Hierarchical Dose-Response Modeling for High-Throughput Toxicity Screening of Environmental Chemicals

Ander Wilson^{1,*}, David M. Reif^{2,**}, and Brian J. Reich^{1,***}

¹Department of Statistics, North Carolina State University, Raleigh, North Carolina, U.S.A.

²Department of Genetics, North Carolina State University, Raleigh, North Carolina, U.S.A.

**email:* ander_wilson@ncsu.edu

***email:* dmreif@ncsu.edu

****email:* brian_reich@ncsu.edu

SUMMARY:

High-throughput screening (HTS) of environmental chemicals is used to identify chemicals with high potential for adverse human health and environmental effects from among the thousands of untested chemicals. Predicting physiologically-relevant activity with HTS data requires estimating the response of a large number of chemicals across a battery of screening assays based on sparse dose-response data for each chemical-assay combination. Many standard dose-response methods are inadequate because they treat each curve separately and under-perform when there are as few as six to ten observations per curve. We propose a semiparametric Bayesian model that borrows strength across chemicals and assays. Our method directly parametrizes the efficacy and potency of the chemicals as well as the probability of response. We use the ToxCast data from the U.S. Environmental Protection Agency (EPA) as motivation. We demonstrate that our hierarchical method provides more accurate estimates of the probability of response, efficacy, and potency than separate curve estimation in a simulation study. We use our semiparametric method to compare the efficacy of chemicals in the ToxCast data to well-characterized reference chemicals on estrogen receptor α (ER α) and peroxisome proliferator-activated receptor γ (PPAR γ) assays, then estimate the probability that other chemicals are active at lower concentrations than the reference chemicals.

KEY WORDS: Bayesian; Computational toxicology; Dose-response; MCMC; Monotonicity; Semiparametric; ToxCast.

1. Introduction

There are thousands of untested chemicals in common use. Comprehensive toxicity testing of all chemicals is infeasible due to high monetary and temporal costs (Judson et al., 2009). To address this problem, a new paradigm in toxicity testing focuses on screening larger numbers of chemicals on a diverse battery of relatively quick and inexpensive high-throughput screening (HTS) assays that measure a variety of cellular and biochemical responses. Each assay measures a single endpoint, such as transcription of a target gene or binding to a specific receptor protein. The aim of HTS is to predict which chemicals are most likely to perturb normal biological processes that lead to adverse human health and environmental effects, and focus scarce testing resources on those chemicals.

To predict potential chemical activity from HTS data requires statistical models to estimate the response of each chemical on each assay and to compare and rank chemicals. There are three main quantities used to compare chemicals: 1) the probability that an active response occurred, 2) the potency or concentration at which a response occurs, and 3) the efficacy or magnitude of the response. These three quantities form the basis of chemical prioritization. With improved estimates of these three quantities, predictive models will be able to better predict which chemicals are most likely to have potentially hazardous effects.

Dose-response modeling for HTS data is unique because there are a large number of curves to estimate but the data for each curve are sparse. For example, the ToxCast project at the US EPA (Dix et al., 2007; Kavlock et al., 2012) has screened nearly 2,000 chemicals on over 700 HTS assays; however, each chemical-assay combination is tested at six to ten unique concentrations and, in most cases, in singlicate at each concentration. Analysis is further complicated by assay and chemical effects, such as assays that are more or less sensitive, correlated assays that measure the same or similar cellular response, and chemicals that are highly active or not active on a variety of assays. Hence, HTS requires a dose-response

method that is robust to the sparsity of the data for each chemical-assay combination, takes advantage of the larger number of chemicals and assays, and accurately estimates the efficacy, potency, and probability of an active response.

A variety of parametric models are used for estimating monotonic dose-response curves (Ritz, 2010). The most common method is the four-parameter log-logistic model (FPLL) which directly parameterizes the efficacy and potency with the Emax (maximal response or upper asymptote) and AC50 (concentration at which the half maximal response occurs), respectively. Figure 1(a) shows an annotated example FPLL dose-response curve. The current release of EPA ToxCast results uses FPLL to fit each dose-response curve with least squares (Judson et al., 2010). Fitting six to ten observations with a four-parameter model using least squares results in poor variance estimates (currently not provided in the ToxCast public release) and no estimates of the probability that an active response occurs.

[Figure 1 about here.]

There are several available frequentist approaches for monotonic curve estimation (e.g. Friedman and Tibshirani, 1984; Mukerjee, 1988; Mammen, 1991; Hall and Huang, 2001; Mammen et al., 2001; Wang and Li, 2008). Recently, several semiparametric Bayesian methods for monotone regression for a single curve have been proposed. Holmes and Heard (2003) proposed using piecewise constant functions with random knots, Neelon and Dunson (2004) utilized a piecewise linear spline model, and Curtis and Ghosh (2011) developed a Bernstein polynomial model. Also, Shively et al. (2009) modeled the monotonic function as the integral of a positive function. None of these general semiparametric regression methods directly parameterize the efficacy or potency of a dose-response relationship, which are widely used for comparing chemicals and predictive modeling.

Our problem is different from these because we are estimating the response for several chemicals on multiple assays. Bayesian hierarchical models have been used in many fields

where the data takes a natural hierarchical structure. Several *in vivo* or developmental toxicity studies have used Bayesian hierarchical models to improve estimates when measuring the response of multiple correlated endpoints tested with a single chemical (e.g Faes et al., 2006; Choi et al., 2010) or used a multivariate model that assumes correlation in residuals for multiple health outcomes (Neelon and Dunson, 2004).

To incorporate dependence in the regression function between four HTS assays and eight nanomaterials, Patel et al. (2012) estimate dose-duration-response surfaces using linear B-splines with two internal knots in both the duration and dose direction. The first knot parameterizes potency with the no observable adverse effect level (NOAEL), an alternative measure to AC50. For each chemical, they model correlation in the knot location and basis coefficients across the assays, but they do not model correlation across chemicals within assay. While the direct parameterization of the NOAEL is appealing, this model does not directly parameterize the probability of a response or the efficacy, two important parameters for prioritization. In addition, the model does not include assay effects which we assume to exist in our data and can potentially improve fitting with the large number of chemicals but small sample size with each chemical-assay. While the simple choice of linear splines with two internal knots provides for a reasonable size model for a generalized additive model, this basis is not realistic for a one-dimensional model (dose only, not dose-duration surfaces).

In this paper, we propose a Bayesian hierarchical model for dose-response curves that is specifically tailored to the high-dimensional, sparse data setting of the ToxCast project, called the zero-inflated piecewise log-logistic model (ZIPLL). ZIPLL is a mixture between a non-active response and an active response that extends FPLL to a more flexible spline formulation while maintaining direct parameterization of the efficacy and potency of each chemical-assay combination. Our Bayesian approach naturally estimates the three key summary statistics and measures of uncertainty for ranks of efficiency and potency which should

allow for decision-makers to use the results appropriately when deciding which chemicals to consider for future, more comprehensive, testing. We use a hierarchical framework that borrows strength across chemicals and assays. This adds robustness, incorporates assay and chemical effects, and allows for estimation of joint distributions of responses across multiple assays. In addition, prior information and covariates can be included to exploit known relationships between chemicals, between assays, and between chemical-assay combinations.

2. The ToxCast Data

The ToxCast project uses a diverse battery of HTS assays and informatic models to rapidly characterize the activity of thousands of chemicals. These chemical activity profiles are used to support decisions regarding prioritization for further testing (Reif et al., 2010), predict *in vivo* activity (Martin et al., 2011), and inform risk assessments (Judson et al., 2011). In support of these goals, the ToxCast project has tested over 2,000 chemicals on over 700 HTS assay endpoints for which analysis is ongoing. The data for the first 309 chemicals tested for Phase I are publicly available (<http://www.epa.gov/ncct/toxcast/data.html>). Figure 7 illustrates the unique structure of the data.

[Figure 2 about here.]

In this paper we use the 309 chemicals in the publicly available data and a subset of 81 assays comprising the multiplexed transcription factor reporter platform (Romanov et al., 2008, www.attagene.com). This platform enables high-content, functional assessment of transcription factor activity, which is a core component of cellular gene regulatory networks. Both cis-regulating response element constructs (CIS) and trans-activating (TRANS) potential of multiple nuclear hormone receptors are measured. These 48 CIS and 25 TRANS assays (plus 8 negative control assays) address relevant cellular processes including response to xenobiotics,

genotoxic stress, hypoxia, oxidative damage, immune-modulation, and endocrine disruption. Martin et al. (2010) evaluated these assays' response to the 309 Phase I chemicals.

Chemicals were diluted in dimethyl sulfoxide (DMSO) at, in general, six to ten unique concentrations on each HTS assay. The concentrations typically ranged from $0.046\mu\text{M}$ to $100\mu\text{M}$ or from $0.091\mu\text{M}$ to $200\mu\text{M}$ with each concentration three times the previous concentration. In cases of overt cytotoxicity, the concentration ranges were shifted up or down by a multiple of 3 in an attempt to recover the concentration range with a chance to show specific assay effects (Martin et al., 2010). Of the 309 Phase I chemicals four were tested in duplicate and one was tested in triplicate. The remaining chemicals were tested once at each concentration. The responses at each concentration are recorded in fold change over DMSO solution; hence, a response of 1 indicate no response. To reduce the inherent heteroskedasticity of the data, we log transformed the data before curve fitting, but return the data to the original scale before analyzing and plotting results.

3. Model Description

Our primary focus is understanding the relationship between the tested doses, x_{ijk} , and the measured responses, y_{ijk} , where i indexes chemical, j indexes assay, and k indexes tested concentrations within a chemical-assay combination. We assume a univariate Gaussian regression model $y_{ijk} = f_{ij}(x_{ijk}) + \epsilon_{ijk}$ and $\epsilon_{ijk} \stackrel{\text{iid}}{\sim} \text{N}(0, \sigma^2)$.

The ZIPLL regression function is a mixture between an active and non-active response

$$f(x_{ijk}; \boldsymbol{\theta}_{ij}, \mathbf{w}_{ij}, Z_{ij}) = \begin{cases} t_{ij} - (t_{ij} - b_{ij}) \times \text{Logit} \{g(x_{ijk}; a_{ij}, \mathbf{w}_{ij})\} & \text{if } Z_{ij} = 1 \\ b_{ij} & \text{if } Z_{ij} = 0. \end{cases} \quad (1)$$

In (1), $\boldsymbol{\theta}_{ij} = (t_{ij}, b_{ij}, a_{ij})$ parameterizes the Emax (upper asymptote), Emin (lower asymptote), and AC50 of the active response, $\mathbf{w}_{ij}$ controls the shape of an active response, and the latent Z_{ij} indicates if an active response occurred.

The function $g(x; a, \mathbf{w})$ in (1) can be any monotone decreasing function in x with location parameter a and shape parameter $\mathbf{w}$, such that $g(a; a, \mathbf{w}) = 0$ for all $\mathbf{w}$. When $t \geq b$, the active response is nondecreasing with upper and lower horizontal asymptotes t and b , respectively. The constraint that $g(a; a, w) = 0$ insures that the response is $(t + b)/2$ when $x = a$. Hence, a is the AC50 of the active response. When $g(x; a, \mathbf{w}) = -w\{\log(x) - \log(a)\}$ active responses follow FPLL (Figure 1(a)), we refer to this model as this is a zero-inflated log-logistic model (ZILL).

In ZIPLL, we replace the log-linear function with a piecewise log-linear spline to add robustness to misspecification of the shape of the dose-response curve. The new function is

$$g(x_{ijk}; a_{ij}, \mathbf{w}_{ij}) = \mathbf{\Psi}(x_{ijk}, a_{ij})^T \mathbf{w}_{ij} = \sum_l w_{ijl} \Psi_l(x_{ijk}, a_{ij}), \quad (2)$$

where $\mathbf{\Psi}(x_{ijk}, a_{ij})$ is a vector of continuous linear basis functions with $\mathbf{\Psi}(a_{ij}, a_{ij}) = \mathbf{0}$ and $\mathbf{w}_{ij}$ is a $(p + 2)$ -vector of unknown basis coefficients.

We construct the linear basis using $2p + 1$ internal knots with one knot at 0. We choose symmetric fixed internal knots at $\{-\xi_{p/2}, \dots, -\xi_1, \xi_0, \xi_1, \dots, \xi_{p/2}\}$. The l^{th} basis function is

$$\Psi_l(x_{ijk}, a_{ij}) = \begin{cases} [\max\{\log(x_{ijk}) - \log(a_{ij}), \xi_l\} - \xi_{l-1}] \mathbf{1}_{(x_{ijk} < a_{ij})} & \text{for } l = -p/2 - 1, \dots, -1 \\ [\min\{\log(x_{ijk}) - \log(a_{ij}), \xi_l\} - \xi_{l-1}] \mathbf{1}_{(x_{ijk} \geq a_{ij})} & \text{for } l = 1, \dots, p/2 + 1, \end{cases}$$

with center knot $\xi_0 = 0$ and external knots $\xi_{-p/2-1} = -\infty$ and $\xi_{p/2+1} = \infty$. Figure 1(b) shows a set of basis functions. This basis ensures a monotone nondecreasing response as long as each element of $\mathbf{w}$ is non-negative, and that a_{ij} is the AC50 (since $\mathbf{\Psi}(a_{ij}, a_{ij}) = \mathbf{0}$). This basis has attractive limiting cases. As $p \rightarrow \infty$ we get the full span of monotonic functions, $g(x)$, through the origin. When the basis coefficients $(w_{ij,1}, \dots, w_{ij,p+2})$ are all equal, ZIPLL reduces to ZILL, and when $Z_{ij} = 1$ it further reduces to FPLL.

3.1 Hierarchical Structure and Prior Specification

We use a Bayesian approach to estimate the regression function and choose a prior that restricts the parameter space to increasing functions and induces a hierarchical structure between curves. ZIPLL is monotone nondecreasing when $t_{ij} \geq b_{ij}$ and $\mathbf{w}_{ij} \geq \mathbf{0}$ and is identifiable if $a_{ij} \in [x_{ij1}, x_{ijn_{ij}}]$. These constraints are met by introducing unconstrained latent parameters $\boldsymbol{\theta}_{ij}^* = (t_{ij}^*, b_{ij}, a_{ij}^*)^T$ and $\mathbf{w}_{ij}^*$, and mapping these unconstrained parameters to their constrained counterparts by $a_{ij} = \min(\mathbf{x}_{ij}) + \{\max(\mathbf{x}_{ij}) - \min(\mathbf{x}_{ij})\} / \{1 + \exp(-a_{ij}^*)\}$, $t_{ij} = \max(t_{ij}^*, b_{ij})$, and $w_{ij} = \exp(w_{ij}^*)$. This formulation allows $\boldsymbol{\theta}_{ij}$ and $\mathbf{w}_{ij}$ to conform with the restricted parameter space but $\boldsymbol{\theta}_{ij}^*$ and $\mathbf{w}_{ij}^*$ to take any real values.

The chemical-assay specific parameters have normal priors

$$\begin{aligned}\boldsymbol{\theta}_{ij}^* &\stackrel{\text{iid}}{\sim} \text{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \mathbf{w}_{ij}^* &\stackrel{\text{iid}}{\sim} \text{N}_{p+2}(\mathbf{S}, \boldsymbol{\Sigma}_s).\end{aligned}$$

To encourage smoothness in the slopes, we put an autoregressive hyperprior on $\mathbf{S}$ similar to Neelon and Dunson (2004). We let S_0 be the *a priori* belief of the average slope and $S_k \sim \text{N}(S_{k-1}, \lambda^{-1})$ for $k = 1, \dots, p+2$. To allow for uncertainty in the smoothness of $\mathbf{S}$ we put a $\text{Gamma}(g_1, g_2)$ hyperprior on λ .

3.2 Assay Effects, Chemical Effects, and Prior Knowledge

It is reasonable to expect that some assays may be more or less sensitive than others and in some cases practitioners may want to incorporate prior knowledge about chemicals and assays such as covariates or known groups of similar chemicals and assays. For example, in the ToxCast data Martin et al. (2010) reported that the number of chemicals active on each assay ranged from 0 to 225 and the expected range of potencies and efficacies varied between assays. This information can be modeled as either fixed or random effects in the prior mean of $\boldsymbol{\theta}_{ij}^*$ or the prior mean of Z_{ij} , $\psi_{ij} = \Pr(Z_{ij} = 1)$. We include both the assay level random

effects and a probit model on the ψ_{ij} in our analysis in Section 5. To account for between assay differences in efficacy and potency we put a random effects model on θ_{ij}^* ,

$$\begin{aligned}\theta_{ij}^* | \mu_j, \Sigma_j &\sim N_3(\mu_j, \Sigma_j) \\ \mu_j &\stackrel{\text{iid}}{\sim} N_3(\mu, \Sigma).\end{aligned}$$

For the analysis of ToxCast data in Section 5 we fit the data with log transformed responses to reduce heterogeneity. Our prior reflects strong confidence that the Emin will be near 1 (or 0 on log scale), the baseline response for DMSO solution, but allows more freedom in the other three parameters. We assume μ is normal with mean $(2, 0, -3)^T$ and variance $\text{diag}(3, .1, 3)$. The prior for Σ is inverse Wishart prior with scale parameter 6 and shape parameter $\text{diag}(60, .2, 60)$.

We also include the chemical level covariate LogP, the log of the partition coefficient. LogP is a measure of solubility and relates to a chemicals' ability to permeate a membrane, a prerequisite to a cellular response. LogP was calculated using Leadscape (Leadscape Inc., Columbus, OH). We use LogP to model the prior for ψ_{ij} with a linear fixed effect and a probit link. To account for the varying sensitivity of the 81 assays, we include an assay level random intercept, $\psi_{ij} = \Phi(\delta_j + \delta \text{LogP}_i)$. The hyperpriors these parameters used to fit the ToxCast data are $\delta_0 \sim N(0, 10^2)$ and $\delta_j \sim N(\delta_0, \sigma_\delta^2)$ with $\delta_0 \sim N(0, 10^2)$ and $\sigma_\delta^{-2} \sim \text{Gamma}(0.1, 0.1)$. The remaining hyperpriors are $\sigma^{-2} \sim \text{Gamma}(1, 0.5)$, $S_0 = 2$ and $\lambda \sim \Gamma(1, .5)$.

3.3 Posterior Computation

Our MCMC algorithm is a hybrid Gibbs and Metropolis-Hastings sampler. The full conditional posterior distributions of μ , Σ , σ^{-2} , Z_{ij} , ψ , λ , Σ_s , and $\mathbf{S}$ have simple conjugate forms. The full conditional distributions for t_{ij}^* and b_{ij} are both mixtures of truncated normals. All full conditionals are detailed in Web Appendix A.

The remaining parameters, a_{ij}^* and $\mathbf{w}_{ij}^*$, do not have closed form posterior distributions. To

reduce autocorrelation we use a resolvable transition kernel based on the Metropolis-Hastings kernel (Robert and Casella, 2004). We provide details on sampling these parameters in Web Appendix A. An R package to implement ZIPLL is provided in the supplemental material.

MCMC sampling returns posterior samples for θ_{ij} and Z_{ij} which provide the estimates of potency (a_{ij}), efficacy (t_{ij}), and activity. The posterior probability of an active response is the proportion of samples with $t_{ij} > b_{ij} + \kappa$ and $Z_{ij} = 1$. We assume that the minimum clinically important response is a one fold increase a baseline measure by using $\kappa = 1$. This is the same assumption used in previous ToxCast analyses (Martin et al., 2010). We use this “clinically important” definition to define active responses in Section 5 in order to be consistent with the current ToxCast practices.

This algorithm performed well on our simulated and real data and by using the resolvable kernel there is a reasonably small level of autocorrelation. For the full 309 chemicals and 81 assays we ran the chain for 50,000 iterations and discarded the first 20,000 as burnin. The smaller simulation with 100 curves was run for 20,000 iterations with 5,000 discarded for burnin. We assessed convergence by inspecting trace plots, and comparing multiple chains.

MCMC sampling is carried out in C called from R (R Development Core Team, 2011) with `.C`. Runtime for simulated data set of 100 curves of eight observations for 20,000 iterations is 42 seconds with ZIPLL. Analysis of all 309 chemicals and 81 assays with 50,000 iterations of ZIPLL including random assay effects and probit model for covariates as specified in Section 5 took 19.9 hours. Both computation times are on a DELL Dual Processor Xeon Six Core 3.6 GHz machine with 60GB RAM.

4. Model Comparisons

4.1 *Simulation Study*

To evaluate the performance of ZIPLL we conducted a simulation study based on the design of the ToxCast data. We summarize the simulation results here and provide a full discussion in Web Appendix B. We fit ZIPLL, ZILL, FPLL using nonlinear least squares, and Bayesian, monotone, piecewise linear spline model proposed by Neelon and Dunson (2004) that fits each chemical-assay combination separately.

When data was simulated from ZILL, the four methods performed similarly, with the hierarchical methods (ZIPLL and ZILL) slightly outperforming the others with respect to pointwise root mean square error (RMSE). When we used an asymmetric response pattern that violated the assumptions of FPLL and ZILL, ZIPLL had smaller pointwise RMSE as well as RMSE on the AC50 and Emax. ZIPLL also had better credible interval coverage with smaller interval widths. Finally, the three Bayesian methods estimated active responses with high probability, while FPLL did not.

4.2 *Cross Validation and Model Fit*

To determine if ZIPLL provides a better fit specifically for the ToxCast data we performed a cross validation study and compared several model fit statistics for the four methods described in Section 4.1 fit to the ToxCast data for 309 chemicals and 81 assays. For ZIPLL, we included the full random effects model and probit model on the probability of response. We determined that one interior knot at 0, two linear segments, performed best based on the cross validation predictive MSE. For cross validation we removed one observation from each chemical-assay combination, fit the remaining data, and predicted the removed response. This was repeated 8 times, leaving out a different observation each time.

ZIPLL had the lower predictive MSE, 0.2177 compared to 0.2200 for ZILL, 0.3041 for monotone, linear splines, and 0.3080 for LS. Whereas ZIPLL had noticeable advantages over

ZILL in the simulation, the two methods were similar in cross validation; however, this small improvement is significant with a paired t -test. We also compared the three Bayesian methods using DIC (Spiegelhalter et al., 2002), log psuedo-marginal likelihood (LPML) (Geisser and Eddy, 1979), continuous ranked probability score (CRPS) (Matheson and Winkler, 1976), and the method of Gelfand and Ghosh (1998) using equal weights for the two components. In each case, ZIPLL outperformed ZILL by a small margin. Therefore, we present the results below assuming the ZIPLL model.

5. ToxCast Data Application

We fit the 309 chemicals and 81 assays with ZIPLL, including assay random effects and probit model as specified in Section 3.2. Figure 3 shows the dose-response estimates for 12 chemicals on pregnane X receptor response element (PXRE) fit with ZIPLL and the reported fits from the ToxCast public use files. The first row shows three chemicals where the ZIPLL posterior mean is similar to the FPLL fits reported in ToxCast. The second and third rows show chemicals where ZIPLL better fits the data by adapting to an asymmetric response pattern.

[Figure 3 about here.]

The bottom row of Figure 3 highlights the importance of probabilistic estimation of an active response. The three chemicals shown have similar response patterns; however, using the current ToxCast methodology one is marked active, having increased by at least one fold change, on PXRE while the other two are not. With ZIPLL, the estimated probabilities of response are between 0.26 and 0.87. Web Appendix C Figure 2 compares the ZIPLL probability with ToxCast indicator for all 309 chemicals on three assays. The majority of chemicals are considered not active or active with both methods. However, using ZIPLL we estimated that several chemicals have posterior probabilities of response between 0.1 and 0.9,

suggesting that there is not conclusive evidence that these chemical responded or not, but are forced to be classified as either active or not in the ToxCast data. The set of chemicals having high non-zero posterior probabilities of response via ZIPLL yet a ToxCast call of no response include several with evidence of PXR activity from other ToxCast assays (e.g. Flumiclorac-pentyl) and/or independent structure-activity models (e.g. Butafenacil) (Kortagere et al., 2010).

5.1 *Summary of Active Responses and Assay and Chemical Effects*

A natural result of the hierarchical analysis is estimation of the joint distribution of responses across assays. Figure 4(a) shows the number of active assay responses for each chemical. The number of assay responses reported in ToxCast tends to be around the lower bound of the ZIPLL posterior interval and is similar to the number of assay responses estimated if we consider anything with a ZIPLL posterior probability of 0.75 to be active. Overall, there are 2667 (2616, 2718) active assay-chemical combinations estimated with ZIPLL compared to 1887 reported in ToxCast. This suggests that some assay responses may be missed using the current ToxCast methods, potentially hindering prioritization efforts.

[Figure 4 about here.]

Figure 4(b) shows the posterior of the assay random intercept for the probit model of probability of response. The most and least sensitive assays had statistically significant random effects. At the lower end, the eight assays with the prefix “M_” are negative controls and all had effects around -10, while more potent assays like PXRE and PPAR γ had large positive effects. The posterior mean of the coefficient for LogP is -0.0005, and this effect was not significant. This may be due to selection bias. Solubility (low logP) was part of the selection criteria for the first 309 chemicals in order to accommodate solubility in Dimethyl sulfoxide (DMSO); however, this restriction was relaxed for chemicals included in forthcoming ToxCast phases, so LogP may prove to be an important factor in future samples.

The simultaneous fitting of all chemicals allows for simple rankings of chemical by potency as well as a measure of uncertainty in the rankings. Figure 5 shows the ranking of chemicals by posterior mean AC50 on three assays: PXRE, peroxisome proliferator-activated receptor γ (PPAR γ), and estrogen receptor α (ER α). Among chemicals with at least a 0.5 posterior probability of being active, the mean AC50 was 32.7 for PXRE, 64.5 for PPAR γ , and 50.6 for ER α . This supports the inclusion of an assay random effect in the model.

5.2 Comparison with Reference Chemicals

A useful way to summarize the results for each assay is to compare chemicals with reference chemicals known to be active on the assay. For example, PPAR γ is a commonly used assay that has a plausible connection with neoplastic pathology (see Peters et al., 1997; Aoki, 2007). Figure 5(b) highlights the response of four reference chemicals for PPAR γ : perfluorooctane sulfonic acid (PFOS), Diethylhexyl phthalate (DEHP), Phthalic acid, mono-2-ethylhexyl ester (PAMEHP), and perfluorooctanoic acid (PFOA) (Casals-Casas and Desvergne, 2011). Because the reference chemicals have known biological effects, other chemicals with a high probability of being more potent than the reference chemicals on a given assay may have greater potential for similar biological effects to the reference chemicals, and thus may be higher priority candidates for additional testing than chemicals that are not as potent as the reference chemicals. The four reference chemicals' posterior mean potencies rank (with 1 being the most potent) 42.9 (35.0, 49.0), 105.7 (57.0,154.0), 114.0 (68.0,156.0), and 122.6 (72.0,159.0), respectively, on this assay among 161 chemicals with at least 0.5 probability of activity, indicating there are many good candidates for further testing.

[Figure 5 about here.]

Another commonly studied assay is ER α . Figure 5(c) shows results for ER α with reference chemicals Bisphenol A (BPA) and Methoxychlor highlighted. These two reference chemicals have mean posterior rank 1 (1.0,1.0) and 7.4 (4.0,10.0), respectively, among the 103 chemicals

with posterior probability of an active response of 0.5 or more. This implies there is at least 0.95 probability that BPA is the most potent chemical among the 309 and very few ToxCast chemicals are more potent than Methoxychlor.

These rankings allow us to estimate the posterior probability that chemicals are more active than the reference chemicals both marginally for each assay and jointly across assays. The ability to estimate this probability jointly across biologically related assays provides an important capability—pathway based prioritization. For assays measuring distinct biological targets, such as $ER\alpha$ and $PPAR\gamma$, no chemical had more than a trivial posterior probability of being more active than the reference chemicals on both, which is expected. As an example of prioritization based upon single assays, Figure 6 shows the 66 chemicals with at least a 0.05 posterior probability of being more potent than the four $PPAR\gamma$ -specific reference chemicals on $PPAR\gamma$. With the more diverse set of reference chemicals available in the forthcoming ToxCast data, comparisons across several assays will be feasible and can provide a probabilistic ranking of chemicals based on the potential for bioactivity on these pathways.

[Figure 6 about here.]

6. Discussion

In this paper we propose a new model for estimating the dose-response relationship for HTS data when a large number of chemicals are tested across several assays. ZIPLL directly parametrizes the AC_{50} and E_{max} . As a result, the efficacy and potency are easily interpretable. Through our simulation study, we demonstrated that ZIPLL accurately estimates the AC_{50} , E_{max} , and the probability of response curve. Further, ZIPLL is robust to assumptions about the shape of the response. Overall, this hierarchical approach to analyzing HTS data outperformed methods that treat each curve as independent and ignore correlation between assays.

Our proposed MCMC algorithm takes about 20 hours to fit 309 chemicals on 81 assays, longer than the published ToxCast method. However, for HTS projects like ToxCast, data are analyzed in large batches, so real-time updates are not necessary. As a result, emphasis is on model performance over efficient computation. In the case that a small number of chemicals were added, all hyperparameters could be fixed based on the full run and the posterior computed for the new chemicals in a few minutes. For larger batches, computation time for ZIPLL scales linearly for both the number of chemicals and number of assays, making runs on larger experiments feasible.

We applied ZIPLL to the ToxCast data and showed that the probabilities of response were largely consistent with the binary classification in the ToxCast public release data. However, in borderline cases ZIPLL added useful information by quantifying the uncertainty in the presence of a response. We also demonstrated the advantage of estimating the posterior distribution of the AC50. This allowed us to rank chemicals and estimate the posterior probability that a chemical is more potent than reference chemicals, which provides a useful tool for prioritization.

Ultimately, a comprehensive risk assessment must include not only coverage of all relevant exposure and hazard factors, but thorough characterization of individual factors as well. The dose-response model provided by ZIPLL will prove especially useful in such a scenario, where the more informative results characterize HTS hazard in a manner that can be quantitatively combined with other risk factors. With the addition of data from future HTS projects having expanded assay coverage and reference chemical sets, these rankings can be extended to estimate the joint probability that chemicals are more active than reference chemicals on multiple assays, thus providing a physiologically relevant, pathway-based hazard assessment.

7. Supplementary Material

The Web Appendix referenced in Section 3, 4, and 5 and the ZIPLL R package are available with this article at the Biometrics website on Wiley Online Library

ACKNOWLEDGEMENTS

Ander Wilson was supported by NIH training grant GM081057: Biostatistics Training in the Omics Era.

REFERENCES

- Aoki, T. (2007). Current status of carcinogenicity assessment of peroxisome proliferator-activated receptor agonists by the US FDA and a mode-of-action approach to the carcinogenic potential. *Journal of Toxicologic Pathology* **20**, 197–202.
- Casals-Casas, C. and Desvergne, B. (2011). Endocrine disruptors: From endocrine to metabolic disruption. *Annu Rev Physiol.* **73**, 135–162.
- Choi, T., Schervish, M., Schmitt, K., and Small, M. (2010). Bayesian hierarchical analysis for multiple health endpoints in a toxicity study. *Journal of Agricultural, Biological, and Environmental Statistics* **15**, 290–307.
- Curtis, S. M. and Ghosh, S. K. (2011). A variable selection approach to monotonic regression with Bernstein polynomials. *J. Appl. Stat.* **38**, 961–976.
- Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., and Kavlock, R. J. (2007). The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicological Sciences* **95**, 5–12.
- Faes, C., Geys, H., Aerts, M., and Molenberghs, G. (2006). A hierarchical modeling approach for risk assessment in developmental toxicity studies. *Computational Statistics & Data Analysis* **51**, 1848 – 1861.

- Friedman, J. and Tibshirani, R. (1984). The monotone smoothing of scatterplots. *Technometrics* **26**, 243–250.
- Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153–160.
- Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85**, 1–11.
- Hall, P. and Huang, L.-S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.* **29**, 624–647.
- Holmes, C. C. and Heard, N. A. (2003). Generalized monotonic regression using random change points. *Statistics in Medicine* **22**, 623–638.
- Judson, R., Richard, A., Dix, D. J., Houck, K., Martin, M., Kavlock, R., Dellarco, V., Henry, T., Holderman, T., Sayre, P., Tan, S., Carpenter, T., and Smith, E. (2009). The toxicity data landscape for environmental chemicals. *Environmental Health Perspectives* **117**, 685–695.
- Judson, R. S., Houck, K. A., Kavlock, R. J., Knudsen, T. B., Martin, M. T., Mortensen, H. M., Reif, D. M., Rotroff, D. M., Shah, I., Richard, A. M., and Dix, D. J. (2010). In vitro screening of environmental chemicals for targeted testing prioritization: The ToxCast project. *Environmental Health Perspectives* **118**, 485–492.
- Judson, R. S., Kavlock, R. J., Setzer, R. W., Cohen Hubal, E. A., Martin, M. T., Knudsen, T. B., Houck, K. A., Thomas, R. S., Wetmore, B. A., and Dix, D. J. (2011). Estimating toxicity-related biological pathway altering doses for high-throughput chemical risk assessment. *Chemical Research in Toxicology* **24**, 451–462.
- Kavlock, R., Chandler, K., Houck, K., Hunter, S., Judson, R., Kleinstreuer, N., Knudsen, T., Martin, M., Padilla, S., Reif, D., Richard, A., Rotroff, D., Sipes, N., and Dix, D. (2012). Update on EPA’s ToxCast program: Proving high throughput decision support tools for

- chemical risk management. *Chemical Research in Toxicology* **25**, 1287–1302.
- Kortagere, S., Krasowski, M. D., Reschly, E. J., Venkatesh, M., Mani, S., and Ekins, S. (2010). Evaluation of computational docking to identify pregnane X receptor agonists in the toxcast database. *Environmental Health Perspectives* **118**, 1412–1417.
- Mammen, E. (1991). Estimating a smooth monotone regression function. *Ann. Statist.* **19**, 724–740.
- Mammen, E., Marron, J. S., Turlach, B. A., and Wand, M. P. (2001). A general projection framework for constrained smoothing. *Statist. Sci.* **16**, 232–248.
- Martin, M. T., Dix, D. J., Judson, R. S., Kavlock, R. J., Reif, D. M., Richard, A. M., Rotroff, D. M., Romanov, S., Medvedev, A., Poltoratskaya, N., Gambarian, M., Moeser, M., Makarov, S. S., and Houck, K. A. (2010). Impact of environmental chemicals on key transcription regulators and correlation to toxicity end points within EPA’s ToxCast program. *Chemical Research in Toxicology* **23**, 578–590.
- Martin, M. T., Knudsen, T. B., Reif, D. M., Houck, K. A., Judson, R. S., Kavlock, R. J., and Dix, D. J. (2011). Predictive model of rat reproductive toxicity from ToxCast high throughput screening. *Biology of Reproduction* **85**, 327–39.
- Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science* **22**, 1087–1096.
- Mukerjee, H. (1988). Monotone nonparametric regression. *The Annals of Statistics* **16**, 741–750.
- Neelon, B. and Dunson, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics* **60**, 398–406.
- Patel, T., Telesca, D., George, S., and Nel, A. (2012). Toxicity profiling of engineered nanomaterials via multivariate dose response surface modeling. *Annals of Applied Statistics* **6**, 1707–1729.

- Peters, J. M., Cattley, R. C., and Gonzalez, F. J. (1997). Role of ppar alpha in the mechanism of action of the nongenotoxic carcinogen and peroxisome proliferator wy-14,643. *Carcinogenesis* **18**, 2029–2033.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Reif, D. M., Martin, M. T., Tan, S. W., Houck, K. A., Judson, R. S., Richard, A. M., Knudsen, T. B., Dix, D. J., and Kavlock, R. J. (2010). Endocrine profiling and prioritization of environmental chemicals using ToxCast data. *Environmental Health Perspectives* **118**, 1714–1720.
- Ritz, C. (2010). Toward a unified approach to dose-response modeling in ecotoxicology. *Environmental Toxicology and Chemistry* **29**, 220–229.
- Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.
- Romanov, S., Medvedev, A., Gambarian, M., Poltoratskaya, N., Moeser, M., Medvedeva, L., Gambarian, M., Diatchenko, L., and Makarov, S. (2008). Homogeneous reporter system enables quantitative functional assessment of multiple transcription factors. *Nat. Methods* **5**, 253–260.
- Shively, T. S., Sager, T. W., and Walker, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 159–175.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**, 583–639.
- Wang, X. and Li, F. (2008). Isotonic smoothing spline regression. *Journal of Computational and Graphical Statistics* **17**, 21–37.

Figure 1. Panel (a) shows an annotated example of the four-parameter log-logistic (FPLL) function. The FPLL model is $f(x; t, b, a, w) = t - (t - b) \times \text{Logit}[-w\{\log(x) - \log(a)\}]$, where x is the tested concentration and (t, b, w, a) parameterizes the Emax, Emin, AC50, and w (rate of increase), respectively. Panel (b) shows six sample basis functions with internal knots $(-4, -2, 0, 2, 4)$ marked with gray circles. This figure appears in color in the electronic version of this article.

Figure 2. Illustration of the ToxCast data structure showing 20 chemicals tested on 8 assays. This is a sample of the 309 chemicals and 81 assays used in Section 5 with curves as reported in the publicly available ToxCast data. All 309 chemicals are tested on all 81 assays, resulting in $309 \times 81 = 25,029$ chemical-assay combinations. This figure appears in color in the electronic version of this article.

Figure 3. Example of ZIPLL and ToxCast estimates for 12 chemicals on the PXRE assay. The ZIPLL posterior mean (thick solid black line) with 95% posterior intervals (dashed black lines), and the ToxCast fit (thin solid gray line) are shown. The legend shows a binary indicator of an active response from ToxCast (1=active) and the posterior probability of an active response using ZIPLL. The FPLL fits are redrawn from the parameter estimated in the ToxCast public release data. This figure appears in color in the electronic version of this article.

Figure 4. Panel (a) plots the number of assays each chemical responded on with 95% posterior intervals. The simultaneous fitting of all chemical-assay combinations allows for the estimation of the joint distribution of assays and naturally propagates the distribution of the total number of assay responses. The $\times$ indicates the number of assay responses reported in the ToxCast data. (The 309 chemical names are omitted on the horizontal axis for readability.) Panel (b) shows posterior mean random intercept for the assay probability of response (δ_j) as discussed in Section 3.2 and 95% posterior intervals. This figure appears in color in the electronic version of this article.

Figure 5. Chemical ranks by potency with 90% credible intervals. The x-axis shows the posterior mean AC50 (more potent to the left). The y-axis ranks the chemicals by potency (most potent, rank one, at top). All chemicals with at least a 0.50 percent probability of active response are plotted. For $PPAR\gamma$ and $ER\alpha$ reference chemicals are marked for comparison. These chemicals have shown documented activity on these assays. This figure appears in color in the electronic version of this article.