Biometrical Journal **0** (0) 0, zzz–zzz / DOI: 10.1002/

Bayesian Decision Theoretic Multiple Comparison Procedures: An Application to Phage Display Data

Luis G. León-Novelo[***1], **Peter Müller** [2], **Wadih Arap** [3], **Jessica Sun** [3], **Renata Pasqualini** [3], and **Kim-Anh Do** [4]

[1] Department of Mathematics, University of Louisiana at Lafayette, LA 70504-1010

[2] Department of Mathematics, University of Texas, Austin, TX 78712

[3] Departments of Genitourinary Medical Oncology and Cancer Biology, M.D. Anderson Cancer Center, Houston, TX 77230-1439

[4] Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX 77230-1402

We discuss a case study that highlights the features and limitations of a principled Bayesian decision theoretic approach to massive multiple comparisons. We consider inference for a mouse phage display experiment with three stages. The data are tripeptide counts by tissue and stage. The primary aim of the experiment is to identify ligands that bind with high affinity to a given tissue. The inference goal is to select a large list of peptide and tissue pairs those with significant increase over stages. The desired inference summary involves a massive multiplicity problem. We consider two alternative approaches to address this multiplicity issue. First we propose an approach based on the control of the posterior expected false discovery rate. We notice that the implied solution ignores the relative size of the increase. This motivates a second approach based on a utility function that includes explicit weights for the size of the increase.

1 Introduction

We discuss a Bayesian decision theoretic approach to control multiplicities in a massive multiple comparison. The discussion is in the context of a particular case study that highlights the features and limitations of such approaches. We analyze data from a mouse phage display experiment. Details of the experiment and the data are discussed later. The experiment is carried out to identify proteins that preferentially bind to specific tissues. Such knowledge could in future be used to develop targeted therapies that deliver a drug to specific tissues and limit side effects (Kolonin et al., 2006; Arap et al., 2006). The data y_{ij} are counts for a large number of tripeptide/tissue pairs, $i = 1, \ldots, n$, across stages $j = 1, 2, 3$. The tripeptides characterize different proteins. For each tripeptide/tissue pair the experiment reports counts over three consecutive stages. The nature of the experiment is such that for proteins that preferentially bind to some type of tissue the counts should be monotone increasing, because the experiment systematically augments counts for preferentially binding protein/tissue pairs. The inference goal is to identify those tripeptide/tissue pairs for

***Corresponding author: e-mail: leonnovelo@gmail.com

which the mean counts, under some suitable probability model, are monotone increasing across the three stages of the experiment.

Let $\alpha_i \in \{0,1\}$, $i = 1,\ldots,n$, denote an indicator for truly increasing mean counts ($\alpha_i = 1$) or not ($\alpha_i = 0$) for the i-th tripeptide/tissue pair. The problem becomes one of deciding about a large number, in our case $n = 257$, of comparisons $\alpha_i = 0$ versus $\alpha_i = 1$. Let $d_i \in \{0,1\}$ denote an indicator for reporting the i-th pair as preferentially binding. Letting $\mathbf{y}$ generically denote the observed data, the decisions are functions $d_i(\mathbf{y})$. The number of falsely reported pairs relative to the number of reported pairs is known as the false discovery proportion,

$$\text{FDP} = \frac{\sum_{i=1}^{n}(1-\alpha_i)d_i}{D+\epsilon}.$$

Here $D = \sum d_i$ is the number of reported decisions, and $\epsilon > 0$ is added to avoid zero division. In our implementation we use $\epsilon = 0.1$. Alternatively one could use $\epsilon = 0$ and define FDP $= 0$ when $D = 0$. At this moment, FDP is neither frequentist nor Bayesian. It is a summary of both, the data, implicitly through $d_i(\mathbf{y})$, and the unknown parameters α_i. Under a Bayesian perspective one would now condition on $\mathbf{y}$ and marginalize with respect to the unknown parameters to define the posterior expected false discovery rate. We run into some good luck when taking the posterior expectation of FDP. The only unknown quantities appear in the numerator, leaving only a trivial expectation of a sum of binary random variables. Let $\overline{a}_i = E(\alpha_i \mid \mathbf{y}) = p(\alpha_i = 1 \mid \mathbf{y})$ denote the posterior probability for the i-th comparison. Then

$$\overline{\text{FDR}} = E(\text{FDP} \mid \mathbf{y}) = \frac{\sum_{i=1}^{n}(1-\overline{a}_i)d_i}{D+\epsilon}.$$

The posterior probabilities $\overline{a}_i$ automatically adjust for multiplicities, in the sense that posterior probabilities are increased (or decreased) when the many (or few) other comparisons seem to be significant. See, for example, Scott and Berger (2006) and Scott and Berger (2010) for a discussion of how $\overline{a}_i$ reflects a multiplicity adjustment. In short, if the probability model includes a hierarchical prior with a parameter that can be interpreted as overall probability of a positive comparison, $\alpha_i = 1$, i.e., as the overall level of noise in the multiple comparison, then posterior inference can learn and adjust for multiplicities by adjusting inference for that parameter. However, Berry and Berry (2004) argue that adjustment of the probabilities alone is only solving half of the problem. The posterior probabilities alone do not yet tell the investigator which comparisons should be reported, in the case of our case study, these are the decisions d_i, $i = 1,\ldots,n$. It is reasonable to use rules that select all comparisons with posterior probability beyond a certain threshold, i.e.,

$$d_i^{\star} = I(\overline{a}_i > t), \tag{1}$$

(Newton; 2004). The threshold can be chosen to control $\overline{\text{FDR}}$ at some desired level. This defines a straightforward Bayesian counterpart to frequentist control of FDR as it is achieved in rules proposed by Benjamini and Hochberg (1995) and others. The Bayesian equivalent to FDR control is the control of posterior expected FDR. See Bogdan et al. (2008) for a recent comparative discussion of Bayesian approaches versus the Benjamini and Hochberg rule.

Alternatives to FDR control have been proposed, for example, in Storey (2007) who introduces the optimal discovery procedure (ODP) that maximizes the number of true positives among all possible tests with the same or smaller

number of false positive results. An interpretation of the ODP as an approximate Bayes rule is discussed in Guindani et al. (2009), Cao et al. (2009) and Shahbaba and Johnson (2011).

In this article we focus on FDR control and apply the rule $d_i^\star$ in a particular case study. The application is chosen to highlight the features and limitations of these rules. In León-Novelo et al. (2012) we report inference for a similar biopanning experiment with much larger human data. The larger sample size makes it possible to consider non-parametric Bayesian extensions.

In section 2 we introduce the case study and the data format. In section 3 we discuss the decision rule. This can be done without reference to the particular probability model. Only after the discussion of the decision rule, in section 4, will we briefly introduce a probability model. In section 5 we validate the proposed inference by carrying out a small simulation study. Section 6 reports inference for the original data. Finally, section 7 concludes with a final discussion.

2 Data

A phage library is a collection of millions of phages, each displaying different peptide sequences. Bacteriophages, for short phages, are viruses. They provide a convenient mechanism to study the preferential binding of peptides to tissues, essentially because it is possible to experimentally manipulate the phages to display various peptides on the surface of the viral particle. In a bio-panning experiment (Ehrlich et al.; 2000) the phage display library is exposed to a target, in our case, injected in a (single) mouse. Later, tissue biopsies are obtained to recover phage from different tissues. Phages with proteins that do not bind to the target tissue are washed away, leaving only those with proteins that are binding specifically to the target. A critical limitation of the described experiment is the lack of any amplification. Some peptides might only be reported with a very small count, making it very difficult to detect any preferential binding. To mitigate this limitation Kolonin et al. (2006) proposed to perform multistage phage display experiments, that is, to perform successive stages of panning (usually three or four) to enrich peptides that bind to the targets. Figure 1 illustrates the design. This procedure allows for the counts of peptides with low initial count to increase in every stage and, therefore, it increases the chance of detecting their binding behavior.

We analyze data from such a bio-panning experiment carried out at M. D. Anderson Cancer Center. The data come from three consecutive mice. At each stage a phage display peptide library was injected into a new animal, and 15 minutes later biopsies were collected from each of the target tissues and the peptide counts were recorded. For the second and third stage the injected phage display peptide library was the already enriched phage display library from the previous stage. The data reports counts for 4200 tripeptides and 6 tissues over 3 consecutive stages. For the analysis we excluded tripeptide-tissue pairs for which the sum of their counts over the three stages was below 5, leaving $n = 257$ distinct pairs. Figure 3 shows the data for these tripeptides/tissue pairs. The desired inference is to identify tripeptide-tissue pairs with an increasing pattern across the three stages, i.e., to mark lines in the figure that show a clear increasing trend from first to third stage. Some lines can be clearly classified as increasing, without reference to any probability model. But for many lines the classification is not obvious. And importantly, some of the seemingly obviously increasing counts might be simply due to chance. Even if none of the peptides were truly preferentially binding to any tissue, among the large number of observed counts some would show an increase, just by

random variation. The purpose of the proposed model-based approach is to define where to draw the line to define a significant increase, and to adjust for the multiplicities.

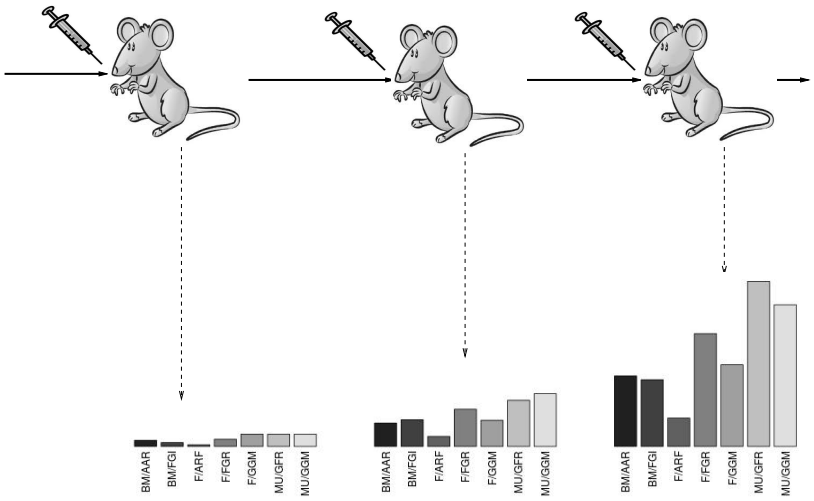

Figure 1 Multi-stage phage display experiment with 3 stages. At each stage a phage display library is injected in a new animal, using for stages 2 and 3 the already enriched phage display library from the previous experiment.

3 The Decision Problem

The proposed approach to select peptide/tissue pairs for reporting is independent of the underlying probability model. It is based on a formalization of the inference problem as a decision problem with a specific utility function. The particular probability model only changes the distribution with respect to which we compute posterior expected utilities.

The only assumptions that we need in the upcoming discussion are that the model includes parameters $\alpha_i \in \{0, 1\}$ that can be interpreted as indicators for increasing mean counts of peptide/tissue pair i across the three stages. Recall that $\overline{a}_i = p(\alpha_i = 1 \mid \mathbf{y})$ denotes the posterior probabilities. We also assume that the model includes parameters $\mu_i \in \Re$ that can be interpreted as the extent of the increase, with $\alpha_i = I(\mu_i > 0)$. We use $\overline{m}_i = E(\mu_i \mid \mathbf{y})$ for the marginal posterior means.

We already introduced $d^\star$ in (1) as a reasonable decision rule to select peptide/tissue pairs for reporting as preferentially binding. Rule $d^\star$ can be justified as control of the false discovery rate (FDR) (Newton, 2004) or, alternatively, as an optimal Bayes rule. To define an optimal rule we need to augment the probability model to a decision problem by introducing a utility function. Let θ and y generically denote all unknown parameters and all observable data. A utility function $u(d, \theta, y)$ formalizes relative preferences for decision d under hypothetical outcomes y and under an assumed truth θ. For example, in our application a utility function could be

$$u(d, \theta, y) = \sum_i \mathrm{d}_i \alpha_i + k \sum_i (1 - \mathrm{d}_i)(1 - \alpha_i) \tag{2}$$

i.e., a linear combination of the number of true positive selections d_i and true negatives. For a given probability model, data and utility function, the optimal Bayes rule is defined as the rule that maximizes u in expectation over all not observed variable, and conditional on all observed variables,

$$d_{opt} = \arg\max_d E(u(d, \theta, y) \mid y). \tag{3}$$

It can be shown that the rule $d^\star$ in (1) arises as Bayes rule under several utility functions that trade off false positive and false negative counts, including the utility in (2) and others. See, for example, Müller et al. (2007), for a discussion.

Alternatively, $d^\star$ can be derived as FDR control. Recall the posterior expected FDR,

$$\overline{\text{FDR}} = E(\text{FDR} \mid \mathbf{y}) = \frac{1}{D+\epsilon} \sum_{i=1}^{n} d_i(1-\overline{a}_i). \tag{4}$$

Similarly, the posterior expected false negative rate (FNR) can be computed as $\overline{\text{FNR}} = E(\text{FNR} \mid \mathbf{y}) = \sum(1 - d_i)\overline{a}_i/(n - D + \epsilon)$. It is easily seen that the pairs selected by $d^\star$ report the largest list for a given value of posterior expected FDR.

Characterizing $d^\star$ as the Bayes rule (3) under (2) highlights a critical limitation of the rule. The utility function (2) weights every true positive, or equivalently, every false negative, equally. Recall that we assume that the model includes a parameter μ_i that can be interpreted as the strength of a true comparison, i.e., in our application, as the level of preferential binding of the i-th peptide/tissue pair. A true positive with small μ_i that is unlikely to lead to any meaningful follow-up experiments is of far less interest to the investigator than a true positive with massively large μ_i. Equivalently, a false negative, i.e., missing to report a truly preferentially binding tripeptide/tissue pair, is less critical when the non-zero μ_i is small than when we miss to report a potentially interesting tripeptide/tissue pair with large μ_i. These considerations lead us to consider a utility function that weights each pair proportional to $\mu_i\alpha_i$. We use the utility function

$$U(d, \theta) = \sum_{i=1}^{n} \mathrm{d}_i\alpha_i\mu_i - k\sum_{i=1}^{n}(1-\mathrm{d}_i)\alpha_i\mu_i - cD. \tag{5}$$

True positives and false negatives are weighted with a positive level of monotonicity. The last term puts a cost c on each reported positive. Without that cost the trivial solution would be $d_i = 1$, for $i = 1, \ldots, n$. Alternatively, the last term can be interpreted as adding a cost for false positives. To see this, write cD as $cD = c\sum d_i\alpha_i + c\sum d_i(1-\alpha_i)$, and include the first term into the first component of (5). This clarifies the role of the term $-cD$. Without a cost for false positives one would set $d_i = 1$ for all comparisons.

Let $\overline{m}_i = E(\mu_i\alpha_i \mid \mathbf{y})$. Straightforward algebra shows that the optimal rule is

$$\mathrm{d}_i{}^B = I(\overline{m}_i \geq c/(k+1)). \tag{6}$$

4 Model

We use $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3})$ to denote the observed counts for tripeptide/tissue pair i across the three stages, for pairs $i = 1, \ldots, n$. Ji et al. (2007) used a model with a Poisson sampling model for y_{ij}, together with a mixture of normal

prior for the parameters. They assumed that the Poisson rates were increasing linear across stages j. For example, consider the pairs with oscillating increase and decrease across the three stages in Figure 2. Although the data for these pairs shows a marked difference in slopes from stages 1 to 2 versus from stages 2 to 3, the parametric model forces one common slope. The selection of the reported tripeptide/tissue pairs in Ji et al. (2007) was based on the posterior posterior probability of that slope being positive. This is a concern when the imputed overall slope is positive like, for example, in the pair marked by A in Figure 3. Outliers like pair A in Figure 3 can inappropriately drive the inference.

We use instead a model with different Poisson rates for all three stages. In anticipation of the inference goal we parameterize the mean counts as $(\gamma_i, \gamma_i\beta_i, \gamma_i\mu_i)$, allowing us to describe increasing mean counts by the simple event $1 < \beta_i < \mu_i$. We write $\text{Poi}(x \mid m)$ to indicate a Poisson distributed random variable x with mean m.

$$p(y_{i1}, y_{i2}, y_{i3} \mid \gamma_i, \beta_i, \mu_i) = \text{Poi}(y_{i1} \mid \gamma_i)\,\text{Poi}(y_{i2} \mid \gamma_i\beta_i)\,\text{Poi}(y_{i3} \mid \gamma_i\mu_i) \tag{7}$$

for $i = 1, \ldots n$. The parameter γ_i can be thought of as the expected mean count of the pair i across the three stages if we were not enriching the tripeptide library at every stage.

We assume gamma random effects distributions for $(\gamma_i, \beta_i, \mu_i)$. Let $\text{Ga}(a, b)$ indicate a gamma distribution with parameters a and b with mean a/b. We assume

$$\gamma_i \sim \text{Ga}(s_\gamma, s_\gamma \cdot t_\gamma),\ \beta_i \sim \text{Ga}(s_\beta, s_\beta t_\beta),\ \mu_i \sim \text{Ga}(s_\mu, s_\mu t_\mu), \tag{8}$$

independently across γ_i, β_i, μ_i, and across $i = 1, \ldots, n$. The model is completed with a prior on the hyperparameters

$$t_\gamma \sim \text{Ga}(t_\gamma | a_{t_\gamma}, b_{t_\gamma}),\ t_\beta \sim \text{Ga}(t_\beta | a_{t_\beta}, b_{t_\beta}),\ t_\mu \sim \text{Ga}(t_\mu | a_{t_\mu}, b_{t_\mu}). \tag{9}$$

Equations (7) through (9) define a sampling model and prior for a multistage phage display experiment. The particular experiment that we analyze in this paper uses three animals for a single replicate of a multistage experiment with three stages, corresponding to mean counts γ_i, $\beta_i\gamma_i$ and $\mu_i\gamma_i$. If desired the model can easily be modified for more stages or for repeat experiments. If multiple, say K, repeat experiments of the three-stage phage display were available, we extend the model by introducing an additional layer in the hierarchy. Let y_{ijk} denote the count for tripeptide/tissue pair i in stage j of the k-th repeat experiment. We replace (7) by

$$p(y_{i1k}, y_{i2k}, y_{i3k} \mid \gamma_{ik}, \beta_i, \mu_i) = \text{Poi}(y_{i1k} \mid \gamma_{ik})\,\text{Poi}(y_{i2k} \mid \gamma_{ik}\beta_i)\,\text{Poi}(y_{i3k} \mid \gamma_{ik}\mu_i) \tag{10}$$

with $\gamma_{ik} \sim \text{Ga}(s_\gamma, s_\gamma \cdot t_\gamma)$ and unchanged priors on (β_i, μ_i).

The conjugate nature of the Poisson sampling model and the gamma random effects distribution and hyperprior simplify posterior inference. All parameters and random effects, including γ_i (or γ_{ik} in model (10)), β_i, μ_i, t_γ, t_β and t_μ have closed form conditional posterior distributions conditional on currently imputed values for all other parameters and latent variables. We implement straightforward Gibbs sampling Markov chain Monte Carlo simulation. Let θ denote the unknown parameters in the sampling model for the observed counts $\mathbf{y}$, and let θ^k denote the imputed parameters after k iterations of a Markov chain Monte Carlo posterior simulation. Recall that α_i was defined as indicator for increasing mean counts, i.e., $\alpha_i = I(\mu_i > \beta_i > 1)$ under model (7). With a Monte Carlo posterior

sample $(\theta_i^1, \ldots, \theta_i^M)$ of size M we estimate $\overline{m}_i$ as

$$\overline{m}_i \approx \frac{1}{M} \sum_{k=1}^{M} \alpha_i^k \mu_i^k \tag{11}$$

and then make the decision $\mathrm{d}_i{}^B$. Posterior probabilities $\overline{a}_i$ are similarly computed as ergodic averages $\frac{1}{M}\sum \alpha_i^k$.

5 A Simulation Study

We carried out a simulation study to validate the proposed approach. We generated $n = 250$ observations of the model described in (7) through (9) with the following parameters

$$\gamma_k \sim \text{Ga}(\gamma \mid s_\gamma^f, s_\gamma^f t_\gamma^f),\ \beta_k \sim \text{Ga}(\beta \mid s_\beta^f, s_\beta^f t_\beta^f) \text{ and } \mu_k \sim \text{Ga}(\mu \mid s_\mu^f, s_\mu^f t_\mu^f), \tag{12}$$

independently. We set the hyper-parameters such that the expected value of γ_i and its variance are small and, besides, β_i and μ_i have both mean 8 and variances 30 and 120 respectively. The motivation for this choice is that γ_i is interpreted as the mean of the counts through the three stages of the pair i if there were no enrichment. Since, initially, the library contains a small amount of the particular tripeptide related with the pair i among the large number of different tripeptides, we expect γ_i to be small. The parameters β_i and μ_i represent the fold increase in mean counts from the first stage to the second and third stages, respectively, due to the library enrichment. We allow these last parameters to have large variances. The Gamma parameters were set to $s_\gamma^f = 3.6, t_\gamma^f = 5/6, s_\beta^f = 13/6, t_\beta^f = 1/8, s_\mu^f = 0.53$ and $t_\mu^f = 0.125$. For 94 out of the $n = 250$ simulated tripeptide/tissue pairs the simulation truth included an increase $\mu_i > \beta_i > 1$, i.e., $\alpha_i = 1$.

The hyper-parameters of the model described in (7) through (9) were chosen taking into account the same considerations and set to $s_\gamma = 0.05$, $a_{t_\gamma} = 3$, $b_{t_\gamma} = 1/2$, $s_\beta = s_\mu = 5/3$, $a_{t_\beta} = a_{t_\mu} = 6$, $b_{t_\beta} = b_{t_\mu} = 25$. Saving every 10^{th} iteration after a 10,000 iteration burn-in, a Monte Carlo posterior sample of size $M = 10,000$ was saved.

Using the criterion (1), we selected the pairs such that, under the assumptions of our model, the expected false discovery rate was $\overline{\text{FDR}} = 0.20$. The implied expected false negative rate was $\overline{\text{FNR}} = 0.17$. Under this rule, we reported 57 pairs for increasing means across the three stages. Of these pairs, 52 truly did in the simulation truth, i.e., 52 were true positives. The observed FDR and FNR were 0.09 and 0.22, respectively.

Alternative to FDR control we considered selection with respect to the utility function (5), with $c/(k+1) = 7.1$. We declare 57 pairs for increasing means. Of these, 53 actually exhibit this pattern in the simulation truth, i.e., 53 of the $D = 57$ reported pairs were true positives. The posterior expected FDR and FNR are $\overline{\text{FDR}} = 0.21$ and $\overline{\text{FNR}} = 0.17$ respectively. Using the known simulation truth α_i we evaluate the observed FDR and FNR as 0.07 and 0.21, respectively. The number of pairs reported by both methods was 53.

To further explore inference under the model we considered a variation of the simulation study with more informative data. We assume $K = 3$ repetitions of the same experiment are available. We simulated data for the additional repeat experiments by generating $\gamma_{ik} \sim \text{Ga}(s_\gamma^f, s_\gamma^f t_\gamma^f)$, β_i, i.i.d., and μ_i as in (12) and y_{ijk} according to the sampling model in (10). For 99 out of n=250 simulated tripeptide pairs $\alpha_i = 1$. As before a Monte Carlo posterior sample of size $M = 10,000$ was saved. We implement inference under the extended model (10) with the same hyperparameter

values as before. Aiming for a posterior expected FDR equal to 0.10, 90 pairs were declared as having increasing means according to the rule (1). Under this rule, of these 90 pairs, 82 truly did in the simulation truth, i.e., 82 were true positives. The observed FDR and FNR were 0.10 and 0.07, respectively. As expected the additional data of the repeat experiment allows to report more pairs at lower FDR.

Alternatively we considered selection with respect to the utility function (5), with $c/(k+1) = 3.9$. This value is chosen for comparison, in order to declare 90 pairs for increasing means. Of these, 84 actually exhibit this pattern in the simulation truth, i.e., 84 of the $D = 90$ reported pairs were true positives. The posterior expected FDR and FNR are $\overline{\text{FDR}} = 0.11$ and $\overline{\text{FNR}} = 0.07$ respectively. Using the known simulation truth α_i we evaluate the observed FDR and FNR as 0.09 and 0.07, respectively. The number of pairs reported by both methods was 86.

In summary, the simulation indicates that for a data set of a sample size comparable to the earlier described phage display data the proposed inference approach is appropriate to detect pairs with increasing mean counts. If there were truly increasing trends for some peptide/tissue pairs, the proposed inference would likely include many in the report. For increases of the size assumed in this simulation most interesting pairs are reported under a modestly stringent control on FDR.

6 Results

6.1 Selecting Tripeptide/Tissue Pairs

In this section we present analysis and results by applying the proposed method to the phage display data described in section 2. The parameter values in our proposed priors are elicited by consulting with the investigators. We fix the hyper-parameters as in the simulation study. In particular, the hyperprior choices imply the following marginal means and variances. The parameter γ_i is interpreted as the expected count if there were no enrichment of the library of tripeptides at every stage. We assume that most of the phage counts are small in the initial state. Therefore, we set the expected value for the first stage counts γ_i to 0.25 and its variance to 2.56. We do not assume any knowledge of the mean increment between the first and the second stage (i.e., β_i) and between the first and the third stage (i.e., μ_i). We center these values around a mean of 5 and allow for a large variance equal to 25. This corresponds to the same hyper-parameters as in the simulation study of Section 5.

We obtained a Monte Carlo posterior sample of size $M = 10,000$, saving the values of the imputed parameters every ten iterations after an initial burn-in of $10,000$ iterations. Similar to the simulation study, we evaluate convergence diagnostics to ensure practical convergence of the MCMC simulation. We found that the Markov chains mixed very well and found no evidence for lack of convergence.

Figure 2 reports the 25 pairs with highest values of $\overline{m}_i$, i.e the pairs chosen according to the optimal rule (6) with a threshold value of $c/(k+1) = 0.9$. We notice that some of the selected pairs have relatively small posterior probability of increasing means, $\overline{a}_i$, for example, $\overline{a}_i = 0.3$ for muscle/AGG. This is also reflected in the high $\overline{\text{FDR}} = 65\%$ and $\overline{\text{FNR}} = 14\%$.

Figure 3 highlights the 25 selected pairs. The utility function selects pairs with large increments across all three stages. Additionally, the criterion selects some pairs with not strictly increasing counts, but with a substantial incre-

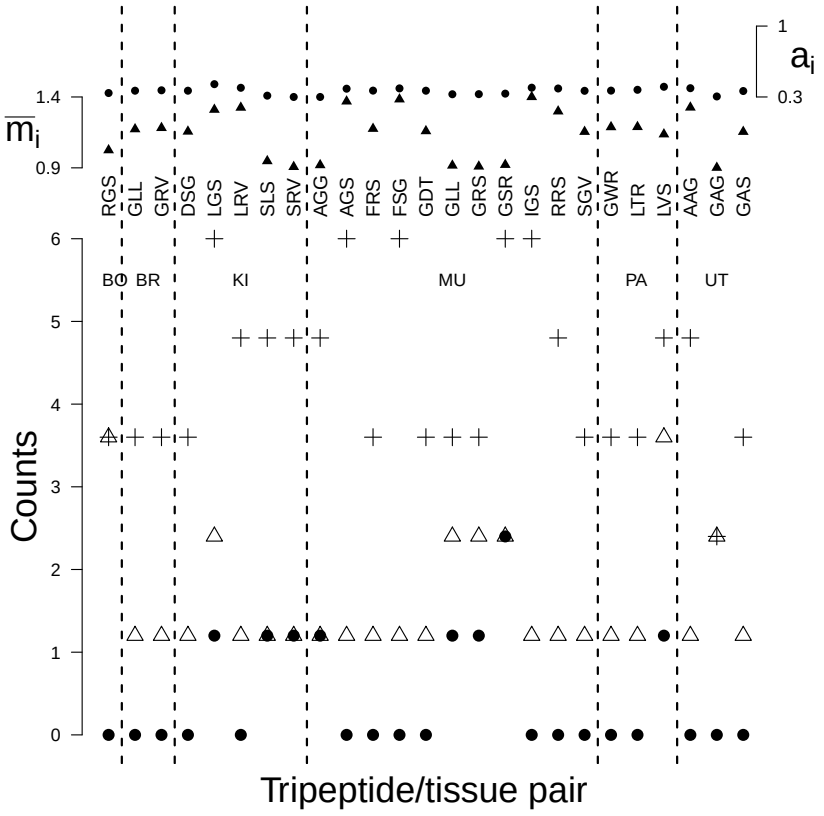

Figure 2 Twenty five tripeptide/tissue pairs with highest estimated $\overline{m}_i$ in (11). Each column represents a tripeptide/tissue pair. The tripeptides in different tissues are separated by dashed vertical lines. The lower section of the plot shows the counts for the three stages: circle, triangle and cross. The middle section shows $\overline{m}_i$ (triangles) with corresponding vertical axis scaled between 0.9 and 1.4. . The upper part plots the posterior probability of increasing means across the three stages, $\overline{a}_i$ (bullets) with corresponding vertical axis scaled 0.3 to 1. The 3-letter codes refer to distinct tripeptides, the two letter codes refer to different tissue types: bowel (BO), brain (BR), kidney (KI), muscle (MU), pancreas (PA) and uterus (UT).

ment between some of the stages. This is in agreement with the underlying utility function (5). The selection also reports some pairs that have small counts over the three stages, but include a large increment in some stage in comparison to the previous count. On the other hand, there are some other pairs that are not selected despite relatively large and nondecreasing counts over all three stages. The model might be detecting that this event can happen because of a high base-line count, and does not necessarily imply a strong binding behavior of the tripeptide to the respective tissue.

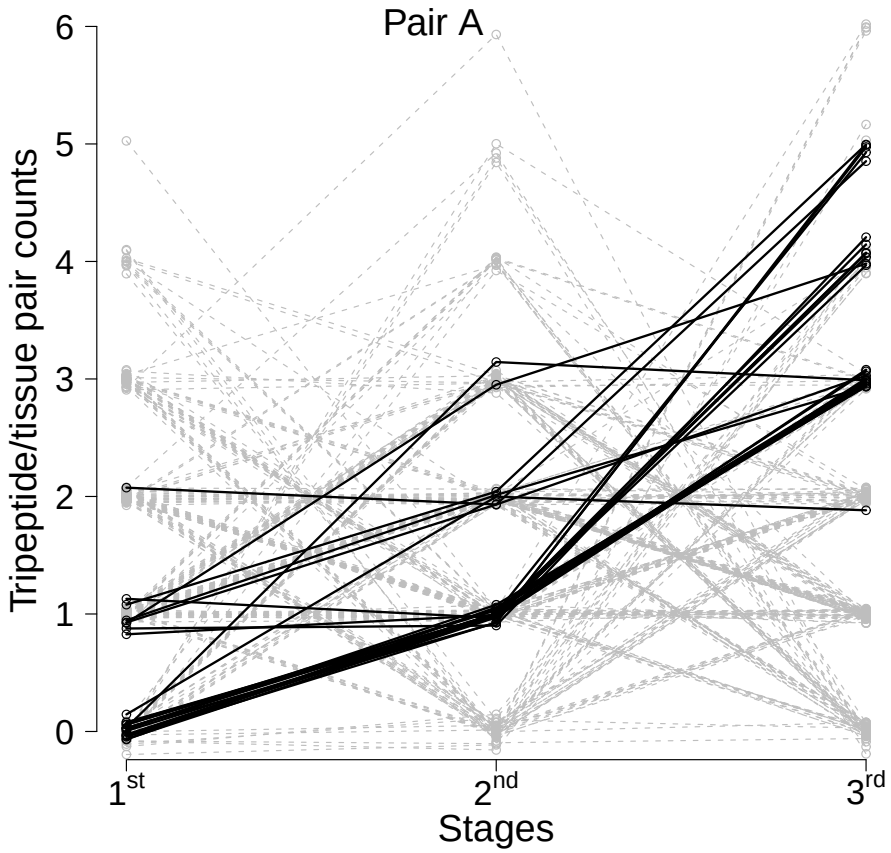

Figure 3 Observed sequence of tripeptide/tissue pair counts across the three stages. Each line connects the three observed counts for one tripeptide/tissue pair. The 25 selected pairs under rule (6) are marked in black (solid line). The non-selected pairs are shaded (dotted line). The threshold for (6) is 0.9. Counts are jittered for better display.

6.2 Multiplicity Adjustment

There is an apparent discrepancy between the posterior probabilities $\overline{a}_i$ reported in Figure 2 and the seemingly obviously increasing counts for the same pairs in Figure 3 (marked as thick lines). In Figure 2, the bullets in the upper part of the figure report the posterior probabilities $\overline{a}_i = p(\alpha_i = 1 \mid \mathbf{y})$ for the selected pairs, with values ranging between 0.3 and 0.4. Comparing with the observed counts in Figure 3 these posterior probabilities seem low. The counts for the selected pairs seem obviously increasing. Figure 4a explains the apparent discrepancy between the two plots. In short, the low posterior probabilities are reasonable because of multiplicity control and high noise. For a quick plausibility argument, focus on the pairs with decreasing counts in Figure 3. If we were to highlight the most strikingly decreasing

trajectories, the selection might look almost as convincing as the currently highlighted increasing counts. However, there is no good biologic reason for decreasing counts. The decreasing trajectories are only due to noise. Honest inference for the increasing trajectories has to adjust for this selection effect and the reported probabilities appear a reasonable summary of the data. Figure 4 shows more details. The plot shows for the top five chosen tripeptide/tissue pairs the observed counts (piecewise linear curves), posterior means (bullets) and 95% credible intervals (vertical line segments) for the Poisson means $\gamma_i, \gamma_i\beta_i, \gamma_i\mu_i$. Note the large posterior uncertainties, due to the small observed counts (ranging from 0 to 4 only). More importantly, note how the posterior means shrink the counts towards an overall mean. This is the posterior adjustment for multiplicities. The displayed pairs are the five pairs with some of the most extreme observed increments across the three stages. The posterior shrinkage reflects an adjustment for the selection bias. We investigated possible sensitivity with respect to the chosen prior model, fearing that the gamma random effects distributions (8) might lead to excessive shrinkage. We considered a model with a non-parametric Dirichlet process prior instead of (8) (results not shown). Posterior probabilities increase slightly, to around 0.45 for the chosen pairs. But substantial shrinkage remains.

Figures 4bcd further elucidate the posterior adjustment for multiplicities. Recall that $E(\beta_i \mid t_\beta) = 1/t_\beta$ is the mean increment in stage 2, and similarly $1/t_\mu$ is the mean increment in stage 3, and $1/t_\gamma$ is the mean baseline count. The figures compare the prior (dashed curves) and marginal posterior distributions (histograms) for the mean baseline count $1/t_\gamma$ and mean increments $1/t_\beta, 1/t_\mu$. The prior was chosen to allow substantial increases. But *a posteriori* the size of the increases is substantially smaller, with the posterior mean $E(1/t_\beta \mid \mathbf{y})$ (increase in stage 2) even slightly higher than $E(1/t_\mu \mid \mathbf{y})$ (increase in stage 3).

Finally, Figure 5 shows posterior estimated $E(\beta_i, \mu_i \mid \mathbf{y})$ for all tripeptide/tissue pairs. We notice clusters of points in this figure. These are pairs with exactly matching triples of counts (y_{i1}, y_{i2}, y_{i3}). For example there are nine pairs with counts $(0, 1, 3)$, and all these pairs are selected.

In summary, the experiment is simply not as informative as it might seem at first glance. It is still useful as a screening experiment to identify possibly interesting tripeptide/tissue pairs that might warrant further investigation. There is a good suggestion of a possible effect for the reported pairs.

7 Discussion

We have shown posterior inference in an application that requires decisions in the face of massive multiplicities. Posterior inference improves in important ways over naive exploratory data analysis of the data. First, posterior inference helps the investigator to decide where to draw the line in selecting pairs with increasing counts. Second, considering the selection as a formal decision problem we recognized that the selection on the basis of FDR only might be inappropriate and were lead to replace statistical significance by a criterion that is closer to biologic significance. Third, posterior probabilities adjust for the massive multiplicity problem by reporting honest posterior probabilities of true positives, i.e., posterior probabilities of the reported pairs being in fact preferentially binding. The adjusted posterior probabilities are far lower than what one might estimate from a first inspection of the data.

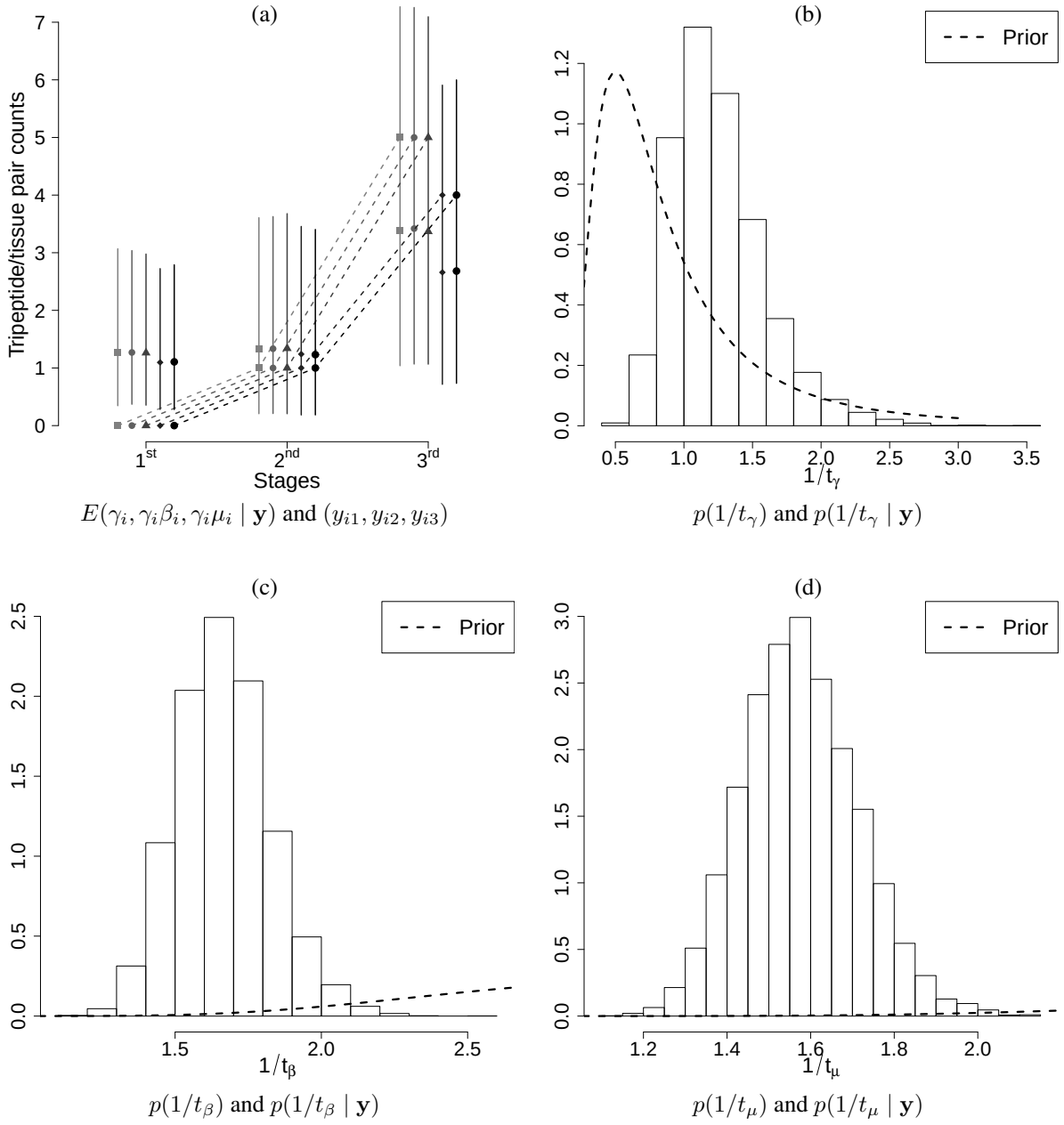

Figure 4 Panel (a) compares posterior means (bullets) and 95% credible intervals (vertical line segments) for the mean counts $(\gamma_i, \gamma_i\beta_i, \gamma_i\mu_i)$ with the observed counts (y_{i1}, y_{i2}, y_{i3}) (piecewise linear curves) for the top five tripeptide/tissue pairs. Panels (b) through (d) compare the prior (dashed lines) and posterior distributions (histograms) for the mean baseline count $1/t_\gamma$ (b) and the increments $1/t_\beta$ (c) and $1/t_\mu$ (d). See the text for a comparison of prior and posterior distributions on t_β and t_μ.

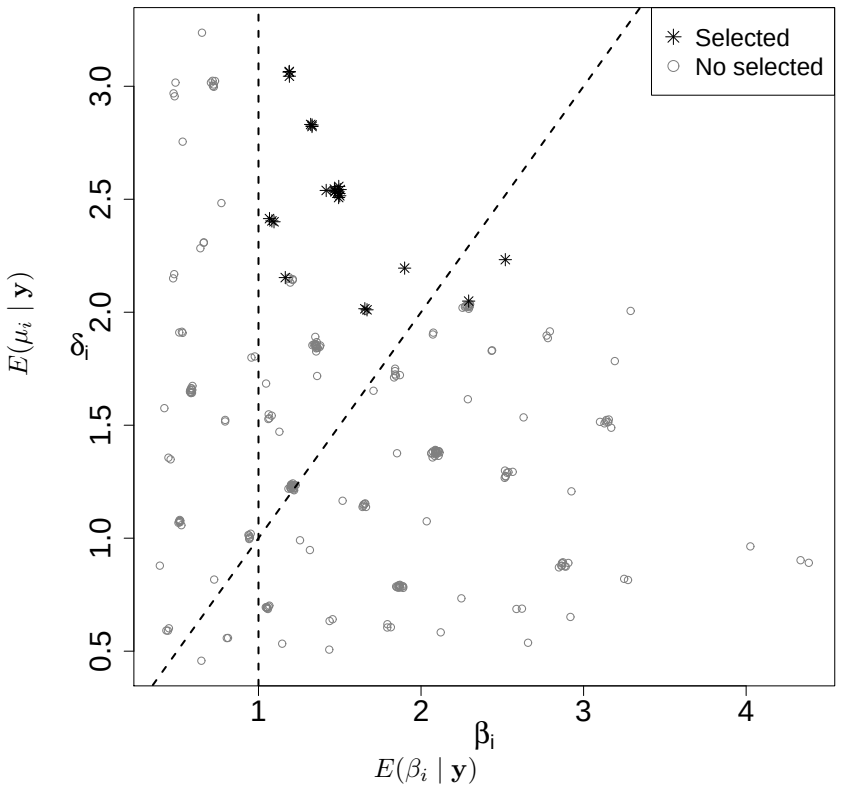

Figure 5 Scatterplot of the posterior means $E(\beta_i, \mu_i \mid \mathbf{y})$. The 25 selected pairs are represented by "$*$". For reference the figure includes a vertical line through 1 and the 45-degree line (dashed lines). The pairs with increasing posterior mean counts fall in the upper right region defined by these lines.

Among the limitations of the proposed approach is the simple structure of the underlying probability model. For a larger data set one could consider semi-parametric extensions to replace the parametric random effects model with a random probability measure G with a nonparametric Bayesian prior on G. Also, the current model entirely ignores dependence structure that might be induced by tissue-specific or protein-specific binding behavior. Increments for tripeptide/tissue pairs that involve the same tissue or protein might be more reasonably represented as a priori correlated.

Acknowledgements The second author was partially supported by grant NIH/R01 CA075981. The last author was supported by the Cancer Center Support Grant (CCSG) (NIH/P30 CA016672) and the M D Anderson Cancer Center Prostate SPORE (NIH/P50 CA140388 02). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

References

Arap, W., Kolonin, M. G., Trepel, M., Lahdenranta, J., Card-Vila, M., Giordano, R. J., Mintz, P. J., Ardelt, P. U., Yao, V. J., Vidal, C. I., Chen, L., Flamm, A., Valtanen, H., Weavind, L. M., Hicks, M. E., Pollock, R. E., Botz, G. H., Bucana, C. D. and MORE (2002). Steps toward mapping the human vasculature by phage display, *Nature Medicine* **8**(2): 121 – 127.

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. Series B. Methodological* **57**(1): 289–300.

Berry, S. and Berry, D. (2004). Accounting for multiplicities in assessing drug safety, *Biometrics* **60**: 418–426.

Bogdan, M., Ghosh, J. K. and Tokdar, S. T. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing, *in* N. Balakrishnan, E. A. P. na and M. J. Silvapulle (eds), *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, Institute of Mathematical Statistics, Beachwood, Ohio, USA, pp. 211–230.

Cao, J., Jie, X., Zhang, S., Whitehurst, A. and White, M. (2009). Bayesian optimal discovery procedure for simultaneous significance testing, *BMC, Bioinformatics* **10**(5).

Ehrlich, G., Berthold, W. and Bailon, P. (2000). Phage display technology. affinity selection by biopanning, *Methods in molecular biology* **147**: 195–208.

Guindani, M., Müller, P. and Zhang, S. (2009). A Bayesian discovery procedure, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(5): 905–925.

Ji, Y., Yin, G., Tsui, K.-W., Kolonin, M. G., Sun, J., Arap, W., Renata, P. and Do, K.-A. (2007). Bayesian mixture models for complex high dimensional count data in phage display experiments, *Journal of the Royal Statistical Society, Series C: Applied Statistics* **56**(2): 139–152.

Kolonin, M. G., Sun, J., Do, K.-A., Vidal, C. I., Ji, Y., Baggerly, K. A., Pasqualini, R. and Arap, W. (2006). Synchronous selection of homing peptides for multiple tissues by in vivo phage display, *FASEB J.* **20**: 979–981.

León-Novelo, L. G., Müller, P., Do, K.-A., Arap, W., Sun, J. and Pasqualini, R. (2012). Semi-parametric Bayesian inference for phage display data, *Biometrics* . To appear.

Müller, P., Parmigiani, G. and Rice, K. (2007). FDR and Bayesian multiple comparisons rules, *in* Bayesian Statistics 8 (ed.), *José Miguel Bernardo and James O. Berger and A. Phillip Dawid and Adrian F. M. Smith*, Oxford University Press, Oxford, pp. 349–370.

Newton, M. A. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method, *Biostatistics (Oxford)* **5**(2): 155–176.

Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing, *Journal of Statistical Planning and Inference* **136**(7): 2144 – 2162.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem, *Annals of Statistics* **38**(5): 2587–2619.

Shahbaba, B. and Johnson, W. (2011). A nonparametric approach for relevance determination, *Technical report*, University of California Irvine.

Storey, J. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(3): 347–368.