Identifying Bio-markers for EcoArray

Ashish Bhan, Keck Graduate Institute
Mustafa Kesir and Mikhail B. Malioutov, Northeastern University

February 18, 2010

1 Introduction

This problem was presented by John Rogers of *EcoArray* for the Math-in-Industry workshop at Harvey Mudd College . *EcoArray* is a company that makes specialized microarrays for environmental and ecological research. The goal of this project was to rigorously identify contaminants that affect fish, and by extension, humans, using microarray data. This is done by finding genes that are differentially expressed in fish under normal conditions and under treatment by 8 different chemical families $PCB-126$, $Bis-A$, Cd, Pb, Hg, $Phenantharene$, $Estradiol$ and $Testosterone$. *EcoArray* uses a proprietary black-box software program called *GeneSpring* to identify differentially expressed genes that could serve as bio-markers for the 8 different contaminants. The goal of this workshop problem was to develop a statistically robust method of identifying differentially expressed genes given replicate measurements with very different levels of variability.

2 Background Information

In the past 10 years there has been an incredible surge in the development of high-throughput methods in molecular biology that are producing genomic data sets of great interest and complexity. The hope is that these new sources of data will help researchers characterize diseases like cancer at a fundamental level that will lead to new methods of treatment that attack the cancer at the level of cellular aberration. One of the most widely used of the high-throughput methods is microarrays. Microarrays allow us to measure the level of expression for thousands of gene transcripts simultaneously on a single experimental glass slide or nylon membrane. Microarrays can be used to measure the dynamic patterns of expression of all the genes in an organism under normal cell activity (for example, the cell cycle in yeast) or in reponse to external stimuli (for example a toxic substance or infection). In this project we consider the exposure of a species of fish widely used in environmental studies called the *Fathead Minnow* to 8 different chemical substances. For each of the 8 treatments we have 4 replicate measurements of gene expression for thousands of genes under normal and treated conditions respectively. Due to the biological nature of the data, the range of variability in the replicates is high and the goal of the project is to identify genes that are highly differentially expressed under a treatment in a robust (low variance) manner.

3 Differential Expression

One of the most important problems in microarray data analysis is the problem of identifying genes that are differentially expressed from a control state to a treatment state - in other words, we are interested in finding out whether the level of expression of a gene is significantly different in the two conditions. An early approach

to this problem [1] is to look at a simple fold change - a gene is considered to be differentially expressed if its average expression level varies by more than a constant factor, usually 2, between the treatment and control conditions. This approach is known to yield rather noisy results since a simple factor of 2 change has quite a different significance if the level of expression is high or low. An attempt to address this problem is using the t-test. Given a set of measurements $x_1^c, ..., x_n^c$ and $x_1^t, ..., x_n^t$ representing replicate measurements for expression levels in the control and treatment respectively, the t-test uses the empirical means m_c and m_t and variances s_c and s_t to compute a normalized distance between the two populations in the form

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

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

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

degrees of freedom. When t exceeds a threshold depending on the confidence level slected, the two populations are considered to be differentially expressed. In the t-test, the difference between the populations means is normalized by the empirical standard deviations, so this addresses some of the limitations associated with the simple fixed-threshold approach outlined above. One problem with this test is that the number of replicates for microarray data, n_c and n_t are usually small due to the costs involved. This leads to poor estimates of sample variance and make the t-test a less than ideal approach to the problem.

4 Bayesian t-test

One way to address the shortcomings of a t-test is using a Bayesian approach that was proposed in [2]. This approach assumes that the expression levels of a gene measured multiple times under the same experimental conditions will have a roughly Gaussian distribution. Given that gene expression (like the height of individuals in the general population) is influenced by many factors this is a reasonable assumption. Now, each gene in each situation (treatment or control) is represented by a normal distribution $\mathcal{N}(x : \mu, \sigma^2)$. For each gene and condition we have a two parameter model $w = (\mu, \sigma^2)$. The Bayesian approach calls for the specification of a prior $P(\mu, \sigma^2)$ and this choice is part of the modeling procedure. In this approach the prior is chosen to be conjugate - the associated distribution called the posterior takes the same form as the prior. If the problem involves estimating only the mean of a normal model of known variance, it is known that the prior and its conjugate are normal distributions. In the case of estimating the standard deviation of a model with known mean, the conjugate prior is a scaled inverse gamma distribution (or $1/\sigma^2$ has a gamma distribution). This leads to a hierarchical model with a vector of four hyperparameters for the prior $\alpha = (\mu_0, \lambda_0, \nu_0)$ and σ_0^2 with the densities

$$P(\mu|\sigma^2) = \mathcal{N}(\mu; \mu_0, \sigma^2/\lambda_0) \tag{3}$$

and

$$P(\sigma^2) = \mathcal{I}(\sigma^2; \nu_0, \sigma_0^2) \tag{4}$$

For data that comes from microarrays, it seems reasonable to assume that μ and σ^2 are *dependent*. This can be verified by looking at the plot in Figure 1. In this figure, for the 4 replicate measurements per gene we plot the standard deviation as a function of the mean expression.

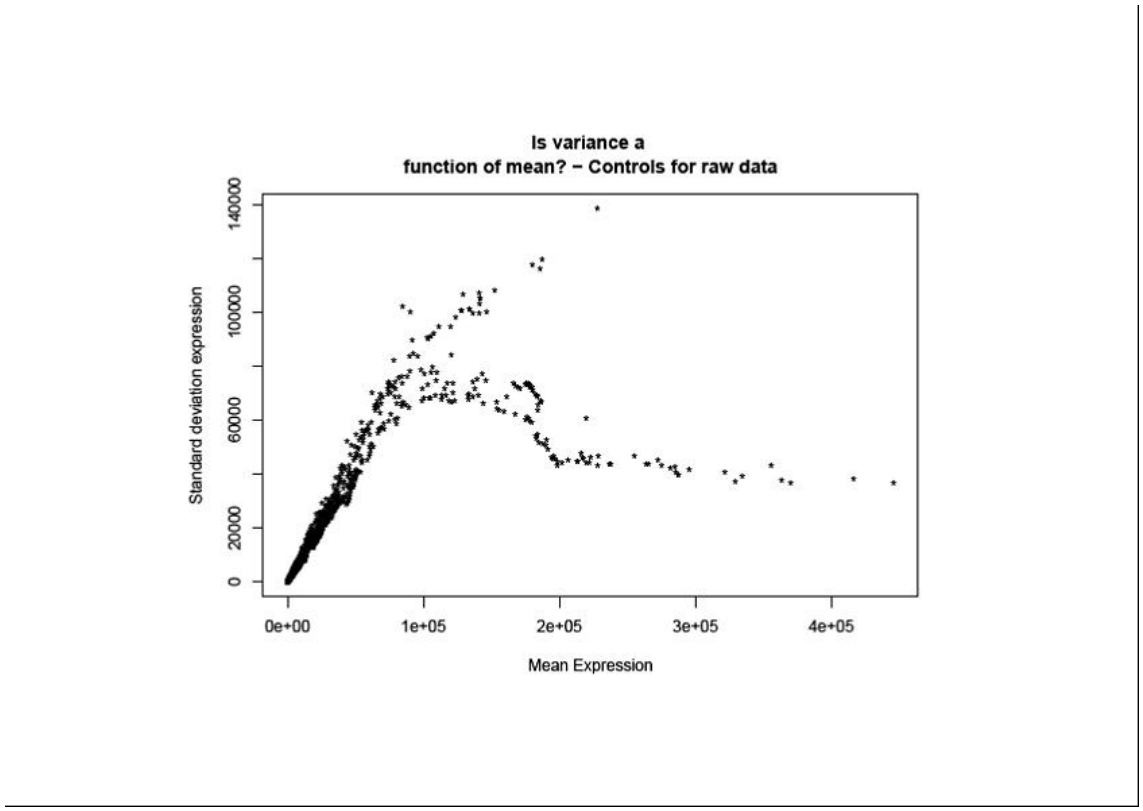

Figure 1: Standard deviation σ as a function of mean expression level μ

The hyperparameters μ_0 and σ^2/λ_0 can be thought of 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. Using some algebra, one obtains the fact that the posterior has the same functional form as the prior

$$P(\mu, \sigma^2|D, \alpha) = \mathcal{N}(\mu; \mu_n, \sigma^2/\lambda_n)\mathcal{I}(\sigma^2; \nu_n, \sigma_n^2) \tag{5}$$

with

$$\mu_n = \frac{\lambda_0}{\lambda_0 + n}\mu_0 + \frac{n}{\lambda_0 + n}m \tag{6}$$

$$\lambda_n = \lambda_0 + n \tag{7}$$

$$\nu_n = \nu_0 + n \tag{8}$$

$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\lambda_0 n}{\lambda_0 + n}(m - \mu_0)^2 \tag{9}$$

The parameters of the posterior distribution combine information from the prior and the data in a principled way. The mean μ_n is a weighted average of the prior mean and the sample mean. The posterior degree of freedom ν_n is the prior degree of freedom ν_0 plus the sample size n and a similar result obtains for the scaling factor λ_n. The posterior sum of squares $\nu_n\sigma_n^2$ is the sum of the prior sum of squares $\nu_o\sigma_0^2$ and the residual uncertainty indicated by the difference between the prior mean and the sample mean. In many cases it is sufficient to use $\mu_0 = m$. The posterior sum of squares is then obtained by adding ν_0 additional observations with deviation σ_0^2. The posterior distribution $P(\mu, \sigma^2|D, \alpha)$ is the most important object of Bayesian analysis and contains the relevant information about *all* possible values of μ and σ^2. Now for each gene we have 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 $P(w_t|D, \alpha_t)$. The posterior distribution is a much richer source of information than simple parameter point estimates or the results of a simple *t*-test. For example, a gene has the same mean expression under control and experimental conditions, but extremely different variances, this difference is undetected by a *t*-test but easily identified from the posterior distributions.

5 Parameter point estimates

In order to compute a modified form of the *t*-test, we need to leverage the posterior distribution with all its richness into single point estimates of the mean and variance of the expression level of a gene in the control and treatment respectively. This can be done in a variety of ways. A robust approach is obtained using the mean of the posterior (MP) estimate. This is given by

$$\mu = \mu_n \text{ and } \sigma^2 = \frac{\nu_n}{\nu_n - 2}\sigma_n^2 \tag{10}$$

If we take $\mu_o = m$ we then get the following MP estimate:

$$\mu = m \text{ and } \sigma^2 = \frac{\nu_n\sigma_n^2}{\nu_n - 2} = \frac{\nu_0\sigma_0^2 + (n-1)s^2}{\nu_0 + n - 2} \tag{11}$$

In the the CyberT software, a modified *t*-test is implemented using the regularized standard deviation of Equation (11). In the simplest case, where we use $\mu_0 = m$, we need to select the values of the background variance σ_0^2 and its strength ν_0. A simple rule of thumb is to assume that $l > 2$ points are needed to estimate

Row#	C_1	C_2	C_3	C_4	E_5	E_6	E_7	E_8	Bayes.p	fold
14775	433316.7	450604.1	474134.8	425568.3	38992.28	31286.5	30647.41	30647.41	0	-13.5561
7701	5.99	100.19	18.49	77.28	15231.19	17421.16	16933.62	16933.62	0	329.3864
8714	1013.33	2499.72	539.29	161.45	22259.42	23827.31	22629.85	22629.85	0	21.67797
3773	412518.8	443820.3	425269.9	383959.4	28504.84	28505.97	28465.74	28465.74	1.11E-16	-14.6176
1973	352545.7	386865.7	358265.7	382413.6	24280.81	24905.21	28238.31	28238.31	8.88E-16	-14.0077
7731	371502.5	384860.5	346664.4	350468.1	22597.7	22838.99	24283.32	24283.32	2.00E-15	-15.4622
5535	336733.1	340067.9	327621.7	313086.5	28150.37	26849.5	26898.9	26898.9	1.78E-14	-12.1097
1864	347102.6	340761.7	352923.2	297238	27209.5	28057.58	28597.34	28597.34	4.06E-14	-11.8976
321	345685.6	360102.1	307911.5	409668.9	25851.4	26239.67	28591.4	28591.4	7.37E-14	-13.0257
6169	12.32	26.72	13.29	6.72	10883.71	11660.43	14425.51	14425.51	8.66E-14	870.3668
13286	298238.6	363482.7	294072.7	330898.4	24475.65	21856.37	22116.16	22116.16	1.68E-13	-14.2075
3879	304442.4	299901.6	273256	272438.8	18395.13	15509.28	19987.74	19987.74	9.40E-13	-15.5663
3188	290847.3	303142.6	274644.3	280838.9	22572.44	21067.54	18886.22	18886.22	1.23E-12	-14.1191
7040	281280	343176.1	266408.4	290957.8	22022.64	22087.29	21059.7	21059.7	1.74E-12	-13.7056
10858	296394.5	268354.4	300916.8	276186.9	22059.22	21640.4	24288.93	24288.93	2.80E-12	-12.3741
8202	299406.9	282612.8	286461.1	271528.7	18016.81	17156.92	20320.97	20320.97	5.60E-12	-15.0366
10747	261879.5	291954.5	278095.1	294571.7	20557.47	18905.92	18870.92	18870.92	6.83E-12	-14.591

Figure 2: The output of CyberT ranked by increasing p-value

the standard deviation properly and then let $n + \nu_0 = l$. A reasonable default value to use is $l = 10$. For σ_0 we could use the standard deviation of the entire set of observations or, more subtly, on categories of similar genes. In CyberT, the expression values of the genes is ranked and the user specifies a window size ws. The default value is $ws = 101$ which corresponds to 50 genes immediately above and below the gene of interest.

6 Results

The top ranking results of the CyberT analysis on the data supplied by EcoArray are shown in Figure 2. These are markedly different from those obtained earlier by EcoArray using the black box routines implemented in GeneSpring. The output of CyberT on the data are genes that have high differential expression *and* low variance among the replicates. This suggests that we have implemented a statistically robust method for identifying differentially expressed genes and used it successfully on the data. A further indicator that our method produces more accurate bio-markers is the following: the biomarkers chosen for Cd and Hg by our method have a high degree of overlap and this makes biochemical sense given the similarity between these two toxins (see `http://en.wikipedia.org/wiki/Cadmium`)- the output produced by GeneSpring did not have this additional signature of consistency.

References

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

[2] Baldi, P., and Long, A.D. *A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes* **Bioinformatics** 17 509-519 (2001).