Efficient Computation of BEAM Estimates

Ron O. Dror and Jonathan G. Murnick

Massachusetts Institute of Technology

This document serves as a supplement to the paper "Bayesian Estimation of Transcript Levels Using a General Model of Array Measurement Noise" by Ron O. Dror, Jonathan G. Murnick, Nicola J. Rinaldi, Voichita D. Marinescu, Ryan M. Rifkin, and Richard A. Young, published in the *Journal of Computational Biology* in 2003. The journal paper describes the Bayesian Estimaton of Array Measurements (BEAM) technique for creating a noise model for gene arrays from experimental data and for using such a noise model to identify significant changes in expression level, combine repeated measurements, or deal with negative expression level measurements. This supplement describes efficient methods to compute the Bayesian estimates discussed in the paper, as well as the associated uncertainties and p-values.*

*This supplement follows the probabilistic notation of the journal paper. Latin characters without serifs (x, $\mathbf{y}$, f) denote random variables, and the corresponding characters with serifs (x, $\mathbf{y}$, f) denote sample values of those random variables. Boldface characters ($\mathbf{y}$) denote vector quantities.

- $p_{\mathsf{x}}(x)$: probability density of x, evaluated at x.
- $p_{\mathsf{x}|\mathsf{y}}(x|y)$: conditional probability density of x given an observation y of y.
- $E(\mathsf{x}|y)$: expected value of x given the observation y of y. This is shorthand for $E(\mathsf{x}|\mathsf{y} = y)$.
- $\hat{x}(y)$: Bayes least squares estimate of x given the observation y of y.
- $\sigma^2_{\hat{x}}(y)$: variance of the posterior distribution of x given the observation y of y.

We assume a noise model of the form

$$y_{ij} = g_{ij} x_{ij} + f_j + e_{ij},$$

where x_{ij} represents the true transcript level of gene j on chip i, and y_{ij} represents the normalized measurement of gene j on chip i (see Equations (1) and (2) of the journal paper). The noise terms g_{ij} and e_{ij} are independent and identically distributed for each chip i and gene j, and the f_j are independent and identically distributed for each gene j. The numerical methods described here apply regardless of the form of the noise probability density functions p_{g}, p_{e}, and p_{f}, or of the prior probability density function p_{x}. For any given noise and prior models, these computations need only be carried out once, to create lookup tables for estimators, uncertainties, and p-values.

1 Estimation of Absolute Transcript Levels

The expectation and variance of the true transcript level given a set of repeated measurements are:

$$\hat{\mathsf{x}}(\mathbf{y}) = E[\mathsf{x}|\mathbf{y}] = \frac{1}{p_{\mathbf{y}}(\mathbf{y})} \int x p_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|x) p_{\mathsf{x}}(x) dx \tag{1}$$

$$\sigma^2_{\hat{x}}(\mathbf{y}) = E[(\mathsf{x} - \hat{x})^2|\mathbf{y}] = E[\mathsf{x}^2|\mathbf{y}] - E[\mathsf{x}|\mathbf{y}]^2. \tag{2}$$

Evaluation of $p_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|x)$ and $p_{\mathbf{y}}(\mathbf{y})$ requires that we consider all combinations of values of error coefficients and true transcript levels that could produce the set of observations $\mathbf{y} =$

$(y_1, y_2, \cdots, y_n)$.

The presence of an unknown bias term f common to all the observations in $\mathbf{y}$ complicates the computation, because $p_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|x)$ is not equivalent to $p_{\mathsf{y}|\mathsf{x}}(y_1|x) \cdots p_{\mathsf{y}|\mathsf{x}}(y_n|x)$. In an earlier version of our work,[1] we assumed that these two quantities were equivalent. In other words, we treated the bias term as a measurement-specific error. Although sometimes reasonable, this simplifying approximation may lead to underestimation of ratios and failure to identify statistically significant differences in the expression level of a particular gene, combined with exaggerations of the increase in confidence due to repeating an experiment. Discarding the bias term entirely would lead to the opposite problems — overestimation of ratios, differences in gene expression labeled falsely as significant, and underestimation of the the increase in confidence due to repeating an experiment. The computational methods described in this supplement treat the bias term rigorously.

Conditioned on the value of the true level x *and* the bias f, the measurements $(y_1, y_2, \cdots, y_n)$ are independent. That is, $p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y}|x, f) = p_{\mathsf{y}|\mathsf{x},\mathsf{f}}(y_1|x, f) \cdots p_{\mathsf{y}|\mathsf{x},\mathsf{f}}(y_n|x, f)$. We could evaluate Equation (1) by computing $E[\mathsf{x}|\mathbf{y}, f]$ for each value of the bias f, and then marginalizing over f. Taking advantage of the fact that the bias term is additive, however, we can evaluate

Equation (1) more efficiently by rewriting it as

$$\begin{aligned} E[\mathsf{x}|\mathbf{y}] &= \frac{\int x p_{\mathsf{x}}(x) p_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|x) dx}{\int p_{\mathsf{x}}(x) p_{\mathbf{y}|\mathsf{x}}(\mathbf{y}|x) dx} \\ &= \frac{\int x p_{\mathsf{x}}(x) \left[\int p_{\mathbf{y},\mathsf{f}|\mathsf{x}}(\mathbf{y}, f|x) df\right] dx}{\int p_{\mathsf{x}}(x) \left[\int p_{\mathbf{y},\mathsf{f}|\mathsf{x}}(\mathbf{y}, f|x) df\right] dx} \\ &= \frac{\int x p_{\mathsf{x}}(x) \left[\int p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y}|x, f) p_{\mathsf{f}} df\right] dx}{\int p_{\mathsf{x}}(x) \left[\int p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y}|x, f) p_{\mathsf{f}} df\right] dx} \\ &= \frac{\int x \left[\int p_{\mathsf{x}}(x) p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y}|x, f) dx\right] p_{\mathsf{f}}(f) df}{\int \left[\int p_{\mathsf{x}}(x) p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y}|x, f) dx\right] p_{\mathsf{f}}(f) df} \\ &= \frac{\int \left[\int x p_{\mathsf{x}}(x) p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y} - f|x, 0) dx\right] p_{\mathsf{f}}(f) df}{\int \left[\int p_{\mathsf{x}}(x) p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y} - f|x, 0) dx\right] p_{\mathsf{f}}(f) df}, \end{aligned} \tag{3}$$

where $\mathbf{y} - f = (y_1 - f, y_2 - f, \cdots, y_n - f)$.

If we define

$$m_k(\mathbf{y}) \equiv \int x^k \, p_{\mathsf{x}}(x) p_{\mathbf{y}|\mathsf{x},f}(\mathbf{y}|x, 0) dx \tag{4}$$

for $k = 0, 1, 2$, we can rewrite Equation (3) as

$$E[x|y] = \frac{\int m_1(\mathbf{y} - f) p_{\mathsf{f}}(f) df}{\int m_0(\mathbf{y} - f) p_{\mathsf{f}}(f) df}. \tag{5}$$

A similar derivation yields

$$E[x^2|\mathbf{y}] = \frac{\int m_2(\mathbf{y} - f) p_{\mathsf{f}}(f) df}{\int m_0(\mathbf{y} - f) p_{\mathsf{f}}(f) df}, \tag{6}$$

allowing us to compute the uncertainty measure of Equation (2).

The integrals in Equations (5) and (6) can be implemented as convolutions of m_k and

p_{f}. Evaluating $m_k(\mathbf{y})$ requires computation of $p_{\mathbf{y}|\mathsf{x},f}(\mathbf{y}|x,0)$, the conditional distribution of $\mathbf{y}$ given x and the fact that the bias is zero. To compute $p_{\mathbf{y}|\mathsf{x},f}(\mathbf{y}|x,0)$, note that $p_{\mathbf{y}|\mathsf{x},f}(\mathbf{y}|x,0) = p_{\mathsf{y}_1|\mathsf{x},f}(y_1|x,0)\cdots p_{\mathsf{y}_n|\mathsf{x},f}(y_n|x,0)$. When $f=0$, our noise model becomes $\mathsf{y} = \mathsf{g}\mathsf{x} + \mathsf{e}$. For each value x, $p_{\mathsf{y}|\mathsf{x},f}(\mathbf{y}|x,0)$ is therefore a convolution of $p_{\mathsf{g}x}$ and p_{e}, where $p_{\mathsf{g}x}(\cdot) = \frac{1}{x}p_g(\frac{\cdot}{x})$.

2 Estimation of Transcript Level Ratios

The log ratio estimator can be written as[†]

$$\begin{aligned}
\hat{r}(\mathbf{y_a},\mathbf{y_b}) &= E\left[\log\frac{\mathsf{x_a}}{\mathsf{x_b}}\middle|\mathbf{y_a},\mathbf{y_b}\right] \qquad (7)\\
&= E[\log\mathsf{x_a} - \log\mathsf{x_b}|\mathbf{y_a},\mathbf{y_b}]\\
&= \frac{1}{p_{\mathbf{y_a},\mathbf{y_b}}(\mathbf{y_a},\mathbf{y_b})}\iint(\log x_a - \log x_b)p_{\mathbf{y_a},\mathbf{y_b}|\mathsf{x_a},\mathsf{x_b}}(\mathbf{y_a},\mathbf{y_b}|x_a,x_b)p_{\mathsf{x}}(x_a)p_{\mathsf{x}}(x_b)dx_a dx_b.
\end{aligned}$$

Unfortunately, $\mathbf{y_a}$ and $\mathbf{y_b}$ are not independent given $\mathsf{x_a}$ and $\mathsf{x_b}$, because of the presence of a common bias. We wish to avoid evaluating $E[\log\mathsf{x_a} - \log\mathsf{x_b}|\mathbf{y_a},\mathbf{y_b},f]$ separately for each value f. Replacing x with $\log\mathsf{x_a} - \log\mathsf{x_b}$ in the derivation of Section 1 gives a variant of Equation (5)

$$\hat{r}(\mathbf{y_a},\mathbf{y_b}) = \frac{\int n_1(\mathbf{y_a}-f,\mathbf{y_b}-f)p_{\mathsf{f}}(f)df}{\int n_0(\mathbf{y_a}-f,\mathbf{y_b}-f)p_{\mathsf{f}}(f)df} \qquad (8)$$

[†]In this section, we omit the base of the logarithm from our notation; the same formulas apply for logarithms of any base.

where

$$\begin{aligned} n_k(\mathbf{y_a}, \mathbf{y_b}) &\equiv \iint (\log \mathsf{x_a} - \log \mathsf{x_b})^k \, p_{\mathbf{y_a},\mathbf{y_b}|\mathsf{x_a},\mathsf{x_b},f}(\mathbf{y_a}, \mathbf{y_b}|x_a, x_b, 0) p_{\mathsf{x_a},\mathsf{x_b}}(x_a, x_b) dx_a dx_b \\ &= \iint (\log \mathsf{x_a} - \log \mathsf{x_b})^k \, p_{\mathbf{y}|\mathsf{x},f}(\mathbf{y_a}|x_a, 0) p_{\mathbf{y}|\mathsf{x},f}(\mathbf{y_b}|x_b, 0) p_{\mathsf{x}}(x_a) p_{\mathsf{x}}(x_b) dx_a dx_b. \end{aligned}$$

To simplify this formula for $k = 0, 1, 2$, define q_k by

$$q_k(\mathbf{y}) \equiv \int (\log x)^k \, p_{\mathbf{y}|\mathsf{x},f}(\mathbf{y}|x, 0) p_{\mathsf{x}}(x) dx.$$

Evaluation of $q_k(\mathbf{y})$ is similar to evaluation of $m_k(\mathbf{y})$, as described in Section 1. Then

$$\begin{aligned} n_0(\mathbf{y_a}, \mathbf{y_b}) &= q_0(\mathbf{y_a}) q_0(\mathbf{y_b}) \\ n_1(\mathbf{y_a}, \mathbf{y_b}) &= q_1(\mathbf{y_a}) q_0(\mathbf{y_b}) - q_0(\mathbf{y_a}) q_1(\mathbf{y_b}) \\ n_2(\mathbf{y_a}, \mathbf{y_b}) &= q_2(\mathbf{y_a}) q_0(\mathbf{y_b}) - 2 q_1(\mathbf{y_a}) q_1(\mathbf{y_b}) + q_0(\mathbf{y_a}) q_2(\mathbf{y_b}) \end{aligned}$$

The numerator and denominator of Equation (8) can then be computed by convolution. To compute the uncertainty $\sigma^2_{\hat{r}}(\mathbf{y_a}, \mathbf{y_b})$ associated with the ratio estimate, we write the variance of the posterior distribution as

$$\sigma^2_{\hat{r}}(\mathbf{y_a}, \mathbf{y_b}) = E[(\mathsf{r} - \hat{r})^2 | \mathbf{y_a}, \mathbf{y_b}] = E[r^2(\mathbf{y_a}, \mathbf{y_b}) | \mathbf{y_a}, \mathbf{y_b}] - E[r(\mathbf{y_a}, \mathbf{y_b}) | \mathbf{y_a}, \mathbf{y_b}]^2.$$

A derivation similar to that of Equation (8) yields

$$E[r^2(\mathbf{y_a}, \mathbf{y_b})|\mathbf{y_a}, \mathbf{y_b}] = \frac{\int n_2(\mathbf{y_a} - f, \mathbf{y_b} - f)p_{\mathsf{f}}(f)df}{\int n_0(\mathbf{y_a} - f, \mathbf{y_b} - f)p_{\mathsf{f}}(f)df}.$$

3 Significance tests

To compute the p-values used to test whether or not two sets of array measurements corresponding to the same gene in two cultures are significantly different, we use the formula

$$p = \int_{x^*=0}^{\infty} F_{x^*}(\hat{r})p_{\mathsf{x}|\mathbf{y}}(x^*|\mathbf{y_a}, \mathbf{y_b})dx^*, \tag{9}$$

where $F_{x^*}(\hat{r}) = \int_{|\hat{r}(\mathbf{y_1},\mathbf{y_2})|\geq|\hat{r}^*|} p_{\mathbf{y}|\mathsf{x}}(\mathbf{y_1}, \mathbf{y_2}|x^*)d\mathbf{y_1}d\mathbf{y_2}$.

We begin by evaluating the function $F_{x^*}(\hat{r})$ for all true levels x^* and ratio estimates $\hat{r}$. For each value of x^*, we compute $p_{\mathbf{y}|\mathsf{x}}(\mathbf{y_1}, \mathbf{y_2}|x^*)$ over a wide range of observations $\mathbf{y_1}$ and $\mathbf{y_2}$.[‡] To do so, note that

$$\begin{aligned} p_{\mathbf{y}|\mathsf{x}}(\mathbf{y_1}, \mathbf{y_2}|x^*) &= \int p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y_1}, \mathbf{y_2}|x^*, f)p_{\mathsf{f}}(f)df \\ &= \int p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y_1} - f, \mathbf{y_2} - f|x^*, 0)p_{\mathsf{f}}(f)df. \end{aligned} \tag{10}$$

We compute $p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y_1}-f, \mathbf{y_2}-f|x^*, 0) = p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y_1}-f|x^*, 0)p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y_2}-f|x^*, 0)$ and then evaluate Equation (10) by convolution. $F_{x^*}(\hat{r})$ amounts to a cumulative distribution of the log ratio estimates $|\hat{r}(\mathbf{y_1}, \mathbf{y_2})|$ over the space of possible measurements $(\mathbf{y_1}, \mathbf{y_2})$. We compute this from

[‡]The vector $\mathbf{y_1}$ represents the same number of measurements as $\mathbf{y_a}$, and $\mathbf{y_2}$ represents the same number of measurements as $\mathbf{y_b}$.

a weighted histogram of the values $|\hat{r}(\mathbf{y_1}, \mathbf{y_2})|$, with the weights given by $p_{\mathbf{y}|\mathsf{x}}(\mathbf{y_1}, \mathbf{y_2}|x^*)$.

Once we have stored lookup tables for the functions $F_{x^*}(\hat{r})$, we can compute the significance value p according to Equation (9). We evaluate $p_{\mathsf{x}|\mathbf{y}}(x^*|\mathbf{y_a}, \mathbf{y_b})$ using an expansion similar to that of Equation (3):

$$p_{\mathsf{x}|\mathbf{y}}(x^*|\mathbf{y_a}, \mathbf{y_b}) = \frac{\int p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y_a}, \mathbf{y_b}|x^*, f) p_{\mathsf{x}}(x^*) p_{\mathsf{f}}(f) df}{\iint p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y_a}, \mathbf{y_b}|x^*, f) p_{\mathsf{x}}(x^*) p_{\mathsf{f}}(f) df\, dx^*} \tag{11}$$

We compute the numerator of Equation (11) using the fact that $p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y_a}, \mathbf{y_b}|x^*, f) = p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y_a} - f, \mathbf{y_b} - f|x^*, 0) = p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y_a} - f|x^*, 0) = p_{\mathbf{y}|\mathsf{x},\mathsf{f}}(\mathbf{y_b} - f|x^*, 0)$. The denominator can be computed in the same manner, or as $\int m_0(\mathbf{y} - f) p_{\mathsf{f}}(f) df$, where m_0 is defined as in Equation (4) and $\mathbf{y}$ represents the observations of both $\mathbf{y_a}$ and $\mathbf{y_b}$.

To accelerate the computations when $\mathbf{y_a}$ and $\mathbf{y_b}$ each represent multiple measurements, one may approximate Equation (9) by $F_{\hat{x}(\mathbf{y_a}, \mathbf{y_b})}(\hat{r}(\mathbf{y_a}, \mathbf{y_b}))$. This amounts to using the expected value of the true transcript level x^* under the null hypothesis, rather than its entire distribution. For our noise and prior models, this approximation produces results close to those of Equation (9).

4 Convergence and sampling

Similar analysis can be used to show that the estimators we propose are guaranteed to converge if the additive noise terms f and e have finite expectation and the prior $p_{\mathsf{x}}(x)$ falls off faster than $\frac{1}{x}$ beyond some transcript level. The uncertainty measures will converge if the additive noise terms have finite variance and the prior falls off faster than $\frac{1}{x^2}$ beyond some

level. Repeated measurements lead to faster convergence under more general conditions. The p-values are well-defined whenever the estimators converge. We found that using our noise model, the integrals converged quickly, with the numerical integrations requiring only a limited sampling range. If one wishes to use an alternative model that lacks desirable convergence properties, one could select estimators based on the median of the probability distribution rather than on its expected value.[2]

References

1. 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. In *Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology*, Washington, D.C., April 2002.

2. J. Theilhaber, S. Bushnell, A. Jackson, and R. Fuchs. Bayesian estimation of fold-changes in the analysis of gene expression: The PFOLD algorithm. *Jour Comp Biol*, 8:585–614, 2001.