Published in final edited form as:

Biometrics. 2013 March ; 69(1): 174–183. doi:10.1111/j.1541-0420.2012.01817.x.

Semi-Parametric Bayesian Inference for Phage Display Data

Luis G. León-Novelo,

Department of Mathematics, University of Louisiana at Lafayette

Peter Müller,

Department of Mathematics, The University of Texas at Austin

Wadih Arap,

Departments of Genitourinary Medical Oncology and Cancer Biology, University of Texas M. D. Anderson Cancer Center

Mikhail Kolonin,

The Brown Foundation Institute of Molecular Medicine for the Prevention of Human Disease, University of Texas. Health and Science Center at Houston

Jessica Sun,

Departments of Genitourinary Medical Oncology and Cancer Biology, University of Texas M. D. Anderson Cancer Center

Renata Pasqualini, and

Department of Genitourinary Medical Oncology - Research, Division of Cancer Medicine, University of Texas M. D. Anderson Cancer Center

Kim-Anh Do

Department of Biostatistics, University of Texas, M.D. Anderson Cancer Center

Luis G. León-Novelo: leonnovelo@gmail.com; Peter Müller: pmueller@math.utexas.edu; Wadih Arap: warap@mdanderson.org; Mikhail Kolonin: Mikhail.G.Kolonin@uth.tmc.edu; Jessica Sun: jesun@mdanderson.org; Renata Pasqualini: @mdanderson.org; Kim-Anh Do: kimdo@mdanderson.org

Summary

We discuss inference for a human 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. We formalize the research question as inference about the monotonicity of mean counts over stages. The inference goal is then to identify a list of peptide-tissue pairs with significant increase over stages. We use a semi-parametric Dirichlet process mixture of Poisson model. The posterior distribution under this model allows the desired inference about the monotonicity of mean counts. However, the desired inference summary as a list of peptide-tissue pairs with significant increase 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.

Supplementary Materials

Table and Figures referenced in Sections 3, 5 and 6 are available with this paper at the Biometrics website on Wiley Online Library. The R and C code, and necessary (simulated and real) data to reproduce the analysis in this paper are located in the folder “analysis.zip”. This folder is also available at the same website.

Keywords

biopanning; decision problem; Dirichlet process mixture; multiple comparison; phage display experiment

1. Introduction

We develop semi-parametric Bayesian inference for data from a human phage display experiment. The experiment is carried out to learn about preferential binding of proteins in certain tissues. The long-term goal is to exploit such knowledge to develop targeted therapies that could deliver a drug to specific tissues and limit side effects such as toxicity (Kolonin *et al.*, 2006; Arap *et al.*, 2002). For example, our previous collaborative work identified the interleukin-11 receptor alpha as a candidate target in the progression of human prostate cancer (Arap *et al.*, 2002 and Zurita AJ, *et al.*, 2004) and served as the basis for an ongoing first-in-man trial. It was identified starting from some of the most amplified protein-tissue pairs in similar phage display experiments. A phage library is a collection of millions to billions of random peptides displayed on filamentous bacteriophage within the pIII capsid protein (Pasqualini, 2000). Bacteriophages, for short phages, are viruses that parasite only bacteria. Phages are one of the most common and diverse biological organisms in the biosphere. They provide a convenient mechanism to study the preferential binding of peptides to tissues. In an *in vivo* phage display experiment an unselected phage display library is administered by intravenous infusion, and 15 minutes later, tissue biopsies are obtained to recover phage from different tissues (Arap *et al.*, 2002). Kolonin *et al.* (2006) proposed a multistage phage display experiment, that is, successive stages of screening (usually three or four) to enrich and amplify tissue specific peptides in parallel. This procedure allows for the counts of peptides to increase in every stage and, therefore, it increases the chance of detecting their selection behavior.

Kolonin *et al.* (2006) use a three step procedure to identify preferentially binding peptide-tissue pairs. In a first step they use a Bayesian beta-binomial model to make a list of the peptide-tissue pairs with relatively large increases from the first to the third stages. The second and third step implement (frequentist) tests in contingency tables comparing counts for each peptide A and tissue T versus non-A and non-T. See the more detailed discussion later, in Subsection 3.1. Any peptide A that passes all three filters is reported as binding with preferential selection to tissue T.

Ji *et al.* (2007) proposed a Bayesian hierarchical model as a way of accounting for correlation between measurements and reducing the number of parameters. They used a false discovery rate (FDR) criterion (see Newton, 2004) to report high-binding peptides for a 3-stage phage display experiment with mouse data. Later, in Section 3, we argue that this parametric hierarchical model is inappropriate for the human data described below. Taking advantage of the large sample size of this data set, we propose instead a semi-parametric Bayesian model that avoids some of the limitations of the fully parametric model described by Ji *et al.* Also, we propose an alternative criterion to select high-binding peptides based on a decision theoretic framework. The proposed model and decision theoretic solution addresses multiplicity issues in several ways. First, a hierarchical model adjusts the posterior probabilities for multiplicities. This is discussed, for example, in Scott and Berger (2010). Second, inference under the particular non-parametric Bayesian model can be interpreted as a Bayesian version of Storey's (2007) optimal discovery procedure (Guindani *et al.* 2009). Third, the use of a problem-specific utility function leads us to use a slight variation of traditional rules. Instead of thresholding p-values or marginal posterior probabilities we threshold marginal posterior mean effects.

The main contributions of this paper are the use of a non-parametric prior to avoid the limitations of specific parametric assumptions, and the use of a decision theoretic framework to address the multiplicity issues arising in the selection of a list of tripeptides-tissue pairs that are reported for significant affinity. The remainder of this paper is organized as follows. Section 2 presents a detailed description of the multistage human data. We propose a Bayesian semi-parametric mixture of Poisson model and briefly describe the Markov Chain Monte Carlo (MCMC) simulation scheme for obtaining random samples of the posterior distribution in Section 3. In Section 4, we discuss the criterion to select the peptides that bind specific to certain tissues. In Section 5, we perform a simulation study to assess further the properties of our model and our peptide selecting criterion. In Section 6, we show the results of applying our model to the human 3-stage screening experiment. Finally, we provide some concluding remarks in Section 7.

2. Data

The data are from cancer patients at M. D. Anderson Cancer Center. The data come from three consecutive human subjects who met the formal criteria for brain-based determination of death (Wijdicks, 2001). See Arap *et al.* (2002) for details of the patient selection and the clinical procedure following strict criteria of clinical ethics. One of the primary goals of the study (Staquicini *et al.*, 2011) is to identify peptides with increasing trend over the consecutive stages.

Figure 1 shows a schematic representation of the experiment. The bacteriophages that are injected in each of the stages, i.e., in each of the three human subjects, provide a convenient mechanism to study the preferential binding of peptides to tissues. In summary, the setup is a 3-stage biopanning experiment with bacteriophages delivering random peptides to tissues and the multiple stages allowing for an augmented and easier to detect signal.

At each stage we record counts for peptide-tissue pairs. Peptides are denoted as CX_7C (here C =cysteine, X = any amino acid, represented by a letter). Tissues included in this study are bone-marrow (BM), fat (F), muscle (MU), prostate (PR) and skin (SK). At each stage a phage display peptide library was injected into a new patient, and 15 minutes later biopsies were collected from each of the target tissues and the phage inserts were sequenced. For the second and third stage the input was the pooled inserts amplified from the previous stage.

The original data are counts for all unique 7-mers X_7 . However, we summarize the data using all implied 3-mers. For example, the 7-mer AGAGADR corresponds to the four unique tripeptides AGA, GAG, DAG and ADR. Note that we do not distinguish between a tripeptide and its mirror (e.g., DAG and GAD are counted as the same) and each tripeptide contained in a 7-mer is counted only once (e.g., the count on AGA is incremented only once, although it is contained twice in the 7-mer). So, an observed 7-mer AGAGADR contributes a count for the four tripeptides AGA, GAG, DAG and ADR. The main reason for recording 3-mers are problems related to sparse counts that would result from recording the 20^7 possible 7-mers (Vendruscolo *et al.*, 2001). In contrast there are only 4200 (20^3 , minus duplicate mirrors) tripeptides. Finally, the data corresponding to the third stage contains two seven-peptides, 'XRGFRAA' in muscle and 'KTXRXARX' in skin where one or more amino-acids, denoted by an 'X', are not identified. We consider the four tripeptides identified in the first one and discard counts for the second seven-peptide.

The data reports counts for all observed tripeptide-tissue pairs over 3 consecutive stages. For the analysis, we excluded tripeptide-tissue pairs for which the cumulative count over all three stages was below 6, leaving $n = 2763$ distinct pairs. Table 1 illustrates the data format by showing 6 rows of the data file.

Figure 4 shows the parallel coordinates plot of raw 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 in Figure 4 can be clearly classified as increasing, without reference to any probability model. But for many lines the classification is not obvious. The purpose of the proposed model-based approach is to define where to draw the line to define a significant increase.

3. Model

3.1 A Semi-parametric Mixture of Poisson Model

We use $N_i = (N_{i1}, N_{i2}, N_{i3})$ to denote the observed counts for tripeptide-tissue pair i across the three stages, for pairs $i = 1, \dots, n$. We cast the desired selection of tissue-specific tripeptides as inference about an increasing trend of mean counts in a probability model for the observed data N_i . Kolonin *et al.* (2006) select the tripeptide A as having high binding affinity for tissue T if the posterior probability under the beta binomial model that the proportion of counts in the third stage is greater equal three times the proportion of counts in the first stage. A second and a third filter consist of constructing two contingency tables and apply the one tailed Fisher exact test to each one independently. Supplemental Web Material (SWM) Table 1 shows the detail setup of these two Fisher exact tests.

Ji *et al.* (2007) used a model-based approach based on a mixture of normal distributions for the parameters of the assumed Poisson distribution of the observations. The strength and attraction of their model is its parsimonious nature. For the relatively small mouse data set, this characteristic is important. Nevertheless, the model has its limitations. The human data analyzed here is ten-fold larger. This allows us to consider a more elaborate model. The model in Ji *et al.* assumes a linear relation on the log-mean scale. For example, consider the pairs that are reported as oscillating in Figure 4b. Although the data shows a marked difference in slopes from stages 1 to 2 versus from stages 2 to 3, the parametric model forces one common slope. This is a concern when the imputed overall slope is positive, e.g., the pair marked by A in the Figure 4b. Outliers like pair A in Figure 4b can inappropriately drive the inference. Taking advantage of the larger sample size, the semiparametric nature of the model that we propose can mitigate this problem. Finally, in Ji *et al.* tripeptides are reported based on statistical significance, formalized as the posterior probability of the overall slope being greater than zero. We will propose an approach that also takes into account the size of the overall slope and is more suitable to incorporate biological significance.

In summary, the choice of an appropriate probability model is driven by the following considerations. First, we wish to limit the impact of specific parametric modeling choices on the inference about monotonicity of the mean counts. The experiment is designed to enlarge in subsequent stages any elevated counts that are due to preferential binding of a peptide to a particular tissue. In contrast, randomly high counts would not be enlarged in later counts. In other words, monotonicity of mean counts is equivalent to preferential binding. This equivalence is central to the experiment. Therefore it is important that inference on monotonicity not be driven by outliers or arbitrary parametric forms. The large number of recorded pairs in the human data allows us to use a semi-parametric approach that reduces dependence on a specific parametric model. Second, we will later build on the probability model to define a formal decision problem for the selection of a final list of tripeptide-tissue pairs. For this, and for a simulation study to validate the model, we have to rely on efficient and fast computations.

These two competing desiderata lead us to consider a semi-parametric mixture model. We will use a mixture of parametric Poisson models, with a nonparametric prior on the mixing measure. Under the Bayesian paradigm, nonparametric priors refer to probability models on probability distributions. In particular, we will use a Dirichlet process (DP) prior for the random mixing measure in the mixture of Poisson model. We will later discuss details of the DP prior. The non-parametric prior on the mixing measure greatly generalizes the underlying parametric model, in much the same way as a mixture of independent normal kernels can approximate arbitrary multivariate distributions in kernel density estimates. Similar semi-parametric mixture models have successfully been applied for Bayesian inference in a variety of other applications, including mixed effects models, for example in Müller and Rosner (1997), Mukhopadhyay and Gelfand (1997) or Kleinman and Ibrahim (1998), case-control studies in Mukherjee *et al.* (2007), or point processes in Kottas and Behseta (2010). The special case of binary outcomes has been discussed, among others, by Basu and Mukhopadhyay (2000).

We start with a sampling model for N_i conditional on assumed mean counts across stages for the peptide-tissue combination i . Conditional on the mean counts, we assume independent Poisson sampling. In anticipation of the final inference goal we parameterize the mean counts as $(\mu_i, \mu_i\beta_i, \mu_i\delta_i)$, allowing us to describe increasing mean counts by the simple event $1 < \beta_i < \delta_i$. We write $\text{Poi}(x | m)$ to indicate a Poisson distributed random variable x with mean m .

$$p(N_{i1}, N_{i2}, N_{i3} | \mu_i, \beta_i, \delta_i) = \text{Poi}(N_{i1} | \mu_i) \text{Poi}(N_{i2} | \mu_i \beta_i) \text{Poi}(N_{i3} | \mu_i \delta_i) \quad (1)$$

for $i = 1, \dots, n$. Model (1) implies independence of the counts across i , and dependence across stages $k = 1, 2, 3$. The dependence across stages is induced by the common base-line random effect μ_i . The sampling model includes different Poisson means for each stage, in contrast to the model proposed by Ji *et al.* (2007). The parameter μ_i can be thought as the expected count mean of the pair i across the three stages if we were not enriching at every stage. We extend (1) to the desired semi-parametric mixture model by assuming a non-parametric prior for a random effects distribution for $\theta_i \equiv (\beta_i, \delta_i)$. Let $G(b, d)$ denote a bivariate random probability measure. We discuss the probability model for G below. A parametric random effects distribution for μ_i keeps computation simple.

$$(\beta_i, \delta_i | G) \sim G, \text{ i.i.d.}, \mu_i \sim \text{Ga}(s_\mu, s_\mu \cdot t_\mu) \quad (2)$$

The parametrization of the Gamma distribution is chosen such that $E(X) = a/b$ for $X \sim \text{Ga}(a, b)$. Choosing a prior for the random probability measure G requires a nonparametric prior. The most commonly used model is the DP prior (Ferguson, 1973; Antoniak, 1974). We write $G \sim \text{DP}(\alpha, G_0)$ for a DP prior on the random probability measure G_0 . The DP prior is indexed with two parameters, a total mass parameter α and a base measure G_0 . The total mass parameter is a precision parameter, and the base measure defines the prior expectation, $E(G) = G_0$. See, for example MacEachern and Müller (1998), Walker *et al.* (1999), and Müller *et al.* (2004) for recent reviews of the DP prior, including posterior inference for DP mixtures similar to the model used here. We assume

$$G \sim \text{DP}(\alpha, G_0) \text{ with } G_0(b, d) = \text{Ga}(b | s_\beta, s_\beta t_\beta) \text{Ga}(d | s_\delta, s_\delta t_\delta). \quad (3)$$

The independence of β and δ in the base measure $G_0(\beta, \delta)$ simplifies posterior inference. It does not reflect any prior information, and does not exclude dependence of the random effects distribution $G(\cdot)$. The non-parametric prior for G allows us to learn about possible

correlations between β_i and δ_i , thus inducing additional correlations between round 2 and round 3 counts, if indicated by the data. The model is completed with a prior on the hyperparameters

$$\alpha \sim \text{Ga}(a_\alpha, b_\alpha), t_\beta \sim \text{Ga}(t_\beta | a_{t_\beta}, b_{t_\beta}), \\ t_\delta \sim \text{Ga}(t_\delta | a_{t_\delta}, b_{t_\delta}), \text{ and, } t_\mu \sim \text{Ga}(t_\mu | a_{t_\mu}, b_{t_\mu}), \quad (4)$$

Later, in the simulation and the data analysis we will fix the hyperparameters for the total mass parameter as $a_\alpha = 5$, $b_\alpha = 0.5$. In general, the specification of the total mass parameter in a DP prior model requires care because posterior inference can be critically sensitive with respect to this prior choice. However, this sensitivity arises mainly in inference about the implied clustering of experimental units, tripeptide-tissue pairs in our case. We will not interpret inference about the implied random partition in our analysis, thus reducing concerns about posterior sensitivity. An alternative approach to fix the prior for α is discussed, for example, in Dorazio *et al.* (2008).

Finally, a natural variation of the model could include μ_i with (β_i, δ_i) and use a similar DP mixture, now for the triple $(\mu_i, \beta_i, \delta_i)$. We use the parametric model for μ_i for reasons of simplicity, and because the main inference goals relate to (β_i, δ_i) only.

3.2 Posterior Simulation

The model (1) through (4) includes the conjugate pair of the Poisson sampling model and the gamma baseline measure of the DP, and conditionally conjugate prior specifications for other parameters. This greatly facilitates posterior simulation by MCMC simulation. The implementation of posterior MCMC follows standard posterior simulation methods for DP mixture models. See, for example, Neal (2000) and MacEachern and Müller (1998). In particular, we use the model augmentation with a latent beta random variable proposed in West (1992) to implement inference for the total mass parameter α . The remaining parameters μ_i , t_μ , t_β and t_δ have closed form conditional posterior distributions, conditional on currently imputed values for all other parameters and latent variables.

4. Selecting Significant Tripeptide-tissue Pairs

Posterior MCMC allows us to carry out essentially all posterior inference of interest. In particular, let $\mathbf{N}$ denote the observed data and $p_i = \Pr(\delta_i > \beta_i > 1 \mid \mathbf{N})$ denote the posterior probability of increasing mean counts for peptide-tissue pair i . The posterior Monte Carlo sample allows easy evaluation of p_i as empirical average of $I(\delta_i > \beta_i > 1)$ over all imputed values (δ_i, β_i) in the Monte Carlo posterior sample. Here $I(A)$ denotes the indicator function of the event A . Let $d_i \in \{0, 1\}$ denote an indicator for reporting significant preferential binding of the peptide-tissue pair i , i.e., increasing mean counts, and let $\mathbf{d}$ denote the vector $(d_1, \dots, d_n)$. A reasonable decision rule is to report all pairs with marginal posterior probability beyond a threshold:

$$d_i^* = I(p_i > t). \quad (5)$$

The rule $\mathbf{d}^*$ can be justified in terms of the false discovery rate (FDR) control (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 $\boldsymbol{\theta}$ and $\mathbf{Y}$ generically denote all unknown parameters and all observable data. A utility function $u(\mathbf{d}, \boldsymbol{\theta}, \mathbf{Y})$ formalizes relative preferences for decision under hypothetical outcomes $\mathbf{Y}$ and under an assumed truth $\boldsymbol{\theta}$. For example, a utility function could be

$$u(\mathbf{d}, \theta, \mathbf{y}) = \sum_i d_i I(\delta_i > \beta_i > 1) + k \sum_i (1 - d_i) [1 - I(\delta_i > \beta_i > 1)], \quad (6)$$

i.e., a linear combination of the number of true positive selections d_i and true negatives. In this case the utility function involves the data $\mathbf{Y}$ only indirectly through the decisions $\mathbf{d}$. 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 variables, and conditional on all observed variables. In our case,

$$\mathbf{d}^* = \arg \max_{\mathbf{d}} E[u(\mathbf{d}, \theta, \mathbf{y}) | \mathbf{y}].$$

It can be shown that $\mathbf{d}^*$ arises as Bayes rule under several utility functions that trade off false positive and false negative counts, including the utility in (6) and others. See, for example, Müller *et. al.* (2007), for a discussion.

Alternatively, $\mathbf{d}^*$ can be derived as FDR control. Let $D = \sum d_i$ denote the number of reported pairs and let $r_i = I(\delta_i > \beta_i > 1)$ denote an indicator for the monotonicity of the true mean counts. Then $\text{FDR} = \{1/(D + \epsilon)\} \sum d_i (1 - r_i)$ is the proportion of false positives in the reported list (with the small offset ϵ to avoid zero division). Similarly, $\text{FNR} = \{1/(n - D + \epsilon)\} \sum (1 - d_i) r_i$ is the false negative rate. We compute the posterior expected FDR as

$$E(\text{FDR} | \mathbf{N}) = \frac{1}{D} \sum_{i=1}^n d_i (1 - p_i). \quad (7)$$

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

Ji *et. al.* (2007) selected the pairs with increasing means across the three stages according to the FDR-based criterion (5) where p_i is the probability of (common) positive slope (across the stages 1 and 2, and 2 and 3) according to their model. As can be seen from (6), the control of posterior expected FDR implies a criterion that does not take into account the size of the increase. In other words, a shortcoming of the rule $\mathbf{d}^*$ is that it implicitly weights all true positives equally. However, it would be desirable to rank their selection strength. To achieve this, we will use a utility function that gives weights to the pairs proportional to the relative increment from the first to the third stages (i.e. δ_i). As alternative to (6) we consider the utility function

$$u_B(\mathbf{d}, \theta) = \sum_{i=1}^n d_i w_i - k \sum_{i=1}^n (1 - d_i) w_i - cD, \quad (8)$$

where $w = (w_1, \dots, w_n)$ is a vector of weights. We use

$$w_i = \delta_i I\{\delta_i > \beta_i > 1\};$$

$c > 0$ represents the cost of declaring that a pair has increasing means; and $k > 0$ is such that $k w_i$ is the cost of falsely not reporting pair i when the corresponding (unknown true) means are actually increasing.

Let $\bar{m}_i = E(w_i | \mathbf{N})$. Straightforward algebra shows that the optimal rule is

$$d_i^B = I(\bar{m}_i \geq c/(k+1)). \quad (9)$$

With a Monte Carlo posterior sample $(\beta_i^1, \delta_i^1), \dots, (\beta_i^M, \delta_i^M)$, we estimate $\bar{m}_i$ as

$$\bar{m}_i \approx \frac{1}{M} \sum_{k=1}^M \delta_i^k I\{\delta_i^k > \beta_i^k > 1\}, \quad (10)$$

and then make the decision d_i^B .

5. A Simulation Study

We carried out a simulation study to validate the proposed approach. We generated $n = 2000$ observations of the model described in (1) through (4), except that the random probability measure G is replaced by a gamma distribution with fixed parameters

$$\mu_k \sim \text{Ga}(\mu | s_\mu^f, s_\mu^f t_\mu^f), \text{ i.i.d. and } (\beta_k, \delta_k) \sim \text{Ga}(\beta | s_\beta^f, s_\beta^f t_\beta^f) \text{Ga}(\delta | s_\delta^f, s_\delta^f t_\delta^f) \quad (11)$$

We set the hyper-parameters such that the expected value of μ_i and its variance are small and, besides, β_i and δ_i have both mean 8 and variances 30 and 120 respectively. The motivation for this choice is that μ_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 μ_i to be small. The parameters β_i and δ_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 large variances. The gamma parameters were set to $s_\mu^f = 3.6$, $t_\mu^f = 5/6$, $s_\beta^f = 13/6$, $t_\beta^f = 1/8$, $s_\delta^f = 0.53$ and $t_\delta^f = 0.125$. For 723 out of the $n = 2,000$ simulated tripeptide-tissue pairs the simulation truth included an increase $\delta_i > \beta_i > 1$.

The hyper-parameters of the model described in (1) through (4) were chosen taking into account the same considerations and set to $s_\mu = 0.05$, $a_{t_\mu} = 3$, $b_{t_\mu} = 1/2$, $s_\beta = s_\delta = 5/3$, $a_{t_\beta} = a_{t_\delta} = 6$, $b_{t_\beta} = b_{t_\delta} = 25$, $a_\alpha = 5$ and $b_\alpha = 0.5$. Saving every 10th iteration after a 10,000 iteration burn-in, a Monte Carlo posterior sample of size $M = 5,000$ was saved.

Using the FDR criterion, we selected the pairs such that, under the assumptions of our model, the expected false discovery rate was 0.2. The implied expected FNR was 0.117. Under this rule, we reported 715 pairs for increasing means across the three stages. Of these 578 truly did have increasing mean counts in the simulation truth, i.e., the observed FDR and FNR were 0.192 and 0.113, respectively.

Next we considered selection with respect to the alternative utility function (8), with $c/(k+1) = 8$. The rule $\mathbf{d}^B$ selects 722 pairs to be reported for increasing means. Of these, 575 actually did exhibit the increasing pattern in the simulation truth. The posterior expected FDR and FNR are 0.206 and 0.117 respectively. Using the known simulation truth we

evaluate the observed FDR and FNR as 0.204 and 0.116, respectively. The number of pairs reported by both methods was 688.

Our model is a particular case of a DP mixture model. Like in any density estimation model we can evaluate the posterior mean estimated distribution as the predictive distribution for a future draw. Specifically, in model (2) we find

$$p(\beta_{n+1}, \delta_{n+1} | \mathbf{N}) = E[p(\beta_{n+1}, \delta_{n+1} | G, \mathbf{N}) | \mathbf{N}] = E[G(\beta_{n+1}, \delta_{n+1}) | \mathbf{N}] \equiv \bar{G},$$

i.e., the predictive distribution for a future pair $(\beta_{n+1}, \delta_{n+1})$ is identical with the posterior mean $\bar{G}$ of the unknown random effects distribution G . Comparing $\bar{G}$ with the simulation truth we find that the estimated distribution matches the true gamma distributions quite closely. See the SWM Figure 1.

For comparison we analyze the same data using the approach proposed in Ji *et al.* (2007). Their model differs in two important ways from the proposed model (1) through (5). First, they assume a simplified sampling model, in our notation, with

$$p(N_{i1}, N_{i2}, N_{i3} | \mu_i^{Ji}, \beta_i^{Ji}) \sim \text{Poi}(\mu_i^{Ji}) \text{Poi}[\mu_i^{Ji} \exp(\beta_i^{Ji})] \text{Poi}[\mu_i^{Ji} \exp(2\beta_i^{Ji})]. \quad (12)$$

We use a superindex Ji to indicate parameter in the Ji *et al.* model. Inference is then based on the posterior probability of β_i^{Ji} being positive. Second, instead of using a nonparametric prior, they model β_i^{Ji} with a mixture of three normal distributions. These were reasonable simplifications for the mouse data (with smaller sample size and counts) that is analyzed in Ji *et al.* (2007), but the same simplifications would unreasonably limit inference for the larger human data with larger counts. For comparison we implemented inference under their model, with the hyperparameter values defined there.

For every i , we computed the ad-hoc estimate $\widehat{\beta}_i^{Ji} = [\log(N_{i3}) - (\log(N_{i1}))]/2$ (replacing $N_{ik} = 0$ by $N_{ik} = 0.1$). When the three counts are different from zero, then $\widehat{\beta}_i^{Ji}$ is the MLE in (12).

We computed posterior probabilities $p_i^{Ji} = P[\beta_i^{Ji} > 0 | \mathbf{N}]$ under the model of Ji *et al.* (2007). We find that all p_i^{Ji} are greater than 0.5, i.e., all pairs are estimated to show preferential binding, including pairs with MLE's $\widehat{\beta}_i^{Ji}$ as low as -2 . The comparison is shown in SWM Figure 2a. Alternatively we compared the same MLE estimates with model-based posterior means $E(\beta_i^{Ji} | \mathbf{N})$, and find little correlation. This comparison is shown in SWM Figure 2b. In summary, the parametric model reports high posterior probability of increasing mean counts even when there is no evidence in the data.

Finally, we compare the proposed semiparametric model with an empirical Bayes (EB) parametric version of it. We consider an alternative parametric model that replaces the random probability measure G in (3) by the parametric model G_0 and we fix t_μ , t_β and t_δ in (4) and assumes $\mu_i \sim \text{Ga}(s_\mu, s_\mu t_\mu)$, $\beta_i \sim \text{Ga}(s_\beta, s_\beta t_\beta)$ and $\delta_i \sim \text{Ga}(s_\delta, s_\delta t_\delta)$. The parameters s and t are chosen to match the empirical moments of $N_{i1} + 0.1$, $N_{i2}/(N_{i1} + 0.1)$ and $N_{i3}/(N_{i1} + 0.1)$. This is, $s_\mu = 0.74$, $t_\mu = 1.43$, $s_\beta = 0.49$, $t_\beta = 0.039$, $s_\delta = 0.20$, $t_\delta = 0.04$.

Figure 2 highlights the contrast between the proposed semi-parametric model and a corresponding EB version of the parametric model (see below). The figure shows posterior estimated (β_i, δ_i) for 100 chosen at random tripeptide-tissue pairs under the proposed semi-

parametric model with inference under the parametric model. The posterior distribution under the EB parametric model significantly shifts inference, especially for many of the outlying pairs. The semiparametric model includes the same marginal prior means, but allows for clustering and more borrowing of strength that can outliers. Additionally, the mean squared errors for β , i.e., $(1/n) \sum_i (E(\beta_i | \mathbf{N}) - \beta_i^T)^2$ with the superindex T indicating the simulation truth, are 22.81 and 64.83 for inference under the semiparametric and the EB model, respectively. Similarly, for δ the average mean squared errors are 54.11 vs. 114.73.

6. Results

We present the analysis of the human phage display data under the proposed method. We fix the hyper-parameters as in the simulation study. In particular, the hyperprior choices imply the following marginal means and variances. The parameter μ_i is interpreted as the expected counts 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 μ_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_j) and between the first and the third stage (i.e., δ_j). We center these values around a mean of 5 and allow for a large variance of 25.

We saved a Monte Carlo posterior sample of size $M = 5,000$, saving the values of the imputed parameters every ten iterations after a 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 no evidence for a practical lack of convergence. We explore the fit of the proposed model and hierarchical shrinkage by plotting the observed tripeptide-tissue counts versus the posterior expected means $E(\mu_i | \mathbf{N})$, $E(\mu_i \beta_j | \mathbf{N})$ and $E(\mu_i \delta_j | \mathbf{N})$. The plot (shown in SWM Figure 3abc) shows the usual hierarchical shrinkage of the mean counts corresponding to smaller counts towards the overall mean.

A scatter plot of $\bar{\beta}_j = E(\beta_j | \mathbf{N})$ and $\bar{\delta}_j = E(\delta_j | \mathbf{N})$ (shown in SWM Figure 3d) includes many points that are close to either the vertical line through 1, indicating no enrichment from the first to the second stages, or close to the identity line, indicating no enrichment from the second to the third stages. This complicates inference by making the decision to report many pairs very sensitive to details of the parametric model. Finally, Figure 3 shows the posterior estimated random effects distribution $\bar{G} = E(G | \mathbf{N})$. This shape of $\bar{G}$ would be very difficult to re-create in a parametric model without overfitting. In particular, the multimodal nature of $\bar{G}$ is intriguing. The secondary modes might allow interesting biologic interpretation.

Table 2 shows the 31 pairs with highest $m\bar{n}_j$, i.e., the pairs chosen by $\mathbf{d}^B$, with a threshold of $c/(k+1) = 9$. We notice that there are some pairs, such as the tripeptide GRM in fat tissue, that present a small posterior probability p_j of increasing means but that are selected under $\mathbf{d}^B$ due to a large $m\bar{n}_j$. Of the top 31 tripeptides all but two (RPR and GWS in bone marrow) have been previously identified through dataset analysis with independent statistical methods in Staquicini *et al.* (2011). Tripeptides progressively enriched in the prostate have not been included in the previously published data because this tissue was excluded from analysis for technical considerations. Two tripeptides (QGS and QGW), identified here as enriched in the bone marrow, were also reported enriched in the same tissue in Staquicini *et al.* (2011). In the same work, the majority of progressively enriched tripeptides (19, including prostate-enriched) were identified as “shared” (enriched in more than one tissue), and, interestingly, four progressively enriched tripeptides were identified as enriched in different tissues.

Figure 4 highlights the selected pairs in a plot of the observed counts across stages. The plot shows the top $D = 62$ tripeptide-tissue pairs that are selected with a threshold value $c(k+1) = 7$ under rule $\mathbf{d}^B$. The criterion selects the pairs with high increasing counts across the three stages. In agreement with the underlying utility function (8), the criterion excludes some pairs with nondecreasing counts, and chooses instead pairs with an oscillating count pattern but with a substantial increment from the first to the second stages or from the second to the third stages. Besides, it picks pairs that have small counts over the three stages but with a large increment in the second and/or third stage count in comparison to the previous count. In contrast, there are some other pairs that are not selected despite relatively large and nondecreasing counts over all three stages. The selection method might be detecting that these events can be explained by the large base-line count of the corresponding tripeptide, rather than strong preferential binding. The expected FDR and FNR are 0.35 and 0.13, respectively.

We investigate some aspects of the nonparametric model. Inspecting marginal posterior distributions $p(\beta_i | N)$ and $p(\delta_i | N)$ we find multimodal densities for many pairs i . Some are shown in SWM Figure 4. Such features cannot be detected by the parametric version of our model. The third pair shown in SWM Figure 4 is an example in which there is no statistical evidence of an increasing pattern (since $\beta_i < 1$). Nevertheless, the decision rule (9) tends to pick this pair due to its large count observed at the third stage.

Although inference on the implied clusters is not a focus of this study, the DP mixture model allows such inference. For the $n = 2763$ peptide-tissue pairs we find a posterior mean of 35.8 clusters.

7. Discussion

We have proposed semi-parametric model-based statistical inference for high dimensional count data arising from phage experiments with parallel biopanning. The probability model is extended to a decision problem by adding a utility function for the choice of reported tripeptide-tissue pairs.

Previously, Ji *et al.* (2007) introduced a model for the analysis of phage experiment mouse data. Analogous to their model, ours accounts for the correlation that exists between the different stages and detects the tripeptides that tend to bind with a specific tissue. Since there is just one observation across the three stages for every tripeptide-tissue pair, the use of a hierarchical model to implement borrowing of strength across all pairs is critical.

The proposed model allows a non-linear mean structure across the three stages, and it includes a non-parametric random effects distribution to accommodate outliers, taking advantage of the large data set. The random effects μ_i , β_i , and δ_i have an easy interpretation: μ_i is the mean count of the tripeptide-tissue pair i if there were no enrichment of the tripeptide library at every stage, while β_i and δ_i are the fold increment at the second and third stages, respectively, due to this enrichment. This parametrization allows for an easy description of the target event of increasing mean counts through the three stages.

The proposed model includes random baseline means μ_i to allow a variation of the amount of tripeptides in the original phage display peptide library. It is important to include such random effects. Large counts at the third stage can be explained by either preferential binding of a certain tripeptide, or simply by high random counts in the original library. The inclusion of the random baseline means allows us to differentiate these two cases.

In selecting the tripeptide-tissue pairs with increasing means across the three stages, we face a massive multiplicity problem. We use a decision theoretic setup with a utility function that

takes into consideration statistical significance as well as biological significance such as the size of the mean increase through the three stages.

The nonparametric nature of our model exploits the large sample size to learn about the distribution of mean patterns across the population of tripeptide-tissue pairs. The semi-parametric model implements multiple shrinkage by implicitly defining clusters of pairs with similar mean pattern. We demonstrated how the semi-parametric model avoids inappropriate overall shrinkage as it would be implied by a comparable parametric model.

Among the remaining limitations of the model is that it fails to allow that different tissues can have different selection behaviors. For example, there may be a tissue that absorbs more tripeptides, or that is associated with different variances of the counts. The problem is mitigated by including the base-line mean count μ_p but the model could be further improved by a priori incorporating a correlation structure of pairs sharing the same tissue. This could be achieved, for example, by adding an additional hierarchical layer in the model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Peter Müller was partially funded by grant NIH/NCI CA075981. Luis León Novelo was supported by NIH 1R01GM081704. The rest of the authors were supported by the Cancer Center Support Grant (CCSG) (P30 CA016672) and the M. D. Anderson Cancer Center Prostate SPORE (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

- Antoniak CE. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*. 1974; 2:1152–1174.
- Arap W, Kolonin MG, Trepel M, Lahdenranta J, Card-Vila M, Giordano RJ, Mintz PJ, Ardelt PU, Yao VJ, Vidal CI, Chen L, Flamm A, Valtanen H, Weavind LM, Hicks ME, Pollock RE, Botz GH, Bucana CD, MORE. Steps toward mapping the human vasculature by phage display. *Nature Medicine*. 2002; 8:121–127.
- Basu S, Mukhopadhyay S. Bayesian analysis of binary regression using symmetric and asymmetric links. *Sankhyā*. 2000; 62:372–387.
- Dorazio R, Mukherjee B, Zhang L, Ghosh M, Jelks H, Jordan F. Modeling unobserved sources of heterogeneity in animal abundance using a dirichlet process prior. *Biometrics*. 2008; 64:635–644. [PubMed: 17680831]
- Ferguson TS. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*. 1973; 1:209–230.
- Guindani M, Müller P, Zhang S. A bayesian discovery procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2009; 71:905–925.
- Ji Y, Yin G, Tsui KW, Kolonin MG, Sun J, Arap W, Pasqualini R, Do KA. Bayesian mixture models for complex high dimensional count data in phage display experiments. *Journal of the Royal Statistical Society, Series C: Applied Statistics*. 2007; 56:139–152.
- Kleinman KP, Ibrahim JG. A semi-parametric bayesian approach to the random effects model. *Biometrics*. 1998; 54:921–938. [PubMed: 9750242]
- Kolonin MG, Sun J, Do KA, Vidal CI, Ji Y, Baggerly KA, Pasqualini R, Arap W. Synchronous selection of homing peptides for multiple tissues by in vivo phage display. *The Journal of the Federation of American Societies for Experimental Biology*. 2006; 20:979–981.
- Kottas A, Behseta S. Bayesian nonparametric modeling for comparison of single-neuron firing intensities. *Biometrics*. 2010; 66:277–286. [PubMed: 19432774]

- MacEachern SN, Müller P. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*. 1998; 7:223–238.
- Mukherjee B, Zhang L, Ghosh M, Sinha S. Bayesian semiparametric analysis of case-control data under conditional gene-environment independence. *Biometrics*. 2007; 63:834–844. [PubMed: 17489972]
- Mukhopadhyay S, Gelfand AE. Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association*. 1997; 92:633–639.
- Müller, P.; Parmigiani, G.; Rice, K. FDR and Bayesian multiple comparisons rules. In: Bernardo, José Miguel; Berger, James O.; Phillip Dawid, A.; Smith, Adrian FM., editors. *Bayesian Statistics*. Vol. 8. Oxford: Oxford University Press; 2007. p. 349-370.
- Müller P, Quintana FA, Rosner GL. Hierarchical meta-analysis over related non-parametric Bayesian models. *Journal of the Royal Statistical Society, Series B (Methodological)*. 2004; 66:735–749.
- Müller P, Rosner G. A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association*. 1997; 92:1279–1292.
- Neal RM. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*. 2000; 9:249–265.
- Newton MA. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics (Oxford)*. 2004; 5:155–176.
- Pasqualini, R.; Arap, W.; Rajotte, D.; Ruoslahti, E. In vivo selection of phage-display libraries. In: Barbass, C., III; Burton, D.; Scott, J.; Silverman, G., editors. *Phage display: a laboratory manual*. Vol. 1. Cold Spring Harbor Laboratory Press; New York: 2000. p. 1-24.
- Scott J, Berger J. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*. 2010; 38:2587–2619.
- Staquicini F, Cardó-Vila M, Kolonin M, Trepel M, Edwards J, Nunes D, Sergeeva A, Efstathiou E, Sun J, Almeida N, et al. Vascular ligand-receptor mapping by direct combinatorial selection in cancer patients. *Proceedings of the National Academy of Sciences*. 2011
- Storey J. The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2007; 69:347–368.
- Vendruscolo M, Paci E, Dobson C, Karplus M. Three key residues form a critical contact network in a protein folding transition state. *Nature*. 2001; 409:641–645. [PubMed: 11214326]
- Walker SG, Damien P, Laud PW, Smith AFM. Bayesian nonparametric inference for random distributions and related functions (Disc: P510-527). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*. 1999; 61:485–509.
- West, M. Technical Report 92-A03. Institute of Statistics and Decision Sciences. Duke University; 1992. Hyperparameter estimation in Dirichlet process mixture models.
- Wijdicks EF. The diagnosis of brain death. *The New England Journal of Medicine*. 2001; 344:1215–1221. [PubMed: 11309637]
- Zurita AJ, Troncoso P, Card-Vila M, Logothetis CJ, Pasqualini R, Arap W. Combinatorial screenings in patients. *Cancer Research*. 2004; 64:435–439. [PubMed: 14744752]

Figure 1.

Schematic representation of the experiment. In the first stage an unselected phage library is injected into the first individual, some phages bind to the tissue of interest, unbound phages are washed out of the tissue and the binding phages are used to create an enriched library, a library with binding peptides, and the process is repeated. Peptides with preferential binding are expected to have increasing counts.

Figure 2.

Scatterplot of 100 posterior means of (β_i, δ_i) under both the proposed semi-parametric model (black bullet, “●”) and under empirical Bayes parametric version of the model (shaded cross, “+”). The posterior means corresponding to each pair under both models are joined by dotted line segments. For display purpose only 100 (out of 2000) chosen at random (β_i, δ_i) are depicted. For reference the vertical line through 1 and the identity line are depicted. The pairs with increasing posterior means fall in the upper right region defined by these lines.

Figure 3.

Estimated random effects distribution $\bar{G} = E(G|\mathbf{N})$ (contours) and random draws (dots). The vertical line through 1 and the identity line are shown for reference. The event $\delta > \beta > 1$ indicates preferential binding for a tripeptide-tissue pair.

Figure 4. Observed tripeptide-tissue pair counts across the three stages. Each line connects the three observed counts for one tripeptide-tissue pair. The three panels arrange the pairs by: (a) non decreasing, (b) oscillating and (c) non increasing observed tripeptide-tissue pair counts across the three stages. The pair A that is discussed in Section 3 is indicated in panel (b). Marked in black are the 62 selected pairs using the optimal rule (9) with a threshold value of 7. Counts are jittered for visual purposes.

Table 1

Structure of the data set. The table shows data for six tripeptide-tissue pairs out of a total $n = 2763$ pairs. For each pair a count is observed in every one of the three stages of the phage display experiment. The three columns correspond to the three stages of the experiment, i.e., the three human subjects.

Tripeptide-tissue	Stage 1	Stage 2	Stage 3
GGH-BM	3	4	70
LGS-F	14	14	21
AGY-MU	1	1	4
AGE-PR	6	1	0
AGG-PR	25	10	45
LGS-SK	14	10	26
...		...	

Table 2

31 tripeptide-tissue pairs selected by using the rule (9) with a threshold value of 9. The expected values of FDR and FNR are 0.273 and 0.135 respectively. The posterior means $\bar{m}_i$ and probabilities p_i are computed under model (1)–(4).

Tissue	Tripeptide	Stage 1	Stage 2	Stage 3	$\bar{m}_i$	p_i
BM	AAR	2	6	96	12.52	0.55
BM	DLL	0	2	9	9.17	0.80
BM	FGI	0	2	12	12.01	0.81
BM	FGR	3	11	101	18.44	0.97
BM	GFR	5	8	110	15.91	0.78
BM	GGH	3	4	70	11.90	0.54
BM	GGM	5	13	72	15.16	0.99
BM	GWG	4	6	72	16.37	0.84
BM	GWS	1	3	21	14.12	0.87
BM	QGS	0	3	15	13.13	0.90
BM	QGW	0	2	13	12.67	0.80
BM	RPR	0	1	13	10.72	0.56
BM	RTS	2	5	25	11.84	0.93
F	GFR	6	15	107	16.77	0.99
F	GGM	5	4	64	13.17	0.67
F	GHG	0	4	60	11.91	0.51
F	GRM	4	5	103	9.65	0.41
F	GWG	2	6	62	17.09	0.89
F	HGW	2	4	61	12.72	0.59
MU	DLL	1	1	13	10.08	0.61
MU	GFR	7	12	171	16.07	0.77
MU	GGM	4	7	104	15.02	0.71
MU	SRT	0	3	13	11.89	0.90
PR	ARF	0	1	21	9.41	0.42
PR	ARR	4	3	44	12.81	0.69
PR	GHG	0	1	16	10.52	0.52
PR	GWG	2	2	17	9.73	0.73
PR	HGW	0	1	16	10.33	0.50

Tissue	Tripeptide	Stage 1	Stage 2	Stage 3	nt_i	p_i
SK	DVR	2	2	20	11.58	0.74
SK	GFR	9	30	183	16.55	1.00
SK	RLR	2	1	20	9.91	0.55