Maximum likelihood estimation of optimal scaling factors for expression array normalization

Alexander J. Hartemink[a], David K. Gifford[a,b], Tommi S. Jaakkola[b], and Richard A. Young[c]

[a]MIT Laboratory for Computer Science
200 Technology Square, Cambridge, MA 02139

[b]MIT Artificial Intelligence Laboratory
200 Technology Square, Cambridge, MA 02139

[c]Whitehead Institute for Biomedical Research
Nine Cambridge Center, Cambridge, MA 02142

ABSTRACT

Data from expression arrays must be comparable before it can be analyzed rigorously on a large scale. Accurate normalization improves the comparability of expression data because it seeks to account for sources of variation obscuring the underlying variation of interest. Undesirable variation in reported expression levels originates in the preparation and hybridization of the sample as well as in the manufacture of the array itself, and may differ depending on the array technology being employed. Published research to date has not characterized the degree of variation associated with these sources, and results are often reported without tight statistical bounds on their significance. We analyze the distributions of reported levels of exogenous control species spiked into samples applied to 1280 Affymetrix arrays. We develop a model for explaining reported expression levels under an assumption of primarily multiplicative variation. To compute the scaling factors needed for normalization, we derive maximum likelihood and maximum a posteriori estimates for the parameters characterizing the multiplicative variation in reported spiked control expression levels. We conclude that the optimal scaling factors in this context are weighted geometric means and determine the appropriate weights. The optimal scaling factor estimates so computed can be used for subsequent array normalization.

Keywords: normalization, scaling, microarray, oligonucleotide array, DNA chip, maximum likelihood, maximum a posteriori, MAP

1. INTRODUCTION

Expression arrays provide a powerful mechanism for measuring genomic expression levels within populations of cells. These arrays permit the simultaneous detection and measurement of tens of thousands of species of mRNA in a single experiment. The vast quantity of data being generated by these arrays presents us with a significant opportunity to transform biology, medicine, and pharmacology using systematic computational methods.[1–10] If it can be suitably leveraged, the impact of this genomic expression data on the understanding of basic cellular processes, the diagnosis and treatment of disease, and the efficacy of targeted therapeutics will be profound.

The effective utilization of large amounts of genomic expression data relies on the data being both available and comparable. Ensuring availability requires that we gather disperse stores of genomic expression data into large, publicly-available databases[11–14] and represent data using standardized exchange formats currently under development.[15,16] However, data availability is of little value unless it is accompanied by data comparability. The need for data comparability is the foundation that undergirds the issue of normalization.

Section 2 of this paper characterizes the different sources of variation in expression array data in order to provide a suitable background and motivation for the problem of normalization. We then formulate the normalization problem in Section 3 and present maximum likelihood (ML) and maximum a posteriori (MAP) estimates for parameters used in modeling the reported expression levels in Sections 4 and 5, respectively. In Section 6, we use these estimates to calculate optimal chip scaling factors for a data set consisting of 1280 Affymetrix GeneChip arrays. We close in Section 7 by discussing these results and offering some directions for further investigation.

Email addresses (in author order): `amink@mit.edu`, `gifford@mit.edu`, `tommi@mit.edu`, `young@wi.mit.edu`

2. SOURCES OF VARIATION IN EXPRESSION ARRAY DATA

We seek to learn how cells variously express their different genes in response to the diverse genetic and environmental environments they encounter. We define these sources of variation collectively as *interesting variation*. Unfortunately, reported expression levels also include other sources of variation that obscure the variation of interest. Sources of *obscuring variation* include variation introduced during the process of sample preparation, during the manufacture of the array, during the hybridization of the sample on the array, and during the scanning and analysis of fluorescent intensity after hybridization. We discuss each of these sources of variation below.

2.1. Sources of interesting variation

Variation in the expression of genes arises at many different levels. At the lowest level, even if we consider a specific gene in a specific cell under a specific environmental condition, there may be variation in the level of gene expression since mRNA transcription and decay are discrete stochastic processes. In practice, most of the variation at this level is hidden by the limitations of current array technology since we cannot measure genetic expression for a single cell but are constrained to measure an ensemble average over a population of cells.

At the next level, if we examine the expression of multiple genes, we introduce more data variation in that different genes are expressed in cells at distinct levels. For a specific genotype and environmental milieu, this variation in the levels of expression across the various genes in a cell's genome gives rise to an *expression profile* for the cell population, acting as a genetic signature of sorts.

At yet the next level, if we observe the expression of genes under a diversity of conditions, we introduce even more data variation because the expression profiles for populations of cells depend dramatically on the genetic and environmental conditions that attain when the cells are observed. For example, knocking out the activity of a protein, altering the temperature, modifying the nutritive environment, or exposing cells to agents that induce infection, mutation, cellular stress, or signaling can all have a significant influence on the expression profile of the population.

2.2. Sources of obscuring variation

Sources of variation introduced during the preparation of sample include variation during mRNA extraction and isolation, variation in the introduction of fluorescent dye, and variation in the rate of dye incorporation. These are influenced by pipette error, temperature fluctuations, and reagent quality.

Sources of variation introduced during the manufacture of the array include variation in the amount of probe present at each feature or spot and variation in the hybridization efficiency of the probes for their mRNA targets. The factors that influence these sources of variation depend upon the type of array being used. In the case of Affymetrix GeneChip oligonucleotide arrays,[17,18] probe concentration and efficiency are influenced by substrate surface characteristics, linker effects, probe design and density, and hybridization kinetics and thermodynamics.[19] In the case of printed microarrays,[20] probe concentration and efficiency are influenced by substrate surface characteristics, cross-linking effects, cDNA library selection and amplification, hybridization kinetics and thermodynamics, and probe deposition technology.

Sources of variation introduced during hybridization of the sample on the array include variation in the amount of sample applied to the array and variation in the amount of target hybridized to the probe. The amount of target-probe hybridization is influenced by the nature and concentrations of the buffers being used, the temperature and duration of the competitive hybridization reaction, the amount of cross-hybridization interference, and the possibility of probe saturation.

Sources of variation introduced after array hybridization include variation in optical measurements, variation in the fluorescent intensity computed from the scan image, and, in the case of printed arrays, variation in the optical response of the different dyes present in the sample. These can be influenced by spot misalignment, discretization effects, imaging algorithms, and scanner lens and laser irregularities.

2.3. Separating interesting variation from obscuring variation

As reported expression levels are a combination of interesting variation and obscuring variation, we need a suitable method for separating the two, where possible. Ideally, given reported levels of expression for a collection of genes across a number of experiments, we would like to develop statistically sound estimates for the levels of gene expression that include interesting variation but exclude, or otherwise account for, obscuring variation. In this paper, we develop a model that serves as a first step for deriving such estimates in the specific context of Affymetrix GeneChip arrays. While there remain sources of obscuring variation that cannot be accounted for in the model, we seek to present a simple model that adequately explains a substantial amount of this variation.

3. PROBLEM FORMULATION

Let the reported expression levels of M spiked controls from a set of N Affymetrix GeneChips be denoted x_{ij} where i indexes the spiked controls and ranges from 1 to M, while j indexes the chips and ranges from 1 to N. The reported spiked control expression levels form an $M \times N$ matrix as shown:

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{MN} \end{bmatrix} \tag{1}$$

We assume that a fixed amount of each spiked control is added to all chips. We denote the true level of expression for each spiked control i to be m_i, for all settings of j.

The reported expression level for each spiked control has a number of sources of variation, as discussed in Section 2. For example, the level reported depends on the actual quantity of control material pipetted into the sample and the actual amount of sample-control mixture injected into the GeneChip. The manufacture of the chip and the density of the probes present on the chip introduce more variation. The temperature of hybridization and variations in the buffer makeup also contribute to differences in reported levels. Because each of these sources of error is multiplicative, we assume that the true expression levels are modified by a multiplicative factor r_j which may (or may not) be different for each chip j and also by a random multiplicative error e_{ij} for each i and j. Under this assumption of purely multiplicative error, we have in formal terms:

$$x_{ij} = m_i \times r_j \times e_{ij} \tag{2}$$

where the e_{ij} factors are assumed to be fairly small and close to 1. For convenience, we transform this equation logarithmically so that the multiplicative errors become additive. Let $y_{ij} = \log(x_{ij})$, $\mu_i = \log(m_i)$, $\rho_j = \log(r_j)$, and $\epsilon_{ij} = \log(e_{ij})$ for all i and all j. The matrix of reported spiked controls after transformation becomes:

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1N} \\ y_{21} & y_{22} & \cdots & y_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ y_{M1} & y_{M2} & \cdots & y_{MN} \end{bmatrix} \tag{3}$$

and the equation describing the error model becomes:

$$y_{ij} = \mu_i + \rho_j + \epsilon_{ij} \tag{4}$$

We assume that ϵ_{ij} is randomly distributed and is drawn from a central normal distribution with variance σ_i^2. We permit the variance σ_i^2 to be different for each spiked control i to account for the fact that different probes on Affymetrix arrays may have different underlying variances in terms of their response to targets. With these assumptions in place, we have a model describing how the (log) reported expression levels for the spiked controls are distributed:

$$y_{ij} \sim N(\mu_i + \rho_j, \sigma_i^2) \tag{5}$$

4. MAXIMUM LIKELIHOOD (ML) ESTIMATION

With a model describing how the (log) reported expression levels for the spiked controls are distributed, we can use maximum likelihood estimation to derive optimal values for the scaling factors necessary to properly normalize each Affymetrix GeneChip. First, we form the log-likelihood $\mathcal{L}$ for observing the data y_{ij} under the assumption of normality outlined in the previous section:

$$\begin{aligned} \mathcal{L} &= \log \left(\prod_{i=1}^{M} \prod_{j=1}^{N} P(y_{ij}|\mu_i, \rho_j, \sigma_i^2) \right) && (6) \\ &= \sum_{i=1}^{M} \sum_{j=1}^{N} -\frac{1}{2} \left(\log(2\pi\sigma_i^2) + \frac{(y_{ij} - \mu_i - \rho_j)^2}{\sigma_i^2} \right) && (7) \end{aligned}$$

Then, we solve for the values of μ_i, ρ_j, and σ_i^2 that maximize the (log) likelihood of observing the data:

$$\arg\max_{\mu_i, \rho_j, \sigma_i^2} \mathcal{L} \tag{8}$$

Setting $\frac{\partial \mathcal{L}}{\partial \mu_i} = 0$, $\frac{\partial \mathcal{L}}{\partial \rho_j} = 0$, and $\frac{\partial \mathcal{L}}{\partial \sigma_i^2} = 0$ in turn yields estimates for the values of the parameters in question:

$$\begin{aligned} \hat{\mu_i} &= \frac{1}{N} \sum_{j=1}^{N} (y_{ij} - \hat{\rho_j}) && (9) \\ \hat{\rho_j} &= \frac{\sum\limits_{i=1}^{M} (\hat{\sigma_i^2})^{-1} (y_{ij} - \hat{\mu_i})}{\sum\limits_{i=1}^{M} (\hat{\sigma_i^2})^{-1}} && (10) \\ \hat{\sigma_i^2} &= \frac{1}{N} \sum_{j=1}^{N} (y_{ij} - \hat{\mu_i} - \hat{\rho_j})^2 && (11) \end{aligned}$$

As the estimates of the $2M + N$ unknown parameters are all coupled, it is necessary to iterate this solution until convergence, which can be done in rounds. Each round monotonically increases the likelihood of the observed values y_{ij} under the model.

Once the iteration of estimates has converged, the N estimates for ρ_j that emerge can be used to derive optimal scaling factors for the N chips. We define the optimal scaling factor for chip j to be s_j and compute it as shown:

$$s_j \equiv \frac{1}{\hat{r_j}} = e^{-\hat{\rho_j}} = \prod_{i=1}^{M} \left(\frac{\hat{m_i}}{x_{ij}} \right)^{w_i} \tag{12}$$

where $\hat{m_i} = e^{\hat{\mu_i}}$ and we have defined the weights w_i to be:

$$w_i \equiv \frac{(\hat{\sigma_i^2})^{-1}}{\sum_{i=1}^{M} (\hat{\sigma_i^2})^{-1}} \tag{13}$$

The optimal scaling factors are simply weighted geometric means of the ratios between $\hat{m_i}$ and x_{ij}, as might be expected, where the weight associated with each spiked control is inversely proportional to the estimated variance for that spiked control.

5. MAXIMUM A POSTERIORI (MAP) ESTIMATION

The fact that the estimates of μ_i, ρ_j, and σ_i^2 are computed iteratively can lead to a problem: as the optimal scaling factors are weighted geometric means where the weights are inversely proportional to the estimated variances, the spiked control with the least variance is weighted increasingly more with each iteration until, in the limit, a single spiked control can become scaled to its mean while the other spiked controls are essentially ignored. This can happen because the variance of the dominant spiked control approaches zero as it is rescaled uniformly to its mean. To avoid this pathological behavior and leverage the information about optimal scaling factors present in each of the spiked controls rather than simply one of the spiked controls, we modify the solution to incorporate a regularization term for the variances. This is accomplished by establishing prior distributions over possible values of the parameters and then estimating the maximum a posteriori (MAP) values of those parameters. In our context, we need only establish a prior for the variances σ_i^2; we can assume a flat prior over the means μ_i and log ratios ρ_j since we do not need regularization terms for these parameters. The assumption of flat priors for μ_i and ρ_j means that the prior terms for these parameters can be set to unity, and therefore the MAP updates for $\hat{\mu_i}$ and $\hat{\rho_j}$ are identical to the ML updates for these parameters.* Formally, we seek to maximize the posterior probability distribution for the parameters given the reported expression levels:

$$P(\mu_i, \rho_j, \sigma_i^2 | y_{ij}) \propto P(y_{ij} | \mu_i, \rho_j, \sigma_i^2) \cdot P(\sigma_i^2) \tag{14}$$

The likelihood term of the previous section reappears in this Bayesian formulation but is now accompanied by the prior distribution over the variances, serving as a regularization term.

As the likelihood is normally distributed, we assume a conjugate form for the prior over the variances, namely, a Wishart distribution. If we further assume that our prior belief about the variances is uninformative in the sense that we have no reason to believe, *a priori*, that the value of σ_i^2 should be different for one value of i than for any other, the multidimensional prior takes a relatively simple factorized form:

$$P(\sigma_i^2) = \prod_{i=1}^{M} C(\alpha, t) \left(\frac{1}{\sigma_i^2}\right)^{\frac{\alpha-3}{2}} e^{-\frac{t}{2\sigma_i^2}} \tag{15}$$

Having defined the likelihood term and the prior term, we can proceed to maximize the *a posteriori* probability by taking partial derivatives with respect to μ_i, ρ_j, and σ_i^2 and setting them equal to zero once again, yielding estimates for the values of the parameters in question. This results in the same equations for $\hat{\mu_i}$ and $\hat{\rho_j}$ as given above in (9) and (10), but a new equation for $\hat{\sigma_i^2}$:

$$\hat{\sigma_i^2} = \frac{\sum_{j=1}^{N} (y_{ij} - \hat{\mu_i} - \hat{\rho_j})^2 + t}{N + \alpha - 3} \tag{16}$$

A non-zero prior setting for t prevents the estimates of σ_i^2 from converging to zero for any i during the iteration process (except perhaps in the limit of infinite data).

6. RESULTS

We collected mRNA expression data from over 320 distinct experiments using Affymetrix *Saccharomyces cerevisiae* GeneChips. As these were low-density Ye6100 chips, four chips were required per experiment to sample the levels of expression for all 6179 ORFs. So we have spiked control measurements from 1280 Affymetrix GeneChips. The collection is comprised of a wide range of different experimental conditions, including various comparisons between wild type *S. cerevisiae* strains and strains with genetic deletions and functional knockouts, as well as time course experiments detailing the response of *S. cerevisiae* to various environmental stresses like heat shock, pH fluctuations, carbon-source shifts, and exposure to reactive oxygen species, for example.

Four different control species (DapX, LysX, PheX, and ThrX) are spiked into the extracted mRNA samples before hybridization. Each Affymetrix Ye6100 GeneChip has a set of three probes for each of the spiked control species. One probe contains features binding near the $3'$ end of the target, one contains features binding near the middle of the target, and one contains features binding near the $5'$ end of the target. Thus, a total of 12 spiked control expression levels are reported for each GeneChip. We use the 12×1280 array of reported spiked control expression levels to produce estimates of the optimal scaling factors for the 1280 GeneChips using the ML and MAP estimation methods shown above. In both cases, results are nearly identical, though we display results below for only the MAP estimates because of their regularization properties. We set $\alpha = 3$ and $t = 1$ in our estimation, but varying these parameters by an order of magnitude has little effect on the results (not shown).

*Although the form of the updates is the same, the actual values of the estimates may be different as the values of $\hat{\mu_i}$ and $\hat{\rho_j}$ depend on the values of $\hat{\sigma_i^2}$.

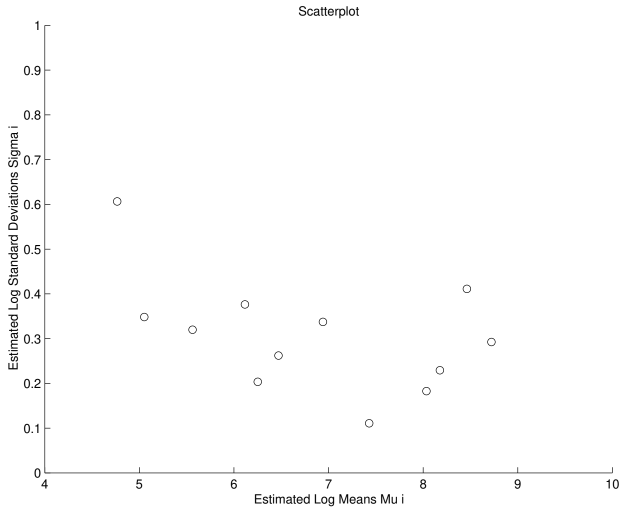

Figure 1. Scatterplot of estimated standard deviation of log expression levels σ_i versus estimated mean of log expression levels μ_i for 12 spiked controls. The estimated standard deviations are generally relatively low and constant, with the exception of the first point. The greater estimated standard deviation associated with the point corresponding to the lowest average level of expression suggests that additive error may be playing a significant role.

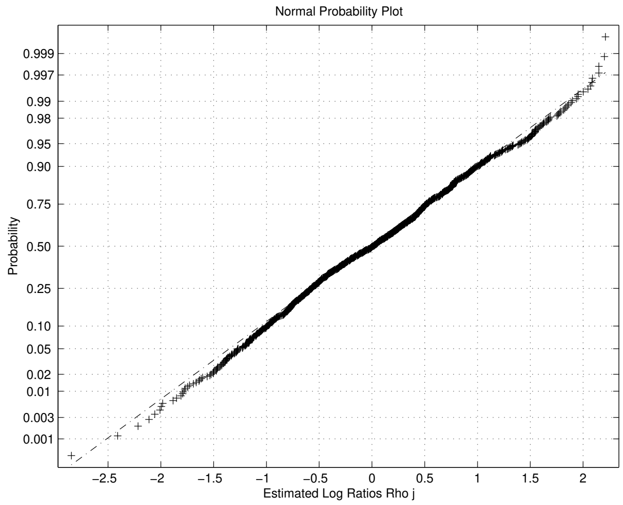

Figure 2. Normal probability plot of estimated log ratios ρ_j. The plot reveals that the estimated log ratios are roughly normally distributed.

Figure 1 is a scatterplot of the estimated standard deviation of log expression levels σ_i versus the estimated mean of log expression levels μ_i for the 12 spiked controls added to the 1280 Affymetrix Ye6100 GeneChips. The estimated standard deviations are generally relatively low and constant, with the exception of the first point. The greater estimated standard deviation associated with the point corresponding to the lowest average level of expression suggests that additive error may be playing a significant role. Although additive error tends to be swamped by multiplicative error for large levels of expression, it should be incorporated in a more complicated model in order to adequately capture sources of variation when expression levels are low.

Figure 2 is a normal probability plot of the estimated log ratios ρ_j. The plot reveals that the estimated log ratios are roughly normally distributed. Recall that we made no assumptions about the form of the distribution of ρ_j in our modeling.

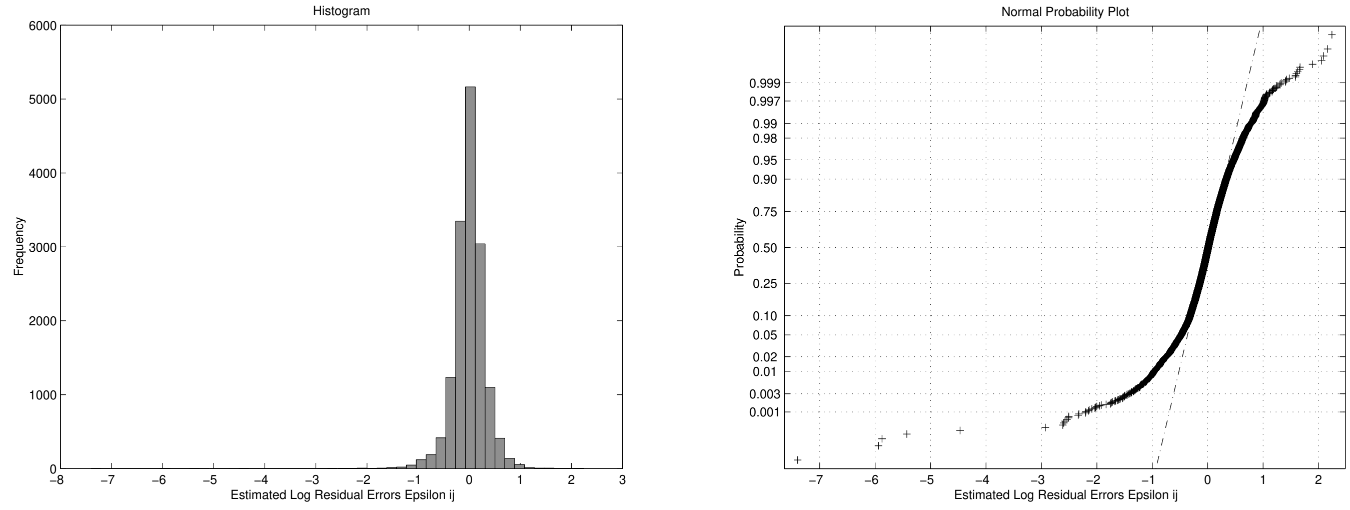

Figure 3. Histogram and normal probability plot of residual errors ϵ_{ij}. The histogram plot on the left appears normal at first glance but the normal probability plot of the same data on the right reveals that the distribution is actually fairly heavy-tailed.

Figure 3 contains both a histogram and a normal probability plot of the residual errors ϵ_{ij}, which represent variations in y_{ij} that remain unexplained after optimally estimating μ_i and ρ_j. The histogram plot on the left appears normal at first glance but the normal probability plot of the same data on the right reveals that the distribution is actually fairly heavy-tailed. We discuss this in greater detail in Section 7.

Once we have computed the estimates of μ_i, ρ_j, and σ_i^2, we can use the estimates of ρ_j to compute the optimal scaling factors for the 320 chips. Figure 4 provides scatterplots of the standard deviation of log expression level versus the mean of log expression level for the 6179 yeast genes with probes on the Ye6100 Affymetrix arrays. The plot in the upper left represents unnormalized expression levels from 320 experiments over widely varying experimental conditions. The plot in the upper right represents unnormalized expression levels from 8 wild type experiments with constant experimental conditions. The lower plots are the same as the corresponding upper plots but are computed from normalized expression levels. Considered column-wise, the plots in Figure 4 reveal that the normalization process is successful in reducing the overall variation in the data. In the case of the 320 experiments, the average standard deviation drops from 0.97 to 0.83, while in the case of the 8 wild type experiments, the average standard deviation drops from 0.73 to 0.54. The fact that points on each plot with low average levels of expression tend to have a much greater standard deviation suggests, consistent with our observations in Figure 1, that additive error is playing a significant role at lower levels of expression.

7. CONCLUSION

In order for data from genomic expression arrays to be comparable, it is necessary that we understand the different sources of variation that are present in reported gene expression levels. To effectively separate the interesting variation in reported expression levels from the obscuring variation, we need statistically sound methods for deriving estimates for the levels of gene expression that include interesting variation but exclude, or otherwise account for, obscuring variation.

In this paper, we attempted to carefully characterize the different sources of variation present in reported gene expression levels. In the context of Affymetrix GeneChips with spiked control probes, we presented an initial model for explaining observed expression levels under the assumption of multiplicative error. We made no assumptions regarding the distributions of the scaling factors applied to each chip, but assumed that the log residual errors were normally distributed with a possibly different variance for each spiked control. Under these assumptions, we developed maximum likelihood (ML) and maximum a posteriori (MAP) estimates of the unknown parameters and used these estimates to compute optimal scaling factors for subsequent array normalization.

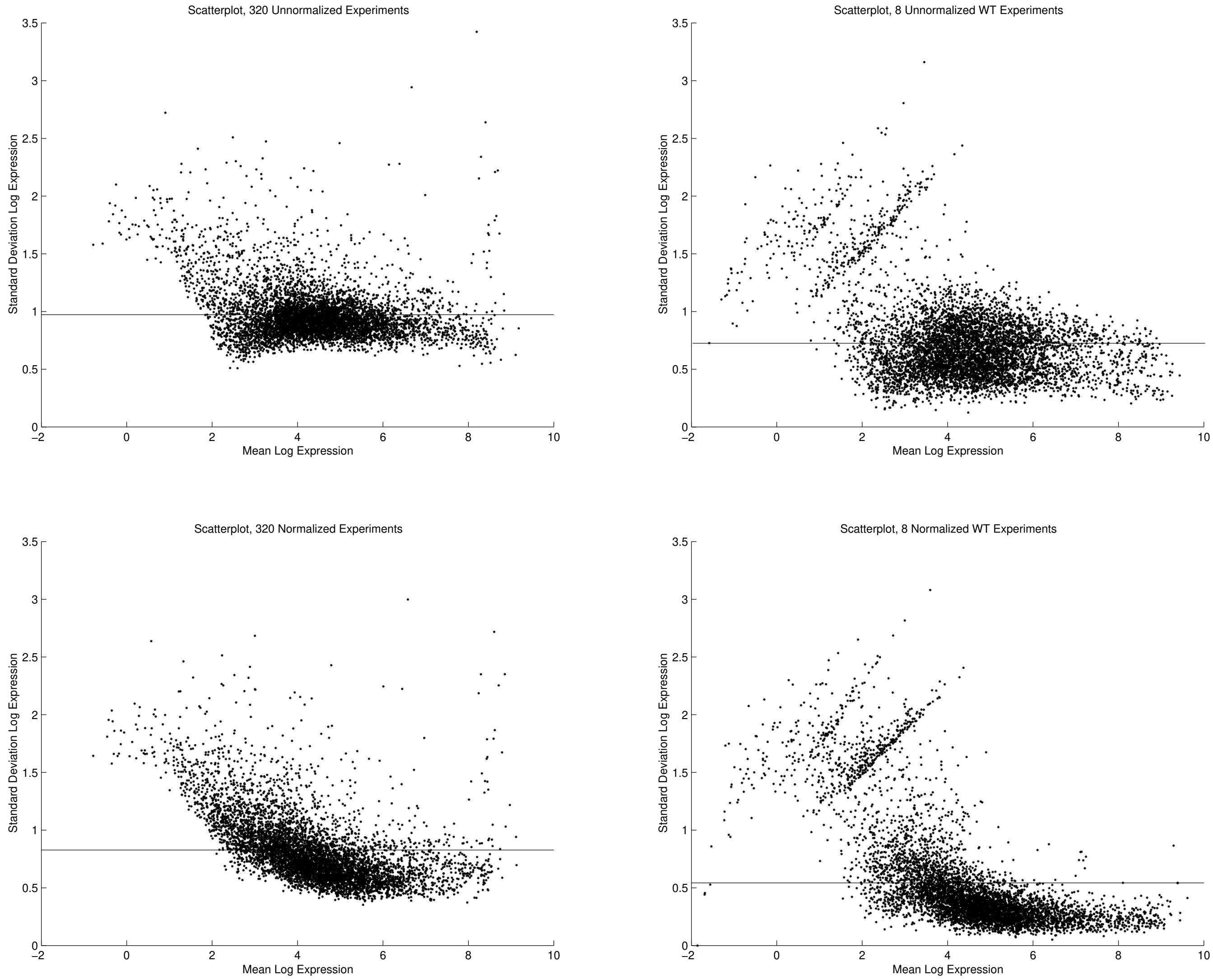

Figure 4. Scatterplots of standard deviation of log expression level versus mean of log expression level for 6179 yeast genes. The plot in the upper left represents unnormalized expression levels from 320 experiments over widely varying experimental conditions. The plot in the upper right represents unnormalized expression levels from 8 wild type experiments with constant experimental conditions. The lower plots are the same as the corresponding upper plots but are computed from normalized expression levels.

There are a number of interesting directions for extending this work. First, the formulation of our initial model is fairly simple in that it is entirely multiplicative and does not incorporate enough terms to adequately model all the sources of variation present in reported expression levels. A more sophisticated model would consider both additive and multiplicative effects, as well as more complicated interaction terms.

Second, the error residuals ϵ_{ij} are clearly not normal as postulated in the context of our initial model. We could consider alternative descriptions of the distribution of these residuals, but the non-normality may be another indication of the simplicity of the model discussed above. It is possible that a more sophisticated model would result in error residuals that are distributed more normally.

Third, although the characterization of different sources of variation presented in Section 2 is applicable to all array technologies, the specific model postulated in this paper is intended only for data gathered on Affymetrix GeneChips that use spiked controls. However, the methodology is general and the ideas should be useful in other settings with suitable modification. Moreover, we are in the process of developing methods for making data from Affymetrix GeneChips comparable with data from printed microarrays, enabling the comparison of data across technology platforms.

ACKNOWLEDGMENTS

The authors would like to thank Barb Cutler and Tarjei Mikkelsen for assistance during the inception of this work, and Ron Dror, Peter Young, and John Barnett for helpful discussions along the way. Hartemink gratefully acknowledges support from the NIH and Merck.

REFERENCES

1. A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young, "Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks," in *Pacific Symposium on Biocomputing*, vol. 6, 2001.
2. 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**, pp. 109–126, 2000.
3. M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr., and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. Natl. Acad. Sci. USA* **97**(1), pp. 262–267, 2000.
4. N. S. Holter, M. M., A. Maritan, M. Cieplak, J. R. Banavar, and N. V. Fedoroff, "Fundamental patterns underlying gene expression profiles: Simplicity from complexity," *Proc. Natl. Acad. Sci. USA* **97**(15), pp. 8409–8414, 2000.
5. N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," in *4th Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, ACM-SIGACT, April 2000.
6. A. A. Alizadeh and et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature* **403**, pp. 503–511, 2000.
7. M. K. Kerr, M. Martin, and G. A. Churchill, "Analysis of variance for gene expression microarray data," *Journal of Computational Biology* **7**, 2000.
8. S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nature Genetics* **22**, pp. 281–285, 1999.
9. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proc. Natl. Acad. Sci. USA* **96**, pp. 2907–2912, March 1999.
10. T. Chen, V. Filkov, and S. S. Skiena, "Identifying gene regulatory networks from experimental data," in *3rd Annual International Conference on Computational Molecular Biology (RECOMB'99)*, ACM-SIGACT, April 1999.
11. "ArrayExpress." *http://www.ebi.ac.uk/arrayexpress/*.
12. "Gene expression omnibus, GEO." *http://www.ncbi.nlm.nih.gov/geo/*.
13. "GeneX." *http://www.ncgr.org/research/genex/*.
14. "Genetic analysis technology consortium, GATC." *http://www.gatconsortium.org/*.
15. "GeneXML." *http://www.ncgr.org/research/genex/genexml.html*.
16. "Gene expression markup language, GEML." *http://www.geml.org/*.
17. A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. A. Fodor, "Light-generated oligonucleotide arrays for rapid DNA sequence analysis," *Proc. Natl. Acad. Sci. USA* **91**, pp. 5022–5026, 1994.
18. G. McGall, J. Labadie, P. Brock, G. Wallraff, T. Nguyen, and W. Hinsberg, "Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists," *Proc. Natl. Acad. Sci. USA* **93**, pp. 13555–13560, 1996.
19. G. McGall. personal communication, 1999.
20. J. DeRisi, V. Iyer, and P. O. Brown, "The MGuide, a complete guide to building your own microarrayer, version 1.1." *http://cmgm.stanford.edu/pbrown/mguide/*, 1998.