Reproductive Toxicology 62 (2016) 92–99

Contents lists available at ScienceDirect

Reproductive Toxicology

journal homepage: www.elsevier.com/locate/reprotox

Aggregate entropy scoring for quantifying activity across endpoints with irregular correlation structure

Guozhu Zhang [a], Skylar Marvel [a], Lisa Truong [c], Robert L. Tanguay [c], David M. Reif [a,b,*]

[a] Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA
[b] Department of Biological Sciences, Center for Human Health and the Environment, North Carolina State University, Raleigh, NC, USA
[c] Department of Environmental and Molecular Toxicology, Sinnhuber Aquatic Research Laboratory, Oregon State University, Corvallis, OR, USA

ARTICLE INFO

Article history:
Received 30 December 2015
Received in revised form 23 March 2016
Accepted 15 April 2016
Available online 27 April 2016

Keywords:
Developmental neurotoxicology
Chemical biology
Morphology
Zebrafish
High throughput screening
ToxCast
Multiplexed assays

ABSTRACT

Robust computational approaches are needed to characterize systems-level responses to chemical perturbations in environmental and clinical toxicology applications. Appropriate characterization of response presents a methodological challenge when dealing with diverse phenotypic endpoints measured using *in vivo* systems. In this article, we propose an information-theoretic method named Aggregate Entropy (AggE) and apply it to scoring multiplexed, phenotypic endpoints measured in developing zebrafish (*Danio rerio*) across a broad concentration-response profile for a diverse set of 1060 chemicals. AggE accurately identified chemicals with significant morphological effects, including single-endpoint effects and multi-endpoint responses that would have been missed by univariate methods, while avoiding putative false-positives that confound traditional methods due to irregular correlation structure. By testing AggE in a variety of high-dimensional real and simulated datasets, we have characterized its performance and suggested implementation parameters that can guide its application across a wide range of experimental scenarios.

1. Introduction

Biological responses in whole animals are the product of coordinated actions (or, in the case of toxic responses, dysregulation) on a systemic level. Accordingly, experimental inquiries into basic biological processes should record multiple phenotypic outcomes when assessing perturbations, from clinical interventions such as drug treatments to environmental stressors such as manufactured chemicals. Innovations in multiplexed endpoint measurement technology and exploratory omics platforms have enabled theoretically comprehensive experiments to be conducted [1]. However, these new, multi-endpoint data present challenges with respect to recapitulating the relevant biological processes: (1) The correlation structure across endpoints is irregular; (2) Individual subjects/samples vary in endpoint presentation; (3) Endpoint measurement methods are imperfect; (4) Experimental questions may depend on subsets and/or recombinations of endpoints. Therefore, analysis methods are needed that can address these challenges while allowing for either focused, *a priori* analysis or data-wide, empirical analysis.

One such area where comprehensive analysis of systemic response is needed is environmental and clinical toxicology, where adverse responses may manifest anywhere from specific abnormalities to collections of several endpoints that count as toxicity in the aggregate. While there is an ever-increasing number of chemicals in commerce and the environment, comprehensive toxicological knowledge is lacking for all but a handful of compounds—mostly pharmaceuticals that have progressed to expensive, late-stage clinical trials. Traditional animal testing is very expensive in terms of labor, time, and money, so high-throughput screening (HTS) is being developed in order to more efficiently assess chemical biocompatibility [2]. Experimental HTS includes both *in vitro* assays that probe molecular action and *in vivo* assays that screen for a

Abbreviations: 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, pigment; CIRC, circulation; TRUN, truncated body; SWIM, swim bladder; NC, notochord & bent tail; TR, touch response; NOAE, no observed adverse effect; AggE, Aggregate Entropy; PP, positive-positive; NN, negative-negative; NP, negative-positive; PN, positive-negative; Any End, any endpoint; ROC, receiver operating characteristic; hpf, hours post fertilization; AOP, adverse outcome pathway; HTS, high throughput screening; MI, mutual information; SE, super endpoint.

* Corresponding author at: Bioinformatics Research Center, Center for Human Health and the Environment, Department of Biological Sciences, North Carolina State University, Raleigh, NC, Box 7566, 1 Lampe Drive, Raleigh NC 27695, USA.
E-mail addresses: gzhang6@ncsu.edu (G. Zhang), swmarvel@ncsu.edu (S. Marvel), lisa.truong@oregonstate.edu (L. Truong), robert.tanguay@oregonstate.edu (R.L. Tanguay), dmreif@ncsu.edu (D.M. Reif).

http://dx.doi.org/10.1016/j.reprotox.2016.04.012

variety of phenotypic endpoints that cover fundamental developmental, structural, and neurological pathways [3–5].

These HTS *in vivo* assays provide an ideal workbench for the development and testing of analysis methods for multiple endpoints, in that the data can be generated on a scale that permits evaluation of an analysis method's ability to address the four challenges presented above. In particular, experimental methods for the zebrafish (*Danio rerio*), a model organism whose fundamental developmental processes are shared across vertebrates and that has high genetic similarity to humans, have exploded in recent years [6,7]. Several endpoints, ranging from specific structural features through outright mortality, have been measured, with a trend toward higher-order assessment of multiple endpoints during embryonic development [8].

Here, we developed an information theory-based method named Aggregate Entropy ("AggE") to consolidate information into classes across endpoints, then tested this method using both simulated and empirical zebrafish data. We characterized the relationship amongst endpoints to identify the biological processes underlying overall developmental assessments; used simulated data to further validate our method across a range of sample sizes; characterized the irregular correlation structure across endpoints using mutual information and normalized information distance; and used this information to reduce noise by collapsing endpoints with similar phenotypic response patterns. Finally, we parameterized AggE distributions to allow for application to new datasets of varying dimensions from multi-endpoint experiments in any model system.

2. Materials and methods

2.1. Empirical data

The empirical data were collected as described in Truong et al. [5] and Noyes et al. [9]. Fig. 1 shows the experimental design and data structure. The data include 1060 unique ToxCast chemicals tested at six concentrations for each chemical (0 μM, 0.0064 μM, 0.064 μM, 0.64 μM, 6.4 μM and 64 μM). There were n = 32 replicates (individual embryo wells) at each concentration. At 120 h post fertilization (hpf), 18 distinct developmental endpoints were evaluated. The data were recorded as binary incidences.

As in Fig. 1(B) and (C), we constructed 19 different biological states, including 18 developmental endpoints plus one NOAE (No Observed Adverse Effect) state. Thus, for each embryo per chemical-per concentration, data were shown as 0 and 1 for 18 binary endpoints with NOAE recorded as $19 - \sum(BinaryEndpoints)$. All analysis was performed using R [13].

2.2. Aggregate Entropy

The traditional Shannon's entropy $H(X)$ [14], in nat units, is:

Let X be a discrete random variable with a possible set of realizations x, thus;

$$H(X) = -\sum_{x} p(x)\log_e p(x)$$

We define a random variable and its realizations as follows:

For each chemical C at a given concentration, let X_i represent embryo i with $i = 1, \ldots, 32$ and B_j represent biological state j with$j = 1, \ldots, 19$. In addition, X_i has realization x_{ij} with its sample value shown in Fig. 1. The probability mass function can be written as:

$$p\left(B_j|C, X_i\right) = \frac{x_{ij}}{19}$$

The Aggregate Entropy (AggE) for chemical C at a given concentration is summarizing the Shannon's entropy of all tested embryos, which is:

$$AggE = -\sum_{i=1}^{32}\sum_{j=1}^{19} p\left(B_j|C, X_i\right)\log_e\left\{p\left(B_j|C, X_i\right)\right\}$$

2.3. Threshold determination

We first used a chi square approximation to the distribution of AggE of each concentration as well as the distribution of the pooled concentration [15,16]. We estimated our chi square degree of freedom by using the Newton algorithm to optimize the logarithm of the full likelihood of a chi square probability density function. Let $(AggE_1, AggE_2, \ldots, AggE_N)$ be a set of AggE, thus the full likelihood can be written as:

$$f(AggE_1, AggE_2, \ldots, AggE_N) = \left(\frac{1}{2^{\frac{k}{2}}\Gamma(k)}\right)^n \times (AggE_1 * \ldots * AggE_N)^{\frac{k}{2}} e^{-\frac{AggE_1+\ldots+AggE_N}{2}}$$

where k is the degree of freedom of a Chi-square distribution and N is the number of chemicals. Since the maximum likelihood estimator is nonlinear, we first took the negative logarithm of the full likelihood. After that, given a start value, we used Newton iteration to optimize the negative logarithm of the full likelihood such that it gave us the optimal estimate of the degree of freedom of our chi square distribution. Our threshold, which depends on the observed incidences of multiple measurements over many individuals, is the critical value of a one-sided chi square test with the significance level of 0.05.

2.4. Endpoint clustering and sensitivity analysis

We next used pairwise mutual information to characterize the relationship among endpoints. Let E_1 andE_2 represent two endpoints with realization $e_1$and $e_2$as observed incidence counts per chemical-per concentration, given the Shannon's entropy defined above, the joint Shannon's entropy forE_1 and E_2 is:

$$H(E_1, E_2) = -\sum_{e_1}\sum_{e_2} p(e_1, e_2)\log_e p(e_1, e_2)$$

And the conditional entropy can be written as:

$$H(E_1|E_2) = -\sum_{e_1}\sum_{e_2} p(e_1, e_2)\log_e p(e_1|e_2)$$

With all these definitions, the mutual information (MI) is:

$$MI(E_1, E_2) = \sum_{e_1}\sum_{e_2} p(e_1, e_2)\log_e \frac{p(e_1, e_2)}{p(e_1)p(e_2)} = H(E_1) - H(E_1|E_2)$$

MI has the following, commutative, property:

$$MI(E_1, E_2) = MI(E_2, E_1)$$

We formed our clusters based on a modified three-step measurement [17]. First, the pairwise mutual information between endpoints, $MI\left(E_i, E_j\right), i, j = 1, \ldots, 18$, is calculated by using R package "infotheo" [18]. Next, the mutual information matrix is transferred to a distance measurement, called normalized information distance [19], which is:

$$d\left(E_i, E_j\right) = 1 - \frac{MI\left(E_i, E_j\right)}{H(E_i) + H\left(E_j\right) + MI\left(E_i, E_j\right)}$$