Classification of Chemicals Based on Concentration-Dependent Toxicological Data Using ToxClust

XIAOWEI ZHANG,*,† JOHN L. NEWSTED,† MARKUS HECKER,†‡ ERIC B. HIGLEY,† PAUL D. JONES,†‡§ AND JOHN P. GIESY†‡§

Toxicology Centre, University of Saskatchewan, 44 Campus Drive, Saskatoon SK S7N 5B3 Canada, ENTRIX, Inc., Okemos, MI, ENTRIX, Inc., Saskatoon, Saskatchewan, Canada, Department of Biomedical Veterinary Sciences, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, Centre for Coastal Pollution and Conservation and Department of Biology and Chemistry, City University of Hong Kong, Kowloon, Hong Kong, SAR China, Department of Zoology, Michigan State University, East Lansing, MI, State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing, and State Key Laboratory of Marine Environmental Science, College of Oceanography and Environmental Science, Xiamen University, Xiamen, P R China

Received November 22, 2008. Revised manuscript received March 24, 2009. Accepted March 26, 2009.

Concentration-dependent response relationships provide essential information on the characteristics of chemical-induced effects on toxicological end points, which include effect (inhibition or induction), potency, and efficacy of the chemical. Recent developments in systems biology and high throughput technologies have allowed simultaneous examination of many chemicals at multiple end point levels. While this increase in the quantity of information generated offers great potential, it also poses a significant challenge to environmental scientists to efficiently manage and interpret these large data sets. Here we present a novel method, ToxClust, that allows clustering of chemicals on the basis of concentration—response data derived with single or multiple end points. This method utilizes a least distance-searching algorithm (LDSA) to measure the pattern dissimilarity of concentration—response curves between chemicals and their relative toxic potency. ToxClust was tested using simulated data and chemical test data collected from the human H295R cell-based in vitro steroidogenesis assay. ToxClust effectively identified similar patterns of simulated data and responses to the exposure with the five model chemicals and separated them into different groups on the basis of their dissimilarities. These observations demonstrate that ToxClust not only provides an effective data analysis and visualization tool, but also has value in hypothesis generation and mechanism-based chemical classification.

Introduction

There is urgent need for methods that allow evaluation of large numbers of chemicals for their potential toxicity and to prioritize these chemicals for further testing (1). Recent developments in toxicogenomics and high-throughput screening techniques, including among others those broadly defined as genomics, transcriptomics, proteomics, and metabolomics utilizing multiple molecular and cellular end points has enabled testing of large numbers of chemicals (2). These new techniques are increasingly being used in priority chemical screening programs such as Tier 1 of the Endocrine Disruptor Screening Program of U.S. Environmental Protection Agency (EPA) (1). During these screening initiatives, researchers collect large amounts of multidimensional data (e.g., gene transcripts, proteins, and enzyme products) for various concentrations of each chemical analyzed (2). One of the major remaining challenges is the prioritization of chemicals for further testing based on the information obtained during Tier 1 (3). However, because of the lack of appropriate data evaluation techniques that effectively handle large quantities of diverse data, progress in the classification of chemicals by mode of action has been slow and is not yet part of many regulatory programs entrusted with this action (4). To accurately classify chemicals based on their critical mechanism of action so that they can be grouped for risk assessments (2) and to take advantage of detailed information on a few model chemicals, we must know the mechanism of action. However, few if any chemicals act through a single mechanism of action but rather cause multiple direct and indirect effects such that the predominant mechanism of toxicity can differ with duration and or magnitude (concentration) of exposure (5). In addition, the pattern of changes in the measurement end points can be different at different doses or durations of exposure. The pattern of response is often restricted to investigations of a single dose and time point, thus, losing valuable information on the complex changes in pattern as a function of duration and intensity of exposure.

Concentration—response relationships of molecular or cellular end points provide essential information for describing the mechanism of a chemical-induced effect. Traditional concentration/dose—response relationship theory is based on the response of a single end point such as metabolite concentration, enzyme activity, hormone level, etc. The concentration—response relationship provides information that is characteristic for the chemical induced toxicity such as no observable adverse effect levels (NOAELs) and effective concentrations (e.g., EC50). The potency of a chemical relative to that of a reference chemical (REP) can be calculated by comparing, for example, doses at which 50% of the maximum effect is observed (EC50) (6). Concentration—response relationships for different end points provide the basis for chemical risk assessment. However, it is often difficult to include all of the possible concentration—response relationships for multiple end points at multiple biological levels of organization in a single index. For these reasons, end points are generally independently considered when calculating indices of relative effects such as the EC50 and NOAEL. However, concentration—response relationships associated with multiple end points can provide useful information as well as possibly identifying a unique
toxicological signature of a chemical, which can then be used to understand its mechanism of action. For this reason, useful information is lost when end points are assessed separately. Conversely, some of the end points will be interrelated such that they represent redundant information relative to classifying the critical mechanism of action. Here, we define the critical mechanism of action as an adverse effect or an indicator of an adverse effect that occurs at the least concentration. While some chemicals can cause different effects at different concentrations, there is one primary or critical effect that occurs at the least concentration. For an example of this concept, see the review by Giesy and Kannan (7).

The primary goal of this research was to develop a novel computational approach to analyze concentration-dependent effect curves for single or multiple end points and to use this information to classify chemicals based on their predicted mechanisms of action. Chemicals with similar effects for certain end points can be depicted as having similar concentration-effect patterns. An inducer (X) is defined as a chemical that increases a biological response in a concentration-dependent manner, while an inhibitor (Y) is defined as a chemical that inhibits a biological response in a concentration-dependent manner (Figure S1A of the Supporting Information). If the dissimilarity of two chemicals is depicted using a Minkowski metric such as the Manhattan distance, the mechanism of the unknown chemical Z can then be determined by comparing the distances between its concentration response curve to that of the chemicals with known mechanisms, i.e., inducer X and inhibitor Y (Figure S1B of the Supporting Information). If chemical Z has a less than average Manhattan distance when compared to X rather than when compared to Y, it would indicate that the concentration–response function of Z is more similar to inducer X than to Y (Figure S1C,D of the Supporting Information). There are situations where chemicals have the same concentration–response curves but different potency (Figure S1E of the Supporting Information). For instance, chemical M has the exact same shape for a concentration–response curve as chemical N, but its EC$_{50}$ is less than the EC$_{50}$ for chemical N. However, because of the differences in potency, the average Manhattan distance between concentration–response curves of chemical M and chemical N would be nonzero (Figure S1F of the Supporting Information). To capture the dissimilarity in concentration–response curves between chemicals based on realistic concentration–response data, we introduced a new algorithm to account for relative potency.

**Materials and Methods**

**Distance Calculating Model.** In an experiment, concentration-dependent response data of m different biological end points can be measured for each chemical, and the dissimilarity (distance) between any chemical 1 and chemical 2 can be defined as a least Manhattan distance of their concentration–response curves, $D = \text{least} (d)$

$$d = \sum_{i=1}^{m} \int_{a}^{b} |f_{i,1}(x) - f_{i,2}(x+c)|dx$$

where $m$ is the total number of measured end points; $f_{i,1}$ is the concentration–response function of the first chemical; $f_{i,2}$ is the concentration–response function of the second chemical; $a$, $b$ is the concentration range; $c$ is the horizontally moved distance of the second concentration response function; and $c$ is the relative potency $C$ of the second chemical when the least distance is achieved.

The least Manhattan distance can then be calculated using a least distance-searching algorithm (LDSA) (Figure 1). For a single end point ($m = 1$), a spline function is fitted for the concentration–response data for each chemical; the splined function for the chemical with the greater tested-dosing range

![FIGURE 1. Least distance searching algorithm. For a single end point the steps are as follows: (1) Fit spline functions for the dose–response relationship for each chemical pair. (2) Fix the spline function (A) of the first chemical. (3) Move the function of the second chemical from the left to right and calculate the average distance between the two functions under the common area (B–D). (4) Identify the least distance and the position achieving this. (5) Calculate the relative potency of the two chemicals.](image-url)
is fixed on the abscissa. The function of the other chemical is then moved from the left to right with the overlapped concentration range equal or greater than half of the lesser dosing range of the two chemicals. The averaged vertical distance \( d \) between the two curves under the common \( x \) axis area is then calculated at each moving step. The bandwidth and step size were 1/20 of log transformed concentration. The overall moved distance \( c \) and the corresponding distance \( d \) were stored in two vectors. The least distance \( D \) between the two functions is then identified and the correspondingly horizontally moved distance is reported as the relative potency \( C \) of the second chemical when compared to the first chemical. Finally, a distance matrix and a relative potency matrix can be calculated for all of the chemicals that have been tested.

The distance matrix (also called dissimilarity matrix) describes pairwise dissimilarity between \( N \) chemicals. It is a square symmetrical \( N \times N \) matrix with the \((i, j)\) element equal to the least distance value \( (D) \) between the \( i \) and the \( j \) chemicals. The diagonal elements are equal to zero, i.e., the distance between an object and itself is postulated as zero. Similarly, the relative potency matrix is a square symmetrical \( N \times N \) matrix but with the \((i, j)\) element equal to the relative potency \((C)\) between the \( i \) and the \( j \) chemicals.

### Simulation of Chemicals with a Single End Point.

For purposes of illustration, 100 different concentration–response curves were simulated using kernel density estimates (Supporting Information). Briefly, the shape of the response curves was based on the cumulative distribution function of a different sample \( X1...X50 \) from the standard normal distribution. The relative potencies \((EC_{50})\) were from a uniform distribution \([-4, 4]\) and the efficacies were simulated using a uniform distribution \([-5, 5]\). The values of \( 4 \) and \( 5 \) were randomly picked to mimic an ordinary situation. The simulated response data were analyzed using the LDSA, principal component analysis (PCA), and hierarchical clustering.

**Chemicals Testing the Experiment Using H295R Cell-Based Steroidogenesis Assays.** To validate the ToxClust approach in the classification of chemicals by mode of action, we tested five chemicals in the H295R cell steroidogenesis assay (Figure S2 of the Supporting Information). The chemicals were chosen based on their known effects on steroid biosynthesis and steroidogenic gene expression and included prochloraz, fadrozole, aminogluthimide, forskolin, and trilostane (Supporting Information). The H295R human adrenocortical carcinoma cell line was obtained from the American type Culture Collection (ATCC CRL-2128; ATCC, Manassas, VA) and was grown at 37 °C with a 5% CO2 atmosphere as previously described (8–11). Cells were exposed to different concentrations of chemicals for 48 h in 24-well plates (COSTAR, Bucks, U.K.). DMSO was used as a carrier solvent and did not exceed 0.1% v/v. Test plates included six chemical concentrations, a solvent control (SC), and a blank control (CTR) in triplicate. Cytotoxic chemical concentrations were not included in the hormone concentration measurements as indicated by the MTT (3-[4,5-dimethylthiazol-2-yl]-2,5-diphenyl tetrazolium bromide) bioassay. Progesterone, androstenedione, estradiol, and testosterone were measured in ether extracts of culture medium by competitive ELISA using the manufacturer’s recommendations [progesterone [Cat # 582601], testosterone [Cat # 582701] and 17β-estradiol [Cat # 582251] were from Cayman Chemical Company, Ann Arbor, MI; androstenedione [Cat # 582-10-3800] was from Diagnostic Systems Laboratories, Inc., Webster, TX (11)]. For gene expression measurement, real-time quantitative RT-PCR was conducted by amplifying 2 \( \mu \)L of cDNA with SYBR Green Supersmix (Applied Biosystems) on an ABI Prism high throughput 7900HT system using the primers previously described (8). Chemical-induced concentration-dependent response curves of hormone production and mRNA levels are provided in Tables S1 and S2 of the Supporting Information.

### Statistical Procedures.

Hierarchical cluster analysis was performed for the chemicals (when \( n > 3 \)) and end points (when \( m > 3 \)) using their corresponding distance matrices. The distance (dissimilarity) matrix between chemicals was calculated using the LDSA described above. In the case of multiple end points \((m > 1)\), the distance between two chemicals was calculated by summing the curve distances of \( m \) end points at every moving step. When clustering of multiple end points was necessary, correlations of the responses between the variables (end points) were used as the distance. Dendrograms of chemicals and end points were vertically plotted against each other in the sample graph. The visualization of the concentration-dependent response curves was implemented using a “striped view” method and plotted in an \( N \times M \) matrix format on the right side of the graph. The \((i, j)\) element of the matrix corresponds to the response curve of the \( i \) chemical at the \( j \) end point levels. The color gradient from left to right displayed the concentration response curves from lower to higher concentration. The relative potency matrices within each chemical cluster provide relative potency estimations among these chemicals. The principal component analysis (PCA) on the potency matrices was conducted to select for display only those few curves that best reflect the important modes of variations using the protocol previously described (12). Three picked curves included the curve corresponding to the median principal component score together with those corresponding to two extreme quantiles, one at each end. To ensure selection of particular curves, we define the 100% % quantile to be \( p(N) + 1 \) where \( p(r) \) is the \( r \)th order statistic of the principal component scores and \( |x| \) is the integer part of \( x \). A program, ToxClust, was written in the \( R \) language (www.r-project.com) to perform the LDSA calculation and hierarchical clustering.

### Results

**Clustering Analysis of Simulated Data.** To illustrate the concept and the performance of ToxClust in a simple condition, we conducted a clustering analysis of 100 simulated concentration-dependent response curves. These 100 different sigmoid curves varied in their shapes, potencies \((EC_{50})\), and efficacies (Figure 2A). Using the LDSA approach, a \( 100 \times 100 \) relative potency matrix and \( 100 \times 100 \) distance matrix were computed using the simulated curves. The first principal component of the potency matrix accounted for 79% of the overall variance of the matrix and displayed a linear relationship with the potency \((EC_{50})\) of the curves (Figure 2B). The first principal component of the distance matrix accounted for 76% of the overall variance of the matrix and displayed a linear relationship with the efficacies of the curves (Figure 2C). The first several principal components are important in the sense of explaining most of the variation. Because these 100 curves were from fairly well-behaved original functions, only three typical curves corresponding to \( p(1) \), \( p(100) \), and \( p(51) \) were displayed. It can be seen from Figure 2D that the contrast between curves is most important; the curve on the top has a concentration dependent “induction”, and the curve at the bottom displays an “inhibition” effect, following the increase of concentration, while the curve in the middle has “no” obvious effects.

Besides PCA, hierarchical clustering is another exploratory analysis tool for the identification of data patterns. Hierarchical clustering analysis based on the calculated distance matrix further separated the 100 curves into two large clusters or five smaller branches (Figure S3A of the Supporting Information). The curves were labeled by the corresponding rank from least to greatest of their efficacies. The first cluster consisted of branch 1 and branch 2, which had curves ranked...
Clustering Analysis of Chemicals Using Four Hormones.

Using the LDSA approach, we first computed a distance and a relative potency matrix using the concentration-dependent hormone production curves for the five model chemicals, and then a divisive hierarchical clustering was computed on the basis of the calculated distance matrix. The response data, expressed as a fold change compared to a control, and the clusters are illustrated in a vertical dendogram (Figure 3). Forskolin was the first chemical separated; it showed concentration-dependent induction of the production of the four hormones. The greatest fold change by forskolin was observed for estradiol. The other four chemicals had very small distances; all had concentration-dependent inhibitory effects on the production of testosterone, estradiol, and androstenedione. Prochloraz was the only chemical beside forskolin that increased progesterone production in a concentration-dependent manner. Androstenedione was the end point observed to have the highest inhibition. The hormone end points were classified using the correlation-based distance metric. The two hormone clusters consisted of progesterone as a single cluster and the other three hormones as a cluster, including testosterone, estradiol, and androstenedione.

Clustering Analysis of Chemicals Using an Concentration-Dependent mRNA Expression. As was observed with hormone production, the use of CYP19A, 3βHSD2, CYP17A, and CYP11B2 gene expression resulted in the chemicals being separated into an inducer cluster and an inhibitor cluster, with the inhibitor cluster further dividing into two subclusters (Figure S4 of the Supporting Information). The inducer-related cluster contained only forskolin and was characterized by the up regulation of all four genes in a concentration-dependent manner. The inhibitor cluster consisted of two major groups with one group containing fadrozole and aminogluthetimide, while the second group contained prochloraz and trilostane. Fadrozole and aminogluthetimide were characterized by slight (less than 3-fold) changes in gene expression that were not monotonic or always concentration dependent. Prochloraz and trilostane were characterized by a general concentration-dependent decrease in the expression of CYP17A, CYP11B2, and 3βHSD2, while the alterations in the expression of CYP19A for these two chemicals were relatively slight; chemical-related changes in gene expression observed in the controls were typically less than 2-fold.

Discussion

Analysis of large sets of molecular and cellular response data is essential for understanding the underlying mechanism of chemical-induced effects in living organisms. The availability...
of high throughput techniques and the sheer amount of raw data obtained using these methods warrants special exploratory computational methods for deriving meaningful information and assisting mechanistic understanding. The concentration and time-dependent response functions provide mechanistic (cause–effect) information for chemical-induced effects. High throughput time-dependent data are routinely analyzed by general multivariate approaches such as PCA, factor analysis, or using Pearson correlation distance metrics in classification analysis (12–14). However, the need for classification and visualization in the analysis of the high throughput concentration-dependent response data type has not been addressed. The main difficulties of analyzing a large set of concentration-dependent response data were due to their distinct characteristics. First, the shapes of the concentration-dependent response curves in many cases differ from each other and are difficult to parametrize. This situation cannot be solved by simply comparing the parameters of the equations. Second, two chemicals with the same response curve shape but different potencies are seen as having the same “effect” or “mechanism” by toxicologists, which suggests that the response (observation) needs to be seen as independent from the specific concentration level (parameter). Finally, response data of chemicals might not have the same concentration range because they could be tested at different concentrations to see effects. For example, chemicals with lower potency are normally reported to be tested within higher concentration ranges. In another situation, some chemicals affect cell viability at higher concentration ranges, and they are normally tested in lower concentration ranges, which do not cause cytotoxicity. In these cases, direct application of more general approaches such as the principal component analysis are inappropriate (12).

In the present study, the LDSA approach used by ToxClust effectively isolated two separate information pools (shape and potency) from the chemical concentration-dependent response curves and stored them in a distance matrix and a potency matrix. The $N \times N$ distance matrix describes the data variation of the original chemical population in such a way that each chemical ($i$) can be seen as a variable and its distance with any other chemical ($j$), $d_{i,j}$ is a measurement of the $i$ variable on the $j$ chemical. When the sample size (number of chemicals) is large, the first of several principal components of the distance matrix shows the variation of the sample efficacy. The projections of the data points onto the principal component directions (principal component scores) are useful to examine the data variation. In the simulation data, a linear relationship was observed between the first PC of the distance matrix and the efficacy of the curves (Figure 2C). The three typical curves corresponding to $p(1)$, $p(100)$ and $p(51)$ effectively displayed the three different major mechanisms: “inhibition”, “induction”, and “no effect” (Figure 2C,D). The hierarchical clustering using the distance matrix also displayed consistent results when compared to the PCA method used on the groups of chemicals (Figure S3 of the Supporting Information). Similarly, the $N \times N$ potency matrix describes the relative potency among chemicals as suggested by the linear relationship between the first PC and the EC50. However, this “potency” information will only be valid when the chemicals examined have similar response curves.

The clustering tree and the compact visualization of the high dimensional data provided insights to the natural organization of data, without requiring any prior knowledge of the chemical mechanisms of action. Because the H295R steroidogenesis screening assay can test chemicals over a wide range of concentrations in a controlled setting, it is a good model for evaluating the utility of ToxClust in classifying chemicals by their effects on the steroidogenic pathway (8, 10). This aspect of the assay is particularly important because a chemical can exhibit multiple modes of action depending on exposure concentration and duration of

FIGURE 3. Clustering analysis of chemicals by ToxClust by use of concentrations of four hormones (progesterone, androstenedione, testosterone, and estradiol) as the classification variables. Calculated chemical and hormone classification trees are shown at the left and the top, respectively. The chemical-induced concentration–response function of hormone production was visualized by the color gradient at the corresponding cross position. Response data were expressed as fold change when compared to the corresponding control. Concentrations of chemicals increase from left to right. Analysis of concentration response relationships was carried out using ToxClust.
exposure, and any model that can evaluate multiple lines of evidence relative to classifying a chemical will reduce the probability of misclassifying a chemical as an endocrine disruptor. Integrating the concentration-dependent response data of the four hormones, ToxClust was able to classify the model chemicals into two clusters: “inducers” and “inhibitors”. Similar clustering results were observed to those derived by use of the concentration-dependent gene expression data. For example, forskolin was the only chemical that induced the production of all four measured hormones in a concentration-dependent manner. Androstenedione is the precursor of testosterone and estradiol and is directly converted from dehydroepiandrosterone by 3βHSD or from 17α-hydroxyprogesterone by CYP17A. The increase of the mRNA level of 3βHSD and CYP17A suggested that the enhanced hormone production caused by forskolin resulted from the transcriptional activation of the steroidogenic pathway. The response profile was also consistent with the previous finding that forskolin acts through stimulating adenyl cyclase and increasing cAMP concentrations in adrenal cells (9, 15, 16). Trilostane was the first inhibitor separated from the remaining three chemicals because it reduced the hormone production of androstenedione, testosterone, and estradiol but not progesterone (Figure 3). From Figure 3 developed by ToxClust, the inhibition on the production of either testosterone or estradiol by trilostane had less potency and efficacy than that on their precursor androstenedione, which is consistent with the previous report that trilostane is a 3βHSD inhibitor (17). In contrast, the inhibition of estradiol production by fadrozole was more potent than that of androstenedione or testosterone. Although fadrozole had been suggested to be an aromatase inhibitor, it might also inhibit the enzyme 3βHSD that can directly convert progesterone, androstenedione, and testosterone. The computational approach ToxClust was able to not only classify chemicals on the basis of multiple response factors, but also examine the relationships between end points used in the classification. For example, androstenedione, testosterone, and estradiol were separated into a cluster from progesterone, which is probably because of their proximity on the metabolic pathway.

ToxClust is a promising new exploratory analysis tool for the analysis of concentration-dependent response data in toxicological research. First, the concept of the LDSA is to identify the dissimilarity between any two chemicals or two end points by comparing their concentration-dependent curve patterns. The dissimilarity calculated by LDSA significantly reduced the data dimension and provided a uniform distance measure that can be adopted by available classification algorithms. ToxClust introduced in this paper utilized an unsupervised algorithm that does not require prior knowledge of the samples. With the use of a LDSA-based chemical distance calculation, supervised, knowledge-based machine learning methods can also be applied together to extract more meaningful information about the experiments. Second, ToxClust allows the visualization of high-dimensional concentration-dependent data in a single graph, which provides a new way to facilitate the study and understanding of large concentration-dependent toxicological data sets. However, the result of the chemical classification is meaningful only when the selection of the end points is relevant to a toxicity pathway. Toxicological pathways have been defined as a subset of adverse outcome pathways extending only to the cellular or extracellular level and easily modeled in vitro. A cellular or extracellular response pathway, when sufficiently perturbed, is expected to result in adverse health effects (18). When these cellular or extracellular end points are coordinate examined in the test of a large number of chemicals, chemicals classified in each cluster tend to have similar patterns of concentration response curves and therefore might hypothetically have the same mode of action.

While ToxClust proved to be an effective exploratory data analysis tool, there are outstanding issues that still need to be addressed. These include (1) developing a system to weight the different data (gene expression, protein expression, hormone production, etc.) according to their impact on the final classification of chemicals by mode of action, and (2) accounting for potency in the derivation distances to better classify chemicals with the same mode of action. Overall, ToxClust, as an exploratory data analysis tool, greatly facilitates analysis of complex data matrices, identifies patterns of selected biological end points, and allows mechanistic understanding of chemical-induced toxicity.

Acknowledgments
Data used in this paper to validate ToxClust had been collected from projects that were conducted with funding from the U.S. Environmental Protection Agency (U.S. EPA), Office of Research and Development (ORD) Service Center/National Health, and Environmental Effects Research Laboratory (NHEERL), Contract GS-10F-0041 L; and from the computational Toxicology Program of the U.S. EPA, ORD, and the Office of Science Coordination and Policy (OSCP); and the U.S. EPA ORD Service Center/NHEERL, Contract GS-10F-0041 L. The research was supported by a Discovery Grant from the National Science and Engineering Research Council of Canada (Project 6807) and from the Western Economic Diversification Canada (Projects 6971 and 6807). The authors acknowledge the support of an instrumentation grant from the Canada Foundation for Infrastructure. John P. Giesy was supported by the Canada Research Chair Program and an at large Chair Professorship at the Department of Biology and Chemistry and Research Centre for Coastal Pollution and Conservation, City University of Hong Kong.

Supporting Information Available
Simulation of concentration-dependent curves, information on the model chemicals, and chemical-induced fold changes on hormone productions and on gene expressions. This material is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited


Supporting Information

Title: Classification of Chemicals Based on Concentration-Dependent Toxicological Data Using “ToxClust©”

Authors: Xiaowei Zhang*1, John L. Newsted2, Markus Hecker1,3, Eric B. Higley1, Paul D. Jones1,5 and John P. Giesy1,4,5,6,7,8
1Toxicology Centre, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
2ENTRIX, Inc., Okemos, MI, USA
3ENTRIX, Inc., Saskatoon, SK, Canada
4Dept. Biomedical Veterinary Sciences, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
5Centre for Coastal Pollution and Conservation and Department of Biology and Chemistry, City University of HongKong, Kowloon, HongKong, SAR China
6Department of Zoology, Michigan State University, East Lansing, MI, USA
7State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing
8State Key Laboratory of Marine Environmental Science, College of Oceanography and Environmental Science, Xiamen University, Xiamen, P R China

Corresponding Author:

Xiaowei Zhang, PhD.
44 Campus Drive,
University of Saskatchewan
Saskatoon SK S7N 5B3 Canada
Tel: 306-966-1204
Fax: 306-966-4796
E-mail: howard50003250@yahoo.com

Submitted to: Environmental Science and Technology

Number of page: 12
Number of table: 2
Number of figure: 4
Simulation of concentration-dependent curves

100 different concentration response curves were simulated using kernel density estimate.

$$\hat{f}(x) = 0.02 \sum_{i=1}^{50} h^{-1} \varphi(h^{-1}(x - X_i))$$

Where $X_1, ..., X_{50}$ from normal distribution $\varnothing$ with mean $\mu \sim$ continuous uniform distribution $[-4, 4]$. The $h$ is the smoothing bandwidth.

The shape of the response curves was based on the cumulative distribution function of a different sample $X_1, ..., X_{50}$ from the standard normal distribution. The relative potencies (EC50) were from a uniform distribution $[-4, 4]$, and the efficacies were simulated using uniform distribution $[-5, 5]$. The value of 4 and 5 were random picked to mimic ordinary situation.
Model chemicals
The model chemicals were chosen based on their known effects on steroid biosynthesis and steroidogenic gene expression and included prochloraz, fadrozole, aminoglutethimide forskolin, and trilostane. Prochloraz is an agricultural imidazol fungicide that inhibits a cytochrome P450 (CYP) enzyme involved in ergosterol synthesis, but has also been reported to inhibit other CYP enzymes, and to act as a potent aromatase inhibitor (1-3). Fadrozole is a non-steroidal reversible and competitive inhibitor of aromatase (CYP19) as well as of enzymatic activities associated with CYP11B (4,5). Aminoglutethimide is an aromatase inhibitor that was also reported to interact with several other steroidogenic protein/enzymes such as steroidogenic acute regulatory protein (StAR), CYP11A (P450scc, cholesterol side-chain cleavage enzyme) and 17 beta-hydroxy-steroid dehydrogenase (17β-HSD) (6,7). Forskolin is a general inducer of steroidogenesis that acts via the activation of cAMP pathways (6). Trilostane a specific inhibitor of 3βHSD and has been shown to alter the production of progesterone in several animal models (8-10). Forskolin, aminoglutethimide and prochloraz were obtained from Aldrich-Sigma Chemical Co. (St. Louis, MO, USA). Fadrazole was obtained from Novartis Pharma AG (Basel, CH). Trilostane was contributed by Dr. Michael A. Parrish from Parrish Business Development (Northumberland, UK)

Reference


Table Legends

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table S1</td>
<td>Chemical-induced fold changes on hormone productions in H295R cells</td>
</tr>
<tr>
<td>Table S2</td>
<td>Chemical-induced fold changes on gene expression in H295R cells</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chemical</th>
<th>Conc. (μM)</th>
<th>Progesterone</th>
<th>Testosterone</th>
<th>Estradiol</th>
<th>Androstenedione</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aminogluthetimide</td>
<td>0.1</td>
<td>1.248</td>
<td>0.979</td>
<td>1.242</td>
<td>1.000</td>
</tr>
<tr>
<td>Aminogluthetimide</td>
<td>0.3</td>
<td>0.985</td>
<td>0.812</td>
<td>0.879</td>
<td>0.911</td>
</tr>
<tr>
<td>Aminogluthetimide</td>
<td>1</td>
<td>1.063</td>
<td>0.912</td>
<td>0.916</td>
<td>0.927</td>
</tr>
<tr>
<td>Aminogluthetimide</td>
<td>3</td>
<td>1.189</td>
<td>0.805</td>
<td>0.776</td>
<td>1.025</td>
</tr>
<tr>
<td>Aminogluthetimide</td>
<td>10</td>
<td>0.792</td>
<td>0.757</td>
<td>0.879</td>
<td>0.884</td>
</tr>
<tr>
<td>Aminogluthetimide</td>
<td>30</td>
<td>0.355</td>
<td>0.366</td>
<td>0.431</td>
<td>0.452</td>
</tr>
<tr>
<td>Aminogluthetimide</td>
<td>100</td>
<td>0.217</td>
<td>0.188</td>
<td>0.289</td>
<td>0.136</td>
</tr>
<tr>
<td>Fadrozole</td>
<td>0.01</td>
<td>0.500</td>
<td>1.331</td>
<td>0.844</td>
<td>1.000</td>
</tr>
<tr>
<td>Fadrozole</td>
<td>0.1</td>
<td>0.524</td>
<td>1.277</td>
<td>0.609</td>
<td>0.979</td>
</tr>
<tr>
<td>Fadrozole</td>
<td>1</td>
<td>0.509</td>
<td>0.861</td>
<td>0.349</td>
<td>0.752</td>
</tr>
<tr>
<td>Fadrozole</td>
<td>3</td>
<td>0.529</td>
<td>0.580</td>
<td>0.332</td>
<td>0.596</td>
</tr>
<tr>
<td>Fadrozole</td>
<td>10</td>
<td>1.229</td>
<td>0.513</td>
<td>0.216</td>
<td>0.186</td>
</tr>
<tr>
<td>Fadrozole</td>
<td>100</td>
<td>0.333</td>
<td>0.086</td>
<td>0.244</td>
<td>0.062</td>
</tr>
<tr>
<td>Forskolin</td>
<td>0.01</td>
<td>0.855</td>
<td>1.520</td>
<td>1.214</td>
<td>1.000</td>
</tr>
<tr>
<td>Forskolin</td>
<td>0.03</td>
<td>1.249</td>
<td>1.874</td>
<td>1.457</td>
<td>0.952</td>
</tr>
<tr>
<td>Forskolin</td>
<td>0.1</td>
<td>1.246</td>
<td>1.448</td>
<td>2.176</td>
<td>1.400</td>
</tr>
<tr>
<td>Forskolin</td>
<td>0.3</td>
<td>1.600</td>
<td>1.709</td>
<td>4.133</td>
<td>1.571</td>
</tr>
<tr>
<td>Forskolin</td>
<td>1</td>
<td>2.105</td>
<td>2.191</td>
<td>6.993</td>
<td>2.324</td>
</tr>
<tr>
<td>Forskolin</td>
<td>3</td>
<td>2.296</td>
<td>2.318</td>
<td>9.865</td>
<td>3.866</td>
</tr>
<tr>
<td>Forskolin</td>
<td>10</td>
<td>2.320</td>
<td>2.775</td>
<td>11.873</td>
<td>4.578</td>
</tr>
<tr>
<td>Prochloraz</td>
<td>0.001</td>
<td>0.912</td>
<td>0.847</td>
<td>1.013</td>
<td>1.000</td>
</tr>
<tr>
<td>Prochloraz</td>
<td>0.01</td>
<td>1.331</td>
<td>0.765</td>
<td>0.940</td>
<td>0.843</td>
</tr>
<tr>
<td>Prochloraz</td>
<td>0.03</td>
<td>1.714</td>
<td>0.677</td>
<td>0.679</td>
<td>0.482</td>
</tr>
<tr>
<td>Prochloraz</td>
<td>0.1</td>
<td>2.843</td>
<td>0.593</td>
<td>0.374</td>
<td>0.243</td>
</tr>
<tr>
<td>Prochloraz</td>
<td>0.3</td>
<td>4.509</td>
<td>0.418</td>
<td>0.263</td>
<td>0.112</td>
</tr>
<tr>
<td>Prochloraz</td>
<td>1</td>
<td>4.878</td>
<td>0.292</td>
<td>0.165</td>
<td>0.038</td>
</tr>
<tr>
<td>Prochloraz</td>
<td>3</td>
<td>4.078</td>
<td>0.307</td>
<td>0.182</td>
<td>0.023</td>
</tr>
<tr>
<td>Trilostane</td>
<td>0.001</td>
<td>1.000</td>
<td>0.906</td>
<td>1.297</td>
<td>0.996</td>
</tr>
<tr>
<td>Trilostane</td>
<td>0.01</td>
<td>0.887</td>
<td>0.648</td>
<td>1.215</td>
<td>0.424</td>
</tr>
<tr>
<td>Trilostane</td>
<td>0.1</td>
<td>0.957</td>
<td>0.260</td>
<td>0.866</td>
<td>0.063</td>
</tr>
<tr>
<td>Trilostane</td>
<td>1</td>
<td>0.991</td>
<td>0.646</td>
<td>0.299</td>
<td>0.085</td>
</tr>
</tbody>
</table>
## Table S2: Chemical-induced fold changes on gene expression in H295R cells

<table>
<thead>
<tr>
<th>Chemical</th>
<th>Conc.(μM)</th>
<th>CYP17A</th>
<th>CYP19A</th>
<th>HSD3b2</th>
<th>CYP11B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aminogluthetimide</td>
<td>0</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Aminogluthetimide</td>
<td>1</td>
<td>1.076</td>
<td>0.738</td>
<td>0.727</td>
<td>0.883</td>
</tr>
<tr>
<td>Aminogluthetimide</td>
<td>3</td>
<td>1.445</td>
<td>0.989</td>
<td>0.943</td>
<td>1.115</td>
</tr>
<tr>
<td>Aminogluthetimide</td>
<td>10</td>
<td>0.968</td>
<td>0.767</td>
<td>0.668</td>
<td>1.701</td>
</tr>
<tr>
<td>Aminogluthetimide</td>
<td>30</td>
<td>1.310</td>
<td>0.900</td>
<td>0.779</td>
<td>0.944</td>
</tr>
<tr>
<td>Aminogluthetimide</td>
<td>100</td>
<td>1.012</td>
<td>0.554</td>
<td>0.408</td>
<td>0.679</td>
</tr>
<tr>
<td>Forskolin</td>
<td>0</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Forskolin</td>
<td>1</td>
<td>3.350</td>
<td>38.058</td>
<td>123.515</td>
<td>19.689</td>
</tr>
<tr>
<td>Forskolin</td>
<td>3</td>
<td>3.646</td>
<td>45.129</td>
<td>144.488</td>
<td>46.877</td>
</tr>
<tr>
<td>Forskolin</td>
<td>10</td>
<td>4.352</td>
<td>52.039</td>
<td>167.827</td>
<td>77.589</td>
</tr>
<tr>
<td>Forskolin</td>
<td>30</td>
<td>4.210</td>
<td>27.000</td>
<td>121.000</td>
<td>32.100</td>
</tr>
<tr>
<td>Prochloraz</td>
<td>0</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Prochloraz</td>
<td>0.3</td>
<td>1.330</td>
<td>1.260</td>
<td>1.380</td>
<td>1.350</td>
</tr>
<tr>
<td>Prochloraz</td>
<td>1</td>
<td>0.819</td>
<td>0.759</td>
<td>0.285</td>
<td>1.714</td>
</tr>
<tr>
<td>Prochloraz</td>
<td>3</td>
<td>0.938</td>
<td>0.872</td>
<td>0.694</td>
<td>11.712</td>
</tr>
<tr>
<td>Prochloraz</td>
<td>10</td>
<td>0.185</td>
<td>0.766</td>
<td>0.143</td>
<td>3.038</td>
</tr>
<tr>
<td>Prochloraz</td>
<td>30</td>
<td>0.057</td>
<td>1.783</td>
<td>0.130</td>
<td>0.103</td>
</tr>
<tr>
<td>Fadrozole</td>
<td>0</td>
<td>1.029</td>
<td>1.001</td>
<td>1.001</td>
<td>1.000</td>
</tr>
<tr>
<td>Fadrozole</td>
<td>1</td>
<td>1.214</td>
<td>0.969</td>
<td>1.054</td>
<td>1.076</td>
</tr>
<tr>
<td>Fadrozole</td>
<td>3</td>
<td>1.164</td>
<td>1.053</td>
<td>1.153</td>
<td>1.132</td>
</tr>
<tr>
<td>Fadrozole</td>
<td>10</td>
<td>1.398</td>
<td>0.966</td>
<td>1.009</td>
<td>1.201</td>
</tr>
<tr>
<td>Fadrozole</td>
<td>30</td>
<td>1.893</td>
<td>0.875</td>
<td>0.702</td>
<td>1.293</td>
</tr>
<tr>
<td>Fadrozole</td>
<td>100</td>
<td>1.897</td>
<td>0.921</td>
<td>0.619</td>
<td>1.258</td>
</tr>
<tr>
<td>Trilostane</td>
<td>0</td>
<td>1.000</td>
<td>0.974</td>
<td>1.001</td>
<td>1.000</td>
</tr>
<tr>
<td>Trilostane</td>
<td>1</td>
<td>0.737</td>
<td>0.919</td>
<td>0.990</td>
<td>0.778</td>
</tr>
<tr>
<td>Trilostane</td>
<td>3</td>
<td>0.551</td>
<td>1.015</td>
<td>0.794</td>
<td>0.639</td>
</tr>
<tr>
<td>Trilostane</td>
<td>10</td>
<td>0.286</td>
<td>0.921</td>
<td>0.428</td>
<td>0.378</td>
</tr>
<tr>
<td>Trilostane</td>
<td>30</td>
<td>0.130</td>
<td>0.710</td>
<td>0.113</td>
<td>0.162</td>
</tr>
<tr>
<td>Trilostane</td>
<td>100</td>
<td>0.143</td>
<td>1.134</td>
<td>0.064</td>
<td>0.048</td>
</tr>
</tbody>
</table>
**Figure Legends**

**Figure S1**  Illustration of Manhattan distance. A) dose-response function of inducer X and inhibitor Y; B) dose-response function of unknown chemical Z; C) average Manhattan distance between dose response curves of chemical C and inducer A: |Z-A| = 0.270; D) average Manhattan distance between dose response curves of chemical Z and inhibitor Y: |Z-Y| = 1.294. |Z-X| < |Z-Y| suggesting dose-response function of Z is more similar to that of chemical X than Y;  E-F) Chemical M and N have the same shape dose-response curve with M being more potent than N. EC50 of chemical M and N can be obtained based on their dose-response functions, which are 2 and 4 respectively. Average Manhattan distance between dose-response functions of chemical M and inhibitor N: |M-N| = 0.673.

**Figure S2**  Steroidogenic pathway in H295R cells. Enzymes are in italics, hormones are bolded and arrows indicate the direction of synthesis.

**Figure S3**  Hierarchical clustering analysis of the 100 simulated concentration dependent curves. The curves were labeled by the corresponding rank of their efficacies. The original tree structure was demonstrated by the upper and lower parts of the dendrogram separately. A) upper dendrogram; B-F) branch 1-5 of the lower dendrograms.

**Figure S4**  Clustering analysis of chemicals by ToxClust©, by use of the four mRNA level (CYP19A, 3βHSD2, CYP17A and CYP11B2) as classification variables. Chemicals were shown in row and endpoints were in column. Chemical induced concentration response curve was visualized by the color gradient at the corresponding cross position. Response data were expressed as fold change comparing to the corresponding control (not shown here). The increase of chemical concentration was from left to right.
Illustration of Manhattan Distance

A

X: inducer
Y: inhibitor

B

Z: unknown

C

Dist: |Z−X| = 0.27

D

Dist: |Z−Y| = 1.294

E

Dist: |M−N| = 0.673
Figure S2

- Cholesterol $\xrightarrow{CYP11A}$ Pregnenolone
  - $3\beta$HSD
  - Progesterone $\xrightarrow{CYP21}$ 11-Deoxycorticosterone
    - $CYP11B2$
  - Corticosterone $\xrightarrow{CYP11B2}$ Aldosterone
  - $CYP17$
  - $3\beta$HSD

- Pregnenolone $\xrightarrow{CYP17}$ 17α-OH Pregnenolone
  - $3\beta$HSD

- 17α-OH Progesterone $\xrightarrow{CYP17}$ Androstenedione
  - $3\beta$HSD

- Androstenedione $\xrightarrow{CYP17}$ Deoxycortisol
  - $17\beta$HSD

- Deoxycorticosterone $\xrightarrow{CYP19}$ Testosterone
  - Estrone
  - $17\beta$-estradiol

- 17α-OH Progesterone $\xrightarrow{CYP17}$ 11-Deoxycortisol
  - $CYP21$
  - Cortisol $\xrightarrow{CYP11B1}$
Figure S3

A: upper dendrogram

B: Branch 1

C: Branch 2

D: Branch 3

E: Branch 4

F: Branch 5
Figure S4

Fold Change

0.03125
0.125
0.5
1
8
32
128

CYP17A
CYP11B2
CYP19A
HSD3b2

Fadrozole (1–100uM)
Aminogluthetimide (1–100uM)
Prochloraz (0.3–30uM)
Trilostane (1–100uM)
Forskolin (1–30uM)