Environmental Pollution 221 (2017) 427–436

Contents lists available at ScienceDirect

Environmental Pollution

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

Prior knowledge-based approach for associating contaminants with biological effects: A case study in the St. Croix River basin, MN, WI, USA☆

Anthony L. Schroeder [a, b], Dalma Martinović-Weigelt [c], Gerald T. Ankley [b], Kathy E. Lee [d], Natalia Garcia-Reyero [e, f], Edward J. Perkins [e], Heiko L. Schoenfuss [g], Daniel L. Villeneuve [b, *]

[a] University of Minnesota – Twin Cities, Water Resources Center, 1985 Lower Buford Circle, St. Paul, MN 55108, USA
[b] U.S. Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Duluth, MN 55804, USA
[c] University of St. Thomas, Department of Biology, Mail OWS 390, 2115 Summit Ave, Saint Paul, MN 55105, USA
[d] U.S. Geological Survey, Toxic Substances Hydrology Program, Grand Rapids, MN 55744, USA
[e] U.S. Army Engineer Research and Development Center – Environmental Laboratory, Vicksburg, MS 39180, USA
[f] Mississippi State University – Institute for Genomics Biocomputing and Biotechnology, Starkville, MS 39762, USA
[g] Aquatic Toxicology Laboratory, WSB-273, St., Cloud State University, St. Cloud, MN 56301, USA

ARTICLE INFO

Article history:
Received 7 June 2016
Received in revised form
14 November 2016
Accepted 3 December 2016
Available online 8 December 2016

Keywords:
Contaminants
Chemical mixtures
Chemical-gene interactions
Comparative toxicogenomics database
Adverse outcome pathway

ABSTRACT

Evaluating potential adverse effects of complex chemical mixtures in the environment is challenging. One way to address that challenge is through more integrated analysis of chemical monitoring and biological effects data. In the present study, water samples from five locations near two municipal wastewater treatment plants in the St. Croix River basin, on the border of MN and WI, USA, were analyzed for 127 organic contaminants. Known chemical-gene interactions were used to develop site-specific knowledge assembly models (KAMs) and formulate hypotheses concerning possible biological effects associated with chemicals detected in water samples from each location. Additionally, hepatic gene expression data were collected for fathead minnows (*Pimephales promelas*) exposed *in situ*, for 12 d, at each location. Expression data from oligonucleotide microarrays were analyzed to identify functional annotation terms enriched among the differentially-expressed probes. The general nature of many of the terms made hypothesis formulation on the basis of the transcriptome-level response alone difficult. However, integrated analysis of the transcriptome data in the context of the site-specific KAMs allowed for evaluation of the likelihood of specific chemicals contributing to observed biological responses. Thirteen chemicals (atrazine, carbamazepine, metformin, thiabendazole, diazepam, cholesterol, p-cresol, phenytoin, omeprazole, ethyromycin, 17β-estradiol, cimetidine, and estrone), for which there was statistically significant concordance between occurrence at a site and expected biological response as represented in the KAM, were identified. While not definitive, the approach provides a line of evidence for evaluating potential cause-effect relationships between components of a complex mixture of contaminants and biological effects data, which can inform subsequent monitoring and investigation.

Published by Elsevier Ltd.

1. Introduction

Evaluating the potential human health and ecological risks associated with exposure to complex chemical mixtures in the ambient environment is one of the central challenges of chemical safety assessment and environmental protection. To assess these risks and take appropriate management actions, there are a number of important questions that need to be addressed through research and/or monitoring efforts. These include: (1) what contaminants are present at a site and what is the potential for exposure to those contaminants; (2) what hazards may be associated with exposure to those contaminants; (3) what evidence exists that

☆ This paper has been recommended for acceptance by Maria Cristina Fossi.

* Corresponding author. US EPA, Mid-Continent Ecology Division 6201 Congdon Blvd, Duluth, MN 55804, USA.
E-mail address: Villeneuve.dan@epa.gov (D.L. Villeneuve).

http://dx.doi.org/10.1016/j.envpol.2016.12.005
0269-7491/Published by Elsevier Ltd.

these hazards are occurring in exposed populations; (4) which contaminant(s) are most likely causing the effects observed; and (5) what is (are) their source(s).

Environmental monitoring has historically relied heavily upon targeted instrumental analysis for chemicals of known or potential concern. While chemical monitoring is well suited to identify and characterize targeted chemicals, it provides little or no information about potential biological effects. Chemical monitoring can be effective as a basis for environmental risk assessment when the hazards associated with detected chemicals are well characterized in terms of potency, effect concentration(s), and/or mode(s) of action, as is the case for many "legacy contaminants" such as PCBs and organochlorine pesticides. However, there are tens of thousands of chemicals for which little or no relevant toxicology data are available (Judson et al., 2009). In the case of these "contaminants of emerging concern" (CECs), chemical monitoring data alone are generally insufficient to support site-specific risk assessment and management.

Effects-based monitoring approaches can provide a useful complement to chemical monitoring. They allow for a direct measurement of biological effects which, if properly anchored to adverse outcomes, can be used to address hazards that may be associated with exposure of extant organisms (Altenburger et al., 2015; Brack et al., 2015; Ekman et al., 2013; Schroeder et al., 2016). Because effects-based monitoring tools measure the integrated biological activity of an entire mixture, they are capable of detecting exposure to chemicals, which investigators may not know to measure, or may not have the analytical methods to detect (Connon et al., 2012; Ekman et al., 2013). While many effects-based monitoring approaches provide a relatively narrow scope of characterization (Altenburger et al., 2015), more open-ended or unsupervised approaches can be employed to cover and evaluate a broader spectrum of biological effects. These include, for example, omics measurements performed on exposed organisms (Berninger et al., 2014; Garcia-Reyero et al., 2008, 2009, 2011; Martinović-Weigelt et al., 2014; Skelton et al., 2014), as well as batteries of pathway-based *in vitro* assays (Escher et al., 2014; Schroeder et al., 2016).

Despite their strengths, effects-based methods have important limitations. Notably, they rarely provide insights into which chemicals are causing the observed biological responses unless coupled with detailed and often costly and time-consuming bioassay-directed fractionation. Without the ability to connect specific chemicals, or at least chemical classes, to a particular effect, it is difficult to determine appropriate management actions. Consequently, approaches that integrate chemical monitoring with biological effects data may be useful to address the questions outlined above and to evaluate risks associated with specific chemicals present in the environment.

Combination of statistical and knowledge-based approaches to data integration can offer efficient means to generate additional lines of evidence that can inform subsequent research, monitoring, or decision-making as appropriate. Existing computational approaches can be used to build network models based on *a priori* knowledge about chemical exposures and biological effects which can allow for integration of these two types of monitoring data (Chindelevitch et al., 2012; Hoeng et al., 2012). For example, Reverse Causal Reasoning, a reverse engineering algorithm, has been used to identify chemicals that provide statistically significant explanations for differential measurements in a molecular profiling data set (Catlett et al., 2013). For this approach, *a priori* knowledge is first used to generate a large network of potential cause and effect relationships, i.e., a Knowledge Assembly Model (KAM). Smaller networks, termed hypotheses (HYPs), are derived from the KAM. For each HYP, the upstream node represents an experimental perturbation such as exposure to a chemical and the downstream nodes represent biological effects, such as a change in mRNA abundance. The edges in the networks specify an "increased", "decreased", or "ambiguous" relationship between chemicals and biological effects. These networks can then be evaluated for richness, which refers to the number of significantly increased or decreased downstream nodes relative to the entire population of nodes, and concordance, which refers to the consistency of the observed state, such as an increase or decrease in mRNA abundance, with the direction of change expected in response to the upstream node (Martin et al., 2012; Laifenfeld et al., 2014).

A KAM can be derived from a knowledge base that provides the cause and effect relationships necessary to develop the network. A number of publically available online resources have assembled, curated, and organized information about chemical-gene and chemical-protein interactions into computationally-accessible databases (Schroeder et al., 2016). For example, both the Search Tool for Interactions of Chemicals (STITCH; Kuhn et al., 2012) and the Comparative Toxicogenomics Database (CTD; Davis et al., 2013) provide information about the impacts of chemicals on biological responses utilizing experimental data from controlled laboratory studies.

Despite unavoidable limitations in terms of chemical and taxonomic coverage (e.g., heavy mammalian bias) and general lack of data for dose-, time-, target-dependency, and route of exposure for many of the chemicals, these sources nonetheless provide a knowledge base suitable for building qualitative KAMs that can be used as a tool for integrated analysis of chemical monitoring and biological effects data. For example, when only chemical monitoring data are available, the KAMs could be a useful first step for identifying contaminants of concern and hypothesizing the potential downstream biological impacts (i.e., perturbed genes or pathways; Schroeder et al., 2016). When both chemical and biological data are available for a site, HYPs derived from KAMs can support statistically guided inference concerning which chemicals are potentially associated with the observed biological responses (Martin et al., 2012). For example, studies utilizing this approach have identified biological effect signatures due to 2-butoxyethanol exposure (Laifenfeld et al., 2010) and drug-induced damage in the liver (Laifenfeld et al., 2014). Thus, KAMs have the potential to identify possible biological effects associated with a particular chemical exposure or, conversely, potential chemical causes associated with a given biological effect.

Recently, we used a KAM based on information in the CTD to predict the biological impacts of chemicals on field-exposed fish when chemical and biological data were not obtained simultaneously (Martinović-Weigelt et al., 2014). The objective of the present study was to further demonstrate the potential utility of knowledge-based approaches. Specifically, we used knowledge from the CTD to develop chemical-gene interaction network models (i.e., KAMs) and applied them, not only to predict potential biological effects of chemicals but also to identify the chemicals in environmental samples that may be associated with observed biological responses (Fig. 1). To achieve this we first measured contaminant concentrations in water collected at five locations associated with two wastewater treatment plants (WWTPs) as well as relative hepatic mRNA transcript abundance in fish exposed *in situ* at each location. The CTD was used to identify genes whose expression had been previously reported to be affected by one or more of the detected chemicals. A KAM was developed for each location to generate location-specific hypotheses about the potential impacts of the chemicals detected in the environmental samples on gene expression. Reverse Causal Reasoning and the KAMs were then used to statistically evaluate HYPs as a means to identify which chemicals in the environmental samples were