2007-07-19

Development of Informative Priors in Microarray Studies

Kassandra M. Fronczyk
Brigham Young University - Provo

DEVELOPMENT OF INFORMATIVE PRIORS IN MICROARRAY STUDIES

by

Kassandra M. Fronczyk

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Statistics

Brigham Young University

August 2007

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Kassandra M. Fronczyk

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	Natalie J. Blades, Chair
Date	Scott D. Grimshaw
Date	Bruce J. Collings

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Kassandra M. Fronczyk in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date	Natalie J. Blades Chair, Graduate Committee

Accepted for the Department

Scott D. Grimshaw
Graduate Coordinator

Accepted for the College

Thomas W. Sederberg
Associate Dean, College of Physical and
Mathematical Sciences

ABSTRACT

DEVELOPMENT OF INFORMATIVE PRIORS IN MICROARRAY STUDIES

Kassandra M. Fronczyk

Department of Statistics

Master of Science

Microarrays measure the abundance of DNA transcripts for thousands of gene sequences, simultaneously facilitating genomic comparisons across tissue types or disease status. These experiments are used to understand fundamental aspects of growth and development and to explore the underlying genetic causes of many diseases. The data from most microarray studies are found in open-access online databases. Bayesian models are ideal for the analysis of microarray data because of their ability to integrate prior information; however, most current Bayesian analyses use empirical or flat priors. We present a Perl script to build an informative prior by mining online databases for similar microarray experiments. Four prior distributions are investigated: a power prior including information from multiple previous experiments, an informative prior using information from one previous experiment, an empirically estimated prior, and a flat prior. The method is illustrated with a two-sample experiment to determine the preferential regulation of genes by tamoxifen in breast cancer cells.

ACKNOWLEDGEMENTS

A sincere thank you to Dr. Blades and all of the professors in the department for providing me with the tools and resources to complete this thesis. Special thanks to all my classmates, friends and family for their encouragement and support.

CONTENTS

CHAPTER

APPENDIX

TABLES

Table

FIGURES

Figure

1. INTRODUCTION

The study of differential expression is important for understanding biological processes because it provides information about which proteins are produced in a cell. Knowledge of protein expression provides clues about the functions of particular genes, allows identification of clusters of related genes, and motivates new hypotheses and experiments. Protein expression is difficult to measure reliably; consequently, mRNA expression levels serve as a reasonable surrogate. For example, recognizing which genes are differentially expressed in cancer cells and normal cells can give some information about cancer.

Until recently, monitoring the simultaneous expression level of thousands of genes in a single experiment was not possible. The Southern blot is a method for searching for a specific DNA molecule. The Southern blot, proposed in 1975, introduced a one-to-one correspondence between clones and hybridization signals. After the invention of the Southern blot, the use of non-porous solid supports and the development of methods for high-density spatial synthesis of oligonucleotides opened up the world of DNA microarray technologies (Lander 1999), which provides expression measurements for thousands of genes at once (Duggan et al. 1999, Shena et al. 1995). Because patterns in which a gene is expressed can be temporal, developmental, and physiological, the factors studied could be different types of tissues, drug treatments, or time points of a biological process.

Considering the microarray community's willingness to share data, the Bayesian framework seems to be a logical approach to the analysis of these experiments; however, most of the current Bayesian analyses do not incorporate biological knowledge. This thesis presents a method for incorporating the information from previous studies into an informative prior for a two-sample Bayesian t-test. Results are compared to the use of an empirical Bayesian-estimated prior.

An overview of the elicitation process is given in Chapter 2. This includes the reasons behind elicitation, a discussion of different approaches to elicitation, and the constraints and limitations of using elicited information. Chapter 3 considers some of the current Bayesian methods for analyzing microarray data. Some of the methods investigated are Efron's empirical Bayes analysis, Lönnstedt and Speed's B-statistic, and Conlon's hierarchical Bayesian model for pooling microarray studies. Chapter 4 presents the case study, and the design and analysis goals of the experiment are explained and the types of prior information available are discussed. Chapter 5 describes the proposed model and the Perl script for automation. Chapter 6 presents the results from the analysis of ten genes and chapter 7 gives the conclusions.

2. ELICITING PRIOR INFORMATION

> We live in an uncertain world, and probability risk assessment deals as directly with that fact as anything we do. Uncertainty arises partly because we are fallible. Mostly, however, uncertainty arises because the world is not as simple as we would have it. The variability of phenomena (including human technology, and often in spite of our pet theories) yields the most uncertainty. —E. M. Dougherty (1993)

Probability has several common interpretations. One of the most common approaches is to interpret probability as an objective long-range frequency. A second interpretation of probability is as a degree of belief, or epistemic probability. This subjective approach suggests probability may be specific to each individual. The elicitation process allows a statistician to quantify individually held beliefs as a number between zero and one. If an individual does not have complete knowledge about a probabilistically well-defined event, that uncertainty can be represented as a probability distribution.

In any statistical analysis, there is some form of background, or prior, knowledge available in addition to the data. Prior knowledge can be elicited from different places, but the most reliable and worthwhile reference is an expert in the field (Kadane and Wolfson 1998). Elicitation is used when estimates are needed on new, rare, or complex phenomena; for forecasting and predictions; to interpret data; and to understand or determine a problem-solving process. It can also include work in selecting or defining the scope of the problem, work in refining the problem, and the processes involved in arriving at a solution to the problem. This information is a representation of the expert's knowledge at a specific point in time; it can and should change when new information becomes available. The knowledge elicitation process has a considerable influence on the quality of the resultant prior knowledge.

2.1 Methods

A substantive expert has opinions and knowledge about his or her field. These opinions can be given in terms of processes, scales, ranks, and countless other forms. Through the elicitation process, an analyst may work with an expert to extract parameters of a family of probability distributions in an organized and logical manner. As Savage (1971) explains, if two experts with the same knowledge are induced to reveal their opinions, then the resulting probability distributions should be the same. Most experts do not know the parameters of these distributions or how to express their knowledge in probabilistic terms. A statistician must be able to pose intuitive questions and discuss the subject matter with the expert to adequately define an intelligible prior distribution that captures the main ideas of the expert's opinion while integrating experience and knowledge of the literature. Kadane and Wolfson (1998) examine the psychology of getting experts to express what they know in distributional form. Meyer and Booker (2001) present a systematic approach for eliciting this expert knowledge.

Knowledge elicitation methods are classified according to how directly information is obtained from the expert. Indirect methods are used to obtain information that cannot be easily expressed directly. Some indirect methods include construct elicitation, document analysis, and laddering. Direct methods elicit the required information directly from the expert, and include interviewing, protocol analysis, and simulation. These techniques are based on the assumption that the expert is able to articulate his or her knowledge. This assumption is not always warranted, as some tasks become automatic after years of repetition.

Construct elicitation methods obtain information about how the expert discriminates between entities in the problem domain. As an example, consider a 90-year-old man who sits at the end of a manufacturing line to remove defective ball bearings. From years of practice, he is able to run his fingers over the ball bearings and remove the defective ball bearings without being able to express his basis for the rejection. He is now retiring and his replacement must be trained. Using a sample of defective and acceptable ball bearings, the inspector asks the elderly man to verbalize the perceptions influencing his acceptance or rejection of each bearing. This practice has been applied with products ranging from clothing and chocolate, to injection molds and machine faults, to ball bearings and steel ladles (Reeve et al. 2004). The most commonly used construct elimination method is Repertory Grid Analysis (Kelley 1955). For this method, the expert is presented with a list of entities and is asked to describe the similarities and differences between them. These similarities and differences are used to determine important attributes of the entities. After evaluating the initial list of attributes, the researcher works with the expert to assign ratings to each pair of entities and attributes.

Document analysis gathers information from existing subject-area literature. This method may or may not involve interaction with a human expert to confirm or enhance this information. For example, literature is integrated in biology studies in different ways: hand-curated pathways have been sufficient for assembling models in numerous studies; literature is frequently accessed for concepts or functional relationships in databases like the Medical Subject Headings (MeSH) and Gene Ontologies (GO); and mining text directly for specific types of information is becoming more popular as text analytics methods become more accurate and accessible (Roberts 2006).

Laddering is a diagramming technique in which the analyst asks the expert questions to systematically build a hierarchy of subject concepts. The analyst begins by stating the name of a seed item from the subject field. Specific questions are used to lead the expert through the task domain or hierarchy. This technique is useful when the subject constructs are known but the interrelationships between them are poorly understood. For example, when interviewing a manager on how to improve team performance, the interviewer starts with the term productivity and the manager is asked a series of standard questions to find terms that are up, down, or lateral in the hierarchy. To move down the hierarchy, the interviewer may ask for examples of productivity; to move up, he may ask about what some same-level items have in common; to move across the hierarchy, he may ask for examples of the upper-level item apart from the same-level item.

These indirect techniques are often effective at obtaining information that is not easily expressed. However, in some situations, these indirect methods do not produce the information needed by the analyst. Instead, the analyst may use a direct method to increase the quality of information and the possibility of error reduction. The most common direct methods include interviewing, protocol analysis, simulation studies, and prototyping.

Interviewing consists of asking the expert questions about the subject of interest and how they perform their tasks. Interviews can be unstructured, semi-structured, or structured. The success of an interview session is dependent on the questions asked and the expert's ability to articulate his or her knowledge; it is difficult to know which questions should be asked, particularly if the interviewer is not familiar with the subject matter.

Protocol analysis (Ericsson and Simon 1984) involves asking the expert to perform a task while "thinking aloud." The intent is to capture both the actions performed and the mental process used to determine these actions. For example, a study examines adults building a lifting device using a child's construction set. Performance evaluation considers specific actions, such as bolting two parts together. The analyst categorizes verbal statements according to reference—the goals of a particular action or the evaluation of the outcome of a test component. On the basis of the references, actions can be grouped into behavioral traits, revealing a pattern of goal decomposition exercised by the problem solver. As with all of the direct methods, the success of the protocol analysis depends on the ability of the expert to describe why he or she is making a decision. In some cases, the expert may not remember that things are done a certain way. In many cases, the verbalized thoughts will only be a subset of the actual knowledge used to perform the task. One method used to augment this information is interruption analysis. For this method, the analyst interrupts the expert at critical points in the task to ask questions about why a particular action is performed.

Simulation methods use a computer system to reproduce a complex task. A simulation attempts to mimic an abstract model of a particular system. These simulations are a useful part of modelling many systems in physics, biology, economics, and engineering to gain insight into the operation of those systems. These techniques are used to study the behavior of objects or systems that cannot be easily or safely tested in reality. In simulation studies, the expert behaves as though a simulation is occuring. For example, with the Wizard-of-Oz technique, people may believe that they are communicating with a piece of software although they are linked to an individual that is trained to respond to the consumer's actions.

In prototyping, the expert evaluates a prototype of the proposed system; this is usually an iterative process as the system is refined. Storyboarding is a type of paper prototyping. For example, customers, users, or developers start a software development project by drawing pictures of the screens, toolbars, and other elements they believe the software should provide. The group continues to evolve these ideas until their requirements and details are finalized.

All of these elicitation methods require iteration. After drawing out parameters for a well-defined and coherent prior, the analyst must give some sort of feedback to the expert. The analyst can ask questions like, "If what you said is true, then ..." and include some information about the properties of the distribution. The expert may agree or disagree and explain further what should happen in those terms. The analyst can continue asking the expert questions to further develop the prior distribution until both the expert and the researcher are satisfied with its characteristics.

The usefulness of any of these elicitation approaches hinges on the analyst's ability to evoke truthful and accurate reports from the experts. Two methods that have been used to encourage trustworthy explanations are scoring rules (Savage 1971) and prediction-based elicitation (Grether 1980a). Scoring rules use incentives to motivate people to state the probability of a random outcome thoughtfully and truthfully. Prediction-based elicitation pays people for accurately predicting random outcomes and then uses these predictions to infer probabilities. For example, to instigate a prompt response to better predict the advance of a potential bird flu epidemic, health experts are being financed to place a wager on the spread of the bird flu. This motivates the experts to give their opinions truthfully and quickly.

2.2 Constraints and Limitations

The elicitation and interpretation of subjective probability is a controversial area of statistics. Nau (2001) discusses whether or not it is even possible to elicit the true probability or probability distribution and whether it makes any difference to the statistical inference. Singpurwalla (2002) compares the Bayesian and frequentist approaches to probability and their consequences. Mosleh and Bier (1996) question whether an individual can be uncertain about a probability, separating uncertainty about the underlying events from that of cognitive imprecision. Benson et al. (1995) express that elicitation of probability requires both the formation of a belief and the assessment of a probability that quantifies that belief. They believe that the former process involves judgement and reasoning, while the latter process is purely judgmental.

Berman (1988) examines some issues about subjective probability or personal opinion probabilities. He suggests three additional reasons for the unreliability of subjective probability estimates:

(1) People tend to extrapolate linearly from existing information. Many things in the world are non-linear.

(2) Breakthroughs in technology or understanding are by definition unpredictable.

(3) People of vision have frequently, if not always, been in the minority.

Berman also suggests that human uncertainty does not necessarily decrease as knowledge increases.

Evans (2000) summarizes some of the philosophical issues in eliciting prior information:

> Engineers who represent their degree-of-belief by probability must be stout-hearted. Once you have gone through the simulations and settled on a realistic expression of your prior beliefs, stick to them and to the resulting afterwards belief, no matter what the actual experimental outcome. Remember, you have already considered that outcome in your extensive simulation. Do not let anyone convince you to be "practical" or "realistic" on their terms. Assert yourself. Say that you have been practical and realistic on a very sound, rational basis.

If there is prior knowledge available, it can and should be used to fully analyze the problem at hand. Many statisticians, including Evans, believe that incorporating this information is a legitimate and logical approach to problem-solving.

2.3 Less Subjective Priors

In theory, the process of inference is simple. Inference involves two steps: the assertion of hypotheses and their proper organization. Thus, there is only one process of validating a conclusion. Many people are uncomfortable with the Bayesian approach to inference because it does not follow a straightforward line of reasoning.

They view the selection of a prior as arbitrary and subjective; however, priors may be chosen to make selection more objective.

A famous example of the use of an objective prior can be found in Mosteller and Wallace (1963). In this paper, Mosteller and Wallace conduct an analysis of the twelve Federalist papers of unknown authorship. The Federalist Papers written by Madison and Hamilton provide weighted prior distributions of word usage. Using these long-range frequencies as the prior distributions for the negative binomial model, the authors conclude that Madison, rather than Hamilton, wrote all twelve of the disputed papers.

There are other examples of using less subjective priors. Many current Bayesian methods for the analysis of microarray experiments assume normality of the log-expression ratios and include reference priors on one or more of the hyperparameters. The authors of such examples claim ignorance of the information about the genome and include broad guesses for the values of the prior distributions. For instance, Conlon et al. (2006) assume that any given gene has a uniform chance of differential expression. Another approach to building a less subjective prior is using information from previous studies.

2.3.1 Power Priors

When existing data is available, a prior may be constructed from this data. Ibrahim and Chen (2000) present a power prior for situations in which historical data are available. The power prior is defined as the likelihood function based on the historical data D_0 raised to a power a_0, where $0 \leq a_0 \leq 1$ is a parameter that controls the influence of the historical data on the current data. Historical data, denoted by $D_0 = (n_0, X_0)$, may be combined with the prior distribution for θ before the historical data D_0 is observed, $\pi(\theta|\cdot)$. A prior distribution for a_0 to obtain the joint power prior distribution for (θ, a_0) is

$$\pi(\theta, a_0|D_0) \propto L(\theta|D_0)^{a_0}\pi_0(\theta|c_0)\pi(a_0|\gamma_0),$$

where c_0 is a specified hyperparameter for the initial prior and γ_0 is a specified hyperparameter vector. In most cases, c_0 is defined to be one (Tsodikov et al. 2003; Fu et al.2005; Ghosh et al. 2004). A natural choice for $\pi(a_0|\gamma_0)$ is a beta prior. Other choices, including a truncated gamma prior or a truncated normal prior, have similar theoretical properties, and similar computational properties can be chosen instead of the beta distribution. According to Ibrahim and Chen, the proposed distributions yield similar results when the hyperparameters are appropriately chosen so the distributions look similar.

Ibrahim and Chen (2000) show that the joint power prior distribution is proper even if $\pi_0(\theta|c_0)$ is chosen to be an improper uniform prior. This power prior can easily be extended to the situation where there are multiple previous studies. If L_0 is the number of historical studies, then $D_{0k} = (n_{0k}, X_{0k})$ is the historical data based on the kth study, $k = 1, \ldots, L_0$ and $D_{0k} = (D_{01}, D_{02}, \ldots, D_{0L_0})$. Then, Ibrahim and Chen define a weight parameter a_{0k} for each historical study, and take the a_{0k} values to be independent and identically distributed beta random variables with hyperparameters $\gamma_0 = (\delta_0, \lambda_0), k = 1, \ldots, L_0$. Letting $a_0 = (a_{01}, \ldots, a_{0L_0})$, the power prior can be expressed as

$$\pi(\theta, a_0|D_0) \propto \prod_{k=1}^{L_0} \left(L(\theta|D_{0k})^{a_{0k}}\pi(a_{0k}|\gamma_0)\right)\pi_0(\theta|c_0). \tag{2.1}$$

This approach enables the analyst to include previous studies in a simple manner. The power prior is informative, yet not necessarily subjective.

Elicitation of reliable prior information is difficult; consequently, many statisticians use uninformative prior distributions in Bayesian analyses. For science to progress, analyses must recognize prior experiments and formalize the information gained for use in subsequent experiments.

3. BAYESIAN METHODS FOR MICROARRAY EXPERIMENTS

Frequentist approaches are commonly used for the analysis of microarray experiments. Inference proceeds by stating a hypothesis and collecting data that will either support or oppose the claim. A suitable model is chosen to fit the data. This model allows the analyst to make inferences about the hypothesis. While frequentist procedures are associated with probability statements about how procedures behave across repeated measurements, Bayesian inference aims instead at making probability statements given a particular measurement or set of measurements.

Bayesian inference is an approach to statistics in which all forms of uncertainty are expressed in terms of probability. A Bayesian approach to a problem starts with the formulation of a model that is hopefully adequate to describe the situation of interest. A prior distribution is formulated over the unknown parameters of the model, which is meant to help form beliefs about the situation before seeing the data. After observing some data, Bayes' Rule is applied to the data to obtain a posterior distribution for the unknown parameters, which takes account of both the prior and the data. From this posterior distribution, probability statements and predictive distributions for future observations can be computed.

In the microarray setting, models often calculate the posterior probability of a gene being differentially expressed. There are many different interpretations of what may form a suitable model and distributions reflecting prior knowledge, resulting in different approaches to the analysis of microarray data.

3.1 Empirical Bayes Models

The large number of genes and small sample size of typical microarray experiments yield inflated t-statistics and a high rate of false discoveries. Table 3.1 provides an example of a microarray experiment with three control and three treatment samples, and the expression levels of three genes are given. While the expression levels in the control and treatment groups are not remarkably different, the variance is very small, making the absolute value of the t-statistic very large (Feingold 2003).

Table 3.1: Calculated t-statistic from replicated study. The three genes displayed have small within-group variability leading to inflated t-statistics.

	Control			Treatment			$\|t\|$
TUBA6	6.84	6.99	6.96	3.87	3.96	4.02	50.2
K-ALPHA-1	6.61	6.79	6.76	5.01	5.06	5.13	25.3
RAB31	5.76	5.88	5.73	4.29	4.24	4.37	23.2

To address the unrealistically small variance estimates, Lönnstedt and Speed (2002) introduce a new statistic for assessing differential expression in microarray datasets with few replicates. Lönnstedt and Speed use an Empirical Bayes approach that uses the data to estimate the hyperparameters and then combines the hyperparameters with statistics taken from the data in the B-statistic, which calculates the log posterior odds of differential expression occurrence.

The expression levels for each gene i in sample j, M_{ij}, are assumed to be independent random variables from an $N(\mu_i, \sigma_i^2)$. These parameters, (μ_i, σ_i^2), are given conjugate priors: normal distributions for the μ_i and inverse gamma distributions for the σ_i^2. The hyperparameters for the priors are estimated by first fixing the unknown proportion of genes that are differentially expressed, p. B_i can then be calculated for each gene i using an explicit formula

$$B_i = \log\left(\frac{p}{1-p}\frac{1}{\sqrt{1+nc}}\left(\frac{a+s_g^2+M_{g.}^2}{a+s_g^2+\frac{M_{g.}^2}{1+nc}}\right)^{v+\frac{n}{2}}\right),$$

where s_g^2 is the gene-specific sum of squares over n and $M_{g.}$ is the average expression level for each gene.

Relative to the t-statistic, the B-statistic decreases the number of false positives and false negatives. The B-statistic also deals with the possible inflation due to small within-group variation and a small number of replicates.

Efron et al. (2001) propose an alternate empirical Bayes approach for detection of differentially expressed genes and estimation of the false discovery rate. Efron et al. assume that the observed gene expression values are a mixture of non-differentially expressed genes and differentially expressed genes. The expression levels of non-differentially expressed genes are characterized by density f_0; the expression levels of differentially expressed genes are characterized by a bimodal distribution, f_1, reflecting the genes which are either turned on or turned off. Neither of these distributions is known; what is observed is the distribution of the scores, f, which is a mixture of f_0 and f_1, shown in Figure 3.1 as a solid line. From the mixture scores, the authors estimate f_0 and f_1 and the posterior probability that a gene is differentially expressed.

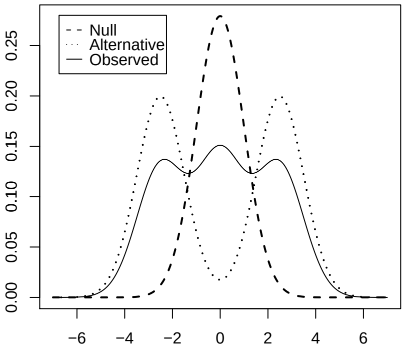


Figure 3.1: Efron's empirically estimated densities of the null and alternative distributions from the observed mixture distribution. The observed f is the mixture of the null f_0 and the alternative f_1.

A modified t-statistic with a fudge factor, a_0, summarizes the expression values of each gene:

$$Z_i = \frac{\bar{D}_i}{a_0 + S_i},$$

where $\bar{D}_i$ is the average of the differences in expression between sample types and S_i is the sample standard deviation for each gene i. The null distribution of Z_i is generated by permuting the sample labels. A logistic regression analysis estimates the ratio of $f_0(Z)/f(Z)$, where $f_0(Z)$ is the density function of the scores for unaffected genes and $f(Z)$ is the mixture density. The posterior probability that a gene with score Z is differentially expressed is calculated by

$$p_1(Z) = 1 - p_0 \frac{f_0(Z)}{f(Z)},$$

where p_0 is the prior probability of differential expression.

Genes with low expression levels have little variance, resulting in very large t-statistics; the smoothing parameter, a_0, in the denominator of the scores prevents such genes from dominating the results of the analysis. The fudge factor, a_0, is obtained by performing the analysis for a range of values for a_0 and then selecting an optimal value. Efron suggests that the a_0 value selected should find the greatest number of differentially expressed genes.

Efron's method is useful in that it handles high-dimensional data robustly when sample sizes are small. Also, a full Bayesian analysis would require prior specification of p_0, p_1, f_0, and f_1, but the authors use the structure of microarray data to estimate an empirical version. The main weakness in Efron's method is that instead of using historical data to estimate p_0, p_1, f_0, and f_1 *a priori*, the authors use the current data. This method uses the data twice, once to estimate the priors and once to perform the analysis using those priors.

Kendziorski et al. (2003) propose an empirical Bayes methodology to improve the estimation of expression fold change by the use of posterior odds for the assessment of differential expression. Fold-change analysis is used to identify genes with expression ratios or differences between a treatment and a control that are outside of a given cutoff or threshold. In this case, inference on each gene uses the information about the fluctuations of expression measurements from all genes.

The authors assume that measurements which share a common mean expression level μ_g appear independently and identically from an observed $f_{obs}(\cdot|\mu_g)$. Two components of the mixture model are specified as the observed $f_{obs}(\cdot|\mu_g)$, which characterizes fluctuations in repeated measurements from a gene having a latent expression level μ_g, and a genome-wide distribution $\pi(\mu_g)$, which represents fluctuations in these means among genes. The authors explore two families for $f_{obs}(\cdot|\mu_g)$. The first family assumes gamma-distributed measurements and the second family uses log-normally distributed measurements. A constant coefficient of variation is assumed in both models. These models also account for differential variation in apparent fold change.

In the gamma-gamma model, the $f_{obs}(\cdot|\mu_g)$ is a gamma distribution with mean value μ_g, shape parameter α, and scale parameter $\lambda_g = \alpha/\mu_g$ for measurements $z > 0$.The genome-wide distribution, $\pi(\mu_g)$, is assumed to be an inverse gamma; fixing α, the quantity $\lambda_g = \alpha/\mu_g$ has a gamma distribution with shape parameter α_0 and scale parameter v. In cases in which there are two conditions, control and treatment, the posterior probability can be calculated by

$$\text{odds}_g = \frac{pv_0^{\alpha}\Gamma(n_1\alpha+\alpha_0)\Gamma(n_2\alpha+\alpha_0)(\sum_{i=1}^{n_1} x_{g,i} + \sum_{i=1}^{n_2} y_{g,i} + v)^{N\alpha+\alpha_0}}{(1-p)\Gamma(\alpha_0)\Gamma(N\alpha+\alpha_0)(\sum_{i=1}^{n_1} x_{g,i} + v)^{n_1\alpha+\alpha_0}(\sum_{i=1}^{n_2} y_{g,i} + v)^{n_2\alpha+\alpha_0}},$$

where x_{ig} and y_{ig} are the measurements from the two conditions for each gene and $N = n_1 + n_2$ is the total number of observations on each gene g. All hyperparameters are estimated by the data.

While this method is dependent on parametric model assumptions, the authors suggest that the method may miss some genes, but a gene that is called differentially expressed is most likely accurately labeled.

Fox and Dimmic (2006) propose a two-sample t-test to determine whether or not a gene is differentially expressed in two different samples. The proposed method explicitly calculates the marginal distribution for the difference in the mean expression of two samples, removing the need for point estimates of the variance that were needed in earlier attempts to construct a t-test.

The authors assume that the likelihood of the observed data for a single gene follows a normal distribution, dependent on the given treatment. That is, the samples from each treatment follow a normal distribution with equal variances and possibly different means, shown by

$$\begin{aligned} y_i &\sim N(\mu, \sigma^2), \\ y_j &\sim N(\mu + \Delta\mu, \sigma^2), \end{aligned}$$

where $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$ represents the number of replicates in the control and treatment samples.

The priors on μ and $\Delta\mu$ are taken to be flat. The prior probability of σ^2 follows a scaled inverse gamma distribution with parameters ν_0 and σ_0^2, where $\nu_0 = 0$ and σ_0^2 is estimated by the data. The priors and likelihood are combined to give the following posterior distribution:

$$\begin{aligned} p(\mu, \Delta\mu, \sigma^2|\mathbf{y}) &\propto p(\mu, \Delta\mu, \sigma^2) \prod_{i=1}^{n_1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) \\ &\quad \times \prod_{j=1}^{n_2} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_j - (\mu + \Delta\mu))^2\right), \end{aligned}$$

where n_1 and n_2 are the number of measurements in each sample. The use of the authors' assumptions and definitions causes the marginal posterior distribution of $\Delta\mu$ to follow a t-distribution. That is,

$$\frac{\Delta\mu - \Delta\bar{y}}{\sigma_n sqrt\frac{1}{n_1} + \frac{1}{n_2}}|\Theta \sim t_{\nu_n},$$

where $\nu_n = n_1 + n_2 - 2$ and $\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2$. A hypothesis test is then performed by asserting the null hypothesis that there is no true difference in expression levels, $\Delta\mu = 0$. When the posterior probability of having no differential expression, $\mathrm{P}(\Delta\mu = 0|\mathbf{y})$, approaches zero, the null is rejected and the gene is called differentially expressed.

3.2 Hierarchical Models

Gottardo et al. (2003) present a hierarchical Bayesian model with independent Gaussian modelling that addresses the two main issues in microarray studies: the small number of replicates and the large number of genes. This model gives rise to four statistics that are useful in different situations.

The first statistic is based on the situation in which there is only treatment data available. The statistic is calculated as follows, where ν_0, τ_0, ν_a and τ_a are the hyperparameters of the inverse gamma priors for the variances of the genes that are not differentially expressed, ν_0 and τ_0, and the genes that are differentially expressed, ν_a and τ_a.

$$B_1 \quad = \quad \left(1 + \frac{1-p}{p}\sqrt{2}\frac{\Gamma(\nu_a)\Gamma(\nu_0 + n_2/2)}{\Gamma(\nu_0)\Gamma(\nu_a + n_2/2)}\frac{\tau_0^{\nu_0}}{\tau_a^{\nu_a}}\frac{(\tau_a + \frac{n_2-1}{2}S_g^2)^{\nu_a + n_2/2}}{(\tau_0 + \frac{n_2}{2}S_{g0}^2)^{\nu_0 + n_2/2}}\right)^{-1},$$

where $S_g^2 = \sum_{i=1}^{n_2}(Y_{gi} - \bar{Y}_g)^2/(n_2 - 1)$ and $S_{g0}^2 = \sum_{i=1}^{n_2}(Y_{gi} - 0)^2/(n_2)$. The B_1 statistic is found to be more powerful than the B-statistic calculated by Lönnstedt and Speed (2002) through simulation studies when the sample size is less than 5. When sample size increases, the variations of the B-statistic are comparable.

The second statistic is used when both control and treatment data are available for each gene; therefore, the analysis is treated as a two-sample problem to determine if there is a difference in mean expression. The calculation of B_2 is essentially the same as the calculation of B_1, with functions of hyperparameters and the expected proportion of differentially expressed genes. The major difference is in the estimates of the variability of gene expression: in B_2 the numerator variability estimate is equivalent to a two-sample pooled s^2, and the denominator is also a pooled estimate of s^2, using a weighted average of the mean from both samples.

A change in the variance of the expression ratios can be attributed to a biological event. Therefore, a third statistic is used to test for a difference in the variance of the expression levels between the control and treatment conditions. The principal difference between calculating B_2 and B_3 is that both the numerator and the denominator include the different functions of the same estimates of gene expression variability for the control and treatment data.

The fourth statistic detects a given gene with different mean ratios and/or different variances in the control and treatment groups. Primarily, B_4 calculation differs from B_3 calculation by using a pooled variance in the denominator, as in B_2.

Gottardo et al. compare B_1 and B_2 to many equivalent statistics and find the B-statistics to be superior in finding correctly differentially expressed genes. There are no current statistics that relate to B_3 and B_4. Lönnstedt and Speed's B-statistic assumes a normal likelihood for the data with normal and inverse gamma priors for the means and variances, respectively. Gottardo's B-statistics account for another level of uncertainty by including normal priors on the mean of the gene expression means and including gamma priors on the parameters of the inverse gamma prior on the variances.

In order to build a suitable parametric model to allow for comparison of normal and tumor tissues and to characterize the behavior of the genes in each group, Ibrahim et al. (2002) develop a class of models with hierarchical priors for the parameters that allow for correlation between the genes.

The expression level for a given gene, x_{jig}, can be described by a mixture random variable with a discrete and a continuous component if $j = 1, 2$ indexes the tissue type, normal vs cancer, and x_{jig} is the mixture random variable for the jth tissue type for the ith individual, $i = 1, 2, ..., n_j$, and the gth gene, $g = 1, 2, ..., G$. The discrete portion is a point mass at some c_0, a threshold value assumed to be the level of expression at which a gene is not differentially expressed. The continuous component, y, is the expression level of the gene and is lognormally distributed. The observed gene expression can be written as

$$x_{jig} = \begin{cases} c_0 & \text{with probability} \quad p \\ c_0 + y & \text{with probability} \quad 1 - p, \end{cases}$$

letting $\delta_{jig} = I(x_{jig} = c_0)$ and $p_{jg} = P(\delta_{jig} = 1)$. The likelihood function based on the data, D, is given by

$$L(\mu, \sigma^{\mathbf{2}}, \mathbf{p}|D) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} \prod_{g=1}^{G} p_{jg}^{\delta_{jig}} (1 - p_{jg})^{1-\delta_{jig}} p(y_{jig}|\mu_{jg}, \sigma_{jg}^2)^{1-\delta_{jig}}.$$

Ibrahim et al. then compute the posterior distribution for each gene of $\xi_g = \psi_{2g}/\psi_{1g}$, where

$$\psi_{jg} \quad = \quad c_0 p_{jg} + (1 - p_{jg}) \left(c_0 + \exp \mu_{jg} + \frac{\sigma_{jg}^2}{2} \right).$$

As in many other Bayesian microarray analyses, Ibrahim et al. (2002) specify a hierarchical prior for μ_{jg} as being independent $N(\mu_{j0}, \tau_0 \sigma_{jg}^2 / \bar{n}_j)$, where $\bar{n}_j = \frac{1}{G} \sum_{g=1}^{G} (n_j - \sum_{i=1}^{n_j} \delta_{jig})$ and $\tau_0 > 0$ is a defined scalar. The μ_{j0} have a prior of $N(m_{j0}, \nu_{j0}^2)$. For σ_{jg}^2, the priors are independent inverse gamma with hyperparameters (a_{j0}, b_{j0}). There is also a gamma prior put on b_{j0} with hyperparameters (q_{j0}, t_{j0}), which allow prior correlation between the genes. Finally, p_{jg} is transformed to e_{jg} by taking the $\log(\frac{p_{jg}}{1-p_{jg}})$. For these values of e_{jg}, a normal prior is designated with mean μ_{j0} and variance $k_{j0} w_{j0}^2$, where μ_{j0} is distributed as $N(\hat{\mu}_{j0}, h_{j0} w_{j0}^2)$ and k_{j0}, h_{j0}, and w_{j0}^2, are the specified hyperparameters. These hyperparameters can either be defined by historical data or expert opinion; if neither is available, Ibrahim and Chen provide some guide values.

Ibrahim and Chen's model is very similar to Kendziorski et al. (2003), with the exception of the prior structure. This approach provides more flexibility for making inferences about the differential expression of genes than other types of clustering algorithms. It distinguishes the pattern of gene expression in the two types of tissue, and can easily be extended to more than two tissue types.

The Bayesian models discussed previously are useful for the analysis of any experiment because of their integration of many levels of uncertainty and because of their possible resolution of the difficulties inherent in microarray data. Bayesian models are also useful when data includes many levels of replication. Oftentimes, many independent, but not necessarily identical, studies are conducted in order to understand a certain biological process. Conlon et al. (2006) introduce a framework for incorporating data from multiple independent microarray experiments with several sources of replication. This framework includes a hierarchical Bayesian model that takes into account each gene on each slide from each experiment.

Conlon et al. (2006) assume that there are only two conditions present in each independent experiment, control and treatment, and that each experiment is conducted using the same assay platform. The model that produces the posterior probability that a gene is differentially expressed based on gene expression levels across $j = 1, ..., J$ independent studies is presented in Figure 3.2, where y_{jges} is the log-expression ratio for gene g in experiment e on slide s. The average expression over all slides within experiment e of study j is given by μ_{jge}. The log-expression ratio for each gene of study j is given by θ_{jg}. There is also an indicator function, I_g, for differential expression of gene g, where p is the percent of differentially expressed genes. The percent of differentially expressed genes, p, has a uniform prior distribution. The posterior distributions for each parameter are simulated using MCMC methods. Finally, the posterior probability, D_g, is calculated for gene g across all studies.

Conlon et al. (2006) use the False Discovery Rate (FDR) defined by Benjamini

$$\begin{aligned}
y_{jges}|\mu_{jge} &\sim N(\mu_{jge}, \tau^2_{jg}) \\
\mu_{jge}|\theta_{jg} &\sim N(\theta_{jg}, \sigma^2_{jg}) \\
\theta_{jg}|I_g = 0 &\sim N(0, \eta^2_{jg0}) \\
\theta_{jg}|I_g = 1 &\sim N(0, c_j \times \eta^2_{jg0}) \\
\eta^2_{jg0} &\sim \frac{as_1^2}{\chi^2_a} \\
c_j &\sim \frac{bs_2^2}{\chi^2_b} \\
I_g &\sim \text{Bernoulli}(p) \\
p &\sim \text{Uniform}(0,1)
\end{aligned}$$

Figure 3.2: Hierarchical model for the probability that a gene is differentially expressed based on gene expression levels across $j = 1, ..., J$ independent studies.

and Hochberg (1995) and the Integration-driven Discovery Rate (IDR) defined by Choi et al. (2003) to evaluate the proposed model. The FDR is the number of false discoveries made divided by the total number of discoveries. The IDR is the number of genes discovered in a meta-analysis that were not discovered in any of the individual studies alone divided by the total number of discoveries. Essentially, the IDR quantifies the gain of information by pooling individual experiments. These studies showed that by using this heirarchical model for pooling data, there was a considerable increase in the IDR for multiple values of γ, while FDR is consistently low; more truly differentially expressed genes were ascertained with a smaller chance of false positives with this model as compared to the individual studies.

4. CASE STUDY

4.1 Tamoxifen and Breast Cancer

This work examines an experiment performed to determine the preferential regulation of genes by tamoxifen in breast cancer cells (Frasor et al. 2006). Estrogens act on target tissues by binding to estrogen receptors. An estrogen receptor is a protein molecule found inside cells that are targets for estrogen action. Estrogen receptors, located in the cell nucleus, contain a site to which only estrogens or closely related molecules can bind. In the absence of estrogen molecules, these estrogen receptors are inactive and have no influence on DNA, but when an estrogen molecule enters a cell and passes into the nucleus the estrogen binds to its receptor and causes the shape of the receptor to change (Parker et al. 1997). This estrogen-receptor complex then binds to specific DNA sites. After the complex binds to the DNA sites, nearby genes become active. The active genes produce molecules of mRNA, which give rise to specific proteins that influence the function of the cell (Hayashi et al. 2003). Estrogen is important in programming the body for sexual reproduction, controlling cholesterol production, and preserving bone strength. Estrogen can also have a deleterious effect on health by advancing the production of epithelial cells in the breast. Although the ability to stimulate cell production is one of estrogen's normal roles, it can also increase a woman's chance of developing breast cancer (Clark et al. 1998).

Although estrogen does not appear to directly cause the DNA mutations that trigger the development of human cancer, estrogen does stimulate cell production. If one or more breast cells already possesses a DNA mutation that increases the risk of developing cancer, these cells, along with normal epithelial cells, will reproduce in response to estrogen stimulation. Thus, estrogen-induced cell production leads to an increase in the total number of mutant cells that exist. These cells have an increased risk of becoming cancerous, so the chance that cancer may develop is increased (Parker et al. 1997). Some drugs that block the action of estrogen in certain tissues can mimic the action of estrogen in other tissues. Differences in chemical structure allow estrogen-like drugs to interact with the estrogen receptors of different tissues. Tamoxifen blocks the action of estrogen in breast tissue by binding to the estrogen receptors of epithelial cells (Swain 2001). The experiment to be investigated assesses an estrogen receptor α (ERα) positive breast cancer cell line (MCF-7) infected with adenovirus-ERβ and treated with tamoxifen. The cells were infected with adenovirus carrying either estrogen receptor β (AdERβ) or no insert (Ad), and treated with trans-hydroxytamoxifen (TOT). The results provide insight into tamoxifen activity in the presence of both ERα and ERβ, which illuminates the potential therapeutic and diagnostic implications of tamoxifen with regard to breast cancer.

4.2 Data Analysis

In a typical microarray experiment, RNA obtained under various conditions (patients, treatments, disease states, etc.) is hybridized to microarrays. By tagging the RNA with a fluorescent marker, intensity values can be obtained that correspond to the amount of labeled RNA bound to the array. On the widely used Affymetrix platform, gene expression is measured using probe sets consisting of 11 to 20 perfect match (PM) probes of 25 nucleotides, which are complementary to a target sequence, and a similar number of mismatch (MM) probes in which the 13th nucleotide has been changed. The MM probe measurements are thought to comprise most of the background cross-hybridization and stray signals affecting the PM probes (Affymetrix 1992). The tamoxifen experiment is conducted using Affymetrix chips.

In performing an exploratory analysis of the six data samples, boxplots of the probe intensities for each chip are created (see Figure 4.1). The box plots of intensities from all arrays should have a similar mean and range. The first sample of the treatment data may be different from the other two samples. This must be corrected for in order to find differences in expression levels due to biological effects rather than slightly different samples. The samples are skewed because the measurements are bounded below by zero.

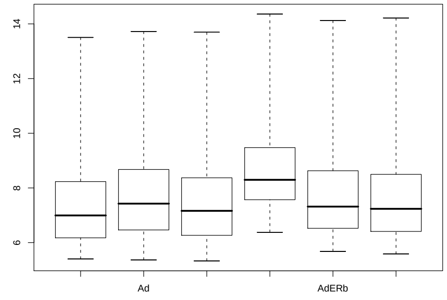


Figure 4.1: Boxplots of the data before normalization; the three control (Ad) replicates are on the left and the three treatment (AdERb) replicates are on the right.

To explore the dependence of the variance of the signal intensities on the strength of the signal, an MVA plot of a pair of chips is examined or the average signals of treatment groups are examined (Heber and Sick 2006). Figure 4.2 is a scatter plot of the average log differences of a pair of chips versus the average mean of their log signals. The MVA plots comparing sample chips within the control do not have any significant abnormalities. There are some problems with the MVA plots comparing the treatment sample chips, as seen in Figure 4.2; while the MVA plots should be linear, this plot has noticable curvature.

To conduct a chip-to-chip analysis, the data must be standardized. The normalization corrects for systematic differences within slides or between slides that do not represent true biological variation between samples.

To normalize the data, a method called IdealMM is used (Bolstad 2001). This approach compares the mismatch and perfect match probe intensities. The quantile

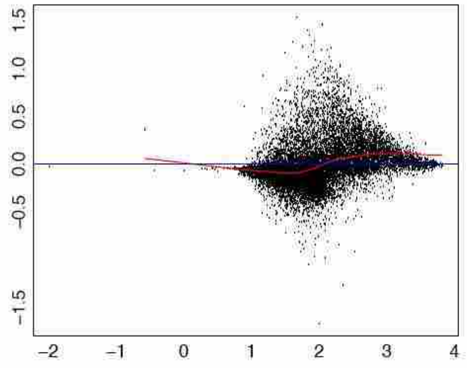


Figure 4.2: MVA plot of two chips within the treatment sample before normalization, where A is the x-axis and M is the y-axis.

normalization method is used to correct for the differences in the distributions of intensities of the chips. This method gives all the chips the same empirical distribution. Finally, the median-polish summary method is used; this method fits a multi-chip linear model to the data from each probe set. The boxplots in Figure 4.3 show that the post-standardization probe intensities are less variable and have a constant median across chips.

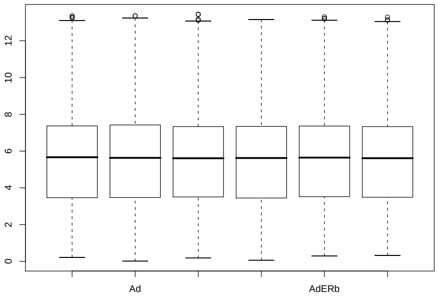


Figure 4.3: Boxplots of the data after normalization; the three control replicates are on the left and the three treatment replicates are on the right.

The normalization process also removes the curvature and other problems in the MVA plots within the treatment samples. The post-normalization MVA plot in

Figure 4.4 displays the same samples observed in Figure 4.2, but no curvature is apparent.

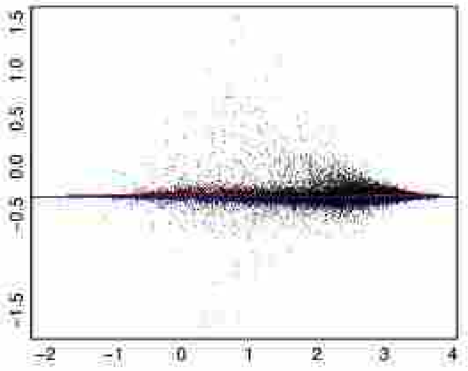


Figure 4.4: MVA plot of two chips within the treatment sample after normalization, where A is the x-axis and M is the y-axis.

The Bayesian t-test proposed by Fox and Dimmic (2006) is performed to determine which genes are differentially expressed using non-informative priors. That is, the mean and the difference in means are given flat priors and the variance prior parameters are estimated by the data. The choice of likelihood and prior distributions results in the marginal posterior distribution of the difference, $\Delta\mu$, following a t-distribution.

Using the marginal posterior distribution of $\Delta\mu$, the t-statistic (see section 3.1) is calculated to determine differential expression. The model is simplistic but effective in finding genes with a high probability of differential expression. The authors use a cutoff p-value of 0.05 to indicate differential expression. Using this bound, more than 3,700 genes are called significant. The top 10 genes and the corresponding t-statistics and p-values are shown in Table 4.1.

Figure 4.5 gives a plot of the difference in mean expression for each gene versus the p-value. From this figure, the genes with large differences have low p-values. The dotted line represents the cutoff p-value to indicate differential expression; in this case, the line is the bound 0.05.

Table 4.1: List of the top 10 genes and their corresponding t-statistics and P-values.

Gene	t-statistic	p-value
202240_at	-38.96389	< 10e-16
211120_x_at	38.00555	< 10e-16
204962_s_at	-38.19704	< 10e-16
202094_at	-29.35590	< 10e-16
209408_at	-27.15826	< 10e-16
221520_s_at	-25.81545	1.110223e-16
211117_x_at	25.10982	1.110223e-16
219978_s_at	-24.61897	2.220446e-16
211118_x_at	21.92268	1.887379e-15
202580_x_at	-21.69021	2.331468e-15

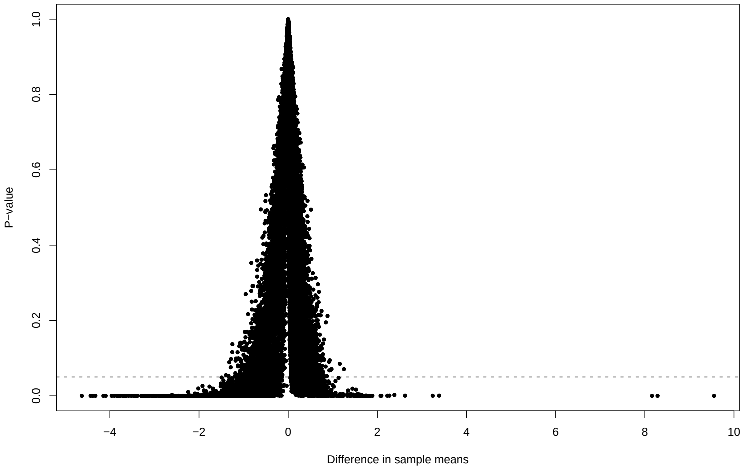


Figure 4.5: Plot of the difference in means versus p-values. The horizontal line represents the 0.05 cutoff p-value indicating differential expression.

The analysis of the data using flat priors for the mean and difference in means and data-driven hyperparameters for the prior on the variance is useful. However, there are many sources of information about the genome that should be incorporated into the analysis.

4.3 Types of Prior Information Available

The gene expression community shares the data collected from experiments through online databases. As a national resource for molecular biology information, the National Center for Biotechnology Information (NCBI) website creates literature, journal, and nucleotide databases for public access and develops tools and software for database mining and data analysis. As of May 1, 2007, Gene Expression Omnibus (GEO) had catalogued more than 150,000 microarray studies (National Center for Biotechnology Information 2002a); the Stanford MicroArray Database had details on over 13,000 experiments (Stanford University 2003); and SAGEMAP included over 600 libraries with tissue and cell information on more than 15 organisms (National Cancer Institute 1996).

PubMed is a database which holds articles from medical and health-related journals. As of May 1, 2007, there were more than 14,500 articles including the keywords *breast cancer* and *estrogen receptors* (National Center for Biotechnology Information 2002b). By inputting these keywords and a specific gene name, more than 350 articles were returned.

The gene Estrogen Receptor 2, or ER beta, is one gene involved in the tamoxifen experiment. As of May 1, 2007, the PubMed library had more than 14,500 articles containing information about Estrogen Receptor 2 and estrogen receptor-related breast cancer (National Center for Biotechnology Information 2002b). The first few articles are cited in Table 4.2. As of May 1, 2007, the Gene Expression Omnibus contained twenty breast cancer experiments involving the gene ER beta, all of which are shown in Table 4.3 (National Center for Biotechnology Information 2002a). One paper presents two separate experiments (Coser et al. 2003), and another paper provides a series of three experiments (Wu et al. 2006). The bolded experiment represents the Tamoxifen experiment examined in this thesis.

Four of the 20 breast cancer experiments found in GEO are very similar to the

Table 4.2: Five of the 14,500 estrogen receptor breast cancer articles mentioning ER beta, as of May 1, 2007.

Author	Experiment
Brama et al. (2007)	Osteoblast-conditioned medium promotes proliferation and sensitizes breast cancer cells to imatinib treatment.
Eakin et al. (2007)	Estrogen receptor α is a putative substrate for the BRCA1 ubiquitin ligase.
Marx et al. (2007)	Proteasome Regulated ERBB2 and Estrogen Receptor Pathways in Breast Cancer.
Poola and Yue (2007)	Estrogen receptor alpha (ERα) mRNA copy numbers in immunohistochemically positive-, and negative breast cancer tissues.
Ray et al. (2007)	Diet-induced obesity and mammary tumor development in relation to estrogen receptor status.

motivating experiment. The first experiment deals with the analysis of the response of estrogen receptor (ER) negative breast cancer cells infected with full-length ER alpha adenoviral constructs to treatment with 17beta-estradiol (E2) (Moggs et al. 2005). The results of this experiment provide insight into the anti-proliferative effect of E2 on breast cancer cells reexpressing ER (see Table 4.4). The second experiment explores the expression profiling of estrogen receptor positive breast cancer cell lines treated with estradiol for 24 hours. MCF-7, T47-D, and BT-474 breast cancer cell lines are examined (Rae et al. 2005). The results identify candidate genes involved in estrogen-stimulated breast cancer growth (see Table 4.5). The third experiment studies the analysis of tumors from 49 breast cancer patients (Farmer et al. 2005). Tumors are classified into a luminal, basal, or novel molecular apocrine class. Apocrine tumors are estrogen receptor negative (ER-) and androgen receptor positive (AR+), while luminal tumors are ER+ and AR+, and basal tumors are ER- and AR-. Summary statistics for gene expression levels are shown in Table 4.6. The fourth experiment is the analysis of estrogen receptor (ER) alpha positive MCF-7 breast cancer cells overexpressing constitutively active c-erbB-2. Results indicate that increased MAPK activation results in loss of ER-alpha expression (see Table 4.7). These four experiments target estrogen receptor breast cancer. The ten excluded experiments either examine estrogen receptors in cancers of other parts of the body or involve breast cancer but not with respect to estrogen receptors.

Instead of the flat priors and data-driven hyperparameters used in the previous analysis, information from these experiments are used to build informative priors.

Table 4.3: The twenty breast cancer experiments involving ER beta, as of May 1, 2007.

Author	Experiment
Coser et al. (2003) (2)	Global analysis of ligand sensitivity of estrogen inducible and suppressible genes in MCF7/BUS breast cancer cells by DNA microarray.
Wu et al. (2003)	DACH1 inhibits transforming growth factor-beta signaling through binding Smad4.
Acevedo et al. (2004)	Selective recognition of distinct classes of coactivators by a ligand-inducible activation domain.
Mecham et al. (2004)	Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements.
Stitziel et al. (2004)	Membrane-associated and secreted genes in breast cancer.
Chen et al. (2005)	Identification of transcriptional targets of HOXA5.
Farmer et al. (2005)	Identification of molecular apocrine breast tumours by microarray analysis.
Itoh et al. (2005)	etrozole-, anastrozole-, and tamoxifen-responsive genes in MCF-7aro cells: a microarray approach.
Moggs et al. (2005)	Anti-proliferative effect of estrogen in breast cancer cells that re-express ERalpha is mediated by aberrant regulation of cell cycle genes.
Poola et al. (2005)	Identification of MMP-1 as a putative breast cancer predictive marker by global gene expression analysis.
Rae et al. (2005)	GREB 1 is a critical regulator of hormone dependent breast cancer growth.
Wonsey and Follettie (2005)	Loss of the forkhead transcription factor FoxM1 causes centrosome amplification and mitotic catastrophe.
Creighton et al. (2006)	Activation of mitogen-activated protein kinase in estrogen receptor alpha-positive breast cancer cells in vitro induces an in vivo molecular phenotype of estrogen receptor alpha-negative human breast tumors.
Dittmer et al. (2006)	Parathyroid hormone-related protein regulates tumor-relevant genes in breast cancer cells.
Frasor et al. (2006)	**Gene expression preferentially regulated by Tamoxifen in breast cancer cells and correlations with clinical outcome.**
Richardson et al. (2006)	X chromosomal abnormalities in basal-like human breast cancer.
Wu et al. (2006)	Glucocorticoid receptor activation signals through forkhead transcription factor 3a in breast cancer cells (Series 1, 2, 3).

Table 4.4: Expression values for Estrogen Receptor 2 (ER beta) in breast cancer cells reexpressing estrogen receptor alpha response to 17beta-estradiol. The expression values for Estrogen Receptor 2 (ER beta) in the analysis of the response of estrogen receptor (ER) negative breast cancer cells infected with full-length ERα adenoviral constructs to treatment with 17beta-estradiol (E2).

Sample	Expression Level
AdlacZ+est	68.5
	15.7
	97.3
AdERa+est	189
	174
	107

Table 4.5: Expression values for Estrogen Receptor 2 (ER beta) in the estrogen effect on estrogen receptor alpha positive breast cancer cell lines. The expression values for Estrogen Receptor 2 (ER beta) in estrogen receptor positive breast cancer cell lines treated with estradiol for 24 hours. MCF-7, T47-D, and BT-474 breast cancer cell lines examined.

Sample	Expression Level
Control	
MCF-7	6.16
	6.22
T47-D	5.80
	5.87
BT-474	5.92
	5.71
Est. Treated	
MCF-7	6.23
	6.27
T47-D	5.89
	5.86
BT-474	5.87
	5.98

Table 4.6: Expression values for Estrogen Receptor 2 (ER beta) in molecular apocrine breast tumors. Analysis of tumors of 49 breast cancer patients. Tumors classified into a luminal, basal, or novel molecular apocrine class. Apocrine tumors are estrogen receptor negative (ER-) and androgen receptor positive (AR+), while luminal tumors are ER+ and AR+, and basal tumors are ER- and AR-.

Sample	Expression Level
Apocrine Tumor ($n = 6$)	6.442 (0.0054)
Basal Tumor ($n = 16$)	6.467 (0.0077)
Luminal Tumor ($n = 27$)	6.325 (0.354)

Table 4.7: Expression values for Estrogen Receptor 2 (ER beta) in ER α positive breast cancer cells response to hyperactivation of MAPK pathway. The expression values for Estrogen Receptor 2 (ER beta) in ER *alpha* positive MCF-7 breast cancer cells overexpressing constitutively active c-erbB-2.

Sample	Expression Level
Control	5.896
	6.016
	5.894
erbB-2	6.022
	6.005
	6.014

5. METHODS

Scientists are rigorously honest about reporting experiment results and how those results are obtained in formal publications. The National Academy of Sciences' report on the responsibilities of authorship in Biological Life Sciences (National Research Council 2003) explain that scientists have an ethical duty to allow free and open access to supporting data. Most scientists agree with this principle because results that cannot be replicated are suspect. This open access to data allows scientists to create models that reflect an increase in genomic knowledge; this increase in knowledge is often disregarded.

5.1 Combining Information Across Studies

Most Bayesian methods for the study of microarray analysis use vague priors or priors with data-driven hyperparameters. Given the vast amount of genomic information available, stronger priors may be constructed. However, the number of genes involved, the variety of gene expression platforms, and the thousands of experiments documented pose some difficulties in building informative priors.

Combining information across multiple studies is challenging. In the case of microarray studies, the expression levels of the same genes have been measured on different array platforms. In addition, technical and biological variability generally lead to measurements of gene expression that may not be comparable across studies. There are few methods that deal with these complications.

To avoid dealing directly with measurements of gene expression that may not be comparable, several approaches have been proposed. Rhodes et al. (2002) compute q-values (Benjamini and Hochberg 1995) for each gene and define a differential expression signature for each experiment as the set of genes with q-values below a pre-defined threshold. The meta-signature is declared to be all genes present in at least J signatures, where J is selected by permutation testing.

In another effort to combine information across studies, Parmigiani et al. (2004) use information on the correlation between gene expression measurements. Rather than providing an aggregate inference, this approach focuses on identifying a set of comparable genes, namely genes for which the correlation of expression values among other genes in the array was similar across studies. This procedure evaluates gene expression consistencies across platforms rather than pooling gene expression values. This method identified genes with reproducible expression patterns across studies and improved correlation across studies.

Gene expression data generated with different microarray platforms are not directly comparable; even within the same platform different protocols for sample preparation, array hybridization, and data analysis can result in variation among datasets. Because the composition of microarrays is regularly updated to incorporate new genes with improved target sequences, it is difficult to combine data from different generations of the same microarray platform. Despite this difficulty, Yuen et al. (2002) compare microarray measurements between Affymetrix GeneChips and two-color cDNA microarrays and find that, although the fold changes of differentially expressed genes showed poor correlation across array platforms, the rank orders of differentially expressed genes are comparable.

In light of the information accumulating about the genome and the ability to combine information across studies, it seems reasonable to believe there is some prior knowledge about the probability that a specific gene will be differentially expressed in a new experiment. Public databases can be queried to obtain information about the expression levels of a gene in different types of tissues. This information can be combined into an informative prior on the probability of differential expression.

5.2 Model Specifications

A two-sample Bayesian t-test will determine if there is a difference in expression between the cells infected with adenovirus carrying either AdERβ or Ad. This test is based on the model used in Fox and Dimmic (2006). The likelihood of the observed data in sample i for a single gene g follows a normal distribution depending on the treatment group. That is,

$$\begin{aligned} y_{ig} &\sim N(\mu_g, \sigma^2) \text{ and} \\ y_{ig} &\sim N(\mu_g + \Delta_g, \sigma^2), \end{aligned}$$

where Δ_g reflects the difference in expression between the treatment groups.

Four sets of priors are explored in this work. First, of the relevant historical studies researched, one of the experiments that is similar to the tamoxifen experiment is used to give estimates for the conjugate prior distributions. We assume normal priors for the mean of each gene, μ_g, normal priors for Δ_g, and inverse gamma priors for σ^2.

A second set of priors applies the power prior approach introduced by Ibrahim and Chen (2000). Each historical experiment has a likelihood that is assumed to follow an $N(\theta_g, \tau^2)$ distribution. The initial priors for μ_g, Δ_g, and σ^2 are taken to be $N(\mu_0, \sigma^2/\lambda_0)$, $N(0, \sigma^2/\lambda_0)$ and $IG(\nu_0, \sigma_0^2)$, respectively. Each a_{0k} has an independent beta distribution with parameters (a_k, b_k), where $k = 1, \ldots, L_0$. The formula shown in section 2.3.1 gives a joint power prior for $(\mu_g, \Delta_g, \sigma^2, a_0)$.

The third set of priors uses an empirical Bayes approach to estimate the parameters of the prior distribution on the three parameters, μ_g, Δ_g, and σ^2.

The fourth set of priors uses an empirical Bayes approach to estimate the parameters of the prior distribution on the variance and assumes flat priors on the mean and difference in means. That is, the joint prior is assumed to be proportional to the inverse gamma prior on the variance with data-driven hyperparameters. The prior distributions lead to the following posterior distribution

$$\begin{aligned} p(\mu_g, \Delta_g, \sigma^2|\mathbf{y}) \quad &\propto \quad p(\mu_g, \Delta_g, \sigma^2) \prod_{i=1}^{n_1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_{ig} - \mu_g)^2\right) \\ &\times \prod_{i=1}^{n_2} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}((y_{ig} - (\mu_g + \Delta_g))^2\right), \end{aligned}$$

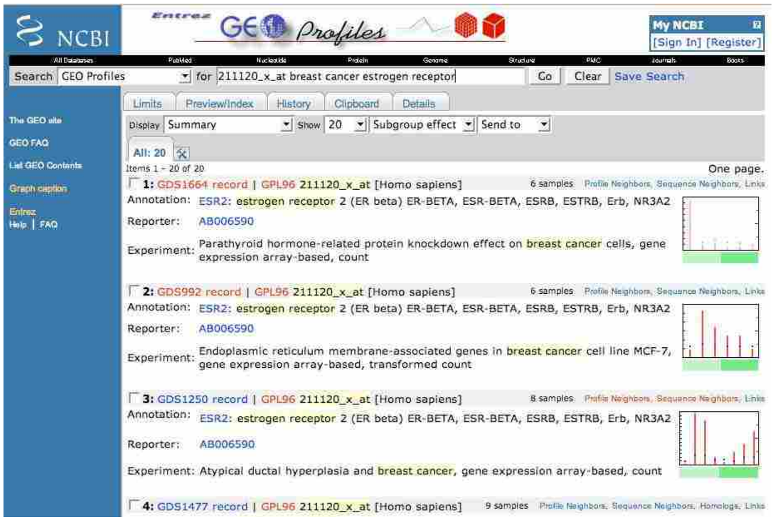
where n_1 and n_2 are the number of measurements in each sample. The marginal posterior for Δ_g can be found and used in a hypothesis test of differential expression. The null hypothesis assumes that the true difference in expression levels is zero; that is, $\Delta_g = 0$ or some other threshold degree. When the posterior probability of no differential expression, $Pr(\Delta_g = 0|\mathbf{y})$, is less than a cutoff value α^*, the null is rejected and the gene is called differentially expressed.

5.3 Perl Script

A Perl script obtains the hyperparameters of the priors for one gene. The script has to search for a full list of previous experiments involving the gene, choose the relevant experiments, and extract information from these experiments to specify the hyperparameters of the prior distributions.

The script accesses the GEO database to search for experiments involving estrogen receptors, breast cancer, and the given gene. The resulting list includes a summary of each experiment, as seen in Figure 5.1. The relevant experiments are chosen by searching the experiment summary, the "Experiment" field in Figure 5.1, for both breast cancer and estrogen receptors. The experiments that are not chosen may examine estrogen receptors in cancers of other parts of the body or involve breast cancer but not with respect to estrogen receptors.

The data set from each of the relevant studies is split into two groups. The web page includes check-boxes that correspond to a specific group of samples. These check-boxes are shown in Figure 5.2 in the "4 assigned subsets" table under the heading "Samples". The subsets may split the samples according to the treatment



Search GEO Profiles for 211120_x_at breast cancer estrogen receptor

Display Summary Show 20 Subgroup effect Send to

Items 1 - 20 of 20 One page.

□ 1: GDS1664 record | GPL96 211120_x_at [Homo sapiens] 6 samples Profile Neighbors, Sequence Neighbors, Links

Annotation: ESR2: estrogen receptor 2 (ER beta) ER-BETA, ESR-BETA, ESRB, ESTRB, Erb, NR3A2

Reporter: AB006590

Experiment: Parathyroid hormone-related protein knockdown effect on breast cancer cells, gene expression array-based, count

□ 2: GDS992 record | GPL96 211120_x_at [Homo sapiens] 6 samples Profile Neighbors, Sequence Neighbors, Links

Annotation: ESR2: estrogen receptor 2 (ER beta) ER-BETA, ESR-BETA, ESRB, ESTRB, Erb, NR3A2

Reporter: AB006590

Experiment: Endoplasmic reticulum membrane-associated genes in breast cancer cell line MCF-7, gene expression array-based, transformed count

□ 3: GDS1250 record | GPL96 211120_x_at [Homo sapiens] 8 samples Profile Neighbors, Sequence Neighbors, Links

Annotation: ESR2: estrogen receptor 2 (ER beta) ER-BETA, ESR-BETA, ESRB, ESTRB, Erb, NR3A2

Reporter: AB006590

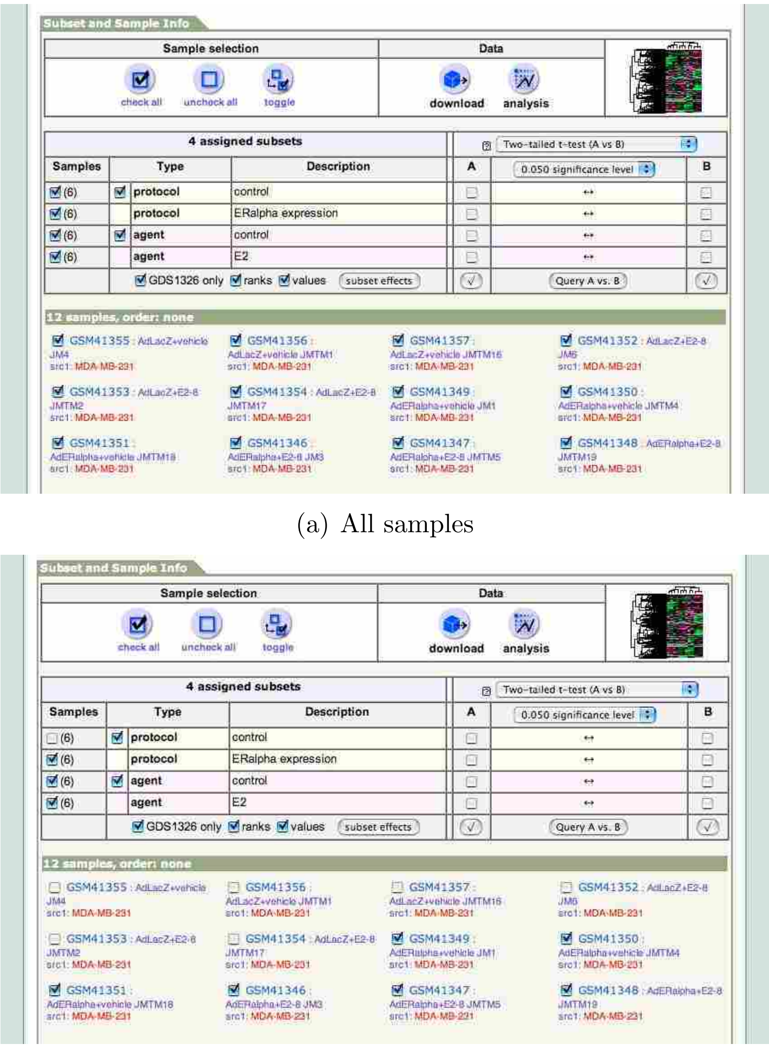
Experiment: Atypical ductal hyperplasia and breast cancer, gene expression array-based, count

□ 4: GDS1477 record | GPL96 211120_x_at [Homo sapiens] 9 samples Profile Neighbors, Sequence Neighbors, Homologs, Links

Figure 5.1: List of experiments for one gene from GEO. The Perl script searches the *Experiment* field for both breast cancer and estrogen receptors to identify the relevant experiments (National Center for Biotechnology Information 2002a).

or by some facet of the tissues. To extract the groups of samples, the HTML source code is searched for the check-box code, and the subset of samples is retrieved from this portion of the HTML code. The script exports the groups of sample numbers to a text file.

The script prints out the labels of the subsets of each data set. The user chooses two of the labels that split the data into two subsets. The files corresponding to the two groups specified by the user are imported. For each sample number within each file, the script retrieves the webpage that includes the table of expression values for all genes in the sample, as seen in Figure 5.3. The gene name is located and the script keeps all information between the gene name and the next gene number. The expression value is extracted from this information, stored in an array, and written to a text file. These files are read into R for further analysis.



4 assigned subsets					? Two-tailed t-test (A vs B)	
Samples		Type	Description	A	0.050 significance level	B
☑ (6)	☑	protocol	control	□	↔	□
☑ (6)		protocol	ERalpha expression	□	↔	□
☑ (6)	☑	agent	control	□	↔	□
☑ (6)		agent	E2	□	↔	□
	☑ GDS1326 only ☑ ranks ☑ values		subset effects	✓	Query A vs. B	✓

12 samples, order: none

☑ GSM41355 : AdLacZ+vehicle JM4
src1: MDA-MB-231

☑ GSM41356 : AdLacZ+vehicle JMTM1
src1: MDA-MB-231

☑ GSM41357 : AdLacZ+vehicle JMTM16
src1: MDA-MB-231

☑ GSM41352 : AdLacZ+E2-8 JM6
src1: MDA-MB-231

☑ GSM41353 : AdLacZ+E2-8 JMTM2
src1: MDA-MB-231

☑ GSM41354 : AdLacZ+E2-8 JMTM17
src1: MDA-MB-231

☑ GSM41349 : AdERalpha+vehicle JM1
src1: MDA-MB-231

☑ GSM41350 : AdERalpha+vehicle JMTM4
src1: MDA-MB-231

☑ GSM41351 : AdERalpha+vehicle JMTM18
src1: MDA-MB-231

☑ GSM41346 : AdERalpha+E2-8 JM3
src1: MDA-MB-231

☑ GSM41347 : AdERalpha+E2-8 JMTM5
src1: MDA-MB-231

☑ GSM41348 : AdERalpha+E2-8 JMTM19
src1: MDA-MB-231

(a) All samples

4 assigned subsets					? Two-tailed t-test (A vs B)	
Samples		Type	Description	A	0.050 significance level	B
□ (6)	☑	protocol	control	□	↔	□
☑ (6)		protocol	ERalpha expression	□	↔	□
☑ (6)	☑	agent	control	□	↔	□
☑ (6)		agent	E2	□	↔	□
	☑ GDS1326 only ☑ ranks ☑ values		subset effects	✓	Query A vs. B	✓

12 samples, order: none

□ GSM41355 : AdLacZ+vehicle JM4
src1: MDA-MB-231

□ GSM41356 : AdLacZ+vehicle JMTM1
src1: MDA-MB-231

□ GSM41357 : AdLacZ+vehicle JMTM16
src1: MDA-MB-231

□ GSM41352 : AdLacZ+E2-8 JM6
src1: MDA-MB-231

□ GSM41353 : AdLacZ+E2-8 JMTM2
src1: MDA-MB-231

□ GSM41354 : AdLacZ+E2-8 JMTM17
src1: MDA-MB-231

☑ GSM41349 : AdERalpha+vehicle JM1
src1: MDA-MB-231

☑ GSM41350 : AdERalpha+vehicle JMTM4
src1: MDA-MB-231

☑ GSM41351 : AdERalpha+vehicle JMTM18
src1: MDA-MB-231

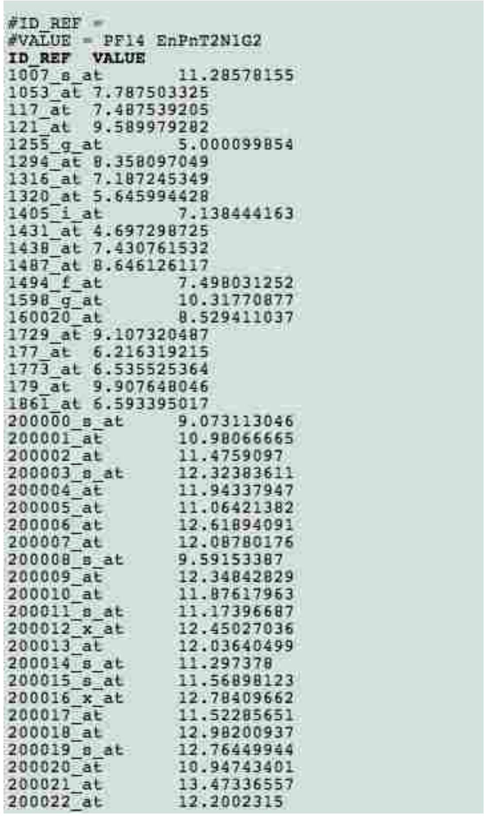
☑ GSM41346 : AdERalpha+E2-8 JM3
src1: MDA-MB-231

☑ GSM41347 : AdERalpha+E2-8 JMTM5
src1: MDA-MB-231

☑ GSM41348 : AdERalpha+E2-8 JMTM19
src1: MDA-MB-231

(b) Subset of samples

Figure 5.2: Possible subsets of samples marked by check-boxes. In (a), all samples are marked by a check. In (b), the subset of samples marked by the first check-box are now un-checked (National Center for Biotechnology Information 2002a).



```
#ID_REF =
#VALUE = PF14 EnPnT2N1G2
ID_REF  VALUE
1007_s_at       11.28578155
1053_at 7.787503325
117_at  7.487539205
121_at  9.589979282
1255_g_at       5.000099854
1294_at 8.358097049
1316_at 7.187245349
1320_at 5.645994428
1405_i_at       7.138444163
1431_at 4.697298725
1438_at 7.430761532
1487_at 8.646126117
1494_f_at       7.498031252
1598_g_at       10.31770877
160020_at       8.529411037
1729_at 9.107320487
177_at  6.216319215
1773_at 6.535525364
179_at  9.907648046
1861_at 6.593395017
200000_s_at     9.073113046
200001_at       10.98066665
200002_at       11.4759097
200003_s_at     12.32383611
200004_at       11.94337947
200005_at       11.06421382
200006_at       12.61894091
200007_at       12.08780176
200008_s_at     9.59153387
200009_at       12.34842829
200010_at       11.87617963
200011_s_at     11.17396687
200012_x_at     12.45027036
200013_at       12.03640499
200014_s_at     11.297378
200015_s_at     11.56898123
200016_x_at     12.78409662
200017_at       11.52285651
200018_at       12.98200937
200019_s_at     12.76449944
200020_at       10.94743401
200021_at       13.47336557
200022_at       12.2002315
```

Figure 5.3: Table of expression values of all genes for one sample.

6. DATA ANALYSIS

The method described in the previous chapter is applied to 10 of the 22,283 genes in the data set. For ease of explanation, the ten genes are given numbers, as shown in Table 6.1. The first five of the genes are randomly chosen from those genes that are differentially expressed, the next two are from the list of genes that are equivalently expressed, and the last three are from the list of genes that are moderately expressed.

Table 6.1: Genes to which the proposed method is applied. These genes are selected from the case study analysis; five from the list of differentially expressed genes, three just beyond the cutoff p-value, and two non-significant genes.

Group	Reference Number	Gene Reference Number
Differentially Expressed	Gene 1	211120_x_at
	Gene 2	218039_at
	Gene 3	200974_at
	Gene 4	202240_at
	Gene 5	61732_r_at
Equivalently Expressed	Gene 6	213570_at
	Gene 7	204773_at
Moderately Expressed	Gene 8	91952_at
	Gene 9	202378_s_at
	Gene 10	31799_at

Table 6.2 gives the four experiments that are returned for all ten genes. This is not a requirement of the script, but a feature of the data. After further investigation of the five data sets, it is determined that all five experiments involve the same genes. The only exceptions are the control genes.

Gene 1 has a very high probability of differential expression. The Perl script is run for Gene 1. The script brings back the list of ways to split the data for four experiments. The control and treatment groups are entered for each experiment. The eight text files with the arrays of gene expression values are imported into R. The two-

Table 6.2: The four breast cancer experiments mined by the Perl script. These four experiments are used in the analysis of all ten genes.

Author	Experiment
Farmer et al. (2005)	Identification of molecular apocrine breast tumours by microarray analysis.
Moggs et al. (2005)	Anti-proliferative effect of estrogen in breast cancer cells that re-express ERalpha is mediated by aberrant regulation of cell cycle genes.
Rae et al. (2005)	GREB 1 is a critical regulator of hormone dependent breast cancer growth.
Creighton et al. (2006)	Activation of mitogen-activated protein kinase in estrogen receptor alpha-positive breast cancer cells in vitro induces an in vivo molecular phenotype of estrogen receptor alpha-negative human breast tumors.

sample t-test model is run four times: once with the power prior distribution using all four previous experiments, "Power Prior," once using the first experiment to estimate the hyperparameters of the prior distributions, "Informative Prior," once using the tamoxifen data to estimate the hyperparameters of the prior distributions, "Empirical Prior," and once with the joint prior proportional to the prior for the variance with data-driven hyperparameters, "Flat Priors". Figure 6.1 gives a plot of the marginal posterior distribution for Δ_g using all four priors. The use of the flat priors on the mean and difference in means and the data-driven hyperparameters of the variance prior and the empirically estimated prior give nearly identical posterior distributions. The spread of these distributions is roughly 40 gene expression units. The power prior, the empirical prior, and the flat priors have posteriors centered at values greater than zero, but they are not as extreme as the posterior of the informative prior. The previous study chosen is one in which Gene 1 has a large probability of differential expression (Moggs et al. 2005). The spread of the posterior distributions decreases as the amount of information in the prior increases; that is, the power prior gives a posterior with about half the spread of the posterior using the empirically estimated prior. The choice of prior distributions also affects the estimated means, as seen in Table 6.3. The P($\Delta_g > 0$) is approximately the same for all four priors.

One common concern about the Bayesian approach to microarray analysis is that prior distributions are not objective. In this case, the choice of priors used does not affect the outcome; namely, Gene 1 is called differentially expressed regardless of the prior distribution. This outcome is expected, as the prior distribution should matter in cases where the difference in expression levels is near, but not equal to zero.

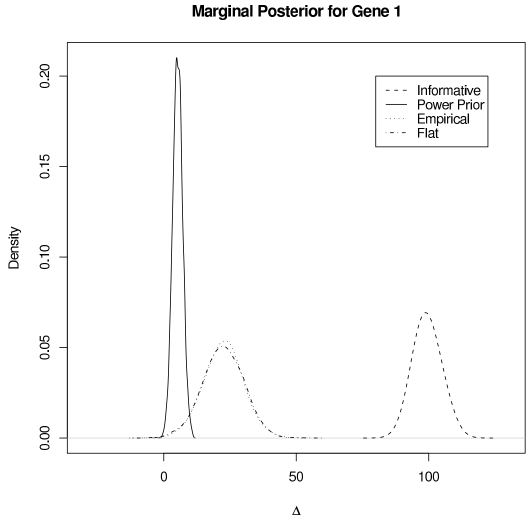


Figure 6.1: Marginal posterior distribution of Δ_g for Gene 1 using four different priors. The power prior has the smallest spread and the empirical prior and flat priors have the largest spread. The informative prior has a spread in between the non-informative priors and power prior and a mean shifted up about 100 gene expression units.

Table 6.3: Expected value of Δ_g and P($\Delta_g > 0$) for Gene 1 using four different priors. The P($\Delta_g > 0$) is approximately the same using any of the prior distributions, regardless of the large differences in the expected values.

Prior Distribution	E(Δ_g)	P($\Delta_g > 0$)
Informative Prior	99.48	> 0.999
Power Prior	2.357	0.996
Empirical Prior	9.547	> 0.999
Flat Priors	9.695	> 0.999

Gene 2 also has a high probability of differential expression. The Perl script brings back the list of ways to split the data for the same four experiments as were used for Gene 1. The control and treatment groups chosen are the same as in Gene 1. The same experiment is used to estimate the hyperparameters for the informative prior. The two-sample t-test model is run four times. Figure 6.2 gives a plot of the marginal posterior distribution for Δ_g using the four priors. The four posterior distributions are roughly centered around zero, though the flat priors pull the posterior slightly to the left. As with Gene 1, the spread of the distributions increases as the amount of knowledge included decreases, with the exception of the empirical prior and the flat priors. The expected value of Δ_g is shown in Table 6.4 using the four priors. The expected values and $\mathrm{P}(\Delta_g > 0)$ are affected by the four priors.

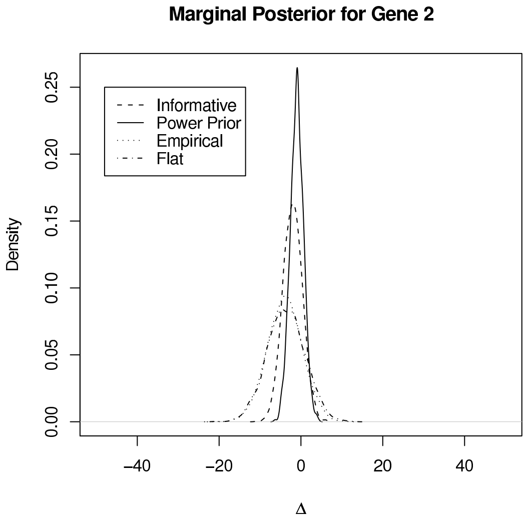


Figure 6.2: Marginal posterior distribution of Δ_g for Gene 2 using four different priors. The power prior has the smallest spread and the empirical prior has the largest spread. The informative and flat priors give posteriors with roughly the same spread.

Gene 3, Gene 4, and Gene 5 are three other differentially expressed genes with approximately the same difference in means. An experiment performed by Moggs et al. (2005) is used to estimate the hyperparameters for the informative prior. The

Table 6.4: Expected value of Δ_g and P($\Delta_g > 0$) for Gene 2 using four different priors. The use of the four priors has a noticeable effect on both the expected values and the P($\Delta_g > 0$).

Prior Distribution	E(Δ_g)	P($\Delta_g > 0$)
Informative Prior	-2.174	0.431
Power Prior	-2.026	0.267
Empirical Prior	-4.508	0.488
Flat Priors	-3.861	0.498

marginal posterior distributions using the four priors of Δ_g for Gene 3, Gene 4 and Gene 5 are shown in Figure 6.3. The marginal posteriors for Gene 3 using the the empirical prior and the flat priors are equivalent. The distributions are centered roughly around zero and the spread increases as the amount of information in the prior decreases. The marginal posteriors of Gene 4 using the empirical and power prior look equivalent to those in Gene 3. The flat priors give a marginal posterior for Gene 4 that is barely shifted to the right. The posterior using the informative prior is shifted down about 40 gene expression units due to the differential expression of the gene in the previous experiment. The posteriors for Gene 5 are close to those in Gene 3 with the exception of the informative prior. The informative prior gives a posterior that is shifted down about 15 gene expression units. The four prior distributions change the expected values and P($\Delta_g > 0$), as shown in Table 6.5.

Gene 4 and Gene 5 are only called differentially expressed using the informative prior. Figure 6.4 gives a plot of the power, informative and empirical priors for both Gene 4 and Gene 5. The empirical prior and the flat priors give equivalent posteriors for all three genes. The power prior has the smallest spread, the empirical prior and flat priors have the largest spread and the informative prior has a spread in between the non-informative priors and the power prior. The informative prior is shifted to the left for both Gene 4 and Gene 5.

The informative prior uses one historical experiment to estimate the prior pa-

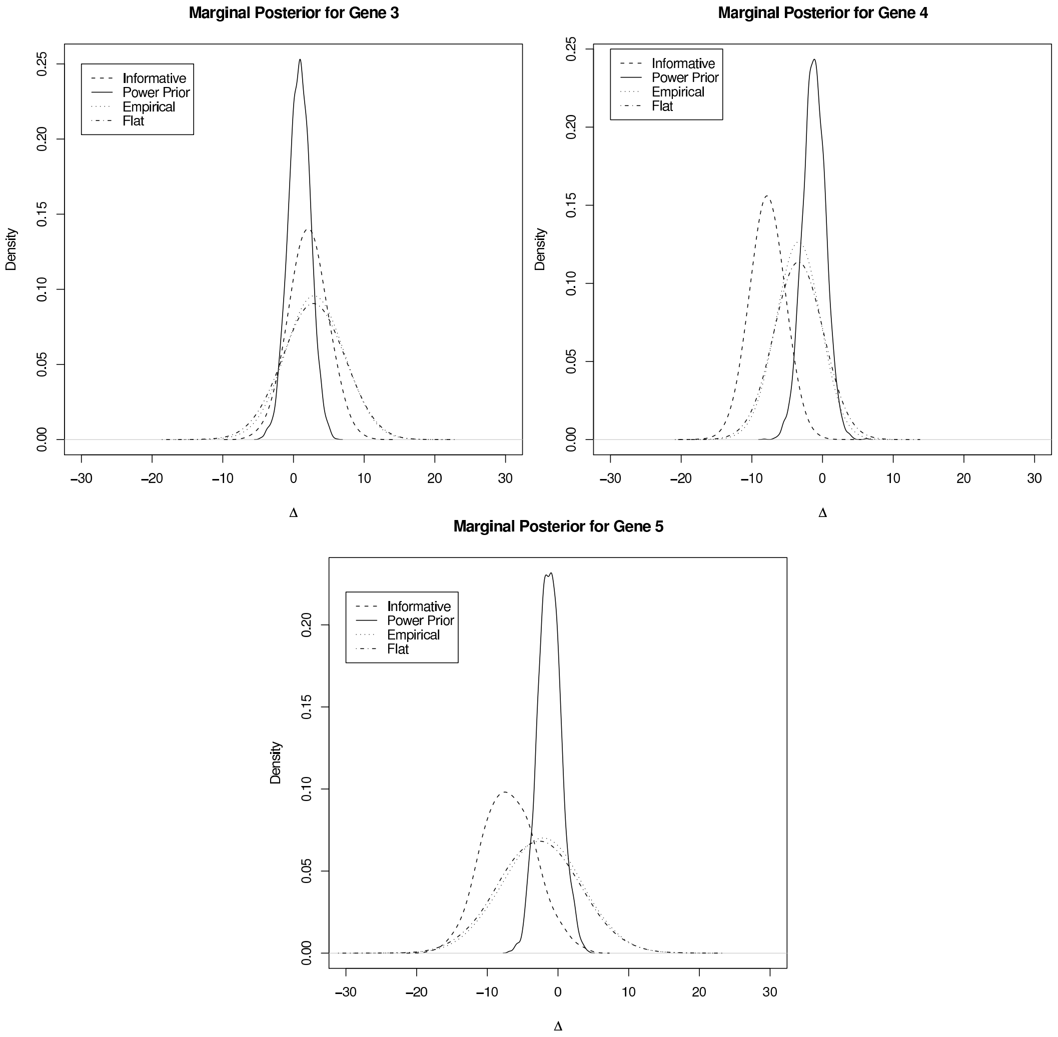


Figure 6.3: Marginal posterior distribution of Δ_g for Gene 3, Gene 4, and Gene 5 using four different priors. The empirical prior and the flat priors give equivalent posteriors for all three genes. The power prior has the smallest spread, the empirical prior and flat priors have the largest spread, and the informative prior has a spread in between the non-informative priors and the power prior. The informative prior is shifted to the left for both Gene 4 and Gene 5.

rameters. The choice of experiment can radically change the prior and, consequently, the posterior. Table 6.6 shows the estimated value of Δ_0 for each of the four previous experiments for Gene 4 and Gene 5. Experiment 3, the strongest historical evidence of differential expression, is used in this analysis. Suppose the expert believed that Experiment 4 is the best reflection of gene expression. With this prior distribution,

Table 6.5: Expected value of Δ_g and $P(\Delta_g > 0)$ for Gene 3, Gene 4, and Gene 5 using four different priors. The expected value of Δ_g varies with the choice of prior distributions for all three genes. The $P(\Delta_g > 0)$ also differs for all three genes, though more so for Gene 5 than for Gene 3 or Gene 4.

Gene 3		
Prior Distribution	$E(\Delta_g)$	$P(\Delta_g > 0)$
Informative Prior	3.455	0.603
Power Prior	0.873	0.718
Empirical Prior	2.061	0.809
Flat Priors	3.802	0.798
Gene 4		
Prior Distribution	$E(\Delta)$	$P(\Delta_g > 0)$
Informative Prior	-38.21	0.013
Power Prior	-1.264	0.203
Empirical Prior	-3.411	0.312
Flat Priors	-3.461	0.321
Gene 5		
Prior Distribution	$E(\Delta)$	$P(\Delta_g > 0)$
Informative Prior	-15.74	0.059
Power Prior	-1.273	0.223
Empirical Prior	-4.366	0.384
Flat Priors	-4.394	0.387

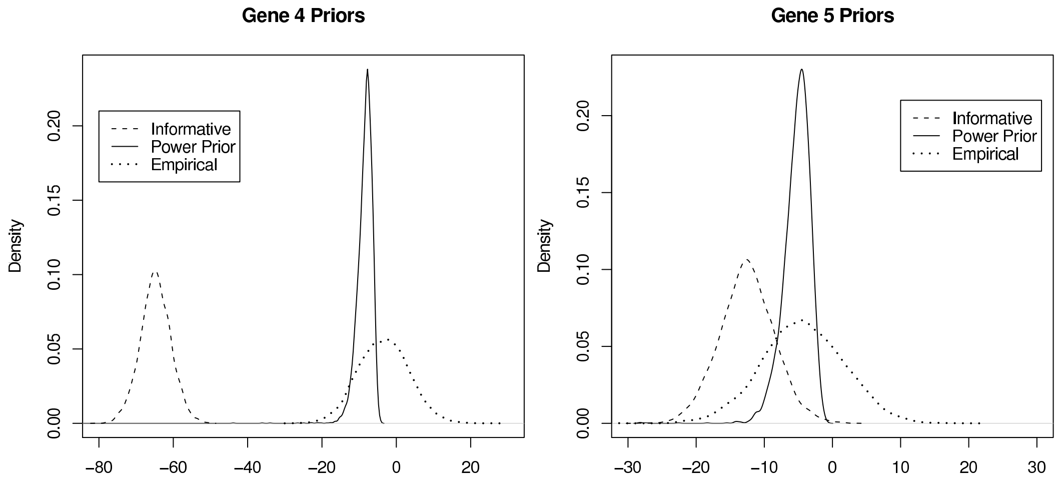


Figure 6.4: Prior distributions for Gene 4 and Gene 5. The informative prior is shifted to the left for both genes, whereas the power and empirical priors are closer to zero.

P($\Delta_g > 0$) is 0.232 and is therefore not called differentially expressed. The power prior combines the information from all four previous experiments. Besides Experiment 3, neither Gene 4 nor Gene 5 has strong evidence for differential expression; therefore, the posterior is shifted closer to zero.

Table 6.6: Estimated value of Δ_0 for each of the four previous experiments for Gene 4 and Gene 5.

Δ_0 (difference in sample means)		
Experiment	Gene 4	Gene 5
1	0.102	0.051
2	-0.599	0.156
3	-64.8	-12.5
4	0.352	-0.108

Gene 6 and Gene 7 are genes with a low probability of differential expression based on the case study analysis. The model is applied in the same fashion as in the previous five genes. The marginal posterior distributions of Δ_g using the four prior distributions for Gene 6 and Gene 7 are shown in Figure 6.5. These distributions have the same center, aside from the informative prior for Gene 7. The informative prior using the experiment by Moggs et al. (2005) shifts the posterior of Δ_g down by about 5 gene expression units. The spread is different across the choice of priors. The power prior has the smallest spread, the empirical prior and flat priors have the largest spread, and the informative prior has a moderate spread. The choice of prior distribution does not affect the mean for Δ_g and P($\Delta_g > 0$) as much as it affects the differentially expressed genes. The means and probabilities are shown in Table 6.7.

Finally, Gene 8, Gene 9, and Gene 10 are moderately expressed genes; that is, these genes are near the boundary for differential expression. The Perl script returns text files of expression values for the same four experiments as in the analysis of all previous genes. The marginal posterior distribution for Δ_g using all four prior distributions for Gene 8, Gene 9, and Gene 10 are similar to each other (see Figure

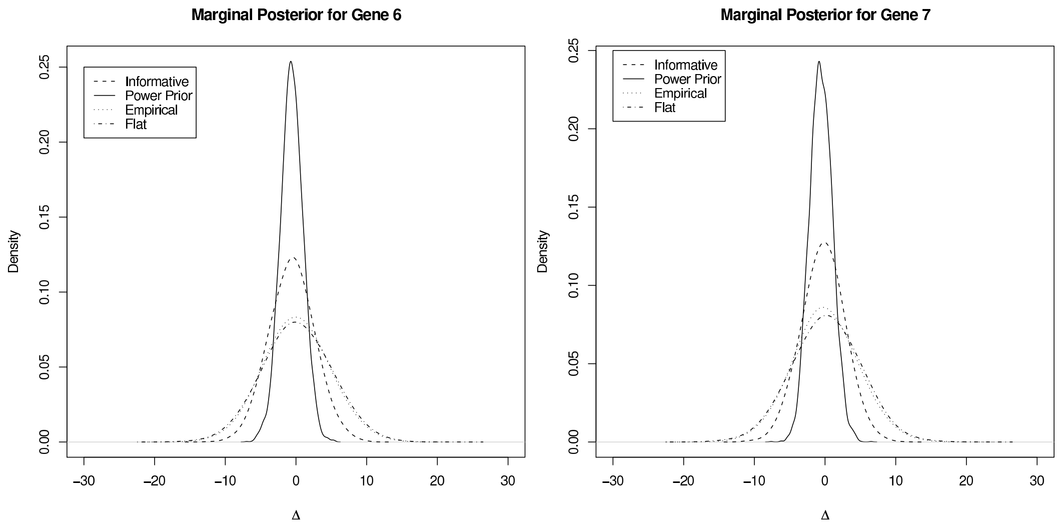


Figure 6.5: Marginal posterior distribution of Δ_g for Gene 6 and Gene 7 using four different priors. The marginal posteriors for both Gene 6 and Gene 7 are as expected: they are centered around zero and the spread increases as the amount of information in the prior decreases.

Table 6.7: Expected value of Δ_g and $\mathrm{P}(\Delta_g > 0)$ for Gene 6 and Gene 7 using three different priors. While the estimates of the expected value of Δ_g and $\mathrm{P}(\Delta_g > 0)$ change with the choice of prior distributions, the difference is not as dramatic as with the differentially expressed genes.

	Gene 6	
Prior Distribution	$\mathrm{E}(\Delta_g)$	$\mathrm{P}(\Delta_g > 0)$
Informative Prior	-0.640	0.465
Power Prior	-0.606	0.357
Empirical Prior	0.027	0.507
Flat Priors	0.072	0.524
	Gene 7	
Prior Distribution	$\mathrm{E}(\Delta_g)$	$\mathrm{P}(\Delta_g > 0)$
Informative Prior	-4.539	0.368
Power Prior	-0.645	0.343
Empirical Prior	0.085	0.505
Flat Priors	0.044	0.523

6.6). The posteriors for Gene 8 and Gene 10 are almost identical. The marginal posterior for Gene 9 using the informative prior is slightly shifted to the left. Also, for both Gene 9 and Gene 10, the empirical prior is visibly more peaked than the posterior using the flat priors. The estimates of the expected value for Δ_g and $P(\Delta_g > 0)$ are affected by the choice of prior distribution, as seen in Table 6.8.

Another common concern about the Bayesian methodology for microarray analysis is that the prior distribution swamps the data. For the three moderately expressed genes, Gene 8, Gene 9, and Gene 10, the informative priors have no greater impact than the flat priors.

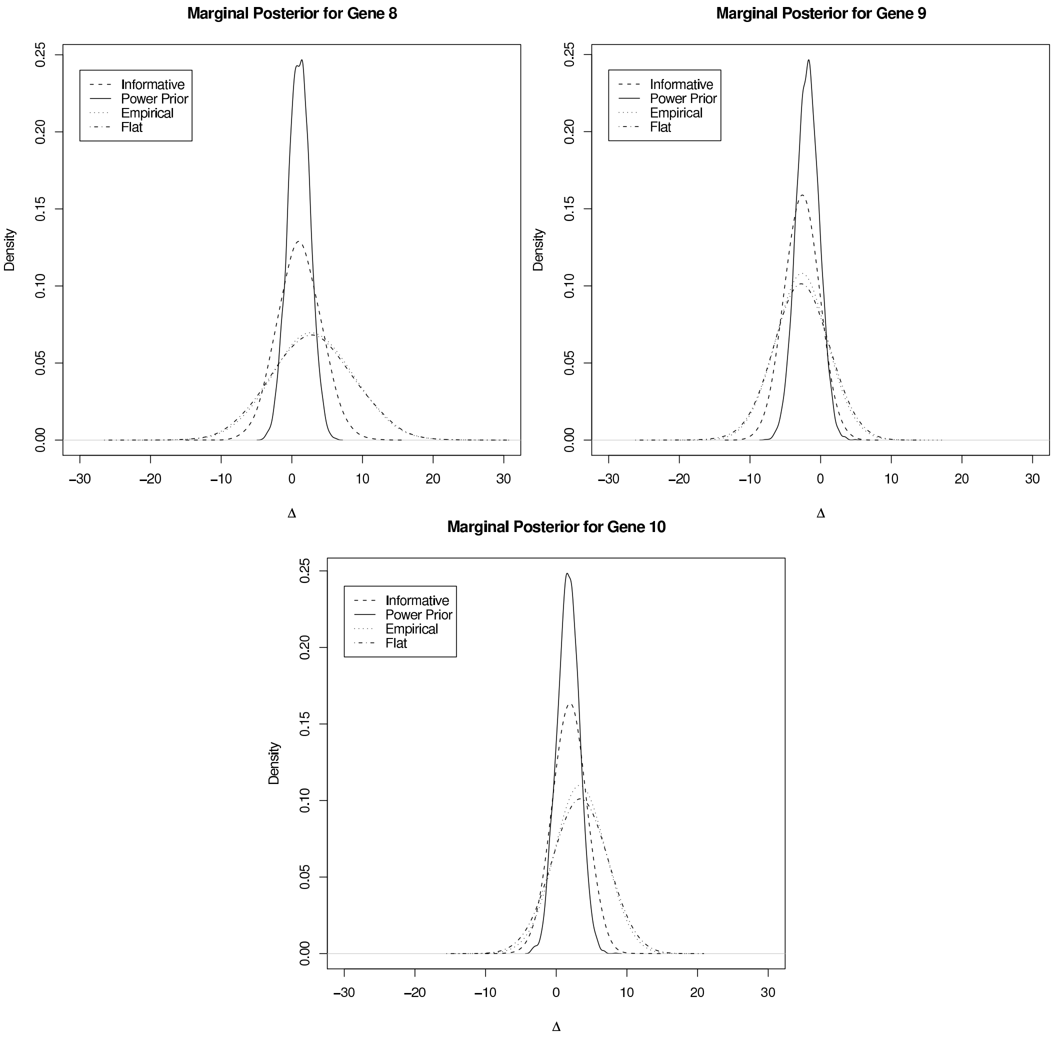


Figure 6.6: Marginal posterior distribution of Δ_g for Gene 8, Gene 9, and Gene 10 using four different priors. The posteriors using the power prior and the empirical prior and flat priors are similar for all three genes. The informative prior shifts slightly for Gene 9 and the spread is much closer to that of the flat priors and empirical prior.

Table 6.8: Expected value of Δ_g and P($\Delta_g > 0$) for Gene 8, Gene 9, and Gene 10 using four different priors. There are more detectable differences in the expected value of Δ_g and the P($\Delta_g > 0$) than with the equivalently expressed genes, but the changes are not as radical as with the differentially expressed genes.

Gene 8		
Prior Distribution	E(Δ_g)	P($\Delta_g > 0$)
Informative Prior	1.114	0.635
Power Prior	0.995	0.737
Empirical Prior	1.029	0.714
Flat Priors	1.019	0.697
Gene 9		
Prior Distribution	E(Δ_g)	P($\Delta_g > 0$)
Informative Prior	-6.739	0.228
Power Prior	-1.949	0.105
Empirical Prior	-0.777	0.317
Flat Priors	-0.787	0.355
Gene 10		
Prior Distribution	E(Δ_g)	P($\Delta_g > 0$)
Informative Prior	2.623	0.679
Power Prior	1.695	0.857
Empirical Prior	1.424	0.818
Flat Priors	1.404	0.839

For a given gene, the Perl script retrieves the expression values for each sample of the historical studies. The model returns marginal posterior distributions for Δ_g using all three proposed prior distributions. The center of the marginal posterior distributions is not as affected by the choice of prior distribution as the spread. For all ten genes, the spread of the posterior increases as the choice of prior moves from the power prior to the informative prior and from the informative prior to the empirical prior. The estimates of the expected value of Δ_g and P($\Delta_g > 0$) are also affected by the choice of priors. The recommended prior distribution for all cases is the power prior. This prior uses information from multiple previous experiments and decreases the variance of the marginal posteriors. The informative prior provides information that is ignored by the flat priors, but it disregards other previous experiments. One possible variation of the informative prior is estimating the parameters of the prior on Δ by the mean and variance of the difference in sample means of the four experiments.

7. CONCLUSIONS

Bayesian models are ideal for the analysis of microarray studies because of the ability to integrate prior knowledge. Most current alternate approaches do not include the information about previous studies in the microarray analysis. The approach proposed in this thesis uses a Perl script to mine GEO for previous experiments to build prior distributions. Four prior distributions are explored: a power prior distribution using historical studies as proposed by Ibrahim and Chen (2000), an informative prior using one historical study to estimate the hyperparameters, a prior with data-driven hyperparameters, and flat priors. A model is proposed similar to the two sample Bayesian t-test presented in Fox and Dimmic (2006) is proposed to detect differentially expressed genes. The process is applied to ten genes from a breast cancer experiment. The script and the model perform as expected.

The model chosen is simple, yet effective. A grand hierarchical model that combines all of the information of the current and the past experiments could have been proposed. While this plan may increase the power to detect differentially expressed genes, there are many drawbacks. One major problem with a hierarchical model combining multiple studies is the difficulty of keeping track of all the levels of replication. Housekeeping can prove to be a daunting task with multiple genes from multiple arrays within each of many studies while also accounting for any missing expression levels, arrays, or replicates. Another difficulty with the hierarchical approach is the choice of historical studies to include. These experiments are likely informative about the same basic biological process, but they may target different populations of people or include some different treatments or replicates. They may also be investigating a different set of genes, which then requires decreasing the number of genes analyzed to a group of common genes from all studies included. That is why, in this work, the historical experiments are used to build a prior distribution.

The results of the analysis imply that the choice of prior distribution does affect the marginal posterior distribution of the difference in means. There are noticable differences in the spread of the posteriors of all three types of genes. The power prior gives the posterior distribution with the smallest spread, and the empirical prior gives the posterior distribution with the largest spread. The informative prior gives posteriors with a spread somewhere between those given by the the power prior and the empirical prior. The estimate of the expected value of the difference also changes between the choices of prior distributions. This change is most apparent in the genes that are differentially expressed. The informative prior, as defined here, is heavily sensitive to the choice of the historical study used to estimate the parameters. The power prior is recommended as the prior distribution because of the incorporation of multiple previous studies.

The prior distributions for the proposed model are more informative than a flat prior on the difference in means. However, elicitation is an iterative process. In an ideal world, this prior would be a starting point. An expert would be consulted and the prior would be modified using elicitation methods. Therefore, in reality, while these priors are one step above a flat prior distribution, there is much room for improvement and further development of this distribution.

There is other information about the genome that could be included in the analysis. Some include the proportion of experiments in which the gene is called differentially expressed, which chromosome the gene is located on, the location of the gene on the chromosome, whether the gene is from the positive or the negative strand, the environment of the gene (co-factors), or information about clusters of genes that work together. This information could be included in the construction of a prior distribution on the probability a given gene is differentially expressed.

The prior distributions are formed with the output from the Perl script. The

Perl script returns the expression values of a given gene from each sample in multiple relevant previous studies. In the case of the ten chosen genes, these previous studies are all the same, though uniformity is not required. The script is also able to overcome obstacles such as different gene accession name forms, different successive genes across experiments, and genes at the end of the file. However, the script does require user input in choosing how to split the data for each experiment. Additionally, the script only works for one gene. Future work will generalize the script to run without user input for all the genes in the breast cancer experiment. Also, because the script is currently specific to the case study experiment, there is future work in creating a script that follows the entire procedure for any given experiment.

This work provides a program to mine previous microarray studies to build informative priors for a Bayesian analysis. It presents a framework for easily incorporating genomic knowledge into an analysis. The informative priors investigated are shown to maintain objectivity and swamp the data no more than the flat priors. The informative prior estimated with data from one historical experiment exhibits an improvement in inference on genes with a moderate difference in expression. Combining information from multiple historical studies, as with the power prior, is preferred over prior parameter estimation using one previous study. With future work, this method can be generalized for use in any given experiment. The resulting list of differentially expressed genes will be more accurate and, consequently, help move genomic research forward.

BIBLIOGRAPHY

Acevedo, M., Lee, K., Stender, J., Katzenellenbogen, B., and Kraus, W. (2004), "Selective recognition of distinct classes of coactivators by a ligand-inducible activation domain." *Molecular Cell*, 13, 725–738.

Affymetrix (1992), "GeneChip Arrays," .

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, 57, 289–300.

Benson, P. G., Curley, S. P., and Smith, G. F. (1995), "Belief assessment: An underdeveloped phase of probability elicitation," *Management Science*, 41, 1639–1653.

Berman, M. (1988), "Subjective probability and expert opinion," *Reliability Engineering and System Safety*, 23, 263–268.

Bolstad, B. (2001), "Bioconductor package: affyPLM," Www.bioconductor.org.

Brama, M., Basciani, S., Cherubini, S., Mariani, S., Migliaccio, S., Arizzi, M., Rosano, G., Spera, G., and Gnessi, L. (2007), "Osteoblast-conditioned medium promotes proliferation and sensitizes breast cancer cells to imatinib treatment," *Endocrine-Related Cancer*, 14, 61–72.

Cech, T. R., Eddy, S. R., Eisenberg, D., Hersey, K., Holtzman, S. H., Poste, G. H., Raikhel, N. V., Scheller, R. H., Singer, D. B., and Waltham, M. C. (2003), *Sharing publication-related data and materials: Responsibilities of authorship in the biological life sciences*, National Academies Press.

Chen, H., Rubin, E., Zhang, H., Chung, S., Jie, C. C., Garrett, E., Biswal, S., and Sukumar, S. (2005), "Identification of transcriptional targets of HOXA5," *Journal of Biological Chemistry*, 280, 19373–19380.

Choi, J., Yu, U., Kim, S., and Yoo, O. (2003), “Combining multiple microarray studies and modeling inter-study variation,” *Bioinformatics*, 19, i84–i90.

Clark, R. A., Snedeker, S., and Devine, C. (1998), “Estrogen and breast cancer,” Cornell University.

Conlon, E. M., Song, J. J., and Liu, J. S. (2006), “Bayesian models for pooling microarray studies with multiple sources of replication,” *Bioinformatics*, 7, 247–259.

Coser, K. R., Chesnes, J., Hur, J., Ray, S., Isselbacher, K. J., and Shioda, T. (2003), “Global analysis of ligand sensitivity of estrogen inducible and suppressible genes in MCF7/BUS breast cancer cells by DNA microarray,” *Proceedings of the National Academy of Science*, 100, 13994–13999.

Creighton, C., Hilger, A., Murthy, S., Rae, J., Chinnaiyan, A., and El-Ashry, D. (2006), “Activation of mitogen-activated protein kinase in estrogen receptor alpha-positive breast cancer cells in vitro induces an in vivo molecular phenotype of estrogen receptor alpha-negative human breast tumors.” *Cancer Research*, 66, 3903–3911.

Dittmer, A., Vetter, M., Schunke, D., Span, P. N., Sweep, F., Thomssen, C., and Dittmer, J. (2006), “Parathyroid hormone-related protein regulates tumor-relevant genes in breast cancer cells,” *Journal of Biological Chemistry*, 281, 563–572.

Duggan, D., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. (1999), “Expression profiling using cDNA microarrays,” *Nature Genetics*, 21, 10–14.

E. M. Dougherty, J. (1993), “Distributed reality,” *IEEE Transactions on Reliability*, 42, 6–9.

Eakin, C., Maccoss, M., Finney, G., and Klevit, R. (2007), "Estrogen receptor alpha is a putative substrate for the BRCA1 ubiquitin ligase." *Proceedings of the National Academy of Sciences of the United States of America*, 105, ePub.

Efron, B. (1986), "Why isn't everyone a Bayesian?" *The American Statistician*, 40, 1–5.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Enmark, E., Pelto-Huikko, M., Grandien, K., Lagercrantz, S., Fried, J. L. G., Nordenskjold, M., and Gustafsson, J. (1997), "Human estrogen receptor beta-gene structure, chromosomal localization, and expression pattern," *Journal of Clinical Endocrinology and Metabolism*, 82, 4258–4265.

Ericsson, K. A. and Simon, H. A. (1984), *Protocol analysis*, MIT Press.

Evans, R. A. (2000), "Subjective probability and prior knowledge," *IEEE Transactions on Reliability*, 49, 249.

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., MacGrogan, G., Bergh, J., Cameron, D., Goldstein, D., Duss, S., Nicoulaz, A.-L., Brisken, C., Fiche, M., Delorenzi, M., and Iggo, R. (2005), "Identification of molecular apocrine breast tumours by microarray analysis." *Oncogene*, 24, 4660–4671.

Feingold, E. (2003), "Microarray basics," University of Pittsburgh lecture.

Fox, R. J. and Dimmic, M. W. (2006), "A two-sample Bayesian t-test for microarray data," *Bioinformatics*, 7, 126–136.

Frasor, J., Chang, E. C., Komm, B., Lin, C.-Y., Vega, V. B., Liu, E. T., Miller, L. D., Smeds, J., Bergh, J., and Katzenellenbogen, B. S. (2006), "Gene expression preferentially regulated by Tamoxifen in breast cancer cells and correlations with clinical outcome," *Cancer Research*, 66, 334–340.

Fu, R., Dey, D. K., and Holsinger, K. E. (2005), "Bayesian models for the analysis of genetic structure when populations are correlated," *Bioinformatics*, 21, 1516–1529.

Ghosh, M., Maiti, T., Kim, D., Chakraborty, S., and Tewari, A. (2004), "Hierarchical Bayesian neural networks: An application to a prostate cancer study," *Journal of the American Statistical Association*, 99, 601–608.

Goodstein, D. (2000), "How science works," Talk at California Institute of Technology.

Gottardo, R., Pannucci, J. A., Kuske, C. R., and Brettin, T. (2003), "Statistical analysis of microarray data: a Bayesian approach," *Biostatistics*, 4, 597–620.

Grether, D. M. (1980a), "Bayes' Rule as a descriptive model: The representativeness heuristic," *Quarterly Journal of Economics*, 95, 537–557.

— (1980b), "Testing Bayes rule and the representativeness heuristic: some experimental evidence," *Journal of Economic Behavior and Organization*, 17, 31–57.

H. E. Kyburg, J. (2001), "Probability as a guide in life," *The Monist*, 84, 135–152.

Hayashi, S., Eguchi, H., Tanimoto, K., Yoshida, T., Omoto, Y., Inoue, A., Yoshida, N., and Yamaguchi, Y. (2003), "The expression and function of estrogen receptor alpha and beta in human breast cancer and its clinical application," *Endocrine-related Cancer*, 10, 193–202.

Heber, S. and Sick, B. (2006), "Quality assessment of Affymetrix GeneChip data," *Journal of Integrative Biology*, 10, 358–368.

Ibrahim, J. G. and Chen, M.-H. (2000), “Power Prior distributions for regression models,” *Statistical Science*, 46, 551–552.

Ibrahim, J. G., Chen, M.-H., and Gray, R. J. (2002), “Bayesian models for gene expression with DNA microarray data,” *Journal of the American Statistical Association*, 97, 88–99.

Itoh, T., Karlsberg, K., Kijima, I., and Yuan, Y. (2005), “etrozole-, anastrozole-, and tamoxifen-responsive genes in MCF-7aro cells: a microarray approach,” *Molecular Cancer Research*, 3, 203–218.

Kadane, J. B. and Wolfson, L. J. (1998), “Experiences in elicitation,” *The Statistician*, 47, 3–19.

Kelley, G. A. (1955), *The psychology of personal constructs*, Norton.

Kendziorski, C. M., Newton, M. A., Lan, H., and Gould, M. N. (2003), “On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles,” *Statistics in Medicine*, 22, 3899–3914.

Lander, E. (1999), “Array of hope,” *Nature Genetics*, 21, 3–4.

Lönnstedt, I. and Speed, T. (2002), “Replicated microarray data,” *Statistica Sinica*, 12, 31–46.

Marx, C., Yau, C., Banwait, S., Zhou, Y., Scott, G., Hann, B., Park, J., and Benz, C. (2007), “Proteasome Regulated ERBB2 and Estrogen Receptor Pathways in Breast Cancer.” *Molecular Pharmacology*, 72, ePub.

Mecham, B. H., Klus, G. T., Strovel, J., Augustus, M., Byrne, D., Bozso, P., Wetmore, D. Z., Mariani, T. J., Kohane, I. S., and Szallasi, Z. (2004), “Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements." *Nucleic Acids Research*, 32, e74.

Meyer, M. A. and Booker, J. M. (2001), *Eliciting and analyzing expert judgement*, Society for Industrial and Applied Mathematics.

Moggs, J. G., Murphy, T. C., Lim, F. L., Moore, D. J., Stuckey, R., Antrobus, K., Kimber, I., and Orphanides, G. (2005), "Anti-proliferative effect of estrogen in breast cancer cells that re-express ERalpha is mediated by aberrant regulation of cell cycle genes." *Journal of Molecular Endocrinology*, 34, 535–551.

Mosleh, A. and Bier, V. M. (1996), "Uncertainty about probability: A reconciliation with the subjectivist viewpoint," *IEEE Transactions on Systems, Man, and Cybernetics*, 26, 202–309.

Mosteller, F. and Wallace, D. L. (1963), "Inference in an authorship problem," *Journal of the American Statistical Association*, 58, 275–309.

National Cancer Institute (1996), "Cancer Genome Anatomy Project," .

National Center for Biotechnology Information (2002a), "Gene Expression Omnibus," May 1, 2007, http://www.ncbi.nlm.nih.gov/geo/index.cgi.

— (2002b), "PubMed," May 1, 2007, http://www.ncbi.nlm.nih.gov/pubmed/index.cgi.

Nau, R. F. (2001), "DeFinetti was right: probability does not exist," *Theory and Decision*, 51, 81–124.

Parker, M., Cowley, S., Heery, D., Henttu, P., Kalkhoven, E., Sjoberg, M., Valentine, J., and White, R. (1997), "Function of estrogen receptors in breast cancer," *Breast Cancer*, 4, 204–208.

Parmigiani, G., Garrett-Meyer, E., Anbazhagan, R., and Gabrielson, E. (2004), "A cross-study comparison of gene expression studies for the molecular classification of lung cancer," *Clinical Cancer Research*, 10, 2922–2927.

Poola, I., DeWitty, R. L., Marshalleck, J. J., Bhatnagar, R., Abraham, J., and Leffall, L. D. (2005), "Identification of MMP-1 as a putative breast cancer predictive marker by global gene expression analysis," *Nature Medicine*, 11, 481–483.

Poola, I. and Yue, Q. (2007), "Estrogen receptor alpha (ERalpha) mRNA copy numbers in immunohistochemically positive-, and negative breast cancer tissues." *BMC Cancer*, 7, 56–61.

Rae, J. M., Johnson, M. D., Scheys, J. O., Cordero, K. E., Larios, J. M., and Lippman, M. E. (2005), "GREB 1 is a critical regulator of hormone dependent breast cancer growth." *Breast Cancer Research and Treatment*, 92, 141–149.

Ray, A., Nkhata, K., Grande, J., and Cleary, M. (2007), "Diet-induced obesity and mammary tumor development in relation to estrogen receptor status." *Cancer Letters*, 250, ePub.

Reeve, J., Owens, R. G., and Neimeyer, G. J. (2004), "Using examples in reperatory grids: The influence on construct elicitation," *Journal of Constructivist Psychology*, 15, 121–126.

Rhodes, D., Barrette, T., Rubin, M., Ghosh, D., and Chinnaiyan, A. (2002), "Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer," *Cancer Research*, 62, 4427–4433.

Rhodes, D., Yu, J.and Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. (2004), "Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression." *Proceedings of National Academy of Science USA*, 101, 9309–9314.

Richardson, A., Wang, Z., Nicolo, A. D., Lu, X., Brown, M., Miron, A., Liao, X., Iglehart, J., Livingston, D., and Ganesan, S. (2006), "X chromosomal abnormalities in basal-like human breast cancer," *Cancer Cell*, 9, 121–132.

Roberts, P. M. (2006), "Mining literature for systems biology," *Briefings in Bioinformatics*, 7, 399–406.

Savage, L. J. (1971), "Elicitation of personal probabilities and expectations," *American Statistical Association*, 66, 783–801.

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

Singpurwalla, N. D. (2002), "Some cracks in the empire of chance: Flaws in the foundations of reliability," *International Statistical Review*, 70, 53–78.

Stanford University (2003), "Stanford MicroArray Database," .

Stitziel, N. O., Mar, B. G., Liang, J., and Westbrook, C. A. (2004), "Membrane-associated and secreted genes in breast cancer," *Cancer Research*, 64, 8682–8687.

Swain, S. M. (2001), "Tamoxifen for patients with estrogen receptor negative breast cancer," *Journal of Clinical Oncology*, 19, 93–97.

Tsodikov, A. D., Ibrahim, J. G., and Yakovlev, A. Y. (2003), "Estimating cure rates from survival data: An alternative to two-component mixture models," *Journal of the American Statistical Association*, 98, 1063–1078.

Wang, P. (1996), "Heuristics and normative models of judgement under uncertainty," *International Journal of Approximate Reasoning*, 14, 221–235.

Wang, P., Coombes, K., Highsmith, W., Keating, M., and Abruzzo, L. (2004), "Differences in gene expression between B-cell chronic lymphcytic leukemia and normal B cells: a meta-analysis of three microarray studies." *Bioinformatics*, 20, 3166–3178.

Wonsey, D. R. and Follettie, M. T. (2005), "Loss of the forkhead transcription factor FoxM1 causes centrosome amplification and mitotic catastrophe," *Cancer Research*, 65, 5181–5189.

Wu, K., Yang, Y., Wang, C., Davoli, M. A., D'Amico, M., Li, A., Cveklova, K., Kozmik, Z., Lisanti——, M. P., Russell, R. G., Cvekl, A., and Pestell, R. G. (2003), "DACH1 inhibits transforming growth factor-beta signaling through binding Smad4," *Journal of Biological Chemistry*, 278, 51673–51684.

Wu, W., Zou, M., Brickley, D., and Pew, T. (2006), "Glucocorticoid receptor activation signals through forkhead transcription factor 3a in breast cancer cells," *Molecular Endocrinology*, 20, 2304–2314.

Yuen, T., Wurmbach, E., Pfeffer, R., Ebersole, B., and Sealfon, S. (2002), "Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays," *Nucleic Acids Research*, 30, 48–52.

```
#!/usr/bin/perl -w
  use WWW::Mechanize;
  use HTML::TokeParser;

my $gene = "211120_x_at";
my $agent = WWW::Mechanize->new();

##Go to GEO Profiles website
$agent->get("http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=geo");

##Search for experiments
$agent->set_visible("GEO Profiles","$gene breast cancer estrogen receptor");
$agent->click_button(number=>1);
my $url = "http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?view=data&acc=GSM";
my $coun=1;

##Get titles and GDS numbers
foreach ($agent->links){
  if ($_->[0] =~ /dataset/){
  $agent->get($_->[0]);
my $stream = HTML::TokeParser->new(\$agent->{content});
 $stream->get_tag("html");
 $stream->get_tag("head");
 $stream->get_tag("meta");
 $stream->get_tag("title");
my $gds = $stream->get_trimmed_text("title","/title");
  if ($gds =~ m/GDS(.*)$gene/){
    $dataset = $1;
   chop($dataset);
   chop($dataset);
   chop($dataset);}
 $stream->get_tag("b");
 $stream->get_tag("b");
 $stream->get_tag("b","/b");
 my $title= $stream->get_text("/b","br");
  if($title =~ /estrogen receptor/ and $title =~ /breast cancer/ and
     $title !~ /tamoxifen/){

##Obtain list of how to split data
my $web = WWW::Mechanize->new();
$web->get("http://www.ncbi.nlm.nih.gov/projects/geo/gds/
                                     gds_browse.cgi?gds=$dataset");
```

```
my @exp;
my $cnt=0;
my $stream = HTML::TokeParser->new(\$web->{content});
  while (my $token = $stream->get_tag("input")) {
      my $type = $token->[1]{type} || "-";
      if ($type =~/HIDDEN/){
      my $text = $token->[1]{name} || "-";
      if ($text =~ /sub/){
if ($text =~ /allsubnames/){
$stream->get_tag("input");
    $stream->get_tag("tr");
    $stream->get_tag("td","/td");
    $stream->get_tag("td");
   $stream->get_tag("td","/td");
   $stream->get_tag("td","/td");
   $stream->get_tag("td","/td");
   $stream->get_tag("td","/td");
    $exp[$cnt] =$stream->get_trimmed_text("td","/td");
if ($exp[$cnt] =~/control/){$exp[$cnt] = "control$cnt";}
$cnt=$cnt+1;}
      my $value = $token->[1]{value} || "-";
    my $file = join("\n","$exp[$cnt-1]","$coun","txt");
      open file, ">>$file";
      print file "$value\n";
      close file;
     $count=$count+1;
}}}
##Choose how to split data
my $temp = WWW::Mechanize->new();
print "@exp\n";
print "Choose first group:\n";
my $group1 =<STDIN>;
print "Choose second group:\n";
my $group2 =<STDIN>;

chomp $group1;
chomp $group2;

my $file1= join(".","$group1","$coun","txt");
my $file2= join(".","$group2","$coun","txt");

##Retrieve expression values for group 1
my @gsm1;
open input1,"$file1";
```

```
@gsm1=<input1>;
close input1;
splice(@gsm1, 0, 1);

my @group1;
my $count=0;
foreach $gsm (@gsm1){
my $url2 = join('',$url,$gsm);
  $temp->get($url2);
  my $values = HTML::TokeParser->new(\$temp->{content});
         #find table values and save them
      if ($values->get_tag("\pre")){
      my $title = $values->get_trimmed_text([$endtag]);
      if($title =~ m/$gene(.*)\n/) {
       my $num = $1;
while($num =~ m/[A-Z]/){
          chop($num);}
      $group1[$count]=$num;
      $count=$count+1;
      }}}
##Print to file
my $outfile1 = join("\n", "$group1","$coun","2","txt");
$grp1=join(",",@group1);
open out1,">$outfile1";
print out1 "$grp1";
close out1;

##Retrieve expression values for group 2
my @gsm2;
open input2,"$file2";
@gsm2=<input2>;
close input2;
splice(@gsm2, 0, 1);

my @group2;
my $count=0;
foreach $gsm (@gsm2){
my $url2 = join('',$url,$gsm);
  $temp->get($url2);
  my $values = HTML::TokeParser->new(\$temp->{content});
      if ($values->get_tag("\pre")){
      my $title = $values->get_trimmed_text([$endtag]);
      if($title =~ m/$gene(.*)\n/) {
       my $num = $1;
while($num =~ m/[A-Z]/){
```

```
          chop($num);}
      $group2[$count]=$num;
      $count=$count+1;
      }}}
##Print to file
my $outfile2 = join(".", "$group2","$coun","2","txt");
$grp2=join(",",@group2);
open out2, ">$outfile2";
print out2 "$grp2";
close out2;}
$agent->back();
$coun=$coun+1;
}}
```