BIOINFORMATICS

Vol. 17 no. 6 2001
Pages 509–519

A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes

Pierre Baldi [1,3,*] *and Anthony D. Long* [2]

[1]*Department of Information and Computer Science and* [2]*Department of Ecology and Evolutionary Biology, Institute for Genomics and Bioinformatics, University of California at Irvine, Irvine, CA 92697-3425, USA*

Received on September 5, 2000; revised and accepted on February 15, 2001

ABSTRACT

Motivation: DNA microarrays are now capable of providing genome-wide patterns of gene expression across many different conditions. The first level of analysis of these patterns requires determining whether observed differences in expression are significant or not. Current methods are unsatisfactory due to the lack of a systematic framework that can accommodate noise, variability, and low replication often typical of microarray data.

Results: We develop a Bayesian probabilistic framework for microarray data analysis. At the simplest level, we model log-expression values by independent normal distributions, parameterized by corresponding means and variances with hierarchical prior distributions. We derive point estimates for both parameters and hyperparameters, and regularized expressions for the variance of each gene by combining the empirical variance with a local background variance associated with neighboring genes. An additional hyperparameter, inversely related to the number of empirical observations, determines the strength of the background variance. Simulations show that these point estimates, combined with a *t*-test, provide a systematic inference approach that compares favorably with simple *t*-test or fold methods, and partly compensate for the lack of replication.

Availability: The approach is implemented in software called Cyber-T accessible through a Web interface at www.genomics.uci.edu/software.html. The code is available as Open Source and is written in the freely available statistical language R.

Contact: pfbaldi@ics.uci.edu; tdlong@uci.edu

*To whom correspondence should be addressed.
[3]Also at Department of Biological Chemistry, College of Medicine, University of California, Irvine.

1 INTRODUCTION

DNA gene expression microarrays allow biologists to study genome-wide patterns of gene expression (DeRisi *et al.*, 1997; Eisen *et al.*, 1998; Holstege *et al.*, 1998). In these arrays, total RNA is reverse-transcribed to create either radioactive- or fluorescent-labeled cDNA which is hybridized with a large DNA library of gene fragments attached to a glass or membrane support. Phosphorimaging or other imaging techniques are used to produce expression measurements for thousands of genes under various experimental conditions. Use of these arrays is rapidly creating terabytes of information, potentially capable of providing fundamental insights into biological processes ranging from gene function, to development, to cancer (Spellman *et al.*, 1998; Alon *et al.*, 1999; Golub *et al.*, 1999; Lee *et al.*, 1999; White *et al.*, 1999; Ly *et al.*, 2000). Unfortunately, data analysis techniques for microarray data are still at an early stage of development (Zhang, 1999). Our goal here is to develop a general Bayesian statistical framework for the analysis of array data.

Gene expression array data can be analyzed on at least three levels of increasing complexity. First, the level of single genes, where one seeks to establish whether each gene in isolation behaves differently in a control versus a treatment situation. The second level considers gene combinations, where clusters of genes are analyzed in terms of common functionalities, interactions, co-regulation, and so forth. The third level attempts to infer the underlying regulatory regions and gene/protein networks that ultimately are responsible for the patterns observed. This paper focuses on the first level of analysis.

For simplicity, we assume that for each gene X we have a set of measurements $x_1^c, \ldots, x_{n_c}^c$ and $x_1^t, \ldots, x_{n_t}^t$ representing expression levels, or rather their logarithms, in both a control and treatment situation. Treatment is of course taken in a broad sense to mean any condition different from the control. For each gene, the fundamental

question we wish to address is whether the level of expression is significantly different in the two situations. While it might seem that standard statistical techniques could easily address such a problem, this is in fact not the case.

One approach commonly used in the current literature is a simple-minded fold approach, in which a gene is declared to have significantly changed if its average expression level varies by more than a constant factor, typically 2, between the treatment and control conditions. Inspection of gene expression data suggests, however, that such a simple '2-fold rule' is unlikely to yield optimal results, since a factor of 2 can have quite different significance depending on expression levels.

A related approach to the same question is the use of a t-test, for instance on the logarithm of the expression levels. This is similar to the fold approach because the difference between two logarithms is the logarithm of their ratio. This approach is not necessarily identical to the first because the logarithm of the mean is not equal to the mean of the logarithms; in fact it is always strictly greater, by convexity of the logarithm function. But with a reasonable degree of approximation, a test of the significance of the difference between the log expression levels of two genes is equivalent to a test of whether or not their fold change is significantly different from 1.

In a t-test, the empirical means m_c and m_t and variances s_c^2 and s_t^2 are used to compute a normalized distance between the two populations in the form:

$$t = (m_c - m_t) \Big/ \sqrt{\frac{s_c^2}{n_c} + \frac{s_t^2}{n_t}} \qquad (1)$$

where, for each population, $m = \sum_i x_i/n$ and $s^2 = \sum_i (x_i - m)^2/(n-1)$ are the well-known estimates for the mean and standard deviation. It is known that t follows approximately a Student distribution, with

$$f = \frac{[(s_c^2/n_c) + (s_t^2/n_t)]^2}{\frac{(s_c^2/n_c)^2}{n_c - 1} + \frac{(s_t^2/n_t)^2}{n_t - 1}} \qquad (2)$$

degrees of freedom. When t exceeds a certain threshold depending on the confidence level selected, the two populations are considered to be different. Because in the t-test the distance between the population means is normalized by the empirical standard deviations, this has the potential for addressing some of the shortcomings of the fixed fold-threshold approach. The fundamental problem with the t-test for microarray data, however, is that the repetition numbers n_c and/or n_t are often small because experiments remain costly or tedious to repeat, even with current technology. Small populations of size $n = 1, 2$ or 3 are still very common and lead, for instance, to significant underestimates of the variances. Thus a better framework is needed to address these shortcomings.

Here we develop a Bayesian probabilistic framework for microarray data, which bears some analogies with the framework used for sequence data (Baldi and Brunak, 2001) and addresses the problem of detecting gene differences. Because a complete Bayesian treatment is computationally demanding, we also develop approximate computational shortcuts to strike a balance between rigor and computational efficiency. In particular, we develop methods for the regularization of the t-test approach.

2 BAYESIAN PROBABILISTIC FRAMEWORK

Several decades of research in sequence analysis and other areas have demonstrated the advantages and effectiveness of probabilistic approaches to biological data. Indeed, DNA microarray data is characterized by a high degree of measurement noise and variability. Biological systems also have very high dimensionality: even in a large array experiment, only a very small subset of relevant variables is measured, or even under control. The vast majority of variables remain hidden and must be inferred or integrated out by probabilistic methods.

The general Bayesian statistical framework codifies how to proceed with data analysis and inference in a rational way. Under a small set of common sense axioms, it can be shown remarkably that subjective degrees of belief must obey the rules of probability and proper induction must proceed in a unique way, by propagation of information through Bayes theorem. In particular, at any given time, any hypothesis or model M can be assessed by computing its posterior probability in light of the data according to Bayes theorem: $P(M|D) = P(D|M)P(M)/P(D)$, where $P(D|M)$ is the data likelihood and $P(M)$ is the prior probability capturing any background information one may have.

Probabilistic modeling of microarray data

In sequence data, the most simple probabilistic model is that of a die (Figure 1), associated with the average composition of the family of DNA, RNA, or protein sequences under study. The next level of modeling complexity is a first-order Markov model with one die per position or per column in a multiple alignment. In spite of their simplicity, these models are routinely used, for instance as background models against which the performances of more sophisticated models can be assessed.

In array data, the simplest model would assume that all data points are independent from each other and extracted from a single continuous distribution, for instance a Gaussian distribution. While trivial, this Gaussian die model still requires the computation of interesting quantities, such as the average level of activity and its standard deviation, which can be useful to calibrate or assess

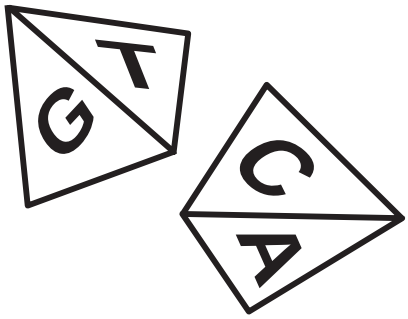

Fig. 1. DNA dice.

global properties of the data. The next equivalent level of modeling is a set of independent distributions, one for each dimension, i.e. for instance each gene. While it is obvious that genes interact with each other in complex ways and therefore are not independent, the independence approximation is still useful and underlies *any* attempt, probabilistic or other, to determine whether expression level differences are significant solely on a *gene-by-gene* basis.

Here we first assume that the expression-level measurements of a gene in a given situation have a roughly Gaussian distribution. In our experience, with common technologies this assumption is reasonable, especially for the *logarithm* of the expression levels, corresponding to log-normal raw expression levels. To the best of our knowledge, large-scale replicate experiments have not been carried out yet to make more precise assessments. It is clear, however, that other distributions, such as gammas or mixtures of Gaussians/gammas, could be introduced at this stage. These would impact the details of the analysis (see also Wiens, 1999), but not the general Bayesian probabilistic framework.

Thus, in what follows we assume that the data has been pre-processed—taking logarithms if needed—to the point where we can model the corresponding measurements of each gene in each situation (treatment or control) with a normal distribution $\mathcal{N}(x; \mu, \sigma^2)$. For each gene and each condition, we have a two-parameter model $w = (\mu, \sigma^2)$, and by focusing on one such model we can omit indices identifying the gene or the condition. Assuming that the observations are independent, the likelihood of the data D is given by:

$$\begin{aligned} P(D|\mu,\sigma^2) &\approx \prod_{i=1}^{n} \mathcal{N}(x_i;\mu,\sigma^2) \\ &= C(\sigma^2)^{-n/2} \mathrm{e}^{-\sum_i (x_i-\mu)^2/2\sigma^2} \\ &= C(\sigma^2)^{-n/2} \mathrm{e}^{-(n(m-\mu)^2+(n-1)s^2)/2\sigma^2}. \end{aligned} \quad (3)$$

Here and everywhere else, we write C to denote the normalizing constant of any distribution. All the information about the sample that is relevant for the likelihood is summarized in the sufficient statistics n, m, and s^2. The case in which either the mean or the variance of the Gaussian model is supposed to be known is of course easier and is well studied in the literature (Box and Tiao, 1973; Pratt *et al.*, 1995).

Priors

A full Bayesian treatment requires introducing a prior distribution $P(\mu, \sigma^2)$. The choice of a prior is part of the modeling process, and several alternatives are possible (Box and Tiao, 1973; Pratt *et al.*, 1995), a sign of the flexibility of the Bayesian approach rather than its arbitrariness. Several kinds of priors for the mean and variance of a normal distribution have been studied in the literature, including the noninformative improper prior and the conjugate prior. For microarray data, the conjugate prior seems to be more suitable and flexible, not only because of its convenient form, but also because it incorporates the basic observation that μ and σ^2 are typically *not* independent.

The conjugate prior. When both the prior and the posterior have the same functional form, the prior is said to be a conjugate prior. When estimating the mean alone of a normal model of known variance, the obvious conjugate prior is also a normal distribution. In the case of dice models for biological sequences, the standard conjugate prior is a Dirichlet distribution (Baldi and Brunak, 2001). The form of the likelihood in equation (3) shows that the conjugate prior density must also have the form $P(\mu|\sigma^2)P(\sigma^2)$, where the marginal $P(\sigma^2)$ is scaled inverse gamma (Appendix) and the conditional distribution $P(\mu|\sigma^2)$ is normal. This leads to a hierarchical model with a vector of four hyperparameters for the prior $\alpha = (\mu_0, \lambda_0, \nu_0, \sigma_0^2)$ with the densities:

$$P(\mu|\sigma^2) = \mathcal{N}(\mu; \mu_0, \sigma^2/\lambda_0) \quad (4)$$

and

$$P(\sigma^2) = \mathcal{I}(\sigma^2; \nu_0, \sigma_0^2). \quad (5)$$

The expectation of the prior is finite if and only if $\nu_0 > 2$. The prior $P(\mu, \sigma^2) = P(\mu, \sigma^2|\alpha)$ is given by:

$$C\sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp\left[-\frac{\nu_0}{2\sigma^2}\sigma_0^2 - \frac{\lambda_0}{2\sigma^2}(\mu_0-\mu)^2\right]. \quad (6)$$

Notice that it makes perfect sense with array data to assume *a priori* that μ and σ^2 are *dependent*, as suggested immediately by visual inspection of typical microarray data sets (Figure 2). The hyperparameters μ_0 and σ^2/λ_0 can be interpreted as the location and scale of μ, and

the hyperparameters ν_0 and σ_0^2 as the degrees of freedom and scale of σ^2. Applying Bayes theorem and after some algebra, the posterior has the same functional form as the prior

$$P(\mu, \sigma^2|D, \alpha) = \mathcal{N}(\mu; \mu_n, \sigma^2/\lambda_n)\mathcal{I}(\sigma^2; \nu_n, \sigma_n^2) \quad (7)$$

with

$$\mu_n = \frac{\lambda_0}{\lambda_0 + n}\mu_0 + \frac{n}{\lambda_0 + n}m \quad (8)$$

$$\lambda_n = \lambda_0 + n \quad (9)$$

$$\nu_n = \nu_0 + n \quad (10)$$

$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\lambda_0 n}{\lambda_0 + n}(m - \mu_0)^2. \quad (11)$$

The parameters of the posterior combine information from the prior and the data in a sensible way. The mean μ_n is a convex weighted average of the prior mean and the sample mean. The posterior degree of freedom ν_n is the prior degree of freedom plus the sample size. The posterior sum of squares $\nu_n\sigma_n^2$ is the sum of the prior sum of squares $\nu_0\sigma_0^2$, the sample sum of squares $(n-1)s^2$, and the residual uncertainty provided by the discrepancy between the prior mean and the sample mean.

While it is possible to use a prior mean μ_0 for gene expression data, in many situations it is sufficient to use $\mu_0 = m$. The posterior sum of squares is then obtained precisely as if one had ν_0 additional observations all associated with deviation σ_0^2. While superficially this may seem like setting the prior after having observed the data (MacKay, 1992), a similar effect is obtained using a preset value μ_0 with $\lambda_0 \to 0$, i.e. with a very broad standard deviation so that the prior belief about the location of the mean is essentially uniform and vanishingly small. The selection of the hyperparameters for the prior is discussed in more detail below.

It can readily be shown that the conditional posterior distribution $P(\mu|\sigma^2, D, \alpha)$ of the mean is normal $\mathcal{N}(\mu_n, \sigma^2/\lambda_n)$, the marginal posterior $P(\mu|D, \alpha)$ of the mean is Student $t(\nu_n, \mu_n, \sigma_n^2/\lambda_n)$, and the marginal posterior $P(\sigma^2|D, \alpha)$ of the variance is scaled inverse gamma $\mathcal{I}(\nu_n, \sigma_n^2)$.

In the literature, semi-conjugate prior distributions are also used where the functional form of the prior distributions on μ and σ^2 are the same as in the conjugate case (normal and scaled inverse gamma, respectively) but independent of each other, i.e. $P(\mu, \sigma^2) = P(\mu)P(\sigma^2)$. However, as previously discussed, this assumption of independence is unlikely to be suitable for DNA microarray data. More complex priors also could be constructed using mixtures, a mixture of conjugate priors leading to a mixture of conjugate posteriors.

3 PARAMETER POINT ESTIMATES

The posterior distribution $P(\mu, \sigma^2|D, \alpha)$ is the fundamental object of Bayesian analysis and contains the relevant information about *all* possible values of μ and σ^2. However, it can be useful to collapse this information-rich distribution into single point estimates. This can be done in a number of ways. In general, the most robust answer is obtained using the mean of the posterior (MP) estimate. An alternative is to use the mode of the posterior, or MAP (maximum *a posteriori*) estimate. For completeness, we derive and compare both kinds of estimates for the conjugate prior. By integration, the MP estimate is given by

$$\mu = \mu_n \quad \text{and} \quad \sigma^2 = \frac{\nu_n}{\nu_n - 2}\sigma_n^2 \quad (12)$$

provided $\nu_n > 2$. If we take $\mu_0 = m$, we then get the following MP estimate:

$$\mu = m \quad \text{and} \quad \sigma^2 = \frac{\nu_n\sigma_n^2}{\nu_n - 2} = \frac{\nu_0\sigma_0^2 + (n-1)s^2}{\nu_0 + n - 2} \quad (13)$$

provided $\nu_0 + n > 2$. This is the default estimate implemented in the Cyber-T software described below. From equation (7), the MAP estimates are:

$$\mu = \mu_n \quad \text{and} \quad \sigma^2 = \frac{\nu_n\sigma_n^2}{\nu_n - 1}. \quad (14)$$

If we use $\mu_0 = m$, these reduce to:

$$\mu = m \quad \text{and} \quad \sigma^2 = \frac{\nu_n\sigma_n^2}{\nu_n - 1} = \frac{\nu_0\sigma_0^2 + (n-1)s^2}{\nu_0 + n - 1}. \quad (15)$$

The modes of the marginal posterior are given by

$$\mu = \mu_n \quad \text{and} \quad \sigma^2 = \frac{\nu_n\sigma_n^2}{\nu_n + 2}. \quad (16)$$

In practice, equations (13) and (15) give similar results and can be used with gene expression arrays. The slight differences between the two closely matches what is seen with Dirichlet priors on sequence data (Baldi and Brunak, 2001), equation (13) generally being a slightly better choice. The Dirichlet prior is equivalent to the introduction of pseudo-counts to avoid setting the probability of any amino acid or nucleotide to zero. In array data, few observation points are likely to result in a poor estimate of the variance. With a single point ($n = 1$), for instance, we certainly want to refrain from setting the corresponding variance to zero; hence the need for regularization, which is achieved by the conjugate prior. In the MP estimate, the empirical variance is modulated by ν_0 'pseudo-observations' associated with a background variance σ_0^2.

4 FULL BAYESIAN TREATMENT AND HYPERPARAMETER POINT ESTIMATES

At this stage of modeling, each gene is associated with two models $w_c = (\mu_c, \sigma_c^2)$ and $w_t = (\mu_t, \sigma_t^2)$; two sets of hyperparameters α_c and α_t; and two posterior distributions $P(w_c|D, \alpha_c)$ and $P(w_t|D, \alpha_t)$. A full probabilistic treatment would require introducing prior distributions over the hyperparameters. These could be integrated out to obtain the true posterior probabilities $P(w_c|D)$ and $P(w_t|D)$, which then could be integrated over all values of w_t and w_c to determine whether the two models are different or not. Notice that this approach is significantly more general than the plain t-test and could in principle detect interesting changes that are beyond the scope of the t-test or fold approaches. For instance, a gene with the same mean but a very different variance between the control and treatment situations goes undetected by these methods, although the change in variance might be biologically relevant. Even if we restrict ourselves to an analysis of the means μ_c and μ_t only, the probability $P(\mu_c \approx \mu_t|D, \alpha_t, \alpha_c)$ must be computed, and would typically require numerical integration. An alternative is to directly model the difference as a parameterized Gaussian with corresponding prior and perform a Bayesian hypothesis test (Baldi and Brunak, 2001). While the latter can be performed easily on today's computers, here we use a simple approximation strategy to the full Bayesian treatment that relies solely on point estimates.

Point estimates, however, require determining hyperparameter values, and this can be addressed in a number of ways (MacKay, 1992, 1999). Here again, one possibility is to define a prior on the hyperparameters and try to integrate them out in order to compute the true posterior $P(w|D)$ and determine the location of its mode, leading to true MAP estimates of w. More precisely, this requires integrating $P(w|\alpha)$ and $P(w|\alpha|D)$ with respect to the hyperparameter vector α. An alternative that avoids the integration of the hyperparameters is the evidence framework described in MacKay (1992). In the evidence framework, we compute point estimates of the hyperparameters by MAP estimation (MP would again require integrating over hyperparameters) over the posterior

$$P(\alpha|D) = \frac{P(D|\alpha)P(\alpha)}{P(D)}. \tag{17}$$

If we take a uniform prior $P(\alpha)$, then this is equivalent to maximizing the evidence $P(D|\alpha)$

$$\begin{aligned} P(D|\alpha) &= P(D|w, \alpha)P(w|\alpha)/P(w|D, \alpha) \\ &= P(D|w)P(w|\alpha)/P(w|D, \alpha). \end{aligned} \tag{18}$$

In principle, computing the evidence requires integrating out the parameters w of the model. Using the expression for the likelihood and the conjugate prior and posterior, however, we can here obtain the evidence without integration, directly from equation (18)

$$P(D|\alpha) = (2\pi)^{-n/2} \frac{\sqrt{\lambda_0}}{\sqrt{\lambda_n}} \frac{(\nu_0/2)^{\nu_0/2}}{(\nu_n/2)^{\nu_n/2}} \frac{\sigma_0^{\nu_0}}{\sigma_n^{\nu_n}} \frac{\Gamma(\nu_n/2)}{\Gamma(\nu_0/2)}. \tag{19}$$

The partial derivatives and critical points of the evidence are discussed in the Appendix, where it is shown, for instance, that the mode is achieved for $\mu_0 = m$.

5 IMPLEMENTATION

For efficiency, we have implemented an intermediate solution in which we use the t-test with the regularized standard deviation of equation (13) and the number of degrees of freedom associated with the corresponding augmented populations of points, which incidentally can be fractional. This solution has been implemented in a Web server called Cyber-T accessible at: http://www.128.200.5.223/CyberT/ (see also Appendix and http://www.genomics.uci.edu for more details). In Cyber-T, plain and Bayesian versions of the t-test can be performed on both the raw data and the log-transformed data.

In the simplest case, where we use $\mu_0 = m$, we must select the values of the background standard deviation σ_0^2, and its strength ν_0. The parameter ν_0 represents the degree of confidence in the background variance σ_0^2 versus the empirical variance. In Cyber-T, the value of ν_0 can be set by the user by clicking on the corresponding button. The smaller n, the larger ν_0 ought to be. A simple rule of thumb is to assume that $K > 2$ points are needed to properly estimate the standard deviation and keep $n + \nu_0 = K$. This allows for a flexible treatment of situations in which the number n of available data points varies from gene to gene. In our current implementation, we use a default of $K = 10$. A special case can be made for genes with activity levels close to the minimal detection level of the technology being used. The measurements for these genes being particularly unreliable, it may be wise to use a stronger prior for them with a higher value of ν_0 (this feature is currently not implemented).

For σ_0, one could use the standard deviation of the entire set of observations or, depending on the situation, of particular categories of genes. We favored a flexible implementation under which the background standard deviation is estimated by pooling together all the neighboring genes contained in a window of size w. Cyber-T automatically ranks the expression levels of all the genes and lets the user choose this window size using the corresponding button. The default is $w = 101$, corresponding to 50 genes immediately above and below the gene under consideration. Adaptive window sizes are briefly discussed in the last section, together with the possibility of deriving regression estimates of σ_0^2.

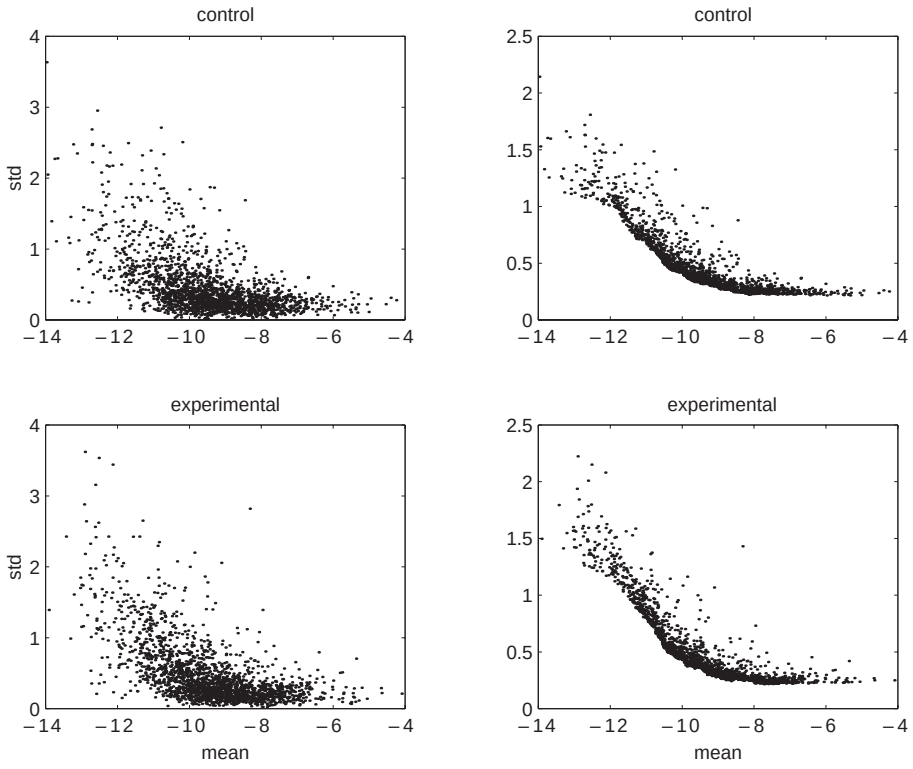

Fig. 2. DNA microarray experiment on *Escherichia coli*. Data obtained from reverse transcribed P^{33} labeled RNA hybridized to commercially available nylon arrays (Sigma Genosys) containing each of the 4290 predicted *E.coli* genes. The sample included a wild-type strain (control) and an otherwise isogenic strain lacking the gene for the global regulatory gene, IHF (treatment). $n = 4$ for both control and experimental situations. The horizontal axis represents the mean μ of the logarithm of the expression levels, and the vertical axis shows the corresponding standard deviations (std $= \sigma$). The left column corresponds to raw data; the right column to regularized standard deviations using equation (13). Window size is $w = 101$ and $K = 10$ (see main text). Data are from Arfin *et al.* (2000).

6 SIMULATIONS

We have used the Bayesian approach and Cyber-T to analyze a number of published and unpublished data sets. In every high density array experiment we have analyzed, we have observed a strong scaling of the expression variance over replicated experiments with the average expression level (on both a log-transformed and raw scale). As a result, a threshold for significance based solely on fold changes is likely to be too liberal for genes expressed at low levels and too conservative for highly expressed genes. While several biologically relevant results are reported elsewhere (Long *et al.*, 2001), we have found that the Bayesian approach compares favorably to a simple fold approach or a straight t-test and partially overcomes deficiencies related to low replication in a statistically consistent way.

One particularly informative data set for comparing the Bayesian approach to simple t-test or fold change is the high density array experiment reported in Arfin *et al.* (2000) comparing *Escherichia coli* cells that were wild type to cells that were mutant for the global regulatory protein Integration Host Factor (IHF). The main advantage of this data set is its four-fold replication for both wild type and mutant alleles. The regularizing effect of the Cyber-T prior based on the background standard deviation is shown for this data in Figure 2 and in the simulation described below. The figure clearly shows that standard deviations vary substantially over the range of expression levels, in this case roughly in a monotonic decreasing fashion, although other behaviors have also been observed. Interestingly, in these plots the variance in log-transformed expression levels is higher for genes expressed at lower levels rather than at higher ones. These plots confirm that genes expressed at low or near background levels may require a stronger value of ν_0, or alternatively could be ignored in expression analyses. The variance in the measurement of genes expressed at a low level is large enough that in many cases it will be difficult to detect significant changes in expression for this class of loci.

In analyzing the data we found that large fold changes in expression were often associated with P-values not indicative of statistical change in the Bayesian analysis, and conversely subtle fold changes were often highly significant as judged by the Bayesian analysis. In these two situations, the conclusions drawn using the Bayesian approach appear robust relative to those drawn from fold change alone, as large non-statistically significant fold changes were often associated with large measurement errors, and statistically significant genes showing less than 2-fold changes were often measured very accurately. As a result of the level of experimental replication seen in Arfin *et al.* (2000), we were able to look at the consistency of the Bayesian estimator relative to the t-test. We found that in independent samples of size 2 drawn from the IHF data set (i.e. two experiments versus two controls) the set of 120 most significant genes identified using the Bayesian approach had approximately 50% of their members in common, whereas the set of 120 most significant genes identified using the t-test had only approximately 25% of their members in common. This suggests that for 2-fold replication the Bayesian approach is approximately twice as consistent as a simple t-test at identifying genes as up- or down-regulated, although with only 2-fold replication there is a great deal of uncertainty associated with high density array experiments.

To further assess the Bayesian approach, here we simulate an artificial data set assuming Gaussian distribution of log expressions, with means and variances in ranges similar to those encountered in the data set of Arfin *et al.* (2000), with 1000 replicates for each parameter combination. Selected means for the log data and associated standard deviations (in brackets) are as follows: -6 (0.1), -8 (0.2), -10 (0.4), -11 (0.7), -12 (1.0). On this artificially generated data, we can compare the behavior of a simple ratio (2-fold and 5-fold) approach, with a simple t-test, with the Bayesian t-test using the default settings of Cyber-T. The main results, reported in Table 1, can be summarized as follows:

- By 5 replications (5 control and 5 treatment) the Bayesian approach and t-test give similar results.
- When the number of replicates is 'low' (2 or 3), the Bayesian approach performs better than the t-test.
- The false positive rate for the Bayesian and t-test approach are as expected (0.05 and 0.01 respectively) except for the Bayesian with very small replication (i.e. 2) where it appears elevated.
- The false positive rate on the ratios is a function of expression level and is much higher at lower expression levels. At low expression levels the false positive rate on the ratios is unacceptably high.
- For a given level of replication the Bayesian approach at $P < 0.01$ detects more differences than a 2-fold change except for the case of low expression levels (where the false positive rate from ratios is elevated).
- The Bayesian approach with 2 replicates outperforms the t-test with 3 replicates (or 2 versus 4 replicates).
- The Bayesian approach has a similar level of performance when comparing 3 treatments to 3 controls, or 2 treatments to 4 controls. This suggests an experimental strategy where the controls are highly replicated and a number of treatments less highly replicated.

7 DISCUSSION AND EXTENSIONS

We have developed a probabilistic framework for array data analysis to address a number of current approach shortcomings related to small sample bias and the fact that fold differences have different significance at different expression levels. The framework is a form of hierarchical Bayesian modeling with Gaussian gene-independent models. Although the Gaussian representation requires further testing, other distributions can easily be incorporated in a similar framework. As a first step, we have implemented a regularized t-test approach that is only a shortcut with respect to the full Bayesian treatment. While there can be no perfect substitute for experimental replication (see also Lee *et al.*, 2000), we have shown nonetheless that this approach is effective and indeed has a regularizing effect on the data. In particular, in controlled experiments, it compares favorably with a standard fold approach or a conventional t-test.

Depending on goals and implementation constraints, the method can be extended in a number of directions. For instance, regression functions could be computed off-line to establish the relationship between standard deviation and expression levels and used to produce background standard deviations. Another possibility is to use adaptive window sizes to compute the local background variance, where the size of the window could depend, for instance, on the derivative of the regression function. In an expression range in which the standard deviation is relatively flat (i.e. between -8 and -4 in Figure 2), the size of the window is less relevant than in a region where the standard deviation varies rapidly (i.e. between -12 and -10 in Figure 2). A more complete Bayesian approach could also be implemented, for instance integrating the marginal posterior distributions, which in the case considered here are Student distributions, to estimate the probability $P(\mu_c \approx \mu_t | D, \alpha_t, \alpha_c)$.

The approach can also be extended to more complex designs and/or designs involving gradients of an experimental variable and/or time series designs. Examples would

Table 1. Number of positives detected out of 1000 genes

	Log expression		Ratio		Plain *t*-test		Bayes	
n	From	To	2-fold	5-fold	$P < 0.05$	$P < 0.01$	$P < 0.05$	$P < 0.01$
2	−8	−8	1	0	38	7	73	9
2	−10	−10	13	0	39	11	60	11
2	−12	−12	509	108	65	10	74	16
2	−6	−6.1	0	0	91	20	185	45
2	−8	−8.5	167	0	276	71	730	419
2	−10	−11	680	129	202	47	441	195
3	−8	−8	0	0	42	9	39	4
3	−10	−10	36	0	51	11	39	6
3	−12	−12	406	88	44	5	45	4
3	−6	−6.1	0	0	172	36	224	60
3	−8	−8.5	127	0	640	248	831	587
3	−10	−11	674	62	296	139	550	261
5	−8	−8	0	0	53	13	39	8
5	−10	−10	9	0	35	6	31	3
5	−12	−12	354	36	65	11	54	4
5	−6	−6.1	0	0	300	102	321	109
5	−8	−8.5	70	0	936	708	966	866
5	−10	−11	695	24	688	357	752	441
2v4	−8	−8	0	0	35	4	39	6
2v4	−10	−10	38	0	36	9	40	3
2v4	−12	−12	446	85	46	17	43	5
2v4	−6	−6.1	0	0	126	32	213	56
2v4	−8	−8.5	123	0	475	184	788	509
2v4	−10	−11	635	53	233	60	339	74

Data was generated using normal distribution on a log scale in the range of Arfin *et al.* (2000), with 1000 replicates for each parameter combination. Means of the log data and associated standard deviations (in brackets) are as follows: −6 (0.1), −8 (0.2), −10 (0.4), −11 (0.7), −12 (1.0). For each value of *n*, the first three experiments correspond to the case of no change and therefore yield false positive rates. Analysis was carried out using Cyber-T with default settings ($w = 101$, $K = 10$) and degrees of freedom $n + \nu_0 - 2$.

include a design in which cells are grown in the presence of different stressors (urea, ammonia, oxygen peroxide), or when the molarity of a single stressor is varied (0, 5, 10 mM). Generalized linear and nonlinear models can be used in this context.

The most challenging problem, however, is to extend the probabilistic framework towards the second level of analysis, taking into account possible interactions and correlations amongst genes. If two or more genes have similar behavior in a given treatment situation, decisions regarding their expression changes can be made more robustly at the level of the corresponding cluster. A number of *ad hoc* clustering procedures have been applied to DNA microarray data (Eisen *et al.*, 1998; Alon *et al.*, 1999; Furey *et al.*, 2000; Heyer *et al.*, 1999; Tamayo *et al.*, 1999) without any clear emerging consensus. Of all clustering algorithms, *k*-means has probably the cleanest probabilistic interpretation as a form of EM (expectation–maximization) on the underlying mixture model. Multivariate normal models and Gaussian processes could provide the starting probabilistic models for this level of analysis.

With a multivariate normal model, for instance, μ is a vector of means and Σ is a symmetric positive definite covariance matrix with determinant $|\Sigma|$. The likelihood has the form

$$C|\Sigma|^{-n/2} \exp\left[-\tfrac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^t \Sigma^{-1}(X_i - \mu)\right]. \quad (20)$$

The conjugate prior, generalizing the normal-scaled-inverse-gamma distribution, is based on the inverse Wishart distribution (Appendix) which generalizes the scaled inverse gamma distribution and provides a prior on Σ. In analogy with the one-dimensional case, the conjugate prior is parameterized by $(\mu_0, \Lambda_0/\lambda_0, \nu_0, \Lambda_0)$. Σ has an inverse Wishart distribution with parameters ν_0 and Λ_0^{-1} (Appendix). Conditioned on Σ, μ has a multivariate normal prior $\mathcal{N}(\mu; \mu_0, \Sigma/\lambda_0)$. The posterior then has the same form, a product of a multivariate normal with an inverse Wishart, parameterized by $(\mu_n, \Lambda_n/\lambda_n, \nu_n, \Lambda_n)$.

The parameters satisfy:

$$\begin{aligned}\mu_n &= \frac{\lambda_0}{\lambda_0+n}\mu_0 + \frac{n}{\lambda_0+n}m \\ \lambda_n &= \lambda_0 + n \\ \nu_n &= \nu_0 + n \\ \Lambda_n &= \Lambda_0 + \sum_1^n (X_i - m)(X_i - m)^t \\ &\quad + \frac{\lambda_0 n}{\lambda_0+n}(m-\mu_0)(m-\mu_0)^t. \qquad (21)\end{aligned}$$

Thus, estimates similar to equation (13) can be derived in this multidimensional case.

Bayesian methods are being applied increasingly to a variety of data-rich domains. Whether or not one subscribes to the axioms and practices of Bayesian statistics (Box and Tiao, 1973; Berger, 1985; Pratt *et al.*, 1995), it is wise to model biological data in general, and microarray data in particular, in a probabilistic manner for the reasons outlined in Section 2. Besides DNA microarrays, there are several other kinds of biological arrays, at different stages of development, that could benefit from a similar probabilistic treatment. By enabling the combinatorial interaction of a large number of molecules with a large library, these high-throughput approaches are rapidly generating terabytes of information, which are overwhelming conventional methods of biological analysis. Going directly to a systematic probabilistic framework may contribute to the acceleration of the discovery process by avoiding some of the pitfalls observed in the history of sequence analysis, where it took several decades for probabilistic models to emerge as the proper framework.

ACKNOWLEDGEMENTS

The work of P.B. was supported by a Laurel Wilkening Faculty Innovation award and a Sun Microsystems award at UCI. The work of A.D.L. was supported by NIH grant GM55073. The UCI Computational Genomics group provided helpful comments and testing during program development.

REFERENCES

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Arfin,S.M., Long,A.D., Ito,E.T., Tolleri,L., Riehle,M.M., Paegle,E.S. and Hatfield,G.W. (2000) Global gene expression profiling in *Escherichia coli* K12: the effects of integration host factor. *J. Biol. Chem.*, **275**, 29 672–29 684.

Baldi,P. and Brunak,S. (2001) *Bioinformatics: The Machine Learning Approach*, 2nd edition. MIT Press, Cambridge, MA.

Berger,J.O. (1985) *Statistical Decision Theory and Bayesian Analysis*. Springer, New York.

Box,G.E.P. and Tiao,G.C. (1973) *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Fodor,S.P., Read,J.L., Pirrung,M.C., Stryer,L., Lu,A.T. and Solas,D. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science*, **251**, 767–773.

Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Heller,R.A., Schena,M., Chai,A., Shalon,D., Bedillon,T., Gilmore,J., Woolley,D.E. and Davis,R.W. (1997) Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl Acad. Sci. USA*, **94**, 2150–2155.

Heyer,L.J., Kruglyak,S. and Yooseph,S. (1999) Exploring expression data: identification and analysis of co-expressed genes. *Genome Res.*, **9**, 1106–1115.

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

Lashkari,D.A., DeRisi,J.L., McCusker,J.H., Namath,A.F., Gentile,C., Hwang,S.Y., Brown,P.O. and Davis,R.W. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl Acad. Sci. USA*, **94**, 13 057–13 062.

Lee,C., Klopp,R.G., Weindruch,R. and Prolla,T.A. (1999) Gene expression profile of aging and its retardation by caloric restriction. *Science*, **285**, 1390–1393.

Lee,M.T., Kuo,F.C., Whitmore,G.A. and Sklar,J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9839.

Lipschutz,R.J., Fodor,S.P., Gingeras,T.R. and Lockhart,D.J. (1999) High-density synthetic oligonucleotide arrays. *Nature Genet.*, **21**, 20–24.

Lipschutz,R.J., Fodor,S.P., Gingeras,T.R. and Lockhart,D.J. (1999) High-density synthetic oligonucleotide arrays. *Nature Genet.*, **21**, 20–24.

Long,A.D., Mangalam,H.J., Chan,B.Y.P., Tolleri,L., Hatfield,G.W. and Baldi,P. (2001) Global gene expression profiling in *Escherichia coli* K12: improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. *J. Biol. Chem.*, **276**, 19937–19944.

Ly,D.H., lockhart,D.J., Lerner,R.A. and Schultz,P.G. (2000) Mitotic misregulation and human aging. *Science*, **287**, 2486–2492.

MacKay,D. (1992) Bayesian interpolation. *Neural Comput.*, **4**, 415–447.

MacKay,D.J.C. (1999) Comparison of approximate methods for handling hyperparameters. *Neural Comput.*, **11**, 1035–1068.

Pratt,J.W., Raiffa,H. and Schlaifer,R. (1995) *Introduction to Statistical Decision Theory*. MIT Press, Cambridge, MA.

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

Schena,M., Shalon,D., Heller,R., Chai,A., Brown,P.O. and Davis,R.W. (1995b) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci. USA*, **93**, 10 614–10 619.

Shalon,D., Smith,S.J. and Brown,P.O. (1996) A DNA microarray system for analysing complex DNA samples usign two-color fluorescent probe hybridization. *Genome Res.*, **6**, 639–645.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

White,K.P., Rifkin,S.A., Hurban,P. and Hogness,D.S. (1999) Microarray analysis of *Drosophila* development during metamorphosis. *Science*, **286**, 2179–2184.

Wiens,B.L. (1999) When log-normal and gamma models give different results: a case study. *The American Statistician*, **53**, 89–93.

Zhang,M.Q. (1999) Large-scale gene expression data analysis: a new challenge to computational biologists. *Genome Res.*, **9**, 681–688.

APPENDIX

Estimating hyperparameters from the evidence

The evidence $P(D|\alpha)$ (equation 19) is continuous and differentiable with respect to the hyperparameters over their corresponding valid ranges. Considering convexity and setting the vector of partial derivatives $\partial P(D|\alpha)\partial\alpha$ to 0 shows that the maximum of the evidence is achieved at a point satisfying

$$\mu_0 = m \tag{22}$$

$$\sigma_0^2 = s^2(n-1)/n\ . \tag{23}$$

Note that the estimate for σ_0^2 leads only to a small upward revision of the standard deviation estimate in equation (13). The relation $\partial P(D|\alpha)/\partial\lambda_0$ can be solved in closed form. It is easy to see, however, that when $\mu_0 = m$, the derivative is always positive and the critical equation has no solutions. The evidence is 0 for $\lambda_0 = 0$ and grows with λ_0 to a computable asymptotic value. In practice, it is sufficient to ensure that λ_0 is large with respect to n, for instance $\lambda_0 = 10n$. In terms of priors, a large value of λ_0 in this case corresponds to a very narrow Gaussian distribution for μ centered on m.

The critical equation for ν_0 cannot be solved in closed form but must be handled numerically. As a function of ν_0, and when $\mu_0 = m$, the evidence has the form:

$$P(D|\alpha) = C\frac{(\nu_0/2)^{\nu_0/2}}{[(\nu_0+n)/2]^{(\nu_0+n)/2}}\frac{\Gamma((\nu_0+n)/2)}{\Gamma(\nu_0/2)} \frac{\sigma_0^{\nu_0}}{[(\nu_0 s^2+(n-1)s^2)/(\nu_0+n)]^{(\nu_0+n)/2}}. \tag{24}$$

As a function of ν_0, the asymptotic value of the evidence $P(D|\alpha)$ with $\mu_0 = m$, $\lambda_0 = +\infty$, and $\sigma_0^2 = s^2(n-1)/n$ is

$$(2\pi)^{-n/2}\frac{(\nu_0/2)^{\nu_0/2}}{[(\nu_0+n)/2]^{(\nu_0+n)/2}}\frac{\Gamma((\nu_0+n)/2)}{\Gamma(\nu_0/2)}\frac{1}{2}s^{-n}. \tag{25}$$

The scaled inverse gamma distribution

The scaled inverse gamma distribution $\mathcal{I}(x;\nu,s^2)$ with $\nu > 0$ degrees of freedom and scale $s > 0$ is given by:

$$\frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}s^{\nu}x^{-(\nu/2+1)}\mathrm{e}^{-\nu s^2/(2x)} \tag{26}$$

for $x > 0$. The expectation is $(\nu/\nu-2)s^2$ when $\nu > 2$; otherwise it is infinite. The mode is always $(\nu/\nu+2)s^2$.

The inverse Wishart distribution

The inverse Wishart distribution $\mathcal{I}(W;\nu,S^{-1})$, where ν represents the degrees of freedom and S is a $k \times k$ symmetric, positive definite scale matrix, is given by

$$\left(2^{\nu k/2}\pi^{k(k-1)/4}\prod_{i=1}^{k}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1} |S|^{\nu/2}|W|^{-(\nu+k+1)/2}\exp\left(-\frac{1}{2}tr(SW^{-1})\right) \tag{27}$$

where W is also positive definite. The expectation of W is $E(W) = (\nu-k-1)^{-1}S$.

The Cyber-T software

Cyber-T is particularly suited to experimental designs in which replicate control cDNA samples are compared to replicate experimental cDNA samples. The program calculates basic summary statistics, performs statistical analyses to determine whether observed differences between the control and experimental values are likely to be real, and automatically produces a number of useful plots of the data.

Cyber-T is designed to accept data in the large data spreadsheet format, which is generated as output by software typically used to analyze array experiment images. An element may correspond to a single spot on the array (typical of membrane- or glass slide-based arrays) or a set of spots (typical of GeneChips; Fodor *et al.*, 1991; Lipschutz *et al.*, 1999) designed to query labeled RNA. We will refer to these elements as genes or gene probes since each element is generally designed to query a gene. For each gene, data consists of background-subtracted expression levels for both experimental and control treatment. It is assumed that data from independent hybridization experiments within a given experiment treatment will be contained in adjacent columns. Each gene will have a number of 'labels' that identify a number of properties of that gene contained in adjacent columns. Examples of labels include: gene name, map location of gene, function of gene, and mRNA length. In order to use Cyber-T, this data matrix should be saved on the user's computer as a tab-delimited text file with column headings removed or prefixed with the hash character (#). Extra blank lines at the end of the file should be removed. These data are uploaded to Cyber-T using the 'Browse' button in the Cyber-T browser window. After uploading the data file, the user defines the columns on which analysis will be performed. Missing data should be coded as 'NA' and data that are at or below background should be coded as '0' and treated as the 'lowest expression level reliably detected', which is defined as the 0.0025 percentile associated with all detected genes. Detailed instructions for using Cyber-T can be found on the corresponding web page.

All statistical analysis is carried out using the hdarray library of functions written in R. R is a freely available statistical analysis environment (http://www.cran.r-project.org) adhering to the Open Source development model (http://www.ci.tuwien.ac.at/R/). The hdarray functions are normally invoked through the Cyber-T Web-based interface, but can also be used directly and extended or modified through an X-Window interface (http://www.x.org/about_x.htm) to R. A brief tutorial on how to analyze data directly in R is available at http://www.genomics.biochem.uci.edu/CyberT/, together with instructions for installing the Cyber-T interface and software. This tutorial lists the functions available as part of the hdarray library and R resources. The library and the Cyber-T Web interface also include routines for analyzing paired samples, which would be produced from two-dye glass-slide microarray experiments (Schena *et al.*, 1995a,b; Shalon *et al.*, 1996; Heller *et al.*, 1997; Lashkari *et al.*, 1997). The Web-based interface is written in Perl (http://www.perl.com) to pass the data and other information to a series of functions. This combination of a hard-wired front-end Web interface to a flexible back end allows users to easily explore their data while simultaneously providing a framework for growth and evolution of nonproprietary analysis routines.

Cyber-T generates three output files, two of which (allgenes.txt and siggenes.txt) can either be viewed in the browser window or downloaded and imported into a spreadsheet application for user-specific formatting. These files return the original data and a number of additional columns containing summary statistics, including the mean and standard deviation of both raw and log-transformed data, estimates of the standard deviations employing the Bayesian prior, t-tests incorporating the Bayesian prior on both the raw and log-transformed data, P-values associated with t-tests, and 'signed fold-change' associated with the experiment. The exact content of these files is detailed online. Cyber-T also generates automatically a postscript file CyberT.ps containing a series of graphs that are useful in visualizing the data.