Building Quantitative Prediction Models for Tissue Residue of Two Explosives Compounds in Earthworms from Microarray Gene Expression Data

ARTICLE *in* ENVIRONMENTAL SCIENCE & TECHNOLOGY · JANUARY 2012

Impact Factor: 5.33 · DOI: 10.1021/es201187u · Source: PubMed

Ping Gong
Engineer Research and Development Center…

George Tucker
Massachusetts Institute of Technology

Barbara Lynn Escalon
Engineer Research and Development Center…

Edward Perkins
Engineer Research and Development Center…

Environmental Science & Technology

ARTICLE

pubs.acs.org/est

Building Quantitative Prediction Models for Tissue Residue of Two Explosives Compounds in Earthworms from Microarray Gene Expression Data

Ping Gong,*,† Po—Ru Loh,‡ Natalie D. Barker,† George Tucker,‡ Nan Wang,§ Chenhua Zhang,‖ B. Lynn Escalon,⊥ Bonnie Berger,*,‡ and Edward J. Perkins⊥

†Environmental Services, SpecPro Inc., San Antonio, Texas, United States

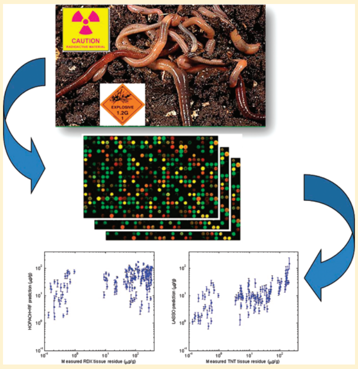
‡Department of Mathematics and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

§School of Computing, University of Southern Mississippi, Hattiesburg, Mississippi, United States

‖Department of Mathematics, University of Southern Mississippi, Hattiesburg, Mississippi, United States

⊥Environmental Laboratory, U.S. Army Engineer Research and Development Center, Vicksburg, Mississippi, United States

S *Supporting Information*


CAUTION
RADIOACTIVE MATERIAL
EXPLOSIVE
1.2G
1
HOPACH+RF prediction (μg/g)
Measured RDX tissue residue (μg/g)
LASSO prediction (μg/g)
Measured TNT tissue residue (μg/g)

ABSTRACT: Soil contamination near munitions plants and testing grounds is a serious environmental concern that can result in the formation of tissue chemical residue in exposed animals. Quantitative prediction of tissue residue still represents a challenging task despite long-term interest and pursuit, as tissue residue formation is the result of many dynamic processes including uptake, transformation, and assimilation. The availability of high-dimensional microarray gene expression data presents a new opportunity for computational predictive modeling of tissue residue from changes in expression profile. Here we analyzed a 240-sample data set with measurements of transcriptomic-wide gene expression and tissue residue of two chemicals, 2,4,6-trinitrotoluene (TNT) and 1,3,5-trinitro-1,3,5-triazacyclohexane (RDX), in the earthworm *Eisenia fetida*. We applied two different computational approaches, LASSO (Least Absolute Shrinkage and Selection Operator) and RF (Random Forest), to identify predictor genes and built predictive models. Each approach was tested alone and in combination with a prior variable selection procedure that involved the Wilcoxon rank-sum test and HOPACH (Hierarchical Ordered Partitioning And Collapsing Hybrid). Model evaluation results suggest that LASSO was the best performer of minimum complexity on the TNT data set, whereas the combined Wilcoxon-HOPACH-RF approach achieved the highest prediction accuracy on the RDX data set. Our models separately identified two small sets of ca. 30 predictor genes for RDX and TNT. We have demonstrated that both LASSO and RF are powerful tools for quantitative prediction of tissue residue. They also leave more unknown than explained, however, allowing room for improvement with other computational methods and extension to mixture contamination scenarios.

1. INTRODUCTION

The exposure concentration used in existing environmental toxicity data and in risk assessment and regulatory activities is a dose metric based on the concentration in the external media to which test organisms are exposed. The use of tissue residue or body burden began 100 years ago, though only in a limited manner, with the publication of the Meyer—Overton theory on the narcotic effects of organic compounds in tissues.[1,2] Recently, the "tissue residue approach for toxicity assessment" or "tissue residue-effects approach" was revisited and the following consensus was reached: when toxicity is defined in terms of tissue concentrations, the variability is often reduced substantially because the toxicokinetics and bioavailability characteristics for that compound are incorporated in the tissue residue determinations.[3] Tissue residues of chemicals, often measured using analytical chemistry methods, can represent the biologically effective dose, i.e., the concentration or dose at the target site, when a direct proportionality exists between them.[4]

Whereas it is the interaction of a toxicant with biomolecules at the target site that triggers the toxic action,[4] there may be thousands of potential targets for any given toxicant, resulting in many different possible scenarios of mode of action. One promising tool for surveying all potential targets is the microarray, after which specific biomolecule(s) can be identified through computational analysis of the cause (dose)—effect relationship.

Special Issue: Ecogenomics: Environmental

Received: April 8, 2011
Accepted: July 21, 2011
Revised: June 23, 2011
Published: July 21, 2011

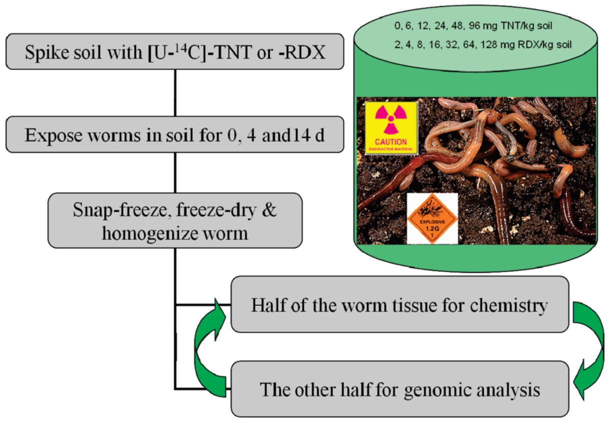

Spike soil with [U-14C]-TNT or -RDX
Expose worms in soil for 0, 4 and14 d
Snap-freeze, freeze-dry & homogenize worm
Half of the worm tissue for chemistry
The other half for genomic analysis
0, 6, 12, 24, 48, 96 mg TNT/kg soil
2, 4, 8, 16, 32, 64, 128 mg RDX/kg soil
CAUTION
RADIOACTIVE MATERIAL
EXPLOSIVE
1.2G

Figure 1. Experimental design.

Transcriptomic analysis bridges the gap between exposure and biochemical/physiological effects. One can discover mechanisms of toxic actions and novel biomarkers by combining transcriptomics with tissue residue analysis. In addition, the resulting toxicogenomic data is being accepted by the U.S. Environmental Protection Agency as part of a weight-of-evidence approach for establishing mechanisms of toxicity for regulated substances.[5] Many other national and international organizations (e.g., the Organization for Economic Cooperation and Development's Environment, Health and Safety Program) have also started to explore and evaluate the use of genomics for risk assessment of chemicals.[6]

To develop earthworm biomarkers of exposure to chemicals in soil, we have previously identified 58 classifier genes from a 15K-probe microarray expression data set. These 58 genes were used in a support vector machine model to separate with high accuracy (83.5%) three groups of earthworm samples: unexposed control, TNT (2,4,6-trinitrotoluene)-exposed and RDX (1,3,5-trinitro-1,3,5-triazacyclohexane)-exposed.[7] In this effort we extend that study to identify a small gene set that can predict tissue residue concentrations. These genes are believed to have great potential to provide insights into toxicokinetics because of their strong ties to total residue levels in earthworm body tissues. Specifically, we applied two different computational approaches to identifying predictor/classification genes and building predictive models: (a) a regression method named LASSO (Least Absolute Shrinkage and Selection Operator) [8] and (b) a decision tree algorithm named Random Forest (RF).[9] Additionally, while both methods are robust to noise, we found that for RDX, prediction accuracy was improved by first reducing the gene set under consideration by applying statistical tests for differential expression followed by HOPACH (Hierarchical Ordered Partitioning And Collapsing Hybrid[10]), an unsupervised clustering algorithm.

2. EXPERIMENTAL SECTION

A 4-day preliminary exposure experiment to identify appropriate chemical concentrations and the most sensitive analytical method revealed that nominal soil amendment concentrations of 0−96 mg [U−^{14}C] TNT/kg or 0−128 mg [U−^{14}C] RDX/kg caused no mortality, and that the Combustion−Scintillation Counting (CSC) method (see Section 2.2 for details) had a consistently higher recovery rate of amended chemicals in both soil and earthworm tissues than the SE-HPLC-ECD (Solvent Extraction followed by High Performance Liquid Chromatography coupled with Electrochemical Detector) method and the SE-HPLC-RFD (Solvent Extraction followed by HPLC equipped with Radioactivity Flow Detector) method (see Supporting Information and ref 11). In good agreement with previous reports,[12,13] our preliminary results also showed that TNT quickly broke down after being amended in soil and little parent compound was recovered in earthworm tissue samples. These results warranted the use of radio-labeled chemicals and the CSC method.

2.1. Experimental Design (see Figure 1).

Earthworms were reared in continuous lab culture as previously described.[14] Mature adults bearing a clear clitellum were used for the entire study. A pristine sandy loam soil amended with TNT or RDX was used for exposure. Three sets of exposure−response experiments were conducted. The first set was a 4-day exposure without purging before takedown. In the second and the third sets, earthworms were allowed to purge gut content overnight on moistened filter paper immediately before takedown after 4-day and 14-day exposures. Nominal exposure concentrations were as follows: 0, 6, 12, 24, 48, 96 mg [U−^{14}C] TNT/kg, or 8, 16, 32, 64, 128 mg [U−^{14}C] RDX/kg for 4 days (the first set) or 14 days (the third set). The 4-day exposure was repeated (the second set) with the same TNT concentrations but different RDX concentrations (2, 4, 8, 16, and 32 mg/kg). Nominal concentrations were verified by HPLC-RFD to be less than 10% variation from target concentrations. Ten worms were added to 325 g (first and third sets) or 100 g (second set) of amended soils per treatment. All takedowns were executed by snap-freezing worms in liquid nitrogen for 3−5 min. Frozen worms were then transferred to −80 °C for storage. Before being split for subsequent chemistry analysis and gene expression profiling, worm body tissue was dried at −40 °C for 48 h in a FreeZone Plus 6 Liter Cascade Console Freeze-Dry System (Labconco, Kansas City, MO) and then homogenized.

2.2. CSC Method (or Radio-Labeled Tracer Method[11]).

Lyophilized and homogenized earthworm tissue was measured into triplicate samples averaging 15 mg. Two instruments were used to measure tissue residues: a Perkin-Elmer model A307 Oxidizer (Waltham, MA) and an R. J. Harvey OX 600 Biological Oxidizer (Tappan, NY). Standards were burned in both instruments for proper calibration. Each sample combusted with the Perkin-Elmer model was placed in an empty Combusto-Cone, loaded into the ignition basket, and burned for 1.5 min. Trapped $^{14}CO_2$ was collected in a mixture of Carbo-Sorb and Permafluor-E+. An empty Combusto-Cone was burned after each sample to prevent carryover. Ultima Gold scintillation cocktail (Perkin-Elmer) was added to the trap before measurement of radioactivity. With the R. J. Harvey oxidizer, the earthworm tissue was transferred into a ceramic boat and loaded into the instrument's combustion chamber, which was heated to 900 °C. Oxygen flow rate was 350 cc/min, and the samples were burned for 3 min. Combustion products then passed through a catalyst bed, and evolved $^{14}CO_2$ was collected in a removable external trap filled with 15 mL of Carbon-14 cocktail (R. J. Harvey). The trap was rinsed with methanol after each use, and a blank was run after each triplicate. Radioactivity was determined using a Packard Tri-Carb 2500TR Liquid Scintillation Counter (Meriden, CT).

2.3. Gene Expression Profiling.

Total RNA was extracted from earthworm tissue using Qiagen RNeasy Mini Kit (Valencia, CA). A custom-designed 15K *E. fetida* oligo array printed by Agilent (Santa Clara, CA) was used for hybridization. These 60-mer oligo probes were selected from a pool of 63 541 validated

Table 1. Earthworm Survival and Number of Earthworms Hybridized for the Three Sets of Toxicity Testing Experiments

treatment[a]	no. of worms survived/hybridized/measured/both hybridized and measured[b]				
	4-day_original (set 1)	4-day_repeat (set 2)	14-day (set 3)	day 0	total
solvent control	10/8/10/8	9/8/9/8	8/8/8/8	10/8/3/2	
6 mg TNT/kg soil	10/8/10/8	10/8/10/8	10/8/10/8		
12 mg TNT/kg soil	10/8/9/8	10/8/10/8	10/8/10/8		
24 mg TNT/kg soil	10/8/10/8	10/8/10/8	10/8/10/8		
48 mg TNT/kg soil	10/8/10/8	0/0/0/0	10/8/9/8		
96 mg TNT/kg soil	10/8/10/8	0/0/0/0	0/0/0/0		
8 (2) mg RDX/kg soil	10/8/10/8	10/8/10/8	10/8/10/8		
16 (4) mg RDX/kg soil	9/8/7/6	10/8/10/8	10/8/10/8		
32 (8) mg RDX/kg soil	10/8/10/8	10/8/10/8	10/8/10/8		
64 (16) mg RDX/kg soil	10/8/10/8	10/8/10/8	10/8/10/8		
128 (32) mg RDX/kg soil	10/8/10/8	10/8/10/8	9/8/9/8		
total worms survived	109	89	97	10	305
total worms hybridized	88	72	80	8	248
total worms measured	106	89	96	3	294
total worms both hybridized and measured[b]	86	72	80	2	240

[a] RDX concentrations in the second set of experiments are given in parentheses. [b] Both hybridized for gene expression profiling and chemically measured for tissue residues.

probes, each targeting a unique *E. fetida* transcript.[15] Sample cRNA synthesis, labeling, hybridization, and microarray processing were performed according to manufacturer's protocol One-Color Microarray-Based Gene Expression Analysis (version 1.0). Agilent One-Color Spike-Mix was diluted 5000-fold and 5 mL of the diluted spike-in mix was added to 500 ng of each of the total RNA samples prior to labeling reactions. After hybridization, the arrays were scanned at PMT level 350 using GenePix 4200AL scanner (Molecular Devices, Sunnyvale, CA). Gene expression data were acquired from scanned array images using Agilent's Feature Extraction Software (v.9.1.3).

2.4. Microarray Data Preprocessing. First signal intensity was converted into relative RNA concentration based on the linear standard curve of spike-in RNAs. Next, tissue residue values and relative RNA concentrations were log-transformed. For cross-array normalization, each array's median log relative RNA concentration was subtracted from the values. For intra-array normalization, the expression value of each probe was centered and rescaled to have mean 0 and variance 1 because LASSO regression is sensitive to the scaling of each feature.

2.5. Data Analyses. *2.5.1. Variable Selection via Differential Expression and HOPACH.* As an optional step prior to model-fitting with LASSO and Random Forest, the number of genes under consideration was reduced by differential expression and clustering. A Wilcoxon rank-sum test was applied to identify differentially expressed genes by comparing each concentration group of RDX- or TNT-treated samples with their control group. Unlike the *t* test or ANOVA, the Wilcoxon test is nonparametric and does not assume a normal distribution of samples. Two lists of significant genes were derived (one for RDX and the other for TNT), and were further downsized by running the HOPACH package v. 2.12.0 (Downloads available at bioconductor.org/packages/release/bioc/html/hopach.html). The HOPACH clustering algorithm builds a hierarchical tree of clusters by recursively partitioning a gene expression data set with the Partitioning Around Medoids algorithm, while ordering and possibly collapsing clusters at each level. The algorithm uses the Mean/Median Split Silhouette criteria to identify the level of the tree with maximally homogeneous clusters.

2.5.2. LASSO Modeling. In our first approach to predicting tissue residue values from gene expression we used LASSO regression, a version of least-squares regression that is regularized so that it selects a small subset of variables to use as predictors. LASSO assumes that the data can be modeled well with a few predictors picked from a large pool of unrelated predictors. Specifically, LASSO minimizes

$$\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

with respect to β, where y are the tissue residue values, X are the preprocessed microarray values, β is a coefficient vector, and λ is a regularization parameter. As λ increases, the model tends to select fewer predictors, so a proper choice of λ can discourage overfitting. We used 10-fold cross-validation to select a value of λ resulting in the least complex model scoring within one standard deviation of the optimum, as typically recommended. We used the glmnet package[16] to fit the LASSO model with cross-validation.

2.5.3. Random Forest Modeling. As a complementary approach, we applied random forest regression to the data (preprocessed and transformed as above). Random Forest is a decision-tree method that works efficiently on large data sets, producing a prediction model along with an importance estimate for each input variable. We used the R package randomForest[17] with the default one-third of predictors considered per split. To ensure a fair estimate of performance, we applied 10-fold cross-validation as above.

3. RESULTS

Each treatment had 10 replicate worms with 8−10 survivors at the end of exposure, except the two highest TNT concentrations (Table 1). A total of 305 worms survived in the three sets of experiments, including 10 worms sampled at the beginning of the experiments (served as Day 0 controls). Eight worms per treatment (except for three TNT treatments where all

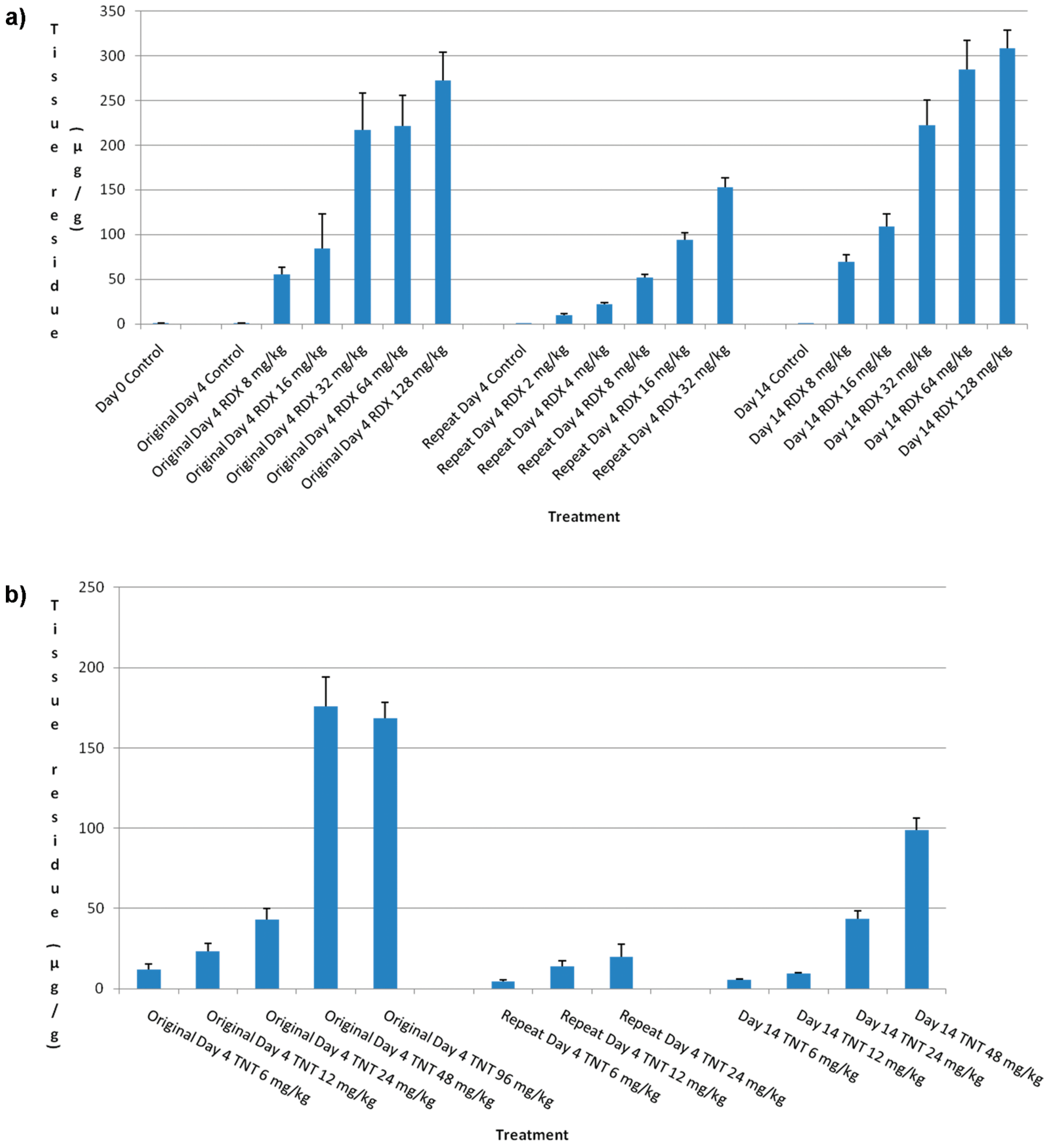


Figure 2. Tissue residue of TNT or RDX in earthworms exposed for 0, 4, or 14 days. Data are shown as mean (column) + standard deviation (error bar) with $n = 7-10$ except for $n = 3$ in Day 0 control.

worms died) were examined for gene expression, resulting in 248 one-color hybridizations. Tissue residues were measured in 294 worm samples. A total of 240 worm samples had both chemistry and gene expression data due to tissue shortage in 8 samples (6 in Day 0 control and 2 in 4-day_original 16 mg RDX/kg soil) caused by repeats of failed or low-quality measurements.

3.1. Tissue Residues. One-third of the worm tissue samples were combusted in the Perkin-Elmer oxidizer, and the remaining two-thirds in the R.J. Harvey oxidizer. No significant difference was observed between instruments in their combustion and trapping efficiency. The average tissue residue concentrations per treatment as recovered radioactivity are shown in Figure 2. At the same nominal concentration, tissue residues were higher in the original 4-day exposure than in the repeat 4-day exposure. This might be attributed to two factors: (1) worms were not purged for the original 4-day exposure, leaving extra radio-labeled TNT or RDX in the gut content; and (2) more soil was used in the original 4-day exposure (325 g) than in the repeat 4-day exposure (100 g), providing a larger TNT/RDX pool for worm uptake. Tissue residues had a positive correlation with nominal soil concentrations. Exposure time was also positively correlated with tissue residue of RDX, but not with that of TNT, probably due to TNT mineralization by soil microbes.[13,18]

3.2. Gene Expression Data. Gene expression was profiled in 248 earthworm samples including 32 controls, 120 RDX-treated, and 96 TNT-treated (Table 1). This MIAME compliant 248-array data set has been deposited in NCBI's Gene Expression Omnibus and is accessible through GEO Series accession number GSE18495 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18495).[7]

3.3. Variable Selection via Differential Expression and HOPACH. The Wilcoxon rank-sum test identified 564 and 4187 genes significantly altered by RDX (alpha = 0.0001) and TNT (alpha = 0.001), respectively. HOPACH further reduced them to 33 (RDX) and 55 (TNT) genes by choosing cluster medoids.

3.4. Model Performance. We evaluated the performance of each tissue residue prediction model—LASSO and Random Forest with and without prior variable selection—according to mean squared error (MSE) of prediction on the log-transformed uptake values. To reduce the variation caused by random cross-validation fold assignments, we averaged the MSE of prediction over 10−50 runs of 10-fold cross-validation (Table 2). On the RDX data set, the method of variable selection using the Wilcoxon rank-sum test and HOPACH followed by model-fitting with Random Forest achieved the best performance, explaining roughly one-quarter of the variance. This method significantly outperformed the next-best, Random Forest alone, with p-value 0.0008 according to Student's t test. In contrast, on the TNT data set, LASSO, Random Forest, and HOPACH+LASSO each explained close to half of the variance, with no significant difference in MSE according to the t test. LASSO, however, produced the least complex model, making it preferable over the other methods. Figure 3 plots the tissue residue values predicted by the best performers (aggregated over several 10-fold cross validation runs) against the true values, showing reasonably good correlations given experimental noise.

Table 2. Comparison of Modeling Approaches: Mean Squared Prediction Error and Model Sizes[a]

	RDX		TNT	
modeling approach	MSE of prediction	model size	MSE of prediction	model size
LASSO	5.02	59	2.55	31
Random Forest	4.56	∼100	2.46	∼50
HOPACH+LASSO	5.26	4	2.48	28
HOPACH+RF	4.26	∼30	2.64	∼30

[a] The MSE of prediction was computed for log uptake values and averaged over several (10−50) runs of 10-fold cross-validation to reduce variability caused by fold assignments. For comparison, the population variances of the log RDX and log TNT uptake values are 5.56 and 4.30, respectively. Model performance for Random Forest is less sensitive to the number of genes used, so approximate optimal model sizes are listed for the RF models.

3.5. Genes of Interest. LASSO and Random Forest each identified a small number of genes of interest from among the 15208 probes assayed and within the subsets selected by HOPACH (Table 2). The output of the LASSO regression model is an estimated coefficient vector $\hat{\beta}$, indicating which genes are in the model and to what extent they are important in predicting the tissue residue vector values. LASSO selected 31 genes important in predicting TNT tissue residue, the data set on which it was a best performer. These are listed in Table 3 with biological annotations, along with the 33 genes selected by HOPACH for RDX, all used in the best-performing HOPACH +Random Forest model for RDX. In addition to choosing a small subset of variables to include in a model, Random Forest also ranks all available features according to their estimated importance. We compared the probes from LASSO to the rank list from Random Forest and found that without prior variable selection, almost all genes used by LASSO were among the top 1% identified by Random Forest and several were among the top 0.1%.

4. DISCUSSION

Microarray gene expression data have been extensively used for discovery of classifiers and biomarkers in phenotypic

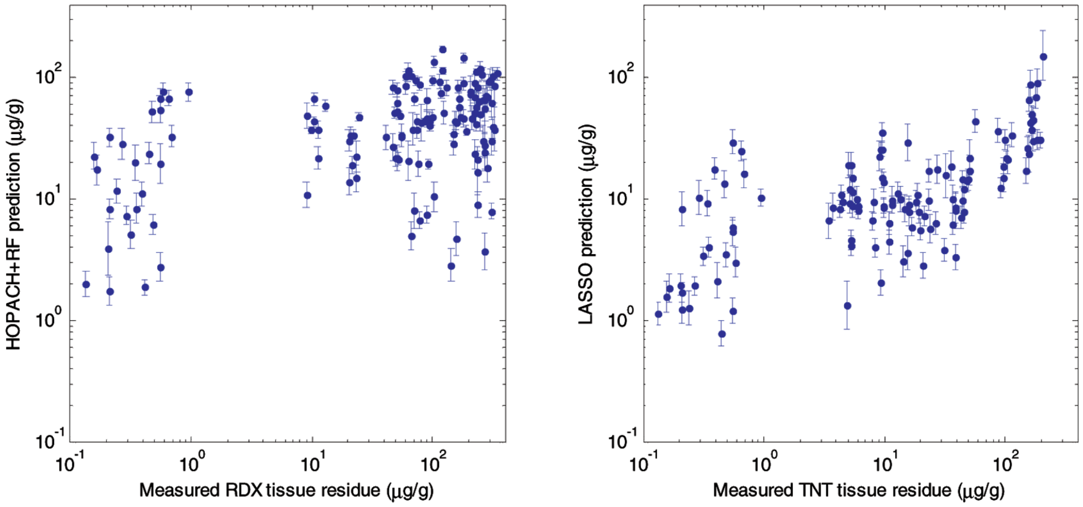


Figure 3. Log−log scatter plots of cross-validated model predictions for RDX and TNT tissue residue versus actual values, using the best-performing models with least complexity: HOPACH+RF for RDX and LASSO for TNT. Prediction values in each plot are aggregated over several runs of 10-fold cross-validation; error bars represent one standard deviation away from the mean prediction.

Table 3. Probe Names and Functional Annotation of Predictor Genes for TNT and RDX Residues in Earthworm Tissue Selected by LASSO (TNT) and HOPACH (RDX)

TNT probe	target gene function	RDX probe	target gene function
TA1-011325	RNA polymerase II	TA1-149667	electron transferring flavoprotein
TA1-019914	olfactory receptor	TA2-058009	acylpeptide hydrolase
TA1-022273	60S ribosome subunit (biogenesis)	TA2-134512	phosphoserine aminotransferase
TA1-048554	EVCP-2	TA2-167807	erythrocyte membrane protein
TA1-053819	nucleoplasmin	TA1-215461	eukaryotic rRNA processing
TA1-059599	unknown	TA1-099490	40S ribosomal protein S8
TA1-082130	general transcription factor	TA2-051630	glutathione S-transferase
TA1-086892	cytochrome oxidase subunit II-like (COX II)	TA1-113395	GRIM-19
TA1-094629	glycosyl hydrolase family 20b	TA2-055248	mitochondrial carrier
TA1-095858	angiotensinogen	TA2-197772	unknown
TA1-098972	rhodopsin kinase	TA2-124905	general transcription factor
TA1-100698	unknown	TA1-058331	I-connectin (TITIN)
TA1-118468	ATP-dependent RNA helicase	TA2-130522	CCR4 NOT-related
TA1-126665	DUF1682	TA1-053314	zinc metallopeptidase
TA1-131169	unknown	TA1-233321	RNA polymerase II associated factor
TA1-144534	zinc ion binding	TA2-018017	testis expressed gene 2
TA1-145408	kinesin	TA2-031252	unknown
TA1-190795	unknown	TA1-129649	EVCP-1
TA1-208940	unknown	TA2-180958	CAAX prenyl protease
TA2-030572	selenoprotein W2	TA1-084735	unknown
TA2-058110	heterogeneous nuclear ribonucleoprotein	TA2-006823	phosphatase 2A
TA2-091957	Unknown	TA2-103803	unknown
TA2-092830	Unknown	TA2-067432	unknown
TA2-095186	MRP-related nucleotide-binding protein	TA2-122279	nucleosome assembly
TA2-108467	aldehyde dehydrogenase	TA2-202612	unknown
TA2-154278	methylcrotonoyl-CoA carboxylase	TA1-034042	unknown
TA2-156504	unknown	TA2-048997	endothelin-converting enzyme-related
TA2-164381	acid phosphatase	TA1-124927	twinfilin
TA2-188381	unknown	TA1-175625	CAAX prenyl protease
TA2-206312	glucose regulated protein	TA2-075017	Ser/Thr protein kinase
TA2-210587	unknown	TA1-111897	YTH domain-containing
		TA1-101100	RNA polymerase III
		TA1-063930	nuclear movement protein related

classification research such as human diseases and environmental contamination. For instance, four drug sensitive genes were identified from genome-wide gene expression data sets, and two linear models built with these predictive marker genes reliably predicted (R^2 = 0.676 or 0.696) clinical response (breast cancer tumor size) to neoadjuvant chemotherapy.[19] However, such studies are rare in quantitative prediction of tissue residue. To the best of our knowledge, the present study is the first attempt at identifying predictor genes and building quantitative models to predict animal uptake of environmental contaminants.

The two computational modeling methods we applied—LASSO and Random Forest—take different approaches to the variable selection problem and have different strengths and weaknesses. LASSO, with its regularization penalty proportional to the sum of the regression coefficients, produces linear models that are sparse (i.e., use few variables) and thus easily interpretable. Furthermore, assuming the process being modeled is in fact linear in a small number of variables and not subject to too much noise, theoretical results exist guaranteeing that LASSO will find approximately the correct model.

When the above assumptions are violated, however, the sparsity of LASSO may be disadvantageous, for example if a continuous measure of variable importance is desired rather than simply the choice of a subset of variables (which may contain inaccuracies). Random Forest is an ensemble decision tree approach that produces such an assessment of variable importance by building a "forest" of many different decision trees; randomness is introduced by restricting each node split to a random subset (typically one-third) of the features available. In contrast to LASSO and unlike traditional decision tree algorithms, Random Forest thus computes a dense model that sacrifices transparency but creates a ranking of all variables according to their estimated importance to the classification or regression.

HOPACH provides another perspective that explicitly seeks to identify and model clusters of features, which are common in genetic data sets. A hybrid between partitioning (top-down) and agglomerative (bottom-up) clustering, HOPACH creates a hierarchical tree of clusters from which representative genes (e.g., medoids) can be chosen to span the original large pool of variables available while avoiding redundancy. This unsupervised clustering approach provides a more direct handle on the variable selection process and can be used as a preprocessing step prior to further modeling on the reduced data as was done here.

In this study, we found that while each technique was slightly better-suited to one data set (LASSO to TNT, Random Forest and HOPACH to RDX), no methodology dominated. Indeed, the similarity in performance of the methods is more striking than the differences, especially considering that LASSO fits a linear model to the data—certainly an oversimplification—whereas Random Forest produces potentially highly nonlinear decision trees. Notably, during our analysis we found that careful preprocessing and normalization of the data to ameliorate microarray artifacts had as large an effect on predictive power as the particular modeling approach taken.

The differences between the gene sets selected by the various models also merit discussion. On a bulk scale, the genes identified by LASSO were clearly also considered to be of high importance by Random Forest; yet, at the same time, out of the top ten genes considered most important by Random Forest, only a few were used by LASSO. We can think of two potential explanations for this behavior. One is redundancy in gene function: for any given biological mechanism relevant to toxicity, it is likely that many genes associated with that mechanism share the same predictive power, and since LASSO selects sparse models, it will tend to choose only one gene from any such set. The second is simply noise. Even within the relatively controlled laboratory setting, any two biological organisms are sure to respond differently despite the same treatment. While in each case our models achieve some success fitting the data, they leave more unknown than explained. There is no perfect set of genes with full predictive power; instead, different model-fitting approaches will take different paths in the struggle to find signal amid the noise.

Tissue residue is a complex phenotype governed by a suite of toxicokinetics processes, including absorption, distribution, metabolism, and excretion (ADME). Hence, many genes of diverse functional groups are potentially involved in ADME and account for the TNT/RDX residue level in earthworm tissue. For instance, our prediction models picked genes putatively coding for enzymes that catalyze a wide range of metabolic processes, including hydrolase, kinase, carboxylase, protease, peptidase, phosphatase, dehydrogenase, cytochrome oxidase, RNA helicase, glutathione S-transferase, and phosphoserine aminotransferase (Table 3). Some of these genes possess opposite functions, e.g., phosphatases that remove phosphate groups from their substrate versus kinases that catalyze the addition of phosphate groups to acceptors.

The functional diversity of model-picked predictor genes well represents the complexity and the interactive and dynamic nature of residue forming processes. In addition, genes in each prediction group have little functional overlap (except for two CAAX prenyl protease in the RDX group). The low redundancy reflects the high efficiency of our computational approaches in selecting gene sets that capture predictive power with low model complexity. On the other hand, it is desirable to view these predictor genes in a pathway/network context and to conduct pathway enrichment analysis so that we may gain an improved understanding about what metabolic pathways are involved in the formation of TNT/RDX residue. However, we are unable to achieve this goal because these genes are distributed in a scattered fashion to many different pathways.

In summary, we have applied LASSO and Random Forest to select predictive genes and build models to predict TNT or RDX residue in earthworm tissues. Each method was tested alone and in combination with prior variable selection using the Wilcoxon rank-sum test and HOPACH. Model evaluation results suggest that LASSO was the best performer of minimum complexity on the TNT data set, whereas the combined Wilcoxon-HOPACH-RF approach achieved the highest prediction accuracy on the RDX data set. Two small sets of ca. 30 predictor genes were identified for RDX and TNT, separately. This study has demonstrated that both LASSO and RF are powerful tools for quantitative prediction of tissue residue. Although our models have achieved some success fitting the data, there is sufficient room left for improvement using other computational methods.

Although this study dealt with soil amended with a single chemical under controlled laboratory conditions, the real-world problem is dominated by mixture scenarios where an approach similar to the one described here may be applied. The mixture scenario certainly warrants further rigorous testing, and we believe that our approach stands a much better chance of achieving the goal of predicting the residue of each contaminant in a mixture scenario compared to the traditional single endpoint approach.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information. Preliminary earthworm toxicity testing results (Supplementary Figure 1); identified predictor gene sets along with their functional annotations and importance rank of all 15208 gene probes estimated by Random Forest-(Supplementary Tables 1−3). This material is available free of charge via the Internet at http://pubs.acs.org. All source code and other details for data analysis and model performance evaluation are available upon request through the corresponding authors.

■ AUTHOR INFORMATION

Corresponding Authors
*Phone: 601-634-3521 (P.G.); 617-253-1827 (B.B.); e-mail: ping.gong@usace.army.mil; bab@mit.edu.

■ ACKNOWLEDGMENT

We thank R. E. Boyd (Engineer Research & Development Center, Environmental Laboratory; ERDC-EL) for maintaining the earthworm culture, X. Guan and C. A. Hayes (ERDC-EL) for caring for the Perkin-Elmer Oxidizer, and Pedro Alvarez and Dong Li (Rice University, Department of Civil and Environmental Engineering) for making the R. J. Harvey Oxidizer available to us when our Perkin-Elmer Oxidizer was down. We also thank the three anonymous reviewers for several suggestions that improved the quality of this manuscript. This work was funded by the U.S. Army Environmental Quality/Installation Applied Research Program as part of projects "Toxicogenomics for Assessment of Munitions and Explosives Related Compounds" and "Impact of Munitions Constituents on Biological Networks". Permission to publish this information was granted by the Chief of Engineers.

REFERENCES

(1) Lipnick, R. L. Structure-activity relationships. In *Fundamentals of Aquatic Toxicology. II. Effects, Environmental Fate, and Risk Assessment*; Rand, G. M., Ed.; Taylor & Francis: Bristol, PA, 1995; pp 609−665.

(2) McCarty, L. S.; Landrum, P. F.; Luoma, S. N.; Meador, J. P.; Merten, A. A.; Shephard, B. K.; van Wezel, A. P. Advancing environmental toxicology through chemical dosimetry: External exposures versus tissue residues. *Integr. Environ. Assess. Manage.* **2011**, 7 (1), 7–27.

(3) Meador, J. P.; Adams, W. J.; Escher, B. I.; McCarty, L. S.; McElroy, A. E.; Sappington, K. G. The tissue residue approach for toxicity assessment: findings and critical reviews from a Society of Environmental Toxicology and Chemistry Pellston Workshop. *Integr. Environ. Assess. Manage.* **2011**, 7 (1), 2–6.

(4) Escher, B. I.; Ashauer, R.; Dyer, S.; Hermens, J. L.; Lee, J. H.; Leslie, H. A.; Mayer, P.; Meador, J. P.; Warne, M. S. Crucial role of mechanisms and modes of toxic action for understanding tissue residue toxicity and internal effect concentrations of organic chemicals. *Integr. Environ. Assess. Manage.* **2011**, 7 (1), 28–49.

(5) U.S. Environmental Protection Agency. *Potential Implications of Genomics for Regulatory and Risk Assessment Applications at EPA*; U.S. Environmental Protection Agency: Washington, DC, 2004.

(6) Van Aggelen, A. G.; Ankley, G. T.; Baldwin, W. S.; Bearden, D. W.; Benson, W. H.; Chipman, J. K.; Collette, T. W.; Craft, J. A.; Denslow, N. D.; Embry, M. R.; Falciani, F.; George, S. G.; Helbing, C. C.; Hoekstra, P. F.; Iguchi, T.; Kagami, Y.; Katsiadaki, I.; Kille, P.; Liu, L.; Lord, P. G.; McIntyre, T.; O'Neill, A.; Osachoff, H.; Perkins, E. J.; Santos, E. M.; Skirrow, R. C.; Snape, J. R.; Tyler, C. R.; Versteeg, D.; Viant, M. R.; Volz, D. C.; Williams, T. D.; Yu, L. Integrating omic technologies into aquatic ecological risk assessment and environmental monitoring: Hurdles, achievements, and future outlook. *Environ. Health Perspect.* **2010**, *118* (1), 1–5.

(7) Li, Y.; Wang, N.; Perkins, E. J.; Zhang, C.; Gong, P. Identification and optimization of classifier genes from multi-class earthworm microarray dataset. *PLoS One* **2010**, *5* (10), e13715.

(8) Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. Royal Stat. Soc. Ser. B (Methodological)* **1996**, *58* (1), 267–288.

(9) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5–32.

(10) van der laan, M. J.; Pollard, K. S. A new algorithm for hybrid clustering of gene expression data with visualization and the bootstrap. *J. Stat. Plan. Inference* **2003**, *117* (2), 275–303.

(11) Gong, P.; Escalon, B. L.; Hayes, C. A.; Perkins, E. J. Uptake of hexanitrohexaazaisowurtzitane (CL-20) by the earthworm *Eisenia fetida* through dermal contact. *Sci. Total Environ.* **2008**, *390* (1), 295–299.

(12) Dodard, S. G.; Powlowski, J.; Sunahara, G. I. Biotransformation of 2,4,6-trinitrotoluene (TNT) by enchytraeids (*Enchytraeus albidus*) in vivo and in vitro. *Environ. Pollut.* **2004**, *131* (2), 263–273.

(13) Stenuit, B. A.; Agathos, S. N. Microbial 2,4,6-trinitrotoluene degradation: Could we learn from (bio)chemistry for bioremediation and vice versa? *Appl. Microbiol. Biotechnol.* **2010**, *88* (5), 1043–1064.

(14) Pirooznia, M.; Gong, P.; Guan, X.; Inouye, L. S.; Yang, K.; Perkins, E. J.; Deng, Y. Cloning, analysis and functional annotation of expressed sequence tags from the earthworm *Eisenia fetida*. *BMC Bioinform.* **2007**, *8* (Suppl 7), S7.

(15) Gong, P.; Pirooznia, M.; Guan, X.; Perkins, E. J. Design, validation and annotation of transcriptome-wide oligonucleotide probes for the oligochaete annelid *Eisenia fetida*. *PLoS One* **2010**, *5* (12), e14266.

(16) Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33* (1), 1–22.

(17) Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2* (3), 18–22.

(18) Robertson, B. K.; Jjemba, P. K. Enhanced bioavailability of sorbed 2,4,6-trinitrotoluene (TNT) by a bacterial consortium. *Chemosphere* **2005**, *58* (3), 263–270.

(19) Sano, H.; Wada, S.; Eguchi, H.; Osaki, A.; Saeki, T.; Nishiyama, M. Quantitative prediction of tumor response to neoadjuvant chemotherapy in breast cancer: Novel marker genes and prediction model using the expression levels. *Breast Cancer* **2011**, (DOI: 10.1007/s12282-011-0263-8).