \theta_i \phi_j + \epsilon_{ij}, \tag{1}$$

where $\mathrm{PM}_{ij} - \mathrm{MM}_{ij}$ is the difference between the PM and MM values for the jth probe pair and the ith array, θ_i is the expression level of the ith array, ϕ_j is a sensitivity factor for the jth probe pair, and $\epsilon_{ij} \sim N(0, \sigma^2)$. This model makes several strong assumptions, whose validity is discussed in Section 3.1.3:

1. The expected response of each probe pair increases linearly with the concentration of the target mRNA, although the rate of increase may vary from probe pair to probe pair.

2. The error (difference between actual and expected response) is additive and Gaussian.

3. The error is independent from probe pair to probe pair and from array to array.

The sensitivity factors ϕ_j could be measured directly from a large set of experiments in which the true concentrations θ_i were known. In practice, such a database is rarely available. Li and Wong therefore fit both the θ_is and the ϕ_js to their data set. Because scaling all the θ_is up by a constant and all the ϕ_js down by the same constant will leave the product $\theta_i \phi_j$ unchanged, they arbitrarily constrain the sum of squares of the phi_is to be equal to the number of probe pairs:

$$\Sigma_j \phi_j^2 = J. \tag{2}$$

They then solve iteratively for the θ_is and ϕ_js by fixing one set of parameters, finding the least squares fit for the other set, and alternating. Figure 1 shows experimental data and model fits for identical probe sets on six arrays.

3.1.2 Model application

The θ_i parameters of the model provide estimates of the relative expression levels of the target gene on the various arrays. The probe-level statistical model of Li and Wong also allows detection of

	Multiplicative	Additive
Whole-chip	• Variation in hybridization time • Variation in reagent concentrations	• Leak of external light during chip reading
Probe-specific	• Inhomogeneities in chip preparation • Variations in laser intensity during chip reading	• Trace contamination with cross-hybridizing oligonucleotides

Table 1: Sources of noise in gene array measurements. A particular source of noise can be broadly categorized as either multiplicative noise, where the noise magnitude is proportional to the signal magnitude, or additive noise, where the noise is independent of the signal. Additionally, each noise source may affect the whole chip similarly or each probe independently (probe specific). Reproduced from [5].

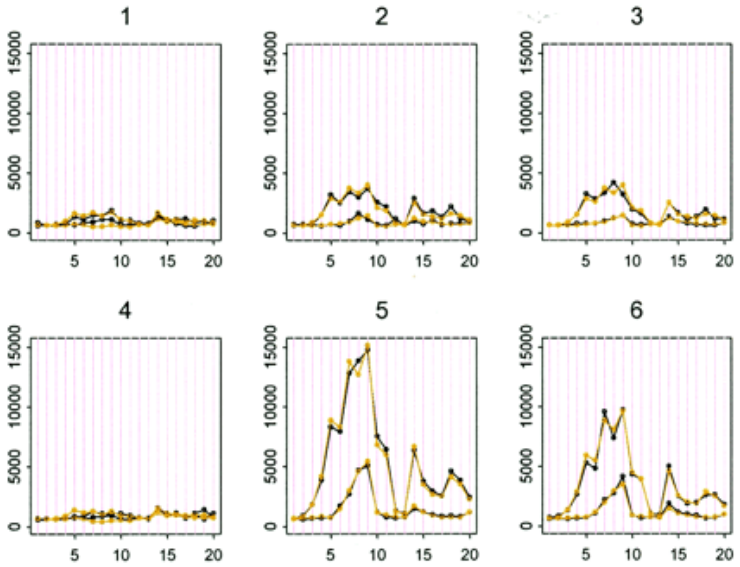

Figure 1: Dark curves represent measured response as a function of probe number for one particular gene on the first six arrays of a database of human Affymetrix chips. Light curves represent the model fit of Equation 1 to the set of probe responses for each array. Reproduced from [20].

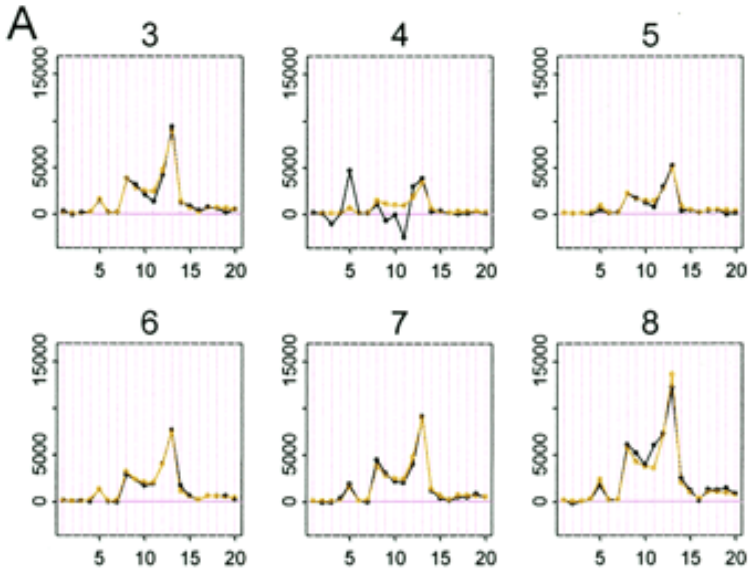

Figure 2: Observed and modeled responses for a probe set targeted at a particular gene on six different arrays. The model fits the observations accurately for all the arrays except array 4. Reproduced from [20].

damaged arrays, imaging artifacts, cross-hybridizing probes, and other anomalous measurements. Once the model has been fit to a set of array measurements, anomalies constitute statistical outliers corresponding to particularly large and improbable values of ϵ_{ij}.

Figure 2 illustrates detection of an outlier array on which an entire probe set is defective. The model provides a close fit to the probe sets on all arrays except array 4. While the average PM$-$MM difference for array 4 is not unusual, the fact that the relative magnitudes of probe responses within this array differs from that on other arrays suggests that the average difference for this array should not be taken seriously. Quantitatively, Li and Wong compute the sum of squares of model residuals for the ith array as $\sum_j (\theta_i \phi_j - y_{ij})^2$. They then discard probe sets on arrays for which this sum is particularly large relative to the others, although the paper does not give a specific threshold for making this decision. Figure 3 shows an example of the application of this outlier detection technique to a chip containing a severe scratch.

A similar technique detects individual probes whose response fits the model poorly across the entire set of arrays. Such a probe pair probably suffers from cross-hybridization that affects the PM and MM probes differently. Li and Wong also discard individual probes on individual arrays with particularly large residuals.

In certain cases, the response of one particular probe in a set dominates the others, perhaps

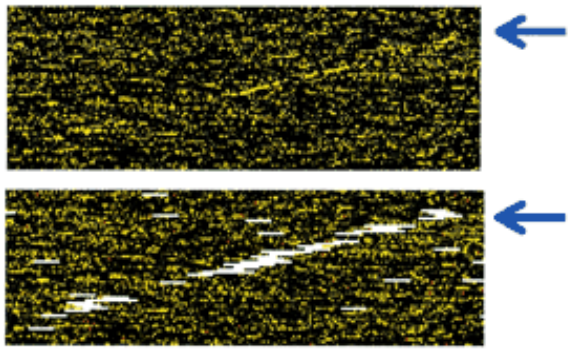

Figure 3: The upper image shows a confocal laser microscopy image of an Affymetrix array containing a scratch that begins near the arrow. The white bars superimposed on this array in the lower image indicate detected array outliers, each consisting of 40 oligonucleotides targeted at a particular gene. Most probe sets affected by the scratch are detected as array outliers. Some probe sets near the scratch are not marked as outliers because the contamination contributed a similar additive perturbation to both PM and MM probes, such that the PM − MM difference remained unchanged. Reproduced from [20].

because of particularly strong cross-hybridization effects. The model will fit such data reasonably well, because the θ_i will be chosen to model the response of that single "high-leverage" probe, and the ϕ_j for all other probes will be close to zero. Li and Wong therefore discard probes which contribute more than 80% of the sum of the squares of the ϕ_js. They discard "high-leverage" arrays in a similar manner.[4]

Li and Wong use an iterative procedure to fit the model to data while identifying outliers which are disregarded in the following model fit. This procedure alternates between fitting the model, identifying anomalous arrays, refitting the model, and identifying anomalous probes. Li and Wong report that the set of outliers typically converges or cycles between a small number of different sets after 5–10 iterations.

3.1.3 Critique

The validity of the techniques of Li and Wong depends largely on the assumptions used to justify the error model itself (Section 3.1.1). Assumption 1, the linear dependence of expected probe response on target mRNA concentration, is a reasonable first-order approximation. An early paper by Affymetrix reported that the mean PM − MM response, averaged over many probes corresponding to the same gene, varies linearly with expression level [21].

[4]Li and Wong do not justify their decision to exclude "high-leverage" arrays from the model fitting process. Such an array might be defective, but it might correspond to the one sample in which a particular gene product is present. In the latter case, discarding it is not desirable.

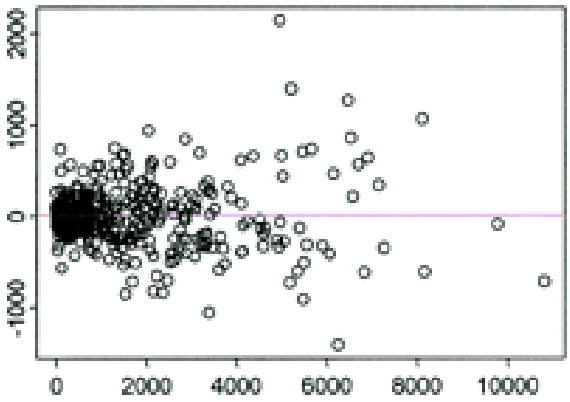

Figure 4: Residuals ϵ_{ij} of the error model versus fitted value θ_i. The variance of residuals increases with θ_i. Reproduced from [20].

Assumption 2, which states that the model residual has an additive Gaussian distribution independent of expression level, is more difficult justify. Through analysis of several hundred Affymetrix yeast chip experiments performed in Richard Young's lab at the Whitehead Institute [1], we have found that the error distribution for averaged probe pair differences depends on true expression level, with error variance increasing significantly at higher expression levels [5]. Model residuals plotted by Li and Wong exhibit a similar trend (Figure 4). We have also found that the fixed additive component of the measurement noise for average PM − MM differences is non-Gaussian, with tails significantly heavier than that of a normal distribution. We fit experimental data with a generalized Laplacian distribution of the form

$$p(x) = \frac{1}{C} e^{-|\frac{x}{s}|^p}, \tag{3}$$

where $p = 0.76$. Assumption 3 is also somewhat suspect, because different probes targeted at the same gene are unlikely to have independent error residuals, if only because they are physically adjacent on Affymetrix chips.

The probe-level approach suffers from the fact that many biologists do not have ready access to individual probe data. This is particularly true for publicly available databases of past experiments. Federally funded labs are required to post their experimental data on the Web, but they typically post average probe pair differences rather than individual differences. Hopefully this practice will change with time.

Nevertheless, the work of Li and Wong represents an important contribution to computational analysis of gene array data. Despite the oversimplification inherent in the error model, it represents a reasonable starting point which achieves useful results and demonstrates the utility of modeling

responses at the individual probe level. Future work may incorporate more accurate models of measurement error (Section 3.3) into this framework. Li and Wong do not address the computation of confidence intervals or significance values for expression level ratios or the use of statistical confidence measures in higher-level analysis (see Sections 3.2 and 3.3), although they suggest they will do so in an upcoming paper.

3.2 Rosetta model

Hughes *et al.* [16] of Rosetta Inpharmatics develop an error model for spotted arrays in order to combine results of multiple experiments and to quantify the statistical significance of observed expression level differences. The Hughes paper focuses on developing novel biological analysis techniques and drawing biological conclusions (Section 4.1) rather than on developing computational methods. In fact, the noise model is detailed only in supplementary material available on the Web [15]. Nevertheless, many biologists perceive the model as the current "state of the art," and it has already achieved widespread use [23]. Unlike the purely additive noise models of Li and Wong [20], Chen *et al.* [3], and Ermolaeva *et al.* [7], the Rosetta model incorporates both additive and multiplicative noise. It takes into account noise due to biological variation as well as measurement noise.

3.2.1 Model derivation

Hughes *et al.* describe their model as "inelegant." This is an understatement. The model is based on questionable assumptions, and the estimation techniques are derived using questionable approximations. Even in the supplementary material, the authors present a cursory and confusing derivation of the model with an abundance of ambiguous notation. This description has led to confusion in other biology labs attempting to apply the model [23]. To make matters worse, the authors refuse to share the value of the multiplicative noise constant f which they derived from their data, claiming this is proprietary information [24]. In this section, we attempt to rederive the model in as clear a manner as possible, making clear the highly questionable assumptions which underlie it. Most notable among these is the implicit assumption that all experimentally observed quantities of interest, as well as logs of their ratios, follow a normal distribution.

Suppose that the distribution of measured values of the two component channels in a two-color spotted array experiment is given by

$$y_r \quad \sim \quad N(\theta_r, \sigma_r^2 + \theta_r^2 f^2) \tag{4}$$

$$y_g \sim N(\theta_g, \sigma_g^2 + \theta_g^2 f^2) \tag{5}$$

where y_r and y_g are measurements of the red and green channels, respectively. The expected values of the two responses, θ_r and θ_g, are assumed to be proportional to the corresponding expression levels. σ_r and σ_g represent the magnitude of additive noise due to background subtraction, while f represents the magnitude of multiplicative noise due to scanner gain fluctuations and variations in hybridization efficiency or dye incorporation efficiency.

Estimating θ_r and θ_g from single instances $\tilde{y}_r$ and $\tilde{y}_g$ of the random variables y_r and y_g, respectively, we have $\hat{\theta}_r = \tilde{y}_r$ and $\hat{\theta}_g = \tilde{y}_g$. Therefore the following distributions hold approximately:

$$y_r \sim N(\tilde{y}_r, \sigma_r^2 + \tilde{y}_r^2 f^2) \tag{6}$$

$$y_g \sim N(\tilde{y}_g, \sigma_g^2 + \tilde{y}_g^2 f^2) \tag{7}$$

$$y_r - y_g \sim N(\tilde{y}_r - \tilde{y}_g, \sigma_r^2 + \sigma_g^2 + (\tilde{y}_r^2 + \tilde{y}_g^2) f^2) \tag{8}$$

Letting $\sigma_{y_r - y_g} = \sqrt{\sigma_r^2 + \sigma_g^2 + (y_r^2 + y_g^2) f^2}$, we define a random variable X as

$$X = \frac{y_r - y_g}{\sigma_{y_r - y_g}} = \frac{y_r - y_g}{\sqrt{\sigma_r^2 + \sigma_g^2 + (y_r^2 + y_g^2) f^2}}. \tag{9}$$

The distribution of X is approximately normal with unit variance. One can perform a hypothesis test for the null hypothesis that the expression levels θ_r and θ_g are identical by assuming that $X \sim N(0, 1)$.[5] The p-value is then given by

$$p = 2(1 - \mathrm{erf}(|\tilde{X}|)), \text{ where } \tilde{X} = \frac{\tilde{y}_r - \tilde{y}_g}{\sqrt{\sigma_r^2 + \sigma_g^2 + (\tilde{y}_r^2 + \tilde{y}_g^2) f^2}}. \tag{10}$$

In compiling their compendium of gene array experiments, Hughes *et al.* repeated each experiment either two or four times (either one or two fluor-reversed pairs). Because they assume that $\log(y_r/y_g)$ is normally distributed for each experiment and that the mean of this distribution is the log ratio of the true concentrations, they combine the results of repeated experiments by averaging log ratios. Hughes *et al.* assert that the standard deviation of $\log_{10}(y_r/y_g)$ is given by $\sigma_{\log_{10}(y_r/y_g)} = \frac{\log_{10}(y_r/y_g)}{X}$. Although they provide no justification for this approximation, it is reasonable when $y_r \approx y_g$. In that case, $y_r/y_g \approx 1$, so a first-order Taylor series expansion gives

$$\log_{10}(y_r/y_g) = \log_{10}\left(1 + \left(\frac{y_r}{y_g} - 1\right)\right) \approx \frac{1}{\ln 10}\left(\frac{y_r}{y_g} - 1\right) = \frac{y_r - y_g}{y_g \ln 10}. \tag{11}$$

[5] When testing the null hypothesis, one might approximate θ_r and θ_g more accurately as $\theta_r = \theta_g = (\tilde{y}_r + \tilde{y}_g)/2$. However, we will follow the formulas of Hughes *et al.* in our derivation.

The assumption that $y_r \approx y_g$ also implies that $y_g \gg |y_r - y_g|$, so the standard deviation of $\log_{10}(y_r/y_g)$ can be approximated by

$$\sigma_{\log_{10}(y_r/y_g)} \approx \frac{\sigma_{y_r - y_g}}{y_g \ln 10} = \left(\frac{\sigma_{y_r - y_g}}{y_r - y_g}\right)\left(\frac{y_r - y_g}{y_g \ln 10}\right) \approx \frac{1}{X}\log_{10}(y_r/y_g) = \frac{\log_{10}(y_r/y_g)}{X}. \quad (12)$$

Letting $r_i = log_{10}(y_r/y_g)$ and $\sigma_i = \sigma_{\log_{10}(y_r/y_g)}$ for the ith repetition of the experiment, the assumed Gaussian distribution of r_i implies that the least squares estimate of the true log ratio is

$$\bar{r} = \sum_i (x_i/\sigma_i^2) / \sum_i \sigma_i^{-2} \quad (13)$$

with associated standard deviation

$$\sigma_{\bar{r},p}^2 = 1/\sum_i \sigma_i^{-2}. \quad (14)$$

At this point, Hughes *et al.* realize that due to the drastic approximations and assumptions made in previous steps, Equation 14 may not be an accurate approximation for the variance of the estimate. They therefore calculate a second estimate of this variance based on the scatter of the individual log ratios r_i:[6]

$$\sigma_{\bar{r},s}^2 = \frac{1}{(n-1)\sum_i \sigma_i^{-2}} \sum_i \frac{(r_i - \bar{r})^2}{\sigma_i^2} \quad (15)$$

where n is the number of replicates of the experiment. The final estimate of the standard deviation of $\bar{r}$ is a somewhat arbitrary weighted combination of the estimates, with the relative weight on $\sigma_{\bar{r},s}$ increasing as the number of replicates increases:

$$\sigma_{\bar{r}} = \frac{\sigma_{\bar{r},p} + (n-1)\sigma_{\bar{r},s}}{n}. \quad (16)$$

Given several repeated experiments, one can determine a p-value for the null hypothesis that that expression levels in the two channels are identical using a formula similar to Equation 10, but with $\frac{\bar{r}}{\sigma_{\bar{r}}}$ substituted for $\tilde{X}$.

The error model developed thus far incorporates measurement noise but not biological noise. Hughes *et al.* observed that the expression level of certain genes varies more than that of others in independently prepared samples of the same yeast strain. In order to take such gene-specific differences in inherent biological variation into account when testing for significant differences between the expression levels of two different yeast strains, Hughes *et al.* adjust the measurement error model of Equation 16 by scaling the estimated variance used to determine the p-value. To

[6] Equation 15 does not give the minimum variance unbiased estimator of the standard deviation of $\bar{r}$; the $\frac{1}{n-1}$ factor produces an unbiased estimator only when the variances σ_i are all equal. However, this approximation is less of a concern than the approximations made earlier in this derivation.

estimate biological noise, they use a series of 63 control experiments in which both channels on the chip correspond to separately grown but genetically identical wild-type yeast cultures. Denote the log ratios of the controls by $r_{i,c}$, with i ranging from 1 to 63. Let $\sigma_{\bar{r},\text{controls}}$ denote the estimated standard deviation of the mean $\bar{r}$ produced by applying Equation (16) to the control data. Because the true log ratios are known to be zero, the variance of $\bar{r}$ can be estimated more accurately as

$$\frac{1}{N}\text{Var}\{r_{i,c}\} = \frac{1}{N}\left(\frac{1}{N-1}\sum_{i=1}^{N} r_{i,c}^2\right). \tag{17}$$

For each gene, define a constant Λ based on the control data by

$$\Lambda = \max\left(\frac{\sqrt{(1/N)\text{Var}\{r_{i,c}\}}}{\sigma_{\bar{r},\text{controls}}}, 1\right). \tag{18}$$

For any combination of experiments, let

$$\sigma_{\bar{r},biological} = \Lambda\sigma_{\bar{r}}. \tag{19}$$

This noise estimate is conservative in that it is no smaller than that of Equation 16. To determine a p-value for the null hypothesis that that expression levels in the two channels are identical, one uses a formula similar to Equation 10, but with $\frac{\bar{r}}{\sigma_{\bar{r},biological}}$ substituted for $\tilde{X}$.

Hughes *et al.* applied this error model extensively in their data analysis, as described in Section 4.1.

3.2.2 Critique

The fact that the estimation techniques embodied by the Rosetta model are not rigorous requires little elaboration. Although the error model includes both additive and multiplicative components, the assumption that measurement noise at a particular expression level is normally distributed is doubtful. Moreover, many of the approximations utilized in deriving practical estimation techniques from the model are poor. In spite of all this, the methods described by Hughes *et al.* have achieved better performance in practice than published alternatives [3, 7]. The Rosetta model may be the only published error model to include gene-specific effects and to specifically consider the effects of biological noise. The model works because it was developed empirically through application to biological data. It is validated to a large extent by its practical utility.

Nevertheless, the estimation techniques of Hughes *et al.* suffer from a number of limitations. They have not been shown to be precise even in simple estimation tasks. They do not generalize to

other array technologies. They lack an ability to incorporate prior information. For example, the Rosetta method lacks a principled manner to deal with the negative measurements which sometimes arise due to background subtraction. These measurements are simply set to a small positive value, and ratios computed from them are capped to the value of 100.

3.3 Bayesian Estimation of Array Measurements (BEAM)

For the sake of comparison, we include a brief discussion of our own work on the Bayesian Estimation of Array Measurements (BEAM) method, a Bayesian approach to estimation of transcript levels, ratios of transcript levels, and associated statistical confidence and significance measurements. Given one or more gene array measurements, a statistical model of measurement noise, and any available prior information about the quantity to be estimated, the BEAM method produces a Bayes least squares estimate and a measure of its uncertainty.

For example, if x represents a true transcript level and y represents the corresponding measurement then the Bayes least squares estimate of x based on y is given by

$$\hat{x}(y) = E(x|y) = \int xp(x|y)dx = \frac{1}{p(y)} \int xp(y|x)p(x)dx, \tag{20}$$

where $p(x)$ captures the prior distribution on the transcript level and $p(y|x)$ captures the noise model. The variance of the posterior distribution of x provides a measure of the uncertainty in the estimate. This variance is simply the expected squared error

$$\sigma^2_{\hat{x}}(y) = E((x-\hat{x})^2|y) = \int (x-\hat{x})^2 p(x|y)dx = \frac{1}{p(y)} \int (x-\hat{x})^2 p(y|x)p(x)dx. \tag{21}$$

One can apply Equations 20 and 21 not only to estimation of a transcript level from a single observation, but also to estimates based on multiple observations. For example, the estimate of x given repeated observations $(y_1, y_2) = \mathbf{y}$ is given by

$$\hat{x}(\mathbf{y}) = E(x|\mathbf{y}) = \frac{1}{p(\mathbf{y})} \int xp(\mathbf{y}|x)p(x)dx = \frac{1}{p(y_1)p(y_2)} \int xp(y_1|x)p(y_2|x)p(x)dx. \tag{22}$$

Similarly, to optimally estimate the ratio of two different transcript levels measured by a gene array as y_a and y_b, one should not take the ratio of the estimated transcript levels corresponding to individual measurements. The optimal estimate is given instead by application of Equation 20, with $r = log_{10}\frac{x_a}{x_b}$ representing the log ratio to be estimated:

$$\begin{aligned} \hat{r}(y_a, y_b) &= E(r|y_a, y_b) = E(log_{10}\frac{x_a}{x_b}|y_a, y_b) \\ &= \frac{1}{p(y_a)p(y_b)} \int_{x_a} p(y_a|x_a)p(x_a) \int_{x_b} log_{10}\frac{x_a}{x_b} p(y_b|x_b)p(x_b)dx_b dx_a. \end{aligned} \tag{23}$$

We illustrate application of the BEAM method using noise and prior models we developed for Affymetrix yeast chip expression data on the basis of 261 Affymetrix gene chip experiments performed in the Young lab (http://web.wi.mit.edu/young/) [5]. This error model includes a multiplicative component with a log normal distribution and an additive component with a generalized Laplacian distribution. The prior encodes the fact that true transcript levels cannot be negative, as well as the decreasing probability of very high transcript levels. Using this model, we address rectification of negative observed measurements, combination of repeated measurements, and identification of changes in expression level. We also produce associated measures of statistical certainty, including p-values for the significance of an observed difference in expression levels. A particularly attractive aspect of the BEAM method is that all of these applications involve a simple lookup or interpolation in a table of precomputed values.

Thus far, we have applied the BEAM technique only to the expression level measurements output by the Affymetrix software. We have not applied the technique to measurements at the individual PM and MM probe level, or to spotted array data. We also have not incorporated gene-specific noise models. Given appropriate error models, however, the Bayesian estimation framework extends naturally to all of these applications. One present challenge is the computation of estimates based on a large number of measurements, which involves integration over a high-dimensional probability distribution.

4 Higher-level processing of expression level data

We consider two examples of higher-level processing of gene array data in order to provide a more complete picture of statistical work in this field and to illustrate the implications of the error models discussed in Section 3. Hughes *et al.* [16] combine their error model with standard clustering methods to facilitate analysis of a compendium of gene array profiles for mutant strains of yeast. Friedman *et al.* [9] infer regulatory relationships between genes by learning Bayesian networks from noisy expression level data. Although Friedman *et al.* do not use an explicit noise model, the performance of their methods might be improved through application of the Rosetta or BEAM methods.

4.1 Analysis of a compendium of expression profiles

Hughes *et al.* used spotted arrays to measure the expression levels of over 6000 genes in 300 different mutated or chemically-treated strains of the yeast S. *cerevisiae* under identical culture

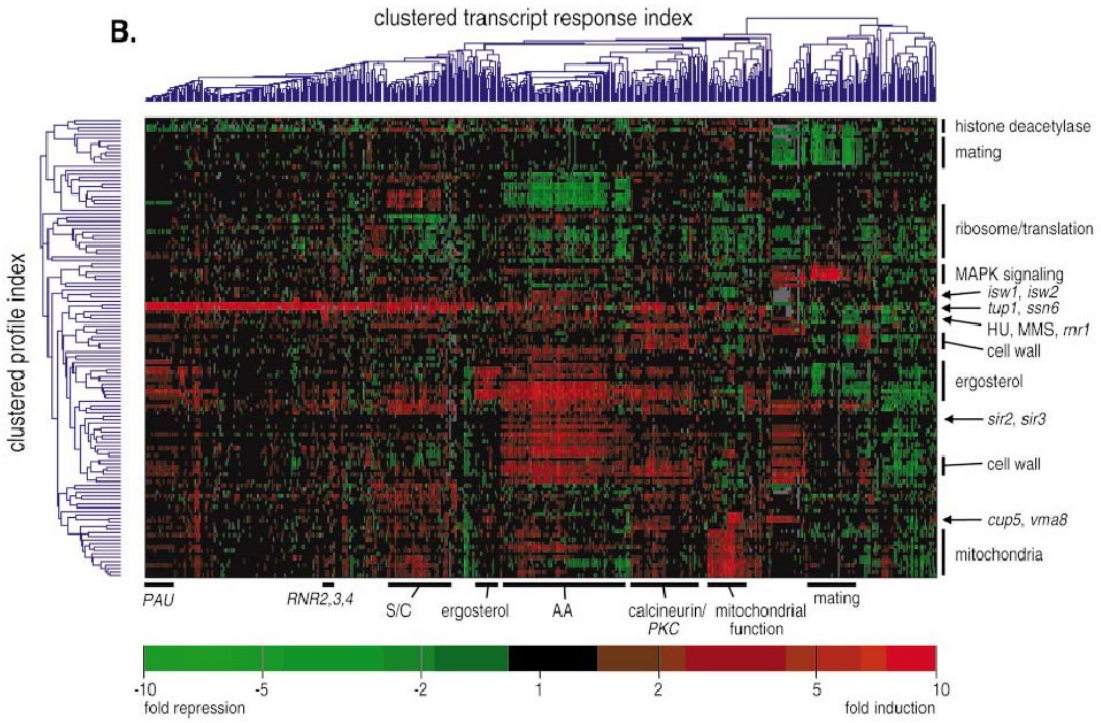

Figure 5: Two-dimensional hierarchical clustering of 127 experiments (rows) and 568 genes (columns). The color of each point indicates the expression level of a particular gene in a particular mutant strain relative to the wild type expression level of that gene. Reproduced from [16].

conditions. On each array, the expression levels of one mutant or chemically-treated strain are compared to those of the wild type strain. The expression levels of a particular strain represent a cellular profile or molecular phenotype of that strain. After constructing a large "compendium" of expression profiles, Hughes *et al.* identify the cellular functions of uncharacterized genes by comparing the profiles of mutant strains in which those genes are deleted to the profiles of strains in which previously characterized genes have been deleted.

To identify groups of yeast strains with similar profiles, Hughes *et al.* use a hierarchical clustering technique similar to that of Eisen *et al.* [6]. They compute a similarity matrix based on the correlation of the profiles of each pair of strains. The clustering algorithm iteratively finds the most similar pair of profiles, merges them into one group, and updates the similarity matrix to reflect the similarity between that group and other existing groups. Monte Carlo randomization techniques assign a significance value to each branch in the resulting binary tree. The tree structure is used to reorder the expression profiles for display as a color-coded matrix in which rows correspond to

different strains and columns to the expression levels of different genes (Figure 5).[7] Strains with similar expression profiles tend to group together on adjacent rows of the matrix. Columns of the matrix, corresponding to the expression levels of individual genes in all compendium experiments, are also ordered by clustering.

Before clustering, Hughes *et al.* filter both genes and experiments by applying significance cuts based on fold change or the p-value for significance of expression level change from the wild type. Specifically, they require each strain undergoing further analysis to possess a certain number of genes whose observed fold change from the wild type is above a certain threshold and whose p-value is below a threshold. Strains whose profiles do not meet this criterion are eliminated from the data matrix before clustering. Similarly, genes which do not change by a threshold amount with a threshold significance value in a minimum number of experiments are eliminated before clustering.

The error model described in Section 3.2.1 plays two roles. First, it is used to produce an estimated log ratio of the expression level of each gene in each yeast strain as compared to the wild type. Correlations of these log ratios across all genes or across all strains serve as similarity measures in the clustering. Second, the noise model is used to produce p-values measuring the statistical significance of observed differences between mutant strains and the wild type. These significance values, along with the estimated log ratios, form the basis of the significance cuts which precede clustering.

Hughes *et al.* identified the function of a number of uncharacterized genes based on the consistent inclusion of these genes in clusters with other genes of known function. For example, Figure 6 shows an expression level matrix representing a subset of profile (strain) and transcript (gene) clusters. The labels on the left-hand side represent clusters of strains whose mutations or chemical treatments are known to affect certain cellular pathways. The strains in which the yeast ergosterol biosynthesis pathway is affected exhibit a unique expression profile characterized by the high expression level of certain transcripts corresponding to the columns labeled "ergosterol." Clustered among these strains is a mutant strain in which the uncharacterized gene *YER044c* has been deleted (row labeled *yer044c*). The fact that the gene *YER044c* (column labeled *YER044c*) falls in the same transcript cluster as other genes whose expression is augmented by disruption of the ergosterol biosynthesis pathway further suggests that *YER044c* is involved in this pathway. After

[7] Hughes *et al.* do not specify how they decide the order of two groups with a common parent for graphical display purposes. Eisen *et al.* [6] use simple measures such as average expression level to decide the ordering, placing the group with the lower measure first in the final ordering.

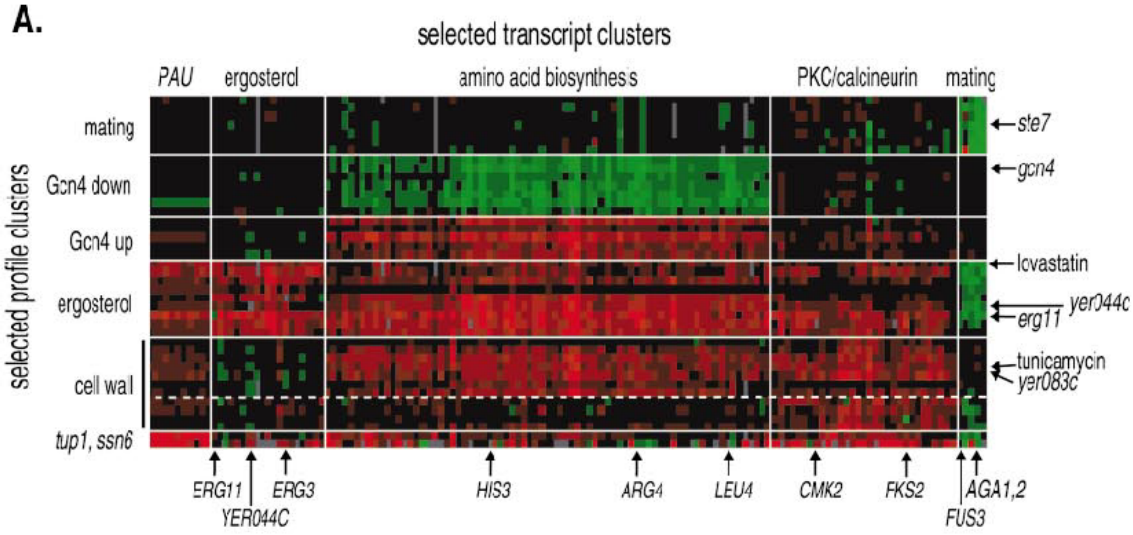

Figure 6: A matrix of selected rows and columns from Figure 5. Groups of columns correspond to prominent gene clusters responding to interference with the ergosterol biosynthesis pathway. Groups of rows correspond to yeast strains in which mutation or chemical treatment has affected a particular pathway. Color scale is identical to Figure 5. Reproduced from [16].

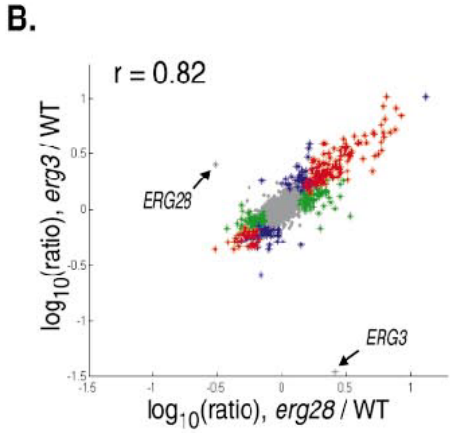

Figure 7: Comparison of profiles of an *erg28*-deleted strain and an *erg3*-deleted strain. Genes that changed significantly from wild type with $p \leq .01$ in both experiments are indicated in red; genes changing significantly in only the *erg3*-deleted strain or the *erg28*-deleted strain are in blue or green, respectively; genes which are anticorrelated with $p \leq .01$ in both experiments are in brown; and genes with $p > .01$ in both experiments are in gray. Reproduced from [16].

confirming its role in the pathway through further biochemical tests, Hughes *et al.* renamed this gene *ERG28*. Figure 7 shows a scatter plot comparing the expression profile of *erg28*, the mutant in which *ERG28* was deleted, with that of *erg3*, the mutant in which known ergosterol synthesis factor *ERG3* was deleted. Hughes *et al.* also verified that the previously uncharacterized human gene homologous to yeast gene *ERG28* functions in sterol biosynthesis.

Using similar techniques, Hughes *et al.* identified an unknown target of the drug dyclonine. The profile of a yeast strain treated with dyclonine proved most highly correlated with the profile of a deletion mutant lacking the gene *erg2*. Further experiments confirmed that dyclonine inhibits Erg2p, the product of the *erg2* gene.

4.2 Learning Bayesian networks from gene array data

Friedman *et al.* [9] attempt to discover regulatory interactions between genes rather than simply identifying genes with similar expression patterns. Gene products such as proteins play a significant role in regulating the transcription of other genes, either by binding to regulatory regions along the DNA or by directly affecting the transcription machinery [14]. Friedman *et al.* represent gene expression levels as nodes in a Bayesian network, and attempt to learn the graph structure of the network from multiple gene array measurements under different conditions.

A Bayesian network consists of a directed acyclic graph G along with a set of distributions of the form $P(X|\mathbf{Pa}^G(X))$, representing the probability distribution of the variable X conditioned on the values of its parents $\mathbf{Pa}^G(X)$. A Bayesian network satisfies the Markov assumption, which states that, conditioned on its parents $\mathbf{Pa}^G(X)$, a variable X is independent of all its nondescendants. Any joint distribution of variables in the graph can therefore be expressed in the form

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\mathbf{Pa}^G(X_i)). \tag{24}$$

An equivalence class of Bayesian network structures can be uniquely represented by a partially directed acyclic graph (PDAG) containing a mixture of directed and undirected edges [22].

Interactions between genes can be described by a causal network, which consists of a directed acyclic graph in which the parents of a variable represent its immediate causes. Perturbing the value of a parent perturbs the value of its children; on the other hand, perturbing the value of a child does not affect the value of its parent. To interpret a causal network as a Bayesian network, Friedman *et al.* rely on the *Causal Markov Assumption*, which states that, conditioned on the values of its immediate causes, a variable is independent of their causes.

The theory of learning Bayesian network from data has been developed extensively over the past decade [13]. Learning the structure of a genetic regulatory network from gene array data, however, poses a number of unique challenges. First, the number of network nodes (genes) is typically much larger than the number of expression profiles. Second, proteins and other compounds whose concentrations are not directly observed in gene array measurements play a crucial role in genetic regulation. Third, the expression level data is highly noisy.

Friedman *et al.* introduce several techniques to address these challenges. They attempt to learn network features rather than uniquely characterizing the network. Because of the shortage of data, no single model dominates the posterior distribution over models, so they analyze a set of high-scoring networks and characterize features common to most of them. These features take the form of Markov relations and order relations. Markov relations specify pairs of variables which are in each other's Markov blanket. The Markov blanket of X is the minimal set of variables $\mathbf{Y}$ such that, conditioned on $\mathbf{Y}$, X is independent of the rest of the variables in the network. Order relations specify pairs of variables X and Y such that X is an ancestor of Y in all the networks of a given equivalence class; under the Causal Markov Assumption, X is a direct or indirect cause of Y.

Friedman *et al.* estimate statistical confidence in identified features using the bootstrap method described in [8]. They select random subsets of the full data set, use each to infer a network structure, and measure the confidence of a feature as the fraction of these networks in which the feature is valid.

To allow reasonable learning times for a graph structure with hundreds of nodes, Friedman *et al.* utilize the "sparse candidate" algorithm described in [10]. This iterative algorithm restricts the set of candidate parents for each node using a local mutual information criterion. It then searches for an optimal network structure within these restrictions. Next, it uses the chosen network structure to select a new set of candidate parents. Specifically, if the node X has parents $\mathbf{Pa}(X)$ in the existing network structure, the algorithm selects as candidate parents those nodes Y giving the highest mutual information between X and $\{Y, \mathbf{Pa}(X)\}$.

Friedman *et al.* apply their Bayesian network learning technique to the data of Spellman *et al.* [28], who measured expression levels of synchronized *S. cerevisiae* (yeast) cell populations at various points in the mitotic cell cycle. This data set contains 76 spotted array expression profiles, each measuring the mRNA concentrations of 6177 genes. Friedman *et al.* analyze a subset of 800 genes

whose expression level varied significantly over the cell cycle. They also analyzed a smaller set of 250 genes. They treat each expression level measurement for a particular gene as an independent sample from a distribution, incorporating the temporal aspect of the measurement only by adding an additional variable denoting the cell cycle phase in all learned networks.

Friedman *et al.* consider two alternative local probability models. The first relies on a continuous Gaussian model for individual nodes, with linear dependencies between nodes such that all joint distributions are Gaussian. The second is a multinomial model, in which each variable takes one of three discrete states, corresponding roughly to under-expressed, normal, and over-expressed genes. A node takes a value 1 if the ratio between the measured expression level and the control expression level is greater than $2^{0.5}$, 0 if the ratio is between $2^{0.5}$ and $2^{-0.5}$, and -1 if the ratio is below $2^{-0.5}$. The control level is determined either as the average expression level of all genes in the experiment, or through measurements of control mRNAs spiked into each test sample in fixed concentrations.

Friedman *et al.* test the robustness of their techniques by analyzing both the linear Gaussian and the multinomial model using sets of 250 and 800 genes. For the multinomial model, they experiment with different expression level thresholds in the discretization process and with a normalization procedure which rescales the expression levels before discretization such that each gene has the same mean and variance. Order relations proved relatively robust to all of these variations (Figure 8). The authors identified several dominant genes which appear to directly or indirectly control the expression levels of a surprising number of other genes. The identity of these genes makes biological sense, as most encode nuclear proteins, or proteins involved in the budding and sporulation process which play an important role in the cell cycle.

The Markov relations proved much less robust to variations in probability models, discretization levels, and normalization method (Figure 8). The highest confidence Markov features identified through each incarnation of the method do represent pairs of genes which biologically might be expected to share a regulatory relationship, but these features differ almost entirely between the linear-Gaussian and multinomial models, and also depend significantly on the choice of threshold for the multinomial model.

Although Friedman *et al.* point out that their algorithm can learn regulatory relationships from noisy measurement data, they do not make explicit use of a noise model. The Rosetta or BEAM methods could be used, at a minimum, to determine which genes were significantly up- or down-regulated in each expression profile, providing more principled thresholds for the multinomial model.

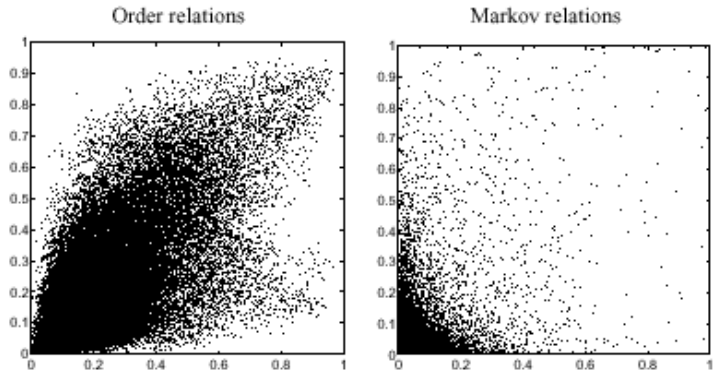

Figure 8: Comparison of feature confidence levels inferred using the multinomial model (x axis) and the linear Gaussian model (y axis). Each point represents one feature. Order relation features are in the left plot, while Markov relation features are on the right. Reproduced from [9].

Even better, an algorithm for learning Bayesian networks might take into account the statistical certainty associated with each input measurement.

A more serious limitation of the Bayesian network method lies in its initial assumptions. The assumption that a gene expression level depends only on the expression levels of other genes *at the same point in time* is doubtful at best for gene regulation. One gene typically regulates another through a protein product, which effectively integrates the activity of the controlling gene over some period of time. This issue is particularly salient for the cell cycle data set chosen by Friedman *et al.*, because cell cycle activators expressed in one stage of the cell cycle are known to promote gene expression at later stages. For example, mRNA expression of the genes *ace2*, *swi5*, and *mcm1* peaks during the M phase of the cell cycle. The protein products of these genes form a complex which binds to the promoter of genes such as *swi4* and *sic1* [27], leading to maximum expression of these genes during the G1 phase of the cycle, approximately 25 minutes later [28]. One might be able to partially overcome this limitation by adding additional network nodes corresponding to protein products of individual genes. The basic assumption that the underlying graph structure is acyclic is also questionable, as genes may form regulatory cycles [27].

5 Conclusions

The computational analyses of Hughes *et al.* and of Friedman *et al.* represent very different philosophies in terms of mathematical rigor and biological accuracy. Hughes *et al.* developed their methods empirically using biological data. The model and its derivation show no semblance

of rigor. However, the model takes into account the effects which are most relevant biologically, at least for the authors' purposes. This model has found immediate — if sometimes incorrect — application by other biologists. On the other hand, Friedman *et al.* use rigorous analysis techniques developed in the graphical analysis community. The most questionable aspect of their work lies not in the computational techniques, but in the biological assumptions they make to apply these techniques to gene array data. Application of their method requires a thorough understanding of advanced tools for graphical model analysis, putting it outside the reach of most biologists. The field of gene array analysis is sorely in need of algorithms which combine statistical rigor with biological accuracy, and which can be implemented as software accessible to biologists. Despite the oversimplification inherent in its underlying noise model, the work of Li and Wong embodies both these characteristics. Perhaps the BEAM method, which is based on Bayesian analysis but can be applied through use of look-up tables precomputed for a particular noise model and prior model, will possess similar appeal.

Acknowledgements

The BEAM work discussed here grew out of a course project in 6.892/7.93 (Functional Computational Genomics), taught by Tommi Jaakkola, David Gifford, and Richard Young. I collaborated with Jon Murnick, Nicola Rinaldi, Voichita Marinescu, Ryan Rifkin, and Richard Young in this research. Nicola Rinaldi and Jon Murnick provided helpful comments on the topics discussed in this area exam paper.

References

[1] ChipDB: A searchable database of gene expression. http://chipdb.wi.mit.edu.

[2] Genechip 3.1 expression analysis algorithm tutorial. Technical report, Affymetrix, 1999.

[3] Yidong Chen, Edward R. Dougherty, and Michael L. Bittner. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, pages 364–374, 1997.

[4] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[5] R. O. Dror, J. G. Murnick, N. J. Rinaldi, V. D. Marinescu, R. M. Rifkin, and R. A. Young. A Bayesian approach to transcript estimation from gene array data: The BEAM technique. Unpublished manuscript.

[6] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95:14863–8, 1998.

[7] Olga Ermolaeva, Mohit Rastogi, Kim D. Pruitt, Gregory D. Schuler, Michael L. Bittner, Yidong Chen, Richard Simon, Paul Meltzer, Jeffrey M. Trent, and Mark S. Boguski. Data management and analysis for gene expression arrays. *Nature Genetics*, 20:19–23, 1998.

[8] N. Friedman, M. Goldszmidt, and A. Wyner. Data analysis with bayesian networks: a bootstrap approach. In *Proc. Fifteenth Conf. on Uncertainty in Artificial Intelligence*, 1999.

[9] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 2000.

[10] N. Friedman, I. Nachman, and D. Pe'er. Learning bayesian network structure from massive datasets: The "sparse candidate" algorithm. In *Proc. Fifteenth Conf. on Uncertainty in Artificial Intelligence*, 1999.

[11] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–7, 1999.

[12] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Maximum likelihood estimation of optimal scaling factors for expression array normalizations. In *SPIE Bios 2001*, 2001.

[13] D. Heckerman. A tutorial on learning with bayesian networks. In M. I. Jordan, editor, *Learning in Graphical Models*. Kluwer, Dordrecht, Netherlands, 1998.

[14] F. C. Holstege, E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. A. Young. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95:717–728, 1998.

[15] T. R. Hughes, M. J. Marton, et al. Addendum to Hughes *et al.*, Cell 102. http://www.rii.com/tech/pubs/cell_hughes.htm, 2000.

[16] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

[17] M. K. Kerr and G. A. Churchill. Experimental design for gene expression microarrays. *Biostatistics*, to appear.

[18] D. A. Lashkari, J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci U S A*, 94:13057–62, 1997.

[19] M. L. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A*, pages 9834–9, 2000.

[20] C. Li and W. H. Wong. Model-based analysis of oligonucelotide arrays: Expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98:31–6, 2001.

[21] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[22] J. Pearl and T. S. Verma. A theory of inferred causation. In *Proceeding of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 441–452, 1991.

[23] Nicola Rinaldi. Personal communication, May 2001.

[24] Nicola Rinaldi. Personal communication, based on phone conversation with Matthew Marton, May 2001.

[25] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[26] J. Schuchhardt, D. Beule, A. Malik, E. Wolski, H. Eickhoff, H. Lehrach, and H. Herzel. Normalization strategies for cDNA microarrays. *Nucleic Acids Res*, 28:E47, 2000.

[27] I. Simon, J. Barnett, N. Hannett, C. Harbison, N. Rinaldi, E. Kanin, J. Schreiber, T. Volkert, J. Wyrick, J. Zeitlinger, and R. A. Young. Genome wide location and function of yeast cell cycle transcription activators. Submitted.

[28] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Compreshensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molelcular Biology of the Cell*, 9:3273–97, 1998.

[29] T. S. Tanaka, S. A. Jaradat, M. K. Lim, G. J. Kargul, X. Wang, M. J. Grahovac, S. Pantano, Y. Sano, Y. Piao, R. Nagaraja, H. Doi, W. H. Wood, K. G. Becker, and M. S. Ko. Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proc Natl Acad Sci U S A*, 1:9127–32, 2000.

[30] J. C. Venter, M. D. Adams, E. W. Myers, et al. The sequence of the human genome. *Science*, 291:1304, 2001.

[31] L. Wodicka, H. Dong, M. Mittmann, M. H. Ho, and D. J. Lockhart. Genome-wide expression monitoring in *saccharomyces cerevisiae*. *Nature Biotechnology*, 15:1359–67, 1997.