EPA Public Access

Author manuscript

Chemosphere. Author manuscript; available in PMC 2018 December 01.

About author manuscripts

Submit a manuscript

Published in final edited form as:

Chemosphere. 2017 December ; 188: 49–59. doi:10.1016/j.chemosphere.2017.08.136.

A statistical framework for applying RNA profiling to chemical hazard detection

Mitchell S. Kostich

USEPA/ORD/NERL/EMMD, 26 West M. L. King Drive, Cincinnati, OH 45268, USA

Abstract

Use of ‘omics technologies in environmental science is expanding. However, application is mostly restricted to characterizing molecular steps leading from toxicant interaction with molecular receptors to apical endpoints in laboratory species. Use in environmental decision-making is limited, due to difficulty in elucidating mechanisms in sufficient detail to make quantitative outcome predictions in any single species or in extending predictions to aquatic communities. Here we introduce a mechanism-agnostic statistical approach, supplementing mechanistic investigation by allowing probabilistic outcome prediction even when understanding of molecular pathways is limited, and facilitating extrapolation from results in laboratory test species to predictions about aquatic communities. We use concepts familiar to environmental managers, supplemented with techniques employed for clinical interpretation of ‘omics-based biomedical tests. We describe the framework in step-wise fashion, beginning with the single test replicates of a single RNA variant, then extending to multi-gene RNA profiling, collections of test replicates, and integration of complementary data. In order to simplify the presentation, we focus on using RNA profiling for distinguishing presence versus absence of chemical hazards, but the principles discussed can be extended to other types of ‘omics measurements, multi-class problems, and regression. We include a supplemental file demonstrating many of the concepts using the open source R statistical package.

Keywords

aquatic; chemical; biomarker; microarray; RNA-seq

Address correspondence to: kostich.mitchell@epa.gov, phone: +1 513 569 7645.

DISCLAIMER

The views expressed in this article are those of the author and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency. Any mention of trade names, products, or services does not imply an endorsement by the U.S. Government or the U.S. Environmental Protection Agency. The EPA does not endorse any commercial products, services, or enterprises.

SUPPLEMENTAL FILE

Demonstrates many concepts described in this paper (variable selection, binary classification, CV, AUCs, pAUCs, performance comparison, likelihood ratios) using the free, open-source statistical programming package R and a publically available dataset generated from fish exposed to an insecticide. We have successfully adapted this generic workflow to a variety of other ‘omics datasets.

1. INTRODUCTION

1.1. Purpose and scope

Use of ‘omics technologies in environmental sciences, particularly for studying chemical toxicity, has expanded dramatically in recent years (Kumar and Denslow, 2016). Nevertheless, incorporation of ‘omics into environmental management decisions has been limited. Use in decision making may be facilitated by a framework for translating ‘omics measurements into traditional endpoints (survival, growth, and reproduction) in species important to environmental managers. This paper introduces a mechanism-agnostic statistical approach for translating biochemical profiling into an environmental decision-support tool. The approach complements the application of ‘omics measurements for elucidating molecular pathways leading from toxicant exposure to apical endpoints of interest (Ankley et al., 2009).

We focus on RNA profiling, due to its widespread use. With slight modification, the discussion applies to other high-throughput biological measurements, such as proteomics, metabolomics, epigenetics, and metagenomics. Many of the principles discussed apply to any ‘diagnostic’ test or predictive model, including mechanistic models. In order to reduce the complexity of the presentation, we focus on distinguishing organisms exposed to a toxicant from non-exposed organisms, however the discussion can be extended to multi-class problems or regression.

Existing reviews describe RNA measurement methods (Bustin, 2002; Hardiman, 2004; Wang et al., 2009), preprocessing, and generating lists of differentially expressed genes (Pan, 2002). We focus on the downstream process of classification (assigning samples to one of several possible classes), classifier (classification algorithm) evaluation, and interpretation. The principles discussed are often used for the development and interpretation of medical diagnostic tests (Shapiro, 1999). This seems appropriate since medical practice, like environmental decision making, requires decisions in the absence of certainty, relying on incomplete information from a variety of sources.

1.2. RNA profiling

‘RNA profiling’ means measuring the abundance of RNA variants (usually distinguished based on originating gene or exons) in a biological sample. RNA profiling is typically performed using either RT-PCR (Bustin, 2002), global microarrays (Hardiman, 2004), or RNA sequencing (Wang et al., 2009). RT-PCR is used to quantify between one and several dozen RNA variants per sample, with costs scaling linearly with the number of RNA variants. When monitoring more variants, global cDNA microarrays and RNA-sequencing, which can measure the abundance of nearly all the RNA variants present in a sample, become cheaper alternatives. Technological advances have rapidly reduced sequencing costs, and sequencing is becoming the most cost effective method for global RNA profiling.

Comparing global RNA profiles between differentially treated biological samples (e.g. unexposed vs. exposed) allows identification of RNAs that change abundance in response to treatment. Toxicological theory suggests that both adaptive and pathological responses are induced by exposure (Fielden and Zacharewski, 2001), and that molecular details can vary

between toxicant classes. Therefore, distinctive biomolecular signatures might be found for different toxicant classes and perhaps distinct chemical structures within each class (Afshari et al., 1999). Even toxicants which do not directly interact with transcription factors, often indirectly induce distinctive changes in RNA profile (Fielden and Zacharewski, 2001). However, transcriptional profiling provides a limited view of physiological changes following stressor exposure, because it only reflects changes in RNA abundance. Gene activity is also often modulated by means not involving RNA abundance (Day and Tuite, 1998), such as translation rate, protein half-life regulation, or protein modification. Therefore, many genes participating in physiological networks affected by stressor exposure may not be identifiable by RNA profiling.

Nevertheless, research shows that RNA profiles can be associated with disease states and toxicant exposures. Biomedical applications include RNA profiling of tumors to predict their aggressiveness and response to treatment (Quackenbush, 2006). Environmental applications include quantification of vitellogenin (VG) RNAs in male fish in order to detect exposure to environmental estrogens (Sumpter and Jobling, 1995). More complex multi-gene signatures have been developed for detecting e.g. hepatotoxins (Bulera et al., 2001; Waring et al., 2001; Thomas et al., 2001; Hamadeh et al., 2002; Steiner et al., 2004), metals (Poynton et al., 2008; Svendsen et al., 2008; Nota et al., 2010; Bougas et al., 2013), endocrine disruptors (Lam et al., 2008; Wang et al., 2008; Wang et al., 2012), explosives (Li et al., 2010; Gong et al., 2012), and many other contaminant classes (Burczynski et al., 2000; Bartosiewicz et al., 2001; Yang et al., 2007; Antczak et al., 2013; Biales et al., 2016). Multi-gene signatures distinguishing contaminated sites from one another (Falciani et al., 2008) and tracking toxicity during chemical fractionation of media containing multiple contaminants (Biales et al., 2013) have also been reported.

1.3. Example scenario

In order to make the discussion concrete, we consider the scenario of screening for chemical hazards in environmental samples. Screening for chemical hazards in environmental media, such as a wastewater, involves exposing test organisms to media samples and identifying RNA profile changes associated with endpoints of concern, such as impairments of growth, reproduction or survival. The goal is estimating the likelihood of adverse apical responses in organisms potentially exposed to the media sampled. This approach can detect a broad range of physiological effects from a much larger variety of chemicals than can be characterized efficiently with analytical chemistry. This is similar to whole effluent toxicity (WET) testing, where chemical hazards are detected based on effects on growth or survival of exposed test organisms (de Vlaming et al., 2000). In both cases, no particular chemical is targeted for monitoring. Instead, biological effects are used to infer the presence of some kind of chemical hazard. RNA profiling can supplement conventional WET testing by allowing measurement of molecular changes that manifest earlier than apical endpoints or at concentrations below those evoking apical effects in test species. The latter is important for protecting environmentally relevant species that are more sensitive than test species. The variety of RNA profiles possible also suggests RNA profiling could provide more specific hints about the class of toxicants potentially involved than non-specific univariate endpoints like survival.

1.4. An example biomarker

Vitellogenin (VG) is the poster-child (Sumpter and Jobling, 1995) for a single-gene biomarker (a set of one or more biochemical features whose measurement can indicate a physiological condition of interest, like response to toxicant exposure). VG is a pre-protein produced in the liver of reproducing females of various species which is transported by the circulation to the ovary where VG is cleaved into pieces that make up the bulk of protein present in egg yolk. Presumably because VG's primary role is in egg production, the VG gene is tightly regulated with very low expression in male fish. However, following estrogenic exposures, VG is strongly induced in males of many fish species, resulting in abundant VG RNAs in the liver.

VG is a useful biomarker, because expression is tightly repressed in unexposed males, the response is fairly specific to estrogenic exposures, and exposure usually results in dramatic, monotonic, concentration-dependent increase in expression (however, see Andersen et al., 2006 for an example of a non-monotonic VG response to a toxicant not normally thought of as estrogenic). This behavior allows researchers to readily distinguish VG signals in unexposed males from VG signals in males exposed to reproductively disruptive estrogen concentrations. These properties are unusual, since more typical genes are 'on' in the unexposed case, and exposure results in more modest changes in activity. Exposure may not be monotonic with increasing exposures. These issues can result in more extensive overlap in RNA levels between exposed and unexposed organisms.

2. TEST CONCENTRATIONS OF TOXICANT

The selection of test concentrations of toxicant used for biomarker development can have an important impact on the eventual utility and interpretation of biomarker results. Biomarker levels in test species can be associated with the likelihood of an apical response in the same species. However, one is usually interested in protecting communities of organisms rather than the species used for testing. Organisms targeted for protection may be threatened (so should not be sacrificed in order to obtain RNA) or difficult to breed in captivity. Therefore, biomarker development is usually conducted in species that are not the target of protection, but selected based on prolific reproduction in captivity, short lifecycle, amenability to use with small environmental samples, and easy experimental manipulation. This means we need to translate between test species biomarker responses and apical effects in environmentally important organisms.

One way to map aquatic concentrations of a chemical to expected effects in an aquatic community uses species sensitivity distributions (SSDs). An SSD (Kooijman, 1987) can be estimated by fitting the concentrations of toxicant eliciting a particular level of apical response (such as LC_{10}) in a variety of species considered representative of the community to be protected, to a statistical distribution, such as log-normal (Figure 1). There are many options for constructing an SSD, selecting test species, and translating species effects into long term community repercussions. This reflects imperfections in the process (Forbes and Calow, 2002). Often, available response levels and species are limited to previously published data (since toxicity testing is expensive). Furthermore, no single SSD distribution fits all data sets well, no consensus exists on how to select test species representative of real-

world aquatic communities, or on levels of interdependence of species within a community. Nevertheless, decisions must be made even in the absence of ideal evidence. Therefore, SSDs continue to be widely used for mapping chemical concentrations onto community effects.

SSDs should be similarly suitable for translating chemical concentrations and evoked biomarker signal levels into potential effect levels in aquatic communities (Fedorenkova et al., 2010). For example, if a biomarker reliably detects toxicant at a detection limit corresponding to the 5th percentile of the SSD, negative biomarker results on new samples suggest toxicant concentrations are no higher than those at which fewer than 5% of species in the community will suffer the level of impact used in constructing the SSD (LC10 in the previous example). This suggests selecting test concentrations for biomarker development corresponding to useful levels of community protection. Positive biomarker results can also be mapped onto an SSD, but multiple concentrations must be used during biomarker development and more complicated regression analysis is required, which is beyond the scope of this work. In the simpler classification setting we focus on, we assume biomarkers are used for screening rather than definitive testing, implying that positive results can be pursued with effects directed fractionation and quantitative analytical chemistry. Results from that latter can be directly mapped onto the SSD.

3. CLASSIFIER DEVELOPMENT

3.1. Transcriptional dose response

We assume RNA levels change in response to increased toxicant exposure, but the response need not have any particular shape, nor be zero anywhere. We do require that biomarker levels at high exposures deviate from levels in unexposed samples at least as much as biomarker levels at the detection limit (see section 2.1). This property should be tested in practice. Exceptions will limit the useful concentration range for the biomarker, and biomarkers that maintain this property should be preferred. In order to simplify discussion, we assume biomarker signals rise after toxicant exposure.

At a given exposure level, transcriptional responses in individual test organisms fluctuate around the test organism population's average response curve, reflecting inter-individual variation in toxicant susceptibility (Bliss, 1935) and stochastic processes (Newman and McCloskey, 2000). Therefore, we expect variation in biomarker levels between replicates at any single toxicant concentration, including unexposed (negative) controls (Figure 2). The degree of overlap in response distributions of unexposed and exposed organisms determines the efficiency with which the biomarker can distinguish these conditions.

3.2. Identifying informative genes

Not all genes are expected to respond to toxicant exposure, and many that do respond, will not respond with specificity to a single toxicant class or apical outcome. Developing classifiers often begins by narrowing the list of RNA variants to those that respond with sensitivity and specificity to the toxicant class of interest. If we know all the mechanistic details (mechanism of action, or MOA) leading from interaction of toxicant with molecular

receptor to manifestation of apical endpoints, as well as the role of involved genes in other physiological processes, we can pick genes for classifier development based on reliable behavior of a gene in the toxicant MOA (favoring sensitivity), and the absence of similar responses of the gene in other physiological processes (improving specificity). If we identify several informative RNAs, we can pick the 'best' one for use in the classifier, or we can combine levels of several RNA variants. Including several variants can provide better performance, because, for any one gene, there will be inaccuracy in measurement of RNA abundance, inter-individual variation in gene responses to a given toxicant concentration, and potential unexpected involvement in other physiological processes (decreasing specificity). Classifiers based on multiple RNA variants not only offset poor behavior from one gene with better behavior of other genes, but also can distinguish more complex patterns, potentially improving specificity. However, including too many variants in the model will increase overfitting problems (section 3.4; Burczynski et al., 2000; Guyon et al., 2002; Steiner et al., 2004; Li et al., 2010).

The most common application of RNA profiling in ecotoxicology is elucidation of MOAs leading from toxicant exposure to apical endpoints (Ankley et al., 2009). This laudable goal facilitates classifier development by identifying genes that are sensitive and specific for a toxicant class. However, most RNA profiling experiments do not efficiently reveal MOA details. One reason was mentioned previously: RNA profiling only detects changes in RNA abundance, but not other regulatory mechanisms, such as post-translational processing. Additionally, profiling is usually conducted on organs or tissues composed of a variety of cell types. Therefore, observing a change in gene A and gene B in the same experiment, does not mean that the changes are happening in the same cell (Brenner, 2010). The two changes can reflect distinct responses in different cell types or cell cycle stages of the same cell type. Even in a single cell, we will typically see responses leading to the toxic outcome, alongside less specific homeostatic processes, including induction of detoxification pathways. In the absence of prior mechanistic information, we cannot assign observed RNA changes to the right process. Furthermore, most experiments only consider samples collected at a single time point after exposure. Therefore, if we see genes A and B changing, even if they reflect a single process in a single cell type, we cannot tell from RNA profiles alone whether gene A caused gene B to change, whether gene B caused gene A to change, or whether a third gene C is the common cause of genes A and B changing. As the number of gene changes observed increases, the number of possible pathway topologies consistent with the data increases dramatically (Brenner, 2010).

These considerations explain why many RNA profiling experiments tend to reiterate known functional relationships between genes, ignore expected relationships not reflected in transcriptional changes, avoid discussing changes in transcription of un-annotated genes, or generate many hypotheses that are seldom pursued. Cheap methods like RNA profiling are insufficient for pursuing the many hypotheses (per the previous paragraph), but must be supplemented with expensive, low throughput, biochemical approaches. These approaches include observing changes after selective alteration of gene activity using genetic knock-in or knock-out, RNAi, interfering antibodies, or small molecule agonists/antagonists. One can also identify direct physical associations between gene products, using two-hybrid assays or co-purification of interacting gene products. In vitro reconstitution or cell transfection can

also be used to corroborate hypotheses by recreating postulated mechanisms under controlled conditions. However, these methods are too slow and expensive to efficiently evaluate every network topology consistent with the large number of transcriptional changes often observed during global RNA profiling.

Facing similar challenges, but motivated by the need for better diagnostic tests, developers of multi-gene classifiers for medical use have gravitated towards mechanism-agnostic approaches (Naoi and Noguchi, 2016). Instead of developing classifiers by first elucidating mechanisms involved in pathology, researchers look for distinctive patterns of RNA changes that can be reliably associated with pathology, without requiring assignment of functional roles to individual changes. This usually begins by identifying RNAs changing in response to treatment (variable selection), using either traditional statistical tests such as t-test or ANOVA (Kerr et al., 2000), Bayesian methods which reduce inferential errors resulting from the use of small sample sizes (McCarthy et al., 2012), or machine learning approaches integrating variable selection with classifier training (Guyon and Elisseeff, 2003).

3.3. A simple classification algorithm

Once informative RNAs are identified, we want to combine their signals into a single signal capable of distinguishing exposed from unexposed test organisms. We can do this with logistic regression, which is an extension of ordinary least squares (OLS) regression. For one independent variable, OLS assumes that for each observation (x_i, y_i) , the dependent variable (Y) follows a normal (Gaussian) distribution $N(m_i, s_i)$, implying y_i and m_i are real numbers between negative infinity and infinity. The conditional mean m_i is assumed a linear function of the independent variable (X) :

$$m_i = c_0 + c_1 * x_i, \quad [1]$$

where ‘*’ signifies multiplication. This is the equation for a straight line. The value of M when X is zero is estimated by c_0 (intercept). The slope c_1 estimates the rate with which M changes with changing X . OLS regression selects the two constant parameters c_0 and c_1 so as to minimize the sum of squared vertical deviations $(y_i - m_i)^2$, or equivalently $(y_i - (c_0 + c_1 * x_i))^2$, of training observations from this line. The variance s is estimated from this sum of squared deviations.

Logistic regression extends OLS to problems where the dependent variable is restricted to the set of two values $\{0, 1\}$, here symbolizing the two classes {unexposed, exposed}. The dependent variable is assumed to have a Bernoulli(p_i) distribution, where the parameter p_i is the probability of the dependent variable Y taking on the value 1 (exposed) given that the independent variable X (signal level) takes on value x_i . This conditional probability can more explicitly be written as $p_i = P(Y=1 | X=x_i)$. For a single informative RNA variant, we can ‘predict’ the exposure probability of a test sample i using the equation:

$$\log(p_i/(1 - p_i)) = c_0 + c_1 * x_i, \quad [2]$$

where $\log()$ indicates the natural logarithm, x_i is the measurement of the RNA variant in sample i , p_i is the estimated probability of exposure given x_i , and $(1 - p_i)$ is the probability of not having been exposed (since the total probability of all possibilities must sum to one). The constant parameters c_0 and c_1 are estimated beforehand from observations of samples with known exposure status (the training set). This equation can be solved for p_i :

$$p_i = 1/(1 + e^{-(c_0 + c_1 * x_i)}), \quad [3]$$

where 'e' is the base of the natural logarithm.

Logistic regression also allows us to combine measurements of p RNA variants from a single test sample i into a predicted probability of exposure:

$$\log(p_i/(1 - p_i)) = c_0 + c_1 * x_{i1} + c_2 * x_{i2} + \dots + c_p * x_{ip}, \quad [4]$$

where x_{i1} through x_{ip} represent the collection of p RNA variant measurements, and c_0 through c_p are constant parameters previously estimated using a set of training samples.

Logistic regression is one of many algorithms available for translating one or more RNA measurements into a probability of exposure. The choice of algorithm is complicated by uncertainty about the true shape of the function relating toxicant exposure to RNA levels, and the large number of RNAs often surviving variable selection. This introduces a trade-off between sufficient flexibility to accommodate various functional relationships and the potential for overfitting.

3.4. Classifier flexibility vs. overfitting

Overfitting describes the condition where an overly flexible (relative to the amount of training data) classifier follows noise in the training data more than it follows the effect of interest. RNA profile-based classifiers are particularly prone to overfitting because the amount of training data (frequently below 10 replicates per condition) is small relative to the number of RNA signals measured (often well over 100). The OLS equation with one independent variable (equation 1) can be used to demonstrate overfitting. For a dataset consisting of only two observations, one can always find a line that perfectly fits (Figure 3 A), running directly through the observations and minimizing the sum of squared deviations to zero. If we extend the equation to:

$$m_i = c_0 + c_1 * x_i + c_2 * x_i^2, \quad [5]$$

where c_2 is the estimate of the rate of change of M in response to changes in the square of X , OLS will always perfectly fit any 3 data points (Figure 3 B), because the extra parameter introduces a bend in our previously straight line. If we include the third power of the independent variable:

$$m_i = c_0 + c_1 * x_i + c_2 * x_i^2 + c_3 * x_i^3, \quad [6]$$

where c_3 is the rate of change of M in response to changes in the third power of X , this introduces another bend, allowing any four data points to be perfectly fit (Figure 3 C). In general, the more parameters (or variables) included, the more flexible the fit becomes. If the number of parameters (or variables) equals or exceeds the number of data points, one can always find a perfectly fitting equation. Therefore, a sufficiently flexible model will not only capture the process of interest, but also random noise in the training data (overfitting), leading to poor predictions with new data.

Logistic regression is particularly liable to overfitting when the number of variables is large relative to the number of training samples. One of the many alternatives far less susceptible to overfitting in this situation is penalized regression (e.g. Zou and Hastie, 2005), where a penalty is imposed on the complexity of the fit (the number and magnitude of the coefficients) in order to reduce overfitting and improve stability (sensitivity of results to small changes in the training data) in the presence of strong correlations between independent variables. Another alternative is random forests (RF) (Breiman, 2001), which uses data resampling to reduce overfitting, while extending the potential of the model to fit more complicated relationships between independent and dependent variables that cannot be captured by logistic or penalized regression. RF makes likelihood ratio predictions using a large ensemble of decision trees trained using randomly varied subsets of the training data. Penalized regression and RF represent complementary approaches based on very different assumptions. One can try both in order to investigate the sensitivity of results to the algorithm employed. Some alternatives (such as support vector machines and naive Bayes classifiers) can produce very good class predictions, but yield biased estimates of p_i poorly suited for estimating the likelihood ratios described in section 5.1. Estimates of p_i from these classifiers can be calibrated using methods such as Platt scaling (Platt, 1999), or isotonic regression (Zadrozny and Elkan, 2002).

For 'easy' discrimination problems, many variable selection and classifier algorithm combinations provide similarly excellent results (Wang et al., 2008). In more difficult cases, different combinations produce substantially different results, but the optimal approach depends on the characteristics of the data (Gong et al., 2012), such as strength of toxicological response or number of replicates. However, several very dissimilar algorithms usually give near-optimal results for the same data (Wang et al., 2008; Li et al., 2010; Gong et al., 2012; Biales et al., 2013). Successful approaches to feature selection for toxicological datasets include ANOVA/t-test (Poynton et al., 2008; Svendsen et al., 2008; Li et al., 2010; Biales et al., 2013), Mann-Whitney test (Gong et al., 2012), Fisher's exact test (Wang et al., 2008), genetic algorithm (Falciani et al., 2008; Wang et al., 2008; Antczak et al., 2013), and other 'integrated' methods (Thomas et al., 2001; Steiner et al., 2004; Li et al., 2010). Successfully employed toxicant classification algorithms include naive Bayes' (Thomas et al., 2001), support vector machines (Steiner et al., 2004; Lam et al., 2008; Wang et al., 2008), penalized regression (Gong et al., 2012; Biales et al., 2013), k-nearest neighbors (Falciani et al., 2008; Lam et al., 2008; Poynton et al., 2008; Wang et al., 2008), shrunken

centroids (Nota et al., 2010), partial least squares discriminants (Svendsen et al., 2008), and tree-based methods (Li et al., 2010; Gong et al., 2012; Antczak et al., 2013; Biales et al., 2013; Biales et al., 2016).

4. CLASSIFIER EVALUATION

4.1. Performance metrics

Once RNA measurements are made at a targeted detection limit concentration (section 2.1) and in unexposed controls, they can be used to train classifiers like those described in the previous section. Many methods are available for evaluating how well a classifier distinguishes these two classes, but we will only describe an instructive subset (Pepe, 2003). One approach selects a p_i cutoff (often defaulting to 0.5) and reports the proportion of samples correctly classified using this cutoff. A shortcoming to this is that results are conditional on the cutoff chosen. A more serious problem is that results change with the ratio of unexposed to exposed samples being evaluated. For example, if we evaluate ‘real world’ performance with a chemical that only occurs at a small proportion of samples (e.g. from 1% of sites), an extreme cutoff that always results in a negative call would be reported to be correct 99% of the time, which is highly misleading. One refinement is to report false positive fractions (FPF: fraction of truly negative samples called positive) and true positive fractions (TPF: fraction of truly positive samples called positive). In the previous case, the FPF is an impressive 0%, but the TPF is a disappointing 0%. However, results still depend on the signal cutoff choice, and both TPF and FPF must be reported.

An alternative approach uses the receiver-operator characteristic (ROC) curve. A ROC curve is the plot (Figure 4) of FPF (or $1 - \text{specificity}$) on the horizontal (x) axis, versus TPF (or sensitivity) on the vertical (y) axis, as a function of cutoff. For most classifiers, an extremely stringent cutoff results in no positive samples (TPF = 0), and all negative samples (FPF = 0), corresponding to the bottom left corner of the plot ($x=0, y=0$). An extremely liberal cutoff results in all positive calls (TPF = 1), and no negative calls (FPF = 1), corresponding to the top right corner of the plot ($x=1, y=1$). Therefore, ROC curves typically begin at the lower left corner and end at the upper right hand corner. Performance is reflected in the path of the curve from the lower left to the upper right. If a classifier assigns samples randomly (an uninformative classifier), the result for a theoretical infinite sample set (the ‘asymptotic’ result) is a straight line on the diagonal between the lower left corner and upper right corner. For finite sized sample sets, estimated ROC curves fluctuate around the asymptotic result, due to sampling variation (variation between any two finite samples from a non-homogenous population). For a perfect classifier (works perfectly for any non-zero cutoff), the curve will look like a single step, rising vertically from the lower left corner to the upper left corner ($x=0, y=1$), then proceeding horizontally to the upper right corner. Typical classifiers fall between these extremes.

The ROC curve reflects both FPF and TPF across all possible biomarker cutoffs, avoiding the issues we identified earlier. However, it is easier to communicate a number than a plot. A common summary statistic is the area under the ROC curve (AUC), which must be between 0 (a straight line extending from the bottom left corner to the bottom right corner, reflecting a classifier that is always wrong) and 1 (the ‘perfect’ classifier). Random classification

asymptotically approaches the diagonal splitting the available plot area into equal halves, resulting in an AUC of 0.5. Informative classifiers will have an asymptotic AUC greater than 0.5.

There are many methods for constructing a ROC curve (Pepe, 2003). At one extreme, a purely empirical approach makes few assumptions, but results in a jagged ROC curve composed of a series of steps, with one potential step for each sample in the test set (the curve smooths out for larger samples). An alternative approach enforces ‘reasonable’ smoothness on the curve regardless of sample size, by assuming the curve belongs to a known family of smooth functions, then estimating function parameters from the data. Many software packages (see supplement) carry out either approach, and generate confidence intervals (CIs) for the AUC or for the difference between AUCs of different classifiers (Pepe, 2003). CIs reflect sampling variation, increasing in width with increasing population variability or decreasing sample number. CIs allow statistically valid classifier comparisons and determinations of whether a classifier performs better than random class assignment.

However, usually we are only interested in a portion of the ROC curve. Screening assumes that applying a method that is cheaper, but less accurate than the gold-standard method is cost effective because most samples are actually unexposed. False positive screening results can be corroborated using the gold-standard method, so the risk of false positives is the extra cost associated with the screening test. By contrast, false negative screening results will usually not be pursued, resulting in incorrect decisions, which is usually a more serious problem. Therefore, screening tools are only useful if the TPF is high. An acceptable lower limit on the TPF can often be defined, based on costs of false negative calls relative to the cost of screening (Ashby and Smith, 2000). In that case, only the area under the ROC curve above the horizontal line representing the TPF lower limit is relevant for estimating classifier performance. This area is called the partial AUC (pAUC). However, theoretical properties of pAUCs are poorly characterized. Fewer software packages support pAUC estimation, and CIs for pAUCs or for the difference between pAUCs usually have to be estimated using data resampling (Pepe, 2003).

4.2. Phases of evaluation

Classifier performance varies between experiments, due to sampling variation and differences between experimental conditions. Differences in conditions stem from genetic changes in test species colonies over time, seasonal effects, differences between test species colonies reared in different labs, differences in lab water composition (e.g. alkalinity and dissolved organic content), differences in measurement protocols, etc. Real world exposures can result in even larger changes in classifier performance due to more substantial differences in background (other than toxicant) composition of the water, interactions with other toxicants present, differences in toxicant concentrations from concentrations used during classifier development, responses of test organisms to caging, etc. Even if experimental protocols are kept constant, evaluation with new samples usually results in worse performance than estimates from the same data used to develop the classifier (training data), due to ‘overfitting’ (section 3.4).

One way to avoid overly optimistic performance estimation during the training experiment is to split data into a training set (for this example, called set 1) used for parameterizing the classifier, and a test set (called set 2) for estimating classifier performance. Since ‘noise’ in the test set will be different than in the training set, overfitting can be evaluated. However, if we use set 2 for training and set 1 for testing, we will get a slightly different answer. More reproducible estimates of real-world performance are obtained by averaging results from both configurations. Subsequently we can train the classifier using pooled data from set 1 and set 2. Since classifier performance tends to improve with larger training samples, the classifier trained with pooled data will usually perform at least as well as the averaged performance estimate from the subsets. Performance estimates will change slightly if we randomly re-assign individual samples to set 1 and set 2, then repeat the analysis. Since larger training sets tend to result in better classifiers, we should make the training set as large as possible. However, the test set must be large enough to provide stable estimates of performance with sufficient granularity to accurately measure the high TPFs expected of screening tools.

A better evaluation method is cross-validation (CV) (Kohavi, 1995), where a series of ‘virtual experiments’ are conducted in silico, in which samples from a single real experiment are split between a larger training set, and a non-overlapping smaller test set. Virtual experiments are repeated with different data subsets assigned to training and test sets. Performance estimates from the virtual experiments are averaged in order to obtain the final estimate. Then all the data are pooled and used to train the final classifier. The final classifier is assumed to have performance at least as good as the averaged performance estimate, because the final classifier is based on the complete set of data points sub-sampled in the virtual experiments. When using a large proportion of data for testing leaves insufficient data for training a good classifier, the pessimistic bias of CV may become unacceptably large. In small experiments, leave-one-out CV (LOOCV), where one data point per iteration is retained for testing, leaving the maximum amount of data possible for classifier training, often provides more reliable performance estimates. However, LOOCV estimates have low precision (Efron, 1983), becoming sub-optimal for larger datasets. For larger datasets (more than ten samples per condition), 5-fold CV (one fifth of data reserved for testing in each iteration) or 10-fold CV is often more accurate than LOOCV.

Although CV is effective for preliminary performance estimates, it fails to account for many variables affecting real-world performance. More realistic evaluation can test performance with an independent experiment conducted in the same lab that developed the biomarker, using larger sample sizes in order to obtain more reliable estimates. Next, we can evaluate performance across several independent laboratories. Subsequently, we can use representative real-world samples whose chemical content is checked using gold-standard methods (presumably analytical chemistry). This sequence accounts for most sources of performance variation, and provides additional data for incremental classifier improvement: after use for testing, the new data can be used for additional training and calibration of the classifier. Real world evaluation also estimates toxicant occurrence rates useful as ‘priors’ (see section 5.1). This sequence is cost-effective, since it proceeds from lower cost experiments to higher cost experiments, allows step-wise classifier improvement using accumulating data, and the sequence can be stopped early if classifiers fail performance

objectives. A similar phased approach is used for developing medical diagnostics, with initial experiments using small sets of well differentiated (severe disease vs. healthy) case-control data, proceeding to larger case-control studies with less severely affected patients, and eventually graduating to very large cohort studies (Pepe et al., 2001).

In the toxicological literature, the proportion of test samples correctly classified is a very common metric, with estimates often based on LOOCV (e.g. Thomas et al., 2001; Falciani et al., 2008; Wang et al., 2008; Wang et al., 2012), or other CV schemes (Antczak et al., 2013; Nota et al., 2010). Steiner et al., 2004 adopted the Matthews correlation coefficient (MCC; Matthews, 1975) as a performance metric, using test sets generated using CV. Like AUC, MCC is relatively insensitive to the ratio of positive to negative samples, but MCC changes with classifier cutoff score. Lam et al., 2008 reported LOOCV-based TPFs and FPFs. AUCs based on 10-fold CV (Li et al., 2010) or independent test sets (Biales et al., 2013 and Biales et al., 2016) have also been employed.

5. DECISION SUPPORT

5.1. Likelihood ratios

In section 3.3 (equations 2 and 4), we showed how $\log(p_i/(1 - p_i))$ could be expressed as a function of a set of independent variables, where p_i is $P(Y=1 | \mathbf{X}=\mathbf{x}_i)$, or the probability of $Y=1$ (exposed) when $\mathbf{X}=\mathbf{x}_i$ is observed, where $\mathbf{x}_i$ (bolded) is the collection of observed values of the independent variables $\{x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip}\}$ in sample i , and $\mathbf{X}$ is the corresponding collection of random variables. The ratio $p_i/(1 - p_i)$ is known as the ‘odds ratio’ (OR). Equation 4 can be used directly to estimate the class probabilities for a new sample based solely on biomarker results. However, in practice we often wish to combine the results of several independent estimates of the probability that $Y=1$, which can be accomplished using ‘likelihood ratios’. The likelihood ratio $L(\mathbf{x}_i)$ is defined as:

$$L(\mathbf{x}_i) = P(\mathbf{X} = \mathbf{x}_i | Y = 1)/P(\mathbf{X} = \mathbf{x}_i | Y = 0), \quad [7]$$

which is the ratio of the probabilities of $\mathbf{X}=\mathbf{x}_i$ given exposure and lack of exposure, reversing the direction of the conditionality expressed by the OR. The difference between OR and $L(\mathbf{x}_i)$ can be thought of in terms of the difference between the probability of rain given the observation of clouds $P(\text{rain}|\text{clouds})$, and the probability of cloudiness given the observation of rain $P(\text{clouds}|\text{rain})$.

Similar to the OR, $L(\mathbf{x}_i)$ can be predicted for new samples (see supplement) using an adjusted form of the logistic regression equation:

$$L(\mathbf{x}_i) = e^{(\log(n_0/n_1) + c_0 + c_1^* x_1 + c_2^* x_2 + \dots + c_p^* x_p)}, \quad [8]$$

where n_0 is the number of unexposed samples in the training set, and n_1 is the number of exposed samples in the training set (Albert, 1982), and $\{c_0, c_1, c_2, \dots, c_p\}$ are estimated

from the training data using logistic regression. This is a standard adjustment for using logistic regression equations trained with case-control or other retrospective studies for class prediction of new samples (Agresti, 2002). Note that the offset term $\log(n_0/n_1)$, drops out of the equation when equal numbers of treated and untreated organisms are used for training, since $\log(1) = 0$. Keeping these numbers equal in the training set allows us to use unmodified logistic regression (and related methods) to directly estimate $L(\mathbf{x}_i)$ in new samples.

For screening new samples, we are interested in $P(Y=1 \mid \mathbf{X}=\mathbf{x}_i)$, or whether the sample was exposed given observation of biomarker levels $\mathbf{x}_i$. Applying Bayes' Theorem, followed by algebraic rearrangement, leads to (Albert, 1982):

$$P(Y = 1 \mid \mathbf{X} = \mathbf{x}_i) = P_0 * L(\mathbf{x}_i) / ((1 - P_0) + P_0 * L(\mathbf{x}_i)), \quad [9]$$

where P_0 is our prior (to making measurement $\mathbf{x}_i$) belief (expressed as a probability) that the sample was exposed to the toxicant class. The added complexity of working with likelihood ratios allows us to incorporate independent information (as the prior probability), together with biomarker screening, into a posterior (to observing $\mathbf{x}_i$) estimated probability $P(Y=1 \mid \mathbf{X}=\mathbf{x}_i)$ of exposure.

Prior probabilities ('priors') can be estimated in a variety of ways depending on available information. If there are abundant representative occurrence data for the class of toxicants, P_0 can be estimated as the frequency with which toxicant has been observed at concentrations at or above the detection limit. If occurrence data are not representative of sites of interest, one can use available occurrence data as a preliminary estimate, though the representational limitation should be described when reporting results. In the absence of sufficient occurrence data, one can use models incorporating estimates of environmental loading rates, transport properties and fate to estimate P_0 . If those estimates are impractical, expert opinion can be used to estimate P_0 . The better the priors, the better we expect posterior estimates (or 'posteriors') to be, but the lack of optimal data should not dissuade us from estimating priors, as long as we report $L(\mathbf{x}_i)$, allowing others with differing opinions to recalculate the posterior using different priors. A more thorough analysis, would generate posteriors across a plausible range of priors, in order to communicate the sensitivity of posteriors to underlying assumptions.

5.2. Replicate test samples

Another advantage of working with likelihood ratios instead of odds ratios is that it simplifies combining data from replicate biomarker tests. If test organisms are chosen at random from a laboratory colony, each replicate test represents a random sample from the tolerance distribution (distribution of sensitivities to a given concentration of toxicant within the population) for members of that colony. Each replicate is 'conditionally independent' from other replicates, that is, an independent test, given the shared tolerance distribution and chemical content in the samples being evaluated. The likelihood ratio based on n conditionally independent measurements is (Albert, 1982):

$$L(\mathbf{x}) = L(\mathbf{x}_1) * L(\mathbf{x}_2) * L(\mathbf{x}_3) * \dots * L(\mathbf{x}_n), \quad [10]$$

where $\mathbf{x}$ is used to represent the collection of n replicate test results $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_n\}$.

The commutative property of multiplication implies that $L(\mathbf{x})$ does not depend on the order of tests and is equally amenable to parallel testing (with all n tests conducted at a single time) and sequential tests (where tests are conducted in sequence, allowing decisions about whether or not to conduct the $n+1^{\text{th}}$ test to be determined after evaluating the first n tests). $L(\mathbf{x})$ can be updated simply by multiplying by $L(\mathbf{x}_{n+1})$.

Converting $L(\mathbf{x})$ into the desired quantity $P(Y=1 \mid \mathbf{X}=\mathbf{x})$ proceeds in the same manner as for a single test (Albert, 1982).

$$P(Y = 1 \mid \mathbf{X} = \mathbf{x}) = P_0 * L(\mathbf{x}) / ((1 - P_0) + P_0 * L(\mathbf{x})) \quad [11]$$

This equation can combine results from completely different types of hazard estimates, as long as the estimates are expressed as probabilities of exceeding the detection limit, and conditional independence of estimates holds. Estimates can come from e.g. biomarker results, chemical measurements, predictions from mechanistic modeling, and expert opinion. This is how these equations are typically applied in medical practice: LRs from a variety of laboratory tests are combined with an LR estimated from physical examination and an LR estimated from patient history. Often, these sources of medical information do not strictly meet the conditional independence assumption, but studies suggest that good estimates result even with moderate departures from conditional independence (Holleman and Simel, 1997).

For positive samples, an arbitrarily high biomarker $L(\mathbf{x})$ can be achieved by combining results from a sufficient number of very modest individual replicate $L(\mathbf{x}_i)$ s. So even if, at the detection limit, the average $L(\mathbf{x}_i)$ from a single test is only 2 (reflecting substantial overlap between biomarker signal distributions in exposed and unexposed organisms), we might get an $L(\mathbf{x})$ of over 1000 from 10 replicate tests (since 2^{10} is 1024), or an $L(\mathbf{x})$ of over 1 million from 20 replicate tests. At higher concentrations, the $L(\mathbf{x}_i)$ s should be higher, reducing the number of replicates needed to achieve a given $L(\mathbf{x})$. Table 1 shows how posteriors change across a range of plausible priors and likelihood ratios from a set of replicate tests. If priors are very low (likely for many toxicant classes), high $L(\mathbf{x})$ are needed to achieve posterior probabilities of over 90%, requiring either high individual test $L(\mathbf{x}_i)$ s or, more realistically, a large number of replicate tests.

6. THE SCOURGE OF SPECIFICITY

6.1. Estimating specificity is challenging

In the previous discussion, performance evaluation, as well as translations of classifier predictions into likelihood ratios, depended on estimating the differential distribution of

signals between unexposed and exposed test organisms. The distribution in laboratory exposed test organisms can be used to explore effects of exposure intensity, as well as effects of controlled changes in background conditions (e.g. pH, temperature, alkalinity, dissolved oxygen, or dissolved organic matter). Follow-on experiments using spiked real-world water samples can examine effects of a wider variety of background conditions on the response to exposure. Although this is not a trivial process, it is more straightforward than exploring the distribution of biomarker signals in organisms not exposed to the toxicant class of interest.

The distribution of signals in organisms not exposed to the toxicant class of interest can be affected not only by the factors that affect signal distributions in exposed organisms, but also by changes in signal distribution caused by other toxic or non-toxic contaminants in the water. For instance, a different chemical class may elicit a similar biomarker response, even though that exposure may not be associated with the same apical endpoint. This could happen if e.g. the overlap in signal patterns reflects a common detoxification pathway, rather than a common mechanism of toxic action. This possibility is difficult to discount without a detailed understanding of the toxic MOA, including involvement of member genes in other physiological pathways, or exhaustive exploration of the behavior of those genes under exposure to every other chemical that could be encountered in practice. Either route requires an enormous amount of experimentation that is unlikely to be affordable in the foreseeable future. Although this possibility may be acceptable for screening, since false positive calls are expected to be followed up with gold-standard methods, it is an important limitation of the usual approach to biomarker development which adversely affects screening efficiency.

6.2. An alternative approach

A spectrum of potential biomarker specificity can be postulated. At one end of the spectrum are biomarkers that distinguish toxicants binding the same molecular receptor in a toxicological pathway, and eliciting the same apical endpoint. Distinctions can be made if e.g. induced detoxification pathways, which tend to reflect physiochemical properties of the chemicals more than their mechanism of toxic action, differed between the toxicants. Less specific biomarkers may still distinguish toxicants affecting the same MOA via interactions with different molecular receptors. In this case, the toxicant binding the receptor further upstream in the MOA might affect a superset of the genes affected by the toxicant binding downstream, providing an opportunity for discrimination. Lower specificity biomarkers might still distinguish one toxicological pathway leading to a given apical endpoint from a different toxicological pathway leading to the same apical endpoint. Even less specific biomarkers might distinguish toxicants affecting entirely different endpoints, since even fewer molecular effects are expected to overlap. Bookending this part of the spectrum would be biomarkers that detect deviations from naturally varying but otherwise healthy conditions.

This last possibility suggests a simplified approach to toxicant detection using RNA profiling: one might use the highly parallel nature of global RNA profiling to characterize RNA abundances of all genes across a wide range of representative ‘clean’ sites in order to estimate a ‘reference’ range for each gene, defining the bounds of ‘normal’ variation. New samples could be evaluated based on whether the resulting RNA profile included expression levels outside the reference range, suggesting unusual conditions potentially due to stressor

exposure. Measurements of the deviating RNAs could be used to guide effects-directed fractionation in order to identify chemicals associated with the unusual expression pattern. Follow-up toxicological investigation could suggest whether the response is related to an endpoint of concern, or whether the response does not indicate toxicity, in which case the 'reference' range for the affected RNAs could be expanded to include the observed expression levels.

The virtue of this iterative approach is that one could quickly develop screening tools for detecting a very wide variety of potential stressor exposures. Instead of having to test hundreds or thousands of different toxicant exposures together with corresponding unexposed controls, one would only examine the range of RNA responses to reference samples collected from across the country, relative to synthetic laboratory water. Perhaps several dozens of sampled sites could suffice as a starting point, requiring less time, money, and sacrificed animals than an approach requiring a large number of toxicant classes to be evaluated sequentially in the laboratory, followed by real-world evaluation. The shortcomings of a reference range approach are that positive findings do not suggest which toxicant class is responsible for unusual readings, what sort of apical endpoints to expect, and probably results in higher FPFs. Also, this approach does not provide an estimate of $P(\mathbf{x}_i | Y=1)$ and $L(\mathbf{x}_i)$ as previously described, but only $P(\mathbf{x}_i | Y = 0)$. Decision making could be based on an acceptable lower limit for $P(\mathbf{x}_i | Y = 0)$, or the probability that the observations are consistent with a clean site. In order to identify responsible toxicants, effects directed fractionation may be needed. Nevertheless, this may be an attractive interim approach to quickly develop global RNA profiling-based screening tests for chemical hazards. In addition, knowledge of the reference range can inform toxicant-targeted biomarker development by allowing developers to screen out biomarkers with responses overlapping those observed in samples from clean real-world sites, allowing more rapid improvement in biomarker specificity.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Afshari CA, Nuwaysir EF, Barrett JC. Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation. *Cancer Research*. 1999; 59(19): 4759–60. <https://www.ncbi.nlm.nih.gov/pubmed/10519378>. [PubMed: 10519378]
- Agresti A. *Categorical Data Analysis*. John Wiley and Sons; Hoboken, New Jersey, USA: 2002. 170–1.
- Albert A. On the use and computation of likelihood ratios in clinical chemistry. *Clinical Chemistry*. 1982; 28(5):1113–9. <https://www.ncbi.nlm.nih.gov/pubmed/7074890>. [PubMed: 7074890]
- Andersen L, Goto-Kazeto R, Trant JM, Nash JP, Korsgaard B, Bjerregaard P. Short-term exposure to low concentrations of the synthetic androgen methyltestosterone affects vitellogenin and steroid levels in adult male zebrafish (*Danio rerio*). *Aquatic Toxicology*. 2006; 76(3):343–52. DOI: 10.1016/j.aquatox.2005.10.008 [PubMed: 16352352]
- Ankley GT, Bencic DC, Breen MS, Collette TW, Conolly RB, Denslow ND, Edwards SW, Ekman DR, Garcia-Reyero N, Jensen KM, Lazorchak JM. Endocrine disrupting chemicals in fish: developing exposure indicators and predictive models of effects based on mechanism of action. *Aquatic Toxicology*. 2009; 92(3):168–78. DOI: 10.1016/j.aquatox.2009.01.013 [PubMed: 19261338]

- Antczak P, Jo HJ, Woo S, Scanlan L, Poynton H, Loguinov A, Chan S, Falciani F, Vulpe C. Molecular toxicity identification evaluation (mTIE) approach predicts chemical exposure in *Daphnia magna*. *Environmental Science & Technology*. 2013; 47(20):11747–56. DOI: 10.1021/es402819c [PubMed: 23875995]
- Ashby D, Smith AF. Evidence-based medicine as Bayesian decision-making. *Statistics in Medicine*. 2000; 19(23):3291–305. DOI: 10.1002/1097-0258(20001215)19:23<3291::aid-sim627>3.0.co;2-t [PubMed: 11113960]
- Bartosiewicz MJ, Jenkins D, Penn S, Emery J, Buckpitt A. Unique gene expression patterns in liver and kidney associated with exposure to chemical toxicants. *Journal of Pharmacology and Experimental Therapeutics*. 2001; 297(3):895–905. <https://www.ncbi.nlm.nih.gov/pubmed/11356909>. [PubMed: 11356909]
- Biales AD, Kostich M, Burgess RM, Ho KT, Bencic DC, Flick RL, Portis LM, Pelletier MC, Perron MM, Reiss M. Linkage of genomic biomarkers to whole organism end points in a toxicity identification evaluation (TIE). *Environmental Science & Technology*. 2013; 47(3):1306–12. DOI: 10.1021/es304274a [PubMed: 23305514]
- Biales AD, Kostich MS, Batt AL, See MJ, Flick RW, Gordon DA, Lazorchak JM, Bencic DC. Initial development of a multigene ‘omics-based exposure biomarker for pyrethroid pesticides. *Aquatic Toxicology*. 2016; 179:27–35. DOI: 10.1016/j.aquatox.2016.08.004 [PubMed: 27564377]
- Bliss CI. The calculation of the dosage-mortality curve. *Annals of Applied Biology*. 1935; 22(1):134–67. DOI: 10.1111/j.1744-7348.1935.tb07713.x
- Bougas B, Normandeau E, Pierron F, Campbell PG, Bernatchez L, Couture P. How does exposure to nickel and cadmium affect the transcriptome of yellow perch (*Perca flavescens*)—results from a 1000 candidate-gene microarray. *Aquatic Toxicology*. 2013; 142:355–64. DOI: 10.1016/j.aquatox.2013.09.009 [PubMed: 24084258]
- Breiman L. Random forests. *Machine Learning*. 2001; 45(1):5–32. DOI: 10.1023/a:1010933404324
- Brenner S. Sequences and consequences. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. 2010; 365(1537):207–12. DOI: 10.1098/rstb.2009.0221 [PubMed: 20008397]
- Bulera SJ, Eddy SM, Ferguson E, Jatke TA, Reindel JF, Bleavins MR, De La Iglesia FA. RNA expression in the early characterization of hepatotoxicants in Wistar rats by high-density DNA microarrays. *Hepatology*. 2001; 33(5):1239–58. DOI: 10.1053/jhep.2001.23560 [PubMed: 11343254]
- Burczynski ME, McMillian M, Ciervo J, Li L, Parker JB, Dunn RT, Hicken S, Farr S, Johnson MD. Toxicogenomics-based discrimination of toxic mechanism in HepG2 human hepatoma cells. *Toxicological Sciences*. 2000; 58(2):399–415. DOI: 10.1093/toxsci/58.2.399 [PubMed: 11099651]
- Bustin SA. Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. *Journal of Molecular Endocrinology*. 2002; 29(1):23–39. DOI: 10.1677/jme.0.0290023 [PubMed: 12200227]
- Day DA, Tuite MF. Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. *Journal of Endocrinology*. 1998; 157(3):361–71. DOI: 10.1677/joe.0.1570361 [PubMed: 9691970]
- de Vlaming V, Connor V, DiGiorgio C, Bailey HC, Deanovic LA, Hinton DE. Application of whole effluent toxicity test procedures to ambient water quality assessment. *Environmental Toxicology and Chemistry*. 2000; 19(1):42–62. DOI: 10.1002/etc.5620190106
- Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *JASA*. 1983; 78(382):316–331. DOI: 10.1080/01621459.1983.10477973
- Falciani F, Diab AM, Sabine V, Williams TD, Ortega F, George SG, Chipman JK. Hepatic transcriptomic profiles of European flounder (*Platichthys flesus*) from field sites and computational approaches to predict site from stress gene responses following exposure to model toxicants. *Aquatic Toxicology*. 2008; 90(2):92–101. DOI: 10.1016/j.aquatox.2008.07.020 [PubMed: 18823667]
- Fedorenkova A, Vonk JA, Lenders HR, Ouborg NJ, Breure AM, Hendriks AJ. Ecotoxicogenomics: Bridging the gap between genes and populations. *Environmental Science & Technology*. 2010; 44(11):4328–33. DOI: 10.1021/es9037287 [PubMed: 20459122]

- Fielden MR, Zacharewski TR. Challenges and limitations of gene expression profiling in mechanistic and predictive toxicology. *Toxicological Sciences*. 2001 Mar 1; 60(1):6–10. DOI: 10.1093/toxsci/60.1.6 [PubMed: 11222867]
- Forbes VE, Calow P. Species sensitivity distributions revisited: a critical appraisal. *Human and Ecological Risk Assessment: An International Journal*. 2002; 8(3):473–92. DOI: 10.1080/20028091057033
- Gong P, Loh PR, Barker ND, Tucker G, Wang N, Zhang C, Escalon BL, Berger B, Perkins EJ. Building quantitative prediction models for tissue residue of two explosives compounds in earthworms from microarray gene expression data. *Environmental Science & Technology*. 2012; 46(1):19–26. DOI: 10.1021/es201187u [PubMed: 21776976]
- Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning*. 2002; 46(1):389–422. DOI: 10.1023/a:1012487302797
- Guyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 2003 Mar.3:1157–82. <http://www.jmlr.org/papers/v3/guyon03a.html>.
- Hamadeh HK, Bushel PR, Jayadev S, Martin K, DiSorbo O, Sieber S, Bennett L, Tennant R, Stoll R, Barrett JC, Blanchard K. Gene expression analysis reveals chemical-specific profiles. *Toxicological Sciences*. 2002; 67(2):219–31. DOI: 10.1093/toxsci/67.2.219 [PubMed: 12011481]
- Hardiman G. Microarray platforms—comparisons and contrasts. *Pharmacogenomics*. 2004; 5(5):487–502. DOI: 10.1517/14622416.5.5.487 [PubMed: 15212585]
- Holleman DR, Simel DL. Quantitative assessments from the clinical examination. *Journal of General Internal Medicine*. 1997; 12(3):165–71. DOI: 10.1007/s11606-006-5024-6 [PubMed: 9100141]
- Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*. 2000; 7(6):819–37. DOI: 10.1089/10665270050514954 [PubMed: 11382364]
- Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *IntJcai*. 1995; 14(2):1137–45. <https://pdfs.semanticscholar.org/0be0/d781305750b37acb35fa187febd8db67bfcc.pdf>.
- Kooijman SA. A safety factor for LC50 values allowing for differences in sensitivity among species. *Water Research*. 1987; 21(3):269–76. DOI: 10.1016/0043-1354(87)90205-3
- Kumar G, Denslow ND. Gene expression profiling in fish toxicology: A review. *Reviews of Environmental Contamination and Toxicology*. 2016; 241:1–38. DOI: 10.1007/398_2016_10
- Lam SH, Mathavan S, Tong Y, Li H, Karuturi RK, Wu Y, Vega VB, Liu ET, Gong Z. Zebrafish whole-adult-organism chemogenomics for large-scale predictive and discovery chemical biology. *PLoS Genetics*. 2008; 4(7):e1000121.doi: 10.1371/journal.pgen.1000121 [PubMed: 18618001]
- Li Y, Wang N, Perkins EJ, Zhang C, Gong P. Identification and optimization of classifier genes from multi-class earthworm microarray dataset. *PloS One*. 2010; 5(10):e13715.doi: 10.1371/journal.pone.0013715 [PubMed: 21060837]
- Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*. 1975; 405(2):442–51. DOI: 10.1016/0005-2795(75)90109-9
- McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation; *Nucleic Acids Research*. 2012. 1–10.
- Naoi Y, Noguchi S. Multi-gene classifiers for prediction of recurrence in breast cancer patients. *Breast Cancer*. 2016; 23(1):12–8. DOI: 10.1007/s12282-015-0596-9 [PubMed: 25700572]
- Newman MC, McCloskey JT. The individual tolerance concept is not the sole explanation for the probit dose-effect model. *Environmental Toxicology and Chemistry*. 2000; 19(2):520–6. DOI: 10.1002/etc.5620190235
- Nota B, Verweij RA, Molenaar D, Ylstra B, van Straalen NM, Roelofs D. Gene expression analysis reveals a gene set discriminatory to different metals in soil. *Toxicological Sciences*. 2010; 115(1):34–40. DOI: 10.1093/toxsci/kfq043 [PubMed: 20133373]
- Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*. 2002; 18(4):546–54. DOI: 10.1093/bioinformatics/18.4.546 [PubMed: 12016052]

- Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*. 2001; 93(14):1054–61. DOI: 10.1093/jnci/93.14.1054 [PubMed: 11459866]
- Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press; USA: 2003.
- Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*. 1999; 10(3):61–74.
- Poynton HC, Zuzow R, Loguinov AV, Perkins EJ, Vulpe CD. Gene expression profiling in *Daphnia magna*, part II: validation of a copper specific gene expression signature with effluent from two copper mines in California. *Environmental Science & Technology*. 2008; 42(16):6257–63. DOI: 10.1021/es800262k [PubMed: 18767696]
- Quackenbush J. Microarray analysis and tumor classification. *New England Journal of Medicine*. 2006; 354(23):2463–72. DOI: 10.1056/nejmra042342 [PubMed: 16760446]
- Shapiro DE. The interpretation of diagnostic tests. *Statistical Methods in Medical Research*. 1999; 8(2):113–34. DOI: 10.1191/096228099666928387 [PubMed: 10501649]
- Steiner G, Suter L, Boess F, Gasser R, de Vera MC, Albertini S, Ruepp S. Discriminating different classes of toxicants by transcript profiling. *Environmental Health Perspectives*. 2004; 112(12):1236.doi: 10.1289/txg.7036 [PubMed: 15345370]
- Sumpter JP, Jobling S. Vitellogenesis as a biomarker for estrogenic contamination of the aquatic environment. *Environmental Health Perspectives*. 1995; 103(Suppl 7):173.doi: 10.2307/3432529
- Svendsen C, Owen J, Kille P, Wren J, Jonker MJ, Headley BA, Morgan AJ, Blaxter M, Stürzenbaum SR, Hankard PK, Lister LJ. Comparative transcriptomic responses to chronic cadmium, fluoranthene, and atrazine exposure in *Lumbricus rubellus*. *Environmental Science & Technology*. 2008; 42(11):4208–14. DOI: 10.1021/es702745d [PubMed: 18589989]
- Thomas RS, Rank DR, Penn SG, Zastrow GM, Hayes KR, Pande K, Glover E, Silander T, Craven MW, Reddy JK, Jovanovich SB. Identification of toxicologically predictive gene sets using cDNA microarrays. *Molecular Pharmacology*. 2001; 60(6):1189–94. <https://www.ncbi.nlm.nih.gov/pubmed/11723225>. [PubMed: 11723225]
- Wang RL, Bencic D, Biales A, Lattier D, Kostich M, Villeneuve D, Ankley GT, Lazorchak J, Toth G. DNA Microarray-based ecotoxicological biomarker discovery in a small fish model species. *Environmental Toxicology and Chemistry*. 2008; 27(3):664–75. DOI: 10.1897/07-192.1 [PubMed: 17990946]
- Wang RL, Bencic D, Biales A, Flick R, Lazorchak J, Villeneuve D, Ankley GT. Discovery and validation of gene classifiers for endocrine-disrupting chemicals in zebrafish (*Danio rerio*). *BMC Genomics*. 2012; 13(1):358.doi: 10.1186/1471-2164-13-358 [PubMed: 22849515]
- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009; 10(1):57–63. DOI: 10.1038/nrg2484
- Waring JF, Ciurlionis R, Jolly RA, Heindel M, Ulrich RG. Microarray analysis of hepatotoxins in vitro reveals a correlation between gene expression profiles and mechanisms of toxicity. *Toxicology Letters*. 2001; 120(1):359–68. DOI: 10.1016/S0378-4274(01)00267-3 [PubMed: 11323195]
- Yang L, Kemadjou JR, Zinsmeister C, Bauer M, Legradi J, Müller F, Pankratz M, Jäkel J, Strähle U. Transcriptional profiling reveals barcode-like toxicogenomic responses in the zebrafish embryo. *Genome Biology*. 2007; 8(10):R227.doi: 10.1186/gb-2007-8-10-r227 [PubMed: 17961207]
- Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates; In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2002. 694–699.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005; 67(2):301–20. DOI: 10.1111/j.1467-9868.2005.00503.x

Figure 1.

Species sensitivity distribution for a pesticide. Each circle is the smallest published acute LC₅₀ for a different genus, the solid curve is a lognormal fit, the dashed curves indicate 95% empirical bootstrap confidence bounds for the fit, and the dotted vertical and horizontal lines intersect at the concentration where 5% of species are expected to suffer at least 50% population loss, assuming exposure durations like those represented in the SSD.

Figure 2. Simulated overlap between probability distributions of biomarker levels in unexposed controls (solid curve A), low toxicant concentration (dotted curve B) and high concentration (dashed curve C).

Figure 3. Simulated extreme overfitting with a 2-parameter OLS fit to two data points (plot A), a 3-parameter fit to three points (B), or a 4-parameter fit to four points (C). OLS fits are depicted by solid curves. The subset of data points used for fitting are triangles. The point generating function is depicted by a dashed line. The remainder of 100 ‘observations’ generated by adding random noise to the generating function are depicted as dots.

Figure 4. Simulated ROC curves for an uninformative (random) classifier (dotted diagonal line), perfect classifier (dashed 'step'), and more typical classifier (solid curve). The vertical axis is the true positive fraction (sensitivity), and the horizontal axis is the false positive fraction (or one minus the specificity).

Table 1

Posterior probability as a function of prior probability and composite likelihood ratio $L(\mathbf{x})$ from a set of replicate tests.

Prior	$L(\mathbf{x})$	Posterior
0.1	10	0.526
0.1	100	0.917
0.1	1,000	0.991
0.01	100	0.503
0.01	1,000	0.910
0.01	10,000	0.990
0.001	1,000	0.500
0.001	10,000	0.909
0.001	100,000	0.990
0.0001	10,000	0.500
0.0001	100,000	0.909
0.0001	1,000,000	0.990