Published in final edited form as:

Toxicol Appl Pharmacol. 2017 January 01; 314: 109–117. doi:10.1016/j.taap.2016.11.010.

A Data-driven Weighting Scheme for Multivariate Phenotypic Endpoints Recapitulates Zebrafish Developmental Cascades

Guozhu Zhang¹, Kyle R. Roell¹, Lisa Truong³, Robert L. Tanguay³, and David M. Reif^{1,2,*}

¹Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA

²Department of Biological Sciences, Center for Human Health and the Environment, North Carolina State University, Raleigh, NC, USA

³Department of Environmental and Molecular Toxicology, Sinnhuber Aquatic Research Laboratory, Oregon State University, Corvallis, OR, USA

Abstract

Zebrafish have become a key alternative model for studying health effects of environmental stressors, partly due to their genetic similarity to humans, fast generation time, and the efficiency of generating high-dimensional systematic data. Studies aiming to characterize adverse health effects in zebrafish typically include several phenotypic measurements (endpoints). While there is a solid biomedical basis for capturing a comprehensive set of endpoints, making summary judgments regarding health effects requires thoughtful integration across endpoints. Here, we introduce a Bayesian method to quantify the informativeness of 17 distinct zebrafish endpoints as a data-driven weighting scheme for a multi-endpoint summary measure, called weighted Aggregate Entropy (wAggE). We implement wAggE using high-throughput screening (HTS) data from zebrafish exposed to five concentrations of all 1,060 ToxCast chemicals. Our results show that our empirical weighting scheme provides better performance in terms of the Receiver Operating Characteristic (ROC) curve for identifying significant morphological effects and improves robustness over traditional curve-fitting approaches. From a biological perspective, our results suggest that developmental cascade effects triggered by chemical exposure can be recapitulated by analyzing the relationships among endpoints. Thus, wAggE offers a powerful approach for analysis of multivariate phenotypes that can reveal underlying etiological processes.

*Corresponding author David M. Reif, Bioinformatics Research Center, Center for Human Health and the Environment, Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA, dmreif@ncsu.edu, (p) 919-513-3812, (o) Ricks Hall 356, (m) Box 7566, 1 Lampe Drive, Raleigh NC 27695.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflicts of Interest

None declared.

Author Contributions

All authors participated in writing and editing the manuscript. GZ conceived the wAggE approach, designed analyses, implemented code, and drafted the manuscript. KRR performed the comparative analysis with curve-fitting. LT and RLT designed the HTS experimentation facilities and carried out all zebrafish experiments. DMR oversaw the methods development and manuscript preparation.

Keywords

Zebrafish; High-dimensional; Bayesian; Developmental Cascade; ToxRefDB; Risk Assessment; Multiple Endpoints; Multivariate; Scoring

---

1. Introduction

There are tens of thousands of compounds currently in commerce and the environment worldwide, and while the number is growing rapidly, the toxicity information for humans or other species is still limited to a relatively small number of chemicals (Wambaugh et al. 2013). A major focus of developing alternative toxicity testing methods is to reduce the cost, complexity, labor, time, throughput, and animal welfare issues in traditional animal assays while retaining useful toxicological profiles (Basketter et al. 2012). High-throughput *in vitro* screening assays, such as ToxCast, were developed for chemicals in order to find targeted receptors and expedite toxicity testing (Judson et al. 2010). However, these assays do not provide systemic organismal responses for outcomes such as developmental toxicity. Thus, developing new cost-effective, high-throughput methods to evaluate the hazard information of these compounds is critical.

Zebrafish (*Danio rerio*), a small, vertebrate organism, has been widely used in toxicological research due to benefits such as *ex vivo* development and optical clarity of the embryo, suitability for high-throughput screening (HTS), cost-effectiveness, and rapid sexual maturation of only 3 months (Delvecchio et al. 2011; Truong et al. 2011). The genomic similarity between zebrafish and humans is approximately 70% (Howe et al. 2013), making it an ideal model to aid in understanding toxicity translatable to human health. Moreover, the developmental stages of zebrafish are characterized in fine detail (Kimmel et al. 1995). This allows studies of developmental progression perturbed by exposure to environmental stressors, where diverse behavioral and morphological endpoints can be assessed (Kokel et al. 2010; Noyes et al. 2015; Truong et al. 2014). Analysis across time points and endpoint types can develop or refine Adverse Outcome Pathways (AOPs), inform risk assessment, and build predictive models for systems toxicology (Reif et al. 2015).

Traditional methods of identifying an inflection point along the curve to determine the effective concentration, such as LD50 (50% lethal dose) or AC50 (half-maximal activity concentration), are concentration-dependent and require major assumptions that are highly sensitive to common sources of noise (Beam and Motsinger-Reif, 2014). For example, the response data are typically expected to be monotonic, which is easier to achieve using *in vitro* cell line models, since the phenotypes are singular measurements of fold-change, percent inhibition, or cell death. Bayesian approaches have been applied to fit curves for scenarios where information can be borrowed across large chemical or assay sets (Wilson et al. 2014); however, curve-fitting may not be appropriate for developmental toxicity *in vivo*, largely because it is difficult to assure homogeneity across doses. This happens for several reasons: 1) manifestation of competing AOPs by different concentrations of chemical; 2) censoring by mortality; and 3) developmental cascade effects. Disentangling these factors is analytically challenging, as evidenced by the high mutual information shared across

endpoints (Zhang et al. 2016). Moreover, the majority of the chemicals remain inactive or in constant response, presenting another challenge in identifying concentration-dependence of potential hazards (Truong et al. 2014; Zhang et al. 2016). In order to address these challenges, Aggregate Entropy (AggE) was designed as a concentration-independent method to interpret the overall effect as a point of departure (POD) without differentially weighting specific endpoints (Zhang et al. 2016).

Although several approaches have been used to aggregate information from multiple endpoints into a summary score, there is no consensus on how endpoints should be weighted (Shaw et al. 2016). Most published weighting schemes are heuristics based upon theoretical biological impact and are heavily weighted toward catastrophic endpoints such as lethality or inability to hatch (Harper et al. 2015; Liu et al. 2013; Padilla et al. 2012). In contrast, we take the opposite approach by deriving weights from observed data, then using empirical wAggE weights to explore biological underpinnings. First, we utilize a Bayesian method to quantify the severity of 17 distinct zebrafish endpoints (YSE, AXIS, EYE, SNOU, JAW, OTIC, PE, BRAI, SOMI, PFIN, CFIN, PIG, CIRC, TRUN, SWIM, NC, and TR). Second, we show that wAggE provides superior performance in terms of the ROC curve in identifying significant morphological effects. Third, we explore whether this weighting scheme reveals developmental cascade effects wherein early phenotypes can predict those occurring at later developmental stages. Fourth, we compare developmental scoring in zebrafish and mammalian results from the U.S. EPA's Toxicity Reference Database (ToxRefDB). Finally, we compare wAggE to a logistic-based curve-fitting method.

2. Materials and methods

2.1. Materials and analysis pipeline

The experimental data are described in Truong et al. 2014. Fig. 1 shows a consensus timeline that includes experimental conditions, key early developmental stages and landmarks (Kimmel et al. 1995), and morphological assessments. The data structure and details about AggE are provided in Zhang et al. 2016. ToxRefDB data were downloaded from <https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data> (toxrefdb_v1, October 2014). All analysis was implemented using custom R code (R core team 2016).

2.2. Weighted Aggregate Entropy

In AggE, each biological state (18 assessed endpoints and No Observed Adverse Outcome) of an embryo is scored independently before summarizing across biological states and screened embryos. Briefly, let $X_1 \dots X_{18}$ represent 18 assessed endpoint of an embryo, with 1 indicating present and 0 indicating absent. represents No Observed Adverse Outcome (NOAE) with a value of $X_{19} = 19 - (X_1 \dots + X_{18})$. The score, which is Shannon's entropy

in nats unit, of this embryo is equal to $E = \frac{X_1}{19} * \log(\frac{X_1}{19}) + \dots + \frac{X_{19}}{19} * \log(\frac{X_{19}}{19})$. Thus, by assigning weight to each biological state, for each chemical at a given concentration, the wAggE can be written as:

$$wAggE = \sum_{\text{All screened embryos}} w1 * \frac{X1}{19} * \log\left(\frac{X1}{19}\right) + \dots + w19 * \frac{X19}{19} * \log\left(\frac{X19}{19}\right)$$

Where $w1 \dots w19$ are the weighting factor for each biological state.

2.3. Bayesian logistic regression model

Fisher's Exact Test was applied to determine if a given concentration of chemical can significantly affect an endpoint compared to the negative control (Reif et al. 2015). The response variable is defined as 1 that indicates if a chemical significantly affects any endpoint at a given concentration; 0 indicates no significant effect being observed. The Bayesian logistic regression model is:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \left(E_{\text{MORT}} + \sum_{i=1}^{17} w_i E_i + E_{\text{NOAE}} \right) + \beta_2 (\text{Con}) + \varepsilon$$

Where π is the probability of success; w_i and E_i are shown in Section 2.2, with the exception of two biological states: Mortality and No Observed Adverse Effect, which are shown as constant E_{MORT} and E_{NOAE} , respectively; Con is the tested concentration; ε is the Gaussian noise; β_0 is the intercept; β_1 is the slope of $wAggE$; β_2 is the slope of concentration. The prior of each parameter is shown in Table 1. Mortality and No Observed Adverse Effect are not weighted, because in this assay mortality assigned at 120 hours post fertilization (hpf) overwrote (i.e. set to zero) all sub-lethal endpoints, and No Observed Adverse Effect represents unobserved biological processes during development. Moreover, $AggE$ has less power to identify chemicals that only cause significant mortality (Zhang et al. 2016). So in this study, mortality remains unweighted in an attempt to increase the power to identify significant mortality. The Bayesian computation was processed using R2OpenBUGS package for R, (Sturtz et al. 2005; Sturtz et al. 2010). Our parameters and model set up for R2OpenBUGS are provided in Appendix A. Trace plots, autocorrelation plots, and the Gelman-Rubin statistic were used to determine the burn-in period to ensure the convergence of the posterior distribution (Gelman et al. 2014). All diagnostic plots and statistics were generated by the Coda package in R (Plummer et al. 2006). Our final Bayesian results were achieved by running 5 independent chains, each with a set of random start values for all parameters, 5,000 total iterations per chain, thinning rate of 2, and burn-in length of 4,000.

2.4. Evaluation of $wAggE$ versus the unweighted alternative

We first compare the ROC curve of two logistic regression models (parameters described above), which are:

$$\text{Model I: } \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 (AggE) + \beta_2 (\text{Con}) + \varepsilon;$$

$$\text{Model III: } \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1(\text{wAggE}) + \beta_2(\text{Con}) + \varepsilon$$

Secondly, we carry out K-fold cross validation with $K=\{5,10,15\}$ for both models and compare the error rate. Because our response variable is unbalanced, containing 487 positive hits versus 4,813 negatives by Fisher's Exact Test from Reif et al. 2015, we need to assure that performance is not driven by this unbalanced sample. We retain the 487 positive hits and randomly selected 10% of the negatives over 500 trials to construct new data sets to evaluate the performances of the two models through ROC curve and K-fold cross validation error rate with $K=\{5, 10, 15\}$. Finally, we use a Chi-square distribution to approximate wAggE, determine the significance level, then evaluate its performance in identifying specific chemical-associated morphological effects determined by Fisher's Exact Test (Reif et al. 2015; Zhang et al. 2016). We compare the ROC curve to the results using unweighted AggE. The ROC curve is calculated by the Verification package in R (NCAR-Research Applications Laboratory, 2015).

2.5. Association between zebrafish developmental assessment and ToxRefDB

The file `toxrefdb_endpoint_matrix_AUG2014_FOR_PUBLIC_RELEASE.csv` was used for statistical enrichment analysis. In this summary file, "1000000" is coded as 0 for negative findings, "NA" represents missing data (untested), and all other numbers are coded as 1 for positive findings. The newest version of the ToxRefDB has 883 chemicals with over 1,000 toxicity endpoints, and the overlap between our chemical set and ToxRefDB is 461. The annotation of ToxRefDB endpoints can be summarized via 6 levels. Level 1 is the type of study (e.g. chronic). Level 2 is the species. Level 3 is the effect category (e.g. developmental reproductive). Level 4 is the endpoint life stage in which effect was observed (e.g. adult). Level 5 is the endpoint type, which is a grouping of effects that could represent the observation of developmental malformations. Since wAggE is a measurement of systematic responses in zebrafish, for the current analysis, we only report level 5 endpoints in ToxRefDB. We performed Fisher's enrichment analysis and calculated relative risk and

concordance rate $\left(\frac{TN+TP}{TN+TP+FN+FP}\right)$ between the wAggE POD and all ToxRefDB level 5 endpoints to evaluate connections between different data streams using different organisms.

2.6. AC50 calculation

wAggE was implemented to identify an overall POD in assessing chemical toxicity. We compared it to standard curve fitting methods that derive an AC50 value for experiments like those analyzed here. We quantified the AC50 values by using different minimum endpoint thresholds for several reasons: many of the endpoints are highly correlated as shown in the correlation structure in Fig. 2 (e.g. snout and jaw); potential noise caused by the endpoint calling method; and endpoints of differential severity and sequence during the developmental course of the experiment. If the total observed endpoint(s) within an individual is greater than or equal to the minimum endpoint threshold, that individual is defined as affected. For example, if the minimum endpoint threshold is "at least 2", only the individual that has more than two annotation endpoints is defined as affected. The

individual with annotated mortality is always counted as affected regardless of the minimum endpoint threshold. Our curve fitting data are percentage affected at each concentration of a chemical. The AC50 dose response curves and values were obtained using a Hill model from the ToxCast curve-fitting R package “tcpl 1.0” (Filer et al. 2015).

3. Results

The results of the posterior estimate of each parameter along with its standard error, 95% credible interval, and Gelman-Rubin diagnostic statistic are shown in Table 2. The summary plots are in Appendix B, trace plots after burn in are in Appendix C, and autocorrelation plots are in Appendix D.

From Table 2, we observe that the weights of AXIS, NC, YSE are significantly higher than those of the rest of the endpoints. Fig. 1 shows that these three endpoints correspond to the earliest observable phenotypes, motivating our hypothesis of developmental cascade effects. We tested whether these three highly weighted, early developmental endpoints could predict endpoints developed later in life by calculating the relative risk and the sensitivity (Fig. 3). In this context, the sensitivity (true positive rate) was defined as the conditional probability of observing a specific endpoint, given one of those three endpoints within a chemical-treated individual. The relative risk was the ratio of the sensitivity and false positive rate that was defined as the conditional probability of observing a specific endpoint without the presence of one of those three endpoints within a chemical-treat individual. The dead samples as well as all negative controls were removed prior to estimating those statistics. The relative risk values indicate extremely high predictive power of those three endpoints, with all values significantly greater than unity ($p < 0.05$). For sensitivity values, we set 0.5 as the baseline of a true positive rate. AXIS (Fig. 3A) was predictive for effects involving EYE, JAW, SNOU, PE, BRAI, PFIN and TRUN. NC (Fig. 3B) is a rare endpoint, observed in only 0.55% of individuals. However, NC represents such a core developmental event that it is highly predictive of other endpoints and shows greater than 0.5 sensitivity for all but 3 specific endpoints (Fig. 3B). We found that YSE (Fig. 3C) was highly predictive of EYE, JAW, SNOU, and especially PE, with a true positive rate = 0.8.

As described in section 2.4, we first compared the performance of two logistic regression models using AggE (Model I) and wAggE (Model II). As shown in Fig. 4A, wAggE shows better prediction in terms of ROC. Both wAggE and concentration have a positive relationship with any adverse outcome, which is the response variable. We next carried out K-fold cross validation for both models, and model II (wAggE) presents a lower error rate in any K-fold validation. We next constructed a balanced data set to further confirm that wAggE shows a better ROC curve and a lower cross validation error rate. Finally, we compared wAggE and AggE to identify specific morphological effects. We followed the procedures described in Zhang et al. 2016 to determine the degree of freedom of chi-square approximation to wAggE, and compared wAggE with Fisher's Exact Test on each individual endpoint. In this study, we used the global threshold (rather than concentration-specific thresholds), since the shift of chi-square distributions across concentration is not very large (Zhang et al. 2016). Moreover, by combining concentrations, we can protect the censoring caused by high mortality rates at the highest concentration. wAggE achieved a better ROC

curve across all adverse outcomes at a significance threshold of 0.05 (Fig. 4B). We found that the performance advantage of wAggE over the unweighted version was maintained even as this significance threshold was tuned to favor either sensitivity or specificity.

ToxRefDB contains *in vivo* systematic toxicity data using mammalian models, such as dogs, rodents, and rabbits, for 883 chemicals (Martin et al. 2009). Here we wanted to address whether there is significant association between integrative developmental assessment using zebrafish and those high-level endpoints in ToxRefDB. We report strong significant associations between the two data sources as endpoints having: 1) a relative risk value plus 95% confidence interval greater than 1; and 2) the p-value of Fisher's enrichment analysis less than 0.05. After filtering, we found that there were 6 (out of 87) level 5 endpoints that met the criteria above. These endpoints and concordance rates (in parentheses) were:

“CHR_mouse_SystemicCarcinogenic_adult_PathologyNonProliferative” (46%),

“CHR_mouse_SystemicCarcinogenic_adult_OrganWeight” (55%),

“CHR_mouse_DevelopmentalReproductive_adult_PathologyGross” (83%),

“CHR_rat_SystemicCarcinogenic_adult_OrganWeight” (50%),

“DEV_rat_DevelopmentalReproductive_fetal_DevelopmentalMalformation” (59%), and

“MGR_rat_SystemicCarcinogenic_juvenile_OtherSystemic” (34%).

These mammalian endpoints represent plausible relationships to zebrafish developmental malformations, with differing levels of concordance due to varying prevalence rates, where many chemicals positive in mammalian assays (with top concentrations chosen to ensure positive responses following initial range-finder studies) were not positive at the highest concentration tested (64 uM) in our assay.

We also compared estimated AC50 values (derived from curve fits) to wAggE using prototypical response patterns (Fig. 5). Lovastatin (TX006301), a statin drug prescribed for lowering cholesterol, caused 100% mortality at 6.4uM, and significantly affects many endpoints at 0.64uM. The AC50 value remains constant regardless of varying adverse outcome thresholds (minimum of specific endpoints). In this case, the AC50 value basically describes the degree of mortality, which is similar to those measurements using *in vitro* cell lines. wAggE is censored at 6.4uM due to 100% mortality and provides the lowest effect level of developmental toxicity using zebrafish at 0.64uM. The chemical 6-{2-[4-(12-benzothiazol-3-yl)piperazin-1-yl]ethyl}-448-trimethyl-34-dihydroquinolin-2(1H)-one methanesulfonate (TX006163) significantly affected all endpoints except mortality at 6.4uM and affected all endpoints at 64uM based on Fisher's Exact Test. The AC50 value remains confined to the space between observed concentrations as the adverse outcome threshold goes up, because there is a generally monotonic response distributed uniformly across endpoints. For this canonical, sigmoidal dose-response, wAggE estimated a POD at a higher concentration than the AC50 method. Tiratricol represents a prototypical chemical that only significantly affects a subset of endpoints; however, these endpoints are highly correlated

(see Fig. 2). As the minimum endpoint threshold increases, there are dramatic shifts in the AC50 value. wAggE defined the POD in the middle of the AC50 value range at 0.64uM.

By setting the minimum endpoint threshold to 1, this zebrafish assay is treated as another cytotoxicity assay, which reduces the strength of this systematic response of this zebrafish assay in reproductive studies. More importantly, by setting the minimum endpoint threshold to 1, one can create many false positives (Zhang et al. 2016). Due to the high correlation of many of the endpoints, such as {SNOU, JAW}, {SNOU PE}, {SNOU, EYE}, {JAW, EYE}, {YSE, PE}, {BRAI, EYE}, {SNOU, AXIS}, each has a spearman correlation of 80% or higher in Fig. 2. In order to apply a robust AC50 measurement for grouping of effects, the minimum endpoint threshold should be increased rather than using 1 as the default. Moreover, wAggE is a dose independent risk assessment, which could potentially address the toxicological variances, such as responses in lower concentration but not in higher concentration. Thus, our method provides a better alternative solution regarding the degree of developmental toxicity *in vivo*. The PODs of all chemicals using wAggE with censoring concentration that causes 100% mortality are in Appendix E. If it is inactive at the highest tested concentration, which is 64uM, it is shown as a miss data point.

4. Discussion

In this study, we designed a Bayesian logistic regression model using data from a multivariate zebrafish HTS developmental study to specify differential weights for each endpoint. Our method improves upon arbitrary or heuristic weighting factors determined *a priori* by experts and better quantifies the sequential course of development in implicating differential severity of endpoints measured simultaneously. These developmental cascade effects highlight the importance of quantifying the weights of specific endpoints for integrative risk assessment using zebrafish. Moreover, this weighting scheme implies the degree of importance of those endpoints and highlights how groups of highly-correlated endpoints may be rooted in early developmental events.

We implemented wAggE to identify an overall POD, rather than an inflection point along a fitted curve, because we assert that the expectation of monotonic concentration-dependence should not necessarily hold for experiments like those analyzed here, considering the complexity arising from chemical perturbation of biological pathways. This is due to endpoints arising of differential severity and sequence during the development course of the experiment where assessment is made periodically (e.g. at 24hpf or 120hpf). Besides the apparent “drop-off” caused by mortality censoring, the developmental consequences of a given chemical may manifest differently at increasing concentration when uniformly assessed at 120hpf. For example, if a chemical caused 25% incidence of a specific endpoint at a low concentration, higher concentrations of that same chemical may elicit more serious endpoints that obscure the observation of the same endpoint, thus creating apparent non-monotonicity. Additional variation can be introduced by technical artifacts, including impurities in chemical stocks or imperfect endpoint calling techniques. Our results demonstrate that wAggE, which does not require homogeneity or monotonicity across doses, provides solid detection power and that the data-driven assignment of endpoint weights can

recapitulate developmental cascade effect that would otherwise confuse methods require monotonicity.

The monotonicity of lethality endpoints (MORT in our study) creates response patterns for which traditional sigmoidal curves are well-suited. This monotonicity arises from the ultimate endpoint that lethality represents, plus the fact that measurement error is minimal in an endpoint such as MORT, generating dose-response pictures that tend to be “well-behaved”. However, for HTS applications, the concentration spacing may be wide. In these situations, wAggE will be superior to curve fits in dealing with apparent toxicological spikes/cliffs in an ultimate lethality endpoint by providing information on the accumulation of sub-lethal endpoints and by not statistically penalizing steep slopes. An extension of wAggE to be explored in future work is the derivation of alternative critical concentrations, which would position the method as a hybrid between curve-fit and strict point-of-departure approaches. Because the AggE statistic presents a continuous value at each tested concentration, fits could be applied across the data range to interpolate between observed values, then compared to the significance threshold to estimate critical concentrations analogous to any AC/IC statistic (e.g. AC10, AC50, AC80).

Comparing wAggE and AggE, we found that wAggE improved the overall prediction as well as prediction of specific morphological effects (Fig. 4). Distributionally, wAggE tends toward right-skewness versus AggE, meaning that it may overestimate the degrees of freedom in the Chi-square approximation for threshold-setting. This may be due to developmental cascade effects that deflate weights for consequent endpoints. Therefore, we suggest using the global threshold (see Methods) when determining significance of wAggE if the goal is detection power (i.e. avoidance of false negatives).

Zebrafish have been shown as a potential alternative model for traditional *in vivo* toxicity testing. Previous research has demonstrated the ability of building predictive models using zebrafish assays to identify teratogenic potential chemicals (Brannen et al. 2010; Selderslaghs et al. 2009; Selderslaghs et al. 2012). A meta-analysis showed that zebrafish can accurately predict many mammalian endpoints, such as rodent developmental defects and lethality (Ducharme et al. 2015). Here, our statistical association analyses between wAggE and high-level ToxRefDB endpoints showed that whole-organism morphological screening in zebrafish provides a useful alternative to traditional animal studies. For instance, Benomyl, which significantly affected zebrafish endpoints including YSE, AXIS, SNOU, EYE, JAW, CFIN, TRUN, and TR, also affected skeletal, cranial, and axial ToxRefDB endpoints in rats.

When combined with new technologies built to speed the pace of toxicity testing and characterize MoA *in vitro* (Collins et al. 2008; Judson et al. 2010; Kleinstreuer et al. 2014), chemical testing using zebrafish provides systematic phenotypic responses that can shed light on etiology of neurotoxicities, teratogenicities, or other adverse outcomes, and perhaps even suggest new targets for *in vitro* assays (Bugel et al. 2014; Garcia et al. 2016; MacRae and Peterson, 2015; Rihel et al. 2010; Tanguay et al. 2014). Given realistic resource limits and the complexity of underlying toxicity mechanisms, bioinformatical approaches for integrating data will be essential to advancing such goals. Developmental concepts such as

the Organization-Activation model suggest that biology itself integrates across multiple scales (Arnold and Breedlove, 1985; Phoenix et al. 1959). Using integrative approaches such as wAggE to appropriately weight combinations of morphological and/or behavioral phenotypes plus targeted *in vitro* data could help identify novel developmental cascade effects resulting from early exposure.

Because the information-theoretic approach underlying Aggregate Entropy is robust to irregular response patterns, wAggE should be applicable to any domain where screening with multiple endpoints is used to derive an overall score, such as nanomaterial testing, ecotoxicology assessments, or HTS applications. For such future applications, expert knowledge could be incorporated into the priors to optimize the balance between *a priori* information and empirically-driven weights. wAggE could be used for *in silico* analysis of similar compounds, where weights could be informative for interpretation of cheminformatic models. In summary, we present weighted Aggregate Entropy as a robust statistical approach for multiple endpoint data that can elucidate developmental cascade effects of endpoints measured simultaneously.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIEHS grants R01 ES19604, R01 ES023788, P42 ES005948, P30 ES025128, RC4 ES019764 P30, P30 ES000210, P42 ES016465, 5T32ES007329, and Environmental Protection Agency (EPA) STAR Grants #835168 and #83579601.

Abbreviations

AC50	half-maximal activity concentration
AOP	Adverse Outcome Pathway
AggE	Aggregate Entropy
EZ	Embryonic Zebrafish
HTS	High-throughput Screening
hpf	hours post fertilization
LD50	50% lethal dose
MoA	Mechanisms of action
POD	Point of Departure
ROC	Receiver Operating Characteristic
ToxRefDB	Toxicity Reference Database
wAggE	Weighted Aggregate Entropy

TN	True Negative
TP	True Positive
FN	False Negative
FP	False Negative
MORT	Mortality
YSE	Yolk sac edema
AXIS	Body axis
EYE	Eye
SNOU	Snout
JAW	Jaw
OTIC	Otic vesicle
PE	Pericardial edema
BRAI	Brain
SOMI	Somite
PFIN	Pectoral fin
CFIN	Caudal fin
PIG	Pigmentation
CIRC	Circulation
TRUN	Truncated body
SWIM	Swim bladder
NC	Notochord & Bent tail
TR	Touch response

References

1. Arnold AP, Breedlove SM. Organizational and activational effects of sex steroids on brain and behavior: a reanalysis. *Horm Behav.* 1985; 19(4):469–498. [PubMed: 3910535]
2. Basketter DR, Clewell H, Kimber I, Rossi A, Blaauboer B, Burrier R, Daneshian M, Goldberg A, Hasiwa N, et al. A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing – t4 report. *Altex.* 2012; 29(1):3–91. [PubMed: 22307314]
3. Beam A, Motsinger-Reif A. Beyond IC₅₀s: Towards Robust Statistical Methods for in vitro Association Studies. *J Pharmacogenomics Pharmacoproteomics.* 2014; 2(120) 2153-0645.10001.
4. Brannen KC, Panzica-Kelly JM, Danberry TL, Augustine-Rauch AA. Development of a zebrafish embryo teratogenicity assay and quantitative prediction model. *Birth Defects Research (Part B).* 2010; 89:66–77.

5. Bugel SM, Tanguay RL, Planchart A. Zebrafish: A marvel of high-throughput biology for 21st century toxicology. *Curr Envir Health*. 2014; 1:341–352.
6. Collins FS, Gray GM, Bucher JR. Toxicology. Transforming environmental health protection. *Science*. 2008; 319(5865):906–907. [PubMed: 18276874]
7. Delvecchio C, Tiefenbach J, Krause HM. The zebrafish: a powerful platform for in vivo, HTS drug discovery. *Assay Drug Dev Technol*. 2011; 9(4):354–364. [PubMed: 21309713]
8. Ducharme NA, Reif DM, Gustafsson J, Bondesson M. Comparison of toxicity values across zebrafish early life stages and mammalian studies: Implications for chemical testing. *Reproductive Toxicology*. 2015; 55:3–10. [PubMed: 25261610]
9. Filer, DL.; Kothiya, P.; Setzer, WR.; Judson, RS.; Martin, MT. The ToxCast™ Analysis Pipeline: An R Package for Processing and Modeling Chemical Screening Data. 2015. https://www.epa.gov/sites/production/files/2015-08/documents/pipeline_overview.pdf
10. Garcia GR, Noyes PD, Tanguay RL. Advancement in zebrafish applications for 21st century toxicology. *Pharmacology & Therapeutics*. 2016; 161:11–21. [PubMed: 27016469]
11. Gelman, A.; Carlin, JB.; Stern, HS.; Dunson, DB.; Vehtari, A.; Rubin, DB. Bayesian Data Analysis. 3rd. ISBN: Chapman & Hall/CRC Press; 2014. 978-1439840955
12. Harper B, Thomas D, Chikkagoudar S, Baker N, Tang K, Heredia-Langner A, Lins R, Harper S. Comparative hazard analysis and toxicological modeling of diverse nanomaterials using the embryonic zebrafish (EZ) metric of toxicity. *J Nanopart Res*. 2015; 17(6):250. [PubMed: 26069453]
13. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*. 2013; 496:498–503. [PubMed: 23594743]
14. Judson RS, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Mortensen HM, Reif DM, Rotroff DM, Shah I, Richard AM, et al. In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environ Health Perspect*. 2010; 118:485–492. [PubMed: 20368123]
15. Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. Stages of embryonic development of the zebrafish. *Dev Dyn*. 1995; 203:253–310. [PubMed: 8589427]
16. Kleinstreuer NC, Yang J, Berg EL, Knudsen TB, Richard AM, Martin MT, Reif DM, Judson RS, Polokoff M, Dix DJ, Kavlock RJ, Houck KA. Phenotypic screening of the ToxCast chemical library to classify toxic and therapeutic mechanisms. *Nature Biotechnology*. 2014; 32:583–591.
17. Kokel D, Bryan J, Laggner C, White R, Cheung CY, Mateus R, Healey D, Kim S, Werdich AA, Haggarty SJ, et al. Rapid behavior-based identification of neuroactive small molecules in the zebrafish. *Nat Chem Biol*. 2010; 6:231–237. [PubMed: 20081854]
18. Liu X, Tang K, Harper S, Harper B, Steevens JA, Xu R. Predictive modeling of nanomaterials exposure effects in biological systems. *International Journal of Nanomedicine*. 2013 <http://dx.doi.org/10.2147/IJN.S40742>.
19. MacRae CA, Peterson RT. Zebrafish as tools for drug discovery. *Nature Reviews Drug Discovery*. 2015; 14:721–731. [PubMed: 26361349]
20. Martin MT, Judson JS, Reif DM, Kavlock RJ, Dix DJ. Profiling chemicals based on chronic toxicity results from the U.S. EPA ToxRef Database. *Environmental Health Perspectives*. 2009; 117(3):392–399. [PubMed: 19337514]
21. NCAR – Research Application Laboratory. verification: Weather Forecast Verification Utilities. 2015. R package version 1.42. <https://CRAN.R-project.org/package=verification>
22. Noyes PD, Haggard DE, Gonnerman GD, Tanguay RL. Advanced Morphological Behavioral Test Platform Reveals Neurodevelopmental Defects in Embryonic Zebrafish 20 Exposed to Comprehensive Suite of Halogenated and Organophosphate Flame Retardants. *Toxicological Sciences*. 2015; 145(1):177–195. [PubMed: 25711236]
23. Padilla S, Corum D, Padnos B, Hunter DL, Beam A, Houck KA, Sipes N, Kleinstreuer N, Knudsen T, Dix DJ, Reif DM. Zebrafish Developmental Screening of the ToxCast™ Phase I Chemical Library. *Reprod Toxicol*. 2012; 33(2):174–187. [PubMed: 22182468]

24. Phoenix CH, Goy RW, Gerall AA, Young WC. Organizing action of prenatally administered testosterone propionate on the tissues mediating mating behavior in the female guinea pig. *Endocrinology*. 1959; 65:369–382. [PubMed: 14432658]
25. Plummer M, Best N, Cowles K, Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*. 2006; 6(1):7–11.
26. R Core Team R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Austria: Vienna; 2016. URL <http://www.R-project.org/>
27. Reif DM, Truong L, Mandrell D, Marvel S, Zhang G, Tanguay RL. High-throughput Characterization of Chemical-associated Embryonic Behavioral Changes Predicts Teratogenic Outcomes. *Arch Toxicol*. 2015
28. Rihel J, Prober DA, Arvanites A, Lam K, Zimmerman S, Jang S, Haggarty SJ, Kokel D, Rubin LL, Peterson RT, Schier AF. “Zebrafish behavioral profiling links drugs to biological targets and rest/wake regulation. *Science*. 2010; 327(5963):348–351. [PubMed: 20075256]
29. Selderslaghs IWT, Rompay ARV, Coen WD, Witters HE. Development of a screening assay to identify teratogenic and embryotoxic chemicals using the zebrafish embryo. *Reproductive Toxicology*. 2009; 28(3):308–320. [PubMed: 19447169]
30. Selderslaghs IWT, Blust R, Witters HE. Feasibility study of the zebrafish assay as an alternative method to screen for developmental toxicity and embryotoxicity using a training set of 27 compounds. *Reproductive Toxicology*. 2012; 33:142–154. [PubMed: 21871558]
31. Shaw BJ, Liddle CC, Windeatt KM, Handy RD. A Critical Evaluation of The Fish Early-life Stage Toxicity Test for Engineered Nanomaterials: Experimental Modifications and Recommendations. *Arch Toxicol*. 2016
32. Sturtz, S.; Ligges, U.; Gelman, A. R2OpenBUGS: a package for running OpenBUGS from R. 2010. URL <http://cran.rproject.org/web/packages/R2OpenBUGS/vignettes/R2OpenBUGS.pdf>
33. Sturtz S, Ligges U, Gelman A. R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*. 2005; 12(3):1–16. <http://hdl.handle.net/10022/AC:P:15341>.
34. Tanguay RL, Truong L, Reif DM, St Mary L, Geier MC, Hao TD. Using embryonic zebrafish and multidimensional screening to evaluate neurobehavioral toxicity and teratology. *Birth Defects Research Part A- Clinical and Molecular Teratology*. 2014; 100(5):377–377.
35. Truong, L.; Harper, SL.; Tanguay, RL. Evaluation of embryotoxicity using the zebrafish model. In: Jean-Gautier, Charles, editor. *Drug safety evaluation: methods and protocols, methods in molecular biology*. 2011. p. 271–279.
36. Truong L, Reif DM, St Mary L, Geier MC, Truong HD, Tanguay RL. Multidimensional In Vivo Hazard Assessment Using Zebrafish. *Toxicological Sciences*. 2014; 137(1):212–233. [PubMed: 24136191]
37. Wambaugh JF, Setzer RW, Reif DM, Gangwal S, Mitchell-Blackwood J, Arnot JA, Joliet O, Frame A, Rabinowitz J, Knudsen TB, Judson RS, Egephy P, Vallero D, Cohen Hubal EA. High-throughput Models for Exposure-based Chemical Prioritization in the ExpoCast Project. *Environ Sci Technol*. 2013; 47(15):8479–8488. [PubMed: 23758710]
38. Wilson A, Reif DM, Reich BJ. Hierarchical dose-response modeling for high-throughput toxicity screening of environmental chemicals. *Biometrics*. 2014; 70:237–246. [PubMed: 24397816]
39. Zhang G, Marvel S, Truong L, Tanguay RL, Reif DM. Aggregate entropy scoring for quantifying activity across endpoints with irregular correlation structure. *Reproductive toxicology*. 2016

Highlights

- Introduced a data-driven weighting scheme for multiple phenotypic endpoints.
- Weighted Aggregate Entropy (wAggE) implies differential importance of endpoints.
- Endpoint relationships reveal developmental cascade effects triggered by exposure.
- wAggE is generalizable to multi-endpoint data of different shapes and scales.

Fig. 1. Zebrafish developing timeline from fertilization to 120hpf

Zebrafish key early developmental stages and associated landmarks. Environmental conditions prior to phenotypic assessments (18 distinct endpoints) are indicated on the top. Timeline of observable phenotypes are listed on the bottom. Only phenotypes that match our data are included.

Fig 2. Correlation structure across 18 endpoints

For each chemical, controls were not included while processing correlation, and incidences were aggregated by concentration by sum function. Upper triangular indicates size proportional Pearson correlation. Lower triangular indicates scatterplot with red line as linear fit.

Fig. 3. Developmental cascade effects: Using AXIS (A), NC (B), and YSE (C) to predict the rest of the endpoints

In each part, sensitivity is presented to the left. The red line is drawn at 0.5 for reference.

Relative risk values and their 95% confidence intervals are shown at the right.

Fig. 4. Performance of AggE versus wAggE in terms of ROC

AggE is plotted in black. wage is plotted in red. **A:** Bayesian models comparison. X axis: False Positive Rate; Y axis: True Positive Rate. **B:** Using AggE and wAggE to predict individual morphological effect. X axis: Endpoint; Y axis: ROC curve.

Fig. 5. Curve fitting method (AC50) vs Point of departure (wAggE)

A: AC50 values using various response types and different thresholds. Log transformed concentrations are -5.05, -2.75, -0.45, 1.86, 4.16 for 0.0064μM, 0.064μM, 0.64μM, 6.4μM, and 64μM, respectively. **B:** Point of departure using wAggE along with individual morphological effects.

Table 1

Parameters and their associated priors.

Parameter	Description	Prior
β_0	Interception	$N(-5, 2)$
β_1	Coefficient of wAggE	$N(0, 2)$
β_3	Coefficient of Concentration	$N(0, 1)$
$w_1 \sim w_{17}$	Weight of Endpoint EYE, SNOU, JAW, AXIS, YSE, PE, SOMI, SWIM, CIRC, TR, FIG, PFIN, BRAI, OTIC, NC, CFIN, TRUN, respectively.	$U(0, 1)$

Table 2

Summary of Bayesian posterior inference.

Parameter	Description	Bayesian Estimate (Empirical mean)	Empirical Standard Deviation	95% Credible Interval	Gelman-Rubin Diagnostic Statistic
β_0	Interception	-6.29	0.18	(-6.66, -5.93)	1.00
β_1	Coefficient of wAggE	1.31	0.06	(1.19, 1.43)	1.00
β_2	Coefficient of Concentration	0.01	0.003	(0.005, 0.016)	1.00
w_1	Weight of EYE	0.07	0.07	(0.002, 0.25)	1.01
w_2	Weight of SNOU	0.20	0.16	(0.008, 0.59)	1.00
w_3	Weight of JAW	0.38	0.22	(0.03, 0.86)	1.00
w_4	Weight of AXIS	0.81	0.14	(0.46, 0.99)	1.00
w_5	Weight of YSE	0.74	0.16	(0.39, 0.99)	1.00
w_6	Weight of PE	0.10	0.09	(0.004, 0.35)	1.01
w_7	Weight of SOMI	0.09	0.09	(0.002, 0.34)	1.01
w_8	Weight of SWIM	0.11	0.11	(0.002, 0.41)	1.02
w_9	Weight of CIRC	0.17	0.16	(0.004, 0.57)	1.01
w_{10}	Weight of TR	0.24	0.18	(0.01, 0.67)	1.01
w_{11}	Weight of PIG	0.11	0.10	(0.003, 0.38)	1.01
w_{12}	Weight of PHIN	0.08	0.08	(0.002, 0.28)	1.02
w_{13}	Weight of BRAI	0.07	0.07	(0.002, 0.25)	1.01
w_{14}	Weight of OTIC	0.08	0.08	(0.002, 0.29)	1.01
w_{15}	Weight of NC	0.73	0.22	(0.20, 0.99)	1.00

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Parameter	Description	Bayesian Estimate (Empirical mean)	Empirical Standard Deviation	95% Credible Interval	Gelman-Rubin Diagnostic Statistic
w_{16}	Weight of CFIN	0.16	0.15	(0.005, 0.56)	1.00
w_{17}	Weight of TRUN	0.17	0.14	(0.008, 0.53)	1.00