Combining Chemodescriptors and Biodescriptors in Quantitative Structure−Activity Relationship Modeling

Douglas M. Hawkins,*,† Subhash C. Basak,‡ Jessica Kraker,† Kevin T. Geiss,§ and Frank A. Witzmann||

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, Natural Resources Research Institute, University of Minnesota−Duluth, 5103 Miller Trunk Highway, Duluth, Minnesota 55811, Human Effectiveness Directorate, Air Force Research Laboratory, 2610 Seventh Street, Wright-Patterson AFB, Ohio 45433, and Department of Cellular & Integrative Physiology and Biochemistry & Molecular Biology, Indiana University School of Medicine, 635 Barnhill Drive, MS405, Indianapolis, Indiana 46202

Received June 20, 2005

In view of the wide distribution of halocarbons in our world, their toxicity is a public health concern. Previous work has shown that various measures of toxicity can be predicted with standard molecular descriptors. In our work, biodescriptors of the effect of halocarbons on the liver were obtained by exposing hepatocytes to 14 halocarbons and a control and by producing two-dimensional electrophoresis gels to assess the expressed proteome. The resulting spot abundances provide additional biological information that might be used in toxicity prediction. QSAR models were fitted via ridge regression to predict eight dependent toxicity measures: d37, arr, EC50MTT, EC50LDH, EC20SH, LECLP, LECROS, and LECCAT. Three predictor sets were used for each—the chemodescriptors alone, the biodescriptors alone, and the combined set of both chemo- and biodescriptors. The results differed somewhat from one dependent to another, but overall it was shown that better results could be obtained by using both chemo- and biodescriptors in the model than by using either chemo- or biodescriptors alone. The library of compounds used was small and quite homogeneous, so our immediate conclusions are correspondingly limited in scope, but we believe the underlying methodologies have broad applicability at the interface of chemical and biological descriptors.

INTRODUCTION

A valuable use of computational chemistry is the prediction of toxicity of chemicals from their structure. The major motivation behind this trend of research is that a large number of chemicals need to be evaluated both in hazard assessment of environmental pollutants and in drug discovery.[1,2] Whereas different groups (Hansch and Leo[3] and Kamlet et al.[4]) have attempted to predict biochemical and toxicological activity of chemicals from their physicochemical properties and physicochemically based substituent constants, for the majority of existing chemicals such properties are unavailable. This is true for the chemicals in commerce currently listed in the Toxic Substances Control Act Inventory (TSCA) and for the various real and virtual combinatorial libraries developed for pharmaceutical drug design. Therefore, structure-based evaluation of chemical toxicity has been a principal method followed by the regulatory agencies for a long time.

Halocarbons serve a wide variety of purposes, including uses as solvents and starting materials in synthetic processes, and are therefore present throughout our environment. Since the earliest observations of toxicity associated with carbon tetrachloride and chloroform, there has been an interest in understanding the molecular basis of halocarbon toxicity.[5−8] This work has led to a large body of QSAR modeling in which experimental measurements of toxicity at the cell level of unicellular organisms and primary hepatocytes of rat are predicted using molecular descriptors. The hierarchical QSAR developed for halocarbons is based on the mechanistic hypothesis that these molecules attach with an electron, go to the anionic state, immediately release an anion, and leave behind a free radical. Under both aerobic and anaerobic conditions the free radical is thought to be responsible for proton extraction, lipid peroxidation, and toxicity.

Our earlier work on the QSAR of a set of 55 halocarbons in *Aspergillus nidulans* and halocarbons in hepatocyte showed that LUMO energy calculated by semiempirical and ab initio methods plays an important role in predicting toxicity of halocarbons.

Molecular descriptors have a number of advantages for use in QSAR work, such as their being obtainable reproducibly and relatively cheaply. Toxicity results from the interaction between the toxicant and its biological target. Chemical structure will be the sole determinant of toxicity only when the biological system reacts to all chemicals of a class in a uniform manner, and when the biological system takes a passive role in chemical−biological interaction.

In the postgenomic era, the role of biology in understanding toxicity of chemicals is being studied using technologies of genomics and proteomics. One such class of information is the pattern of proteins expressed by cells after exposure to different compounds. It is reasonable to assume that

* Corresponding author phone: (612)624-4166; e-mail: doug@stat.umn.edu.
† University of Minnesota.
‡ University of Minnesota−Duluth.
§ Wright-Patterson AFB.
|| Indiana University School of Medicine.

10.1021/ci050252p CCC: $33.50
Published on Web 12/03/2005

Table 1. Compounds, Their Gel Count, and Toxicity

no.	formula	gels	d37	ARR	EC50MTT	EC50LDH	EC20SH	LECLP	LECROS	LECCAT
1	control	27								
2	C_2Cl_4	5	1.2	1.5	0.19	0.43	0.5	1	1	1
3	CH_2Br_2	2	21.4	25.7	6.4	14.7	5	25	5	25
4	CH_2BrCl	2	61.6	69.3	127	34.5	100	10	100	50
5	CH_2Cl_2	2	93.6	135.7	32	49.9	24.99	25	25	50
6	$CHCl_3$	2	24.8	23.5	1.2	5.36	2	2	2	5
7	$CHBr_3$	6	5.2	5.5	0.6	2.72	0.29	0.29	0.29	1.15
8	1,2-$C_2H_4Br_2$	5			0.31	1.35	0.07	0.36	0.07	8.63
9	1,1,2-$C_2H_3Br_3$	5			0.23	0.44	0.12	0.25	0.06	0.12
10	1,2-C_2H_4BrCl	5	24	20.4	0.49	1.41	0.42	0.42	3.36	1.68
11	1,2-$C_2H_4Cl_2$	6	50.8	44.4	4.37	50	3.16	9.5	9.5	38
12	C_2HCl_3	6	11.1	10.6	1.05	1.5	0.83	1.66	0.83	1.66
13	1,1,1-$C_2H_3Cl_3$	7	10	11	0.38	1.11	0.31	0.31	0.61	0.61
14	1,1,2-$C_2H_3Cl_3$	6	10.7	11.8	2.54	5	1.08	1.08	2.16	4.32
15	1,1,2,2-$C_2H_2Cl_4$	6	2.8	3.3	1.73	3.31	1.19	2.38	1.19	4.76

toxicity in the body has a signature of proteins that are over- or underexpressed and that this signature could be assessed in vitro using isolated cells.

In our work, for each run, a hepatocyte culture was exposed to a chemical, and a two-dimensional electrophoresis gel was created. A total of 92 gels were made, covering 14 halocarbons as well as a control. Table 1 lists the compounds tested, the numbers of 2D gels generated for each, and the eight toxicity values. Note that two toxicity measures, d37 and ARR, were not available for two of the halocarbons. The control should not be toxic by any measure. For purposes of toxicity prediction, the control was assigned a toxicity of the lowest figure reported for any of the 14 halocarbons.

Trohalaki et al.[9] studied the toxic effects of a set of 20 halocarbons in the rat hepatocytes as measured by different cellular toxicity indicators. The exposed hepatocytes were analyzed using 2-D gel electrophoresis technology in the lab of Frank Witzmann. The perturbation in the 2-D gel pattern can be looked upon as the toxic effects. However, the principal problem is that there is data on over 1400 spots, so a visual inspection of these proteomics maps does not easily lead to any discernible patterns. Therefore, our research team has resorted to the development of compact biodescriptors and individual protein biomarkers using methods of discrete mathematics and statistics, respectively, as outlined below:

(1) invariants of graphs/matrices associated with proteomics maps[10]

(2) information theoretic biodescriptors[11]

(3) biodescriptors from spectrum-like representations of proteomics maps[12]

(4) statistical approaches to discover *critical protein biomarkers.*

Whereas the compact biodescriptors in items 1−3 above give an overall characterization of proteomics maps[1] exposed to such toxicants as well-known peroxisome proliferators, the statistical method is designed to find critical individual proteins and characteristic 'signatures' involving multiple proteins which are related to toxicity and toxic mode of action.

In this paper we have carried out a comparative study of chemodescriptors (topological, geometrical, and quantum chemical parameters) vis-à-vis biodescriptors (individual proteins associated with halocarbon toxicity in hepatocytes) in predicting the overall cellular toxicity of halocarbons. With the motivation of describing the methodology in detail, the small number of compounds (20 with chemodescriptors and 14 of those with biodescriptors) avoids focusing on the calculations and can instead emphasize the methods incorporated into the analysis. Though the overall body of halocarbon compounds is quite small (so the domain in which toxicities can be predicted is quite limited), there do exist some additional compounds for which the predictive ability of the model could be tested, when the chemo- and biodescriptors are available.

MATERIALS AND METHODS

Presented below are the methods and materials used in the isolation of rat F-344 hepatocytes, exposure of the cells to the toxicants, and the determination of cell-level toxicity for the halocarbons after exposure. Scientists at Wright Patterson Air Force Base generated the data used in this study.

Chemicals. Chemicals and biological reagents were purchased from Sigma Chemical Co. (St. Louis, MO) unless otherwise stated; this includes most of the halogenated test chemicals ($CHBrCl_2$, $CHBr_2Cl$, $CHBr_3$, CH_2Br_2, CH_2BrCl, CH_2Cl_2, CCl_4, $CBrCl_3$, CBr_2Cl_2, CBr_4, C_2HCl_3, 1,1,2,2-$C_2H_2Cl_4$, 1,2-$C_2H_4Cl_2$, 1,2-C_2H_4BrCl, 1,1,2-$C_2H_3Br_3$, 1,2-$C_2H_4Br_2$, 1,1,2-$C_2H_3Cl_3$, 1,1,1-$C_2H_3Cl_3$). C_2Cl_4 and $CHCl_3$ were purchased from Fisher Scientific (Pittsburgh, PA). Chemicals were used as provided by the manufacturer, without further purification. The halogenated chemicals are summarized in Table 2, shown with their CAS number, catalog number, purity/grade, and their experimental dose ranges. Ketamine (Injectable, U.S.P. grade) was purchased from Parke-Davis (Moris Plains, NJ). Xylazine (Injectable, U.S.P. grade) was purchased from Mobay Corporation (Shawnee, KS). Collagenase (Type D) and Protein Assay ESL reagents were purchased from Roche (formerly Boehringer-Mannheim) Biochemicals (Indianapolis, IN). Type I rat tail collagen was purchased from Upstate Biotechnology (Lake Placid, NY). CHEE's modified Eagle Medium (Formula No. 88-5046EA) and Hank's balanced salt solution (HBSS) were purchased from GibcoBRL/Life Technologies (Rockville, MD).

Animals. Fischer $CD^{+2}F^R$(F344)/CrlBR (F-344) male rats (225−300 g) were obtained from Charles River Laboratories. All animals used in this study were handled in accordance

Table 2. Twenty Experimental Halocarbons for QSAR[a]

formula	CAS no.	cat no. (S/F)	purity/grade	doses (mM)
CCl_4	56−23−5	C5331(S)	99.9%	0.043−1.380
CBr_4	558−13−4	C1,108−1(S)	99%	0.075−0.200
$CHBrCl_2$	75−27−4	13,918−1(S)	98+%	0.500−10.000
$CHBr_2Cl$	124−48−1	20,632−6(S)	98%	0.290−2.350
CBr_2Cl_2	594−18−3	32,995−9(S)	95%	0.100−1.000
$CBrCl_3$	75−62−7	B8,225−1(S)	99%	0.310−2.500
CH_2Br_2	74−95−3	D4,168−6(S)	99%	2.500−25.000
$CHCl_3$	67−66−3	BP1145−1(F)	99%	1.000−10.000
C_2Cl_4	127−18−4	O4586−4(F)	certified reagent	0.100−2.000
C_2HCl_3	79−01−6	25,642−0(S)	99.5%	0.277−2.660
1,2-$C_2H_4Cl_2$	107−06−2	27,057−1(S)	99.8%	3.160−38.000
1,1,1-$C_2H_3Cl_3$	71−55−6	23,557−1(S)	99%	1.080−6.480
1,1,2-$C_2H_3Cl_3$	79−00−5	46,621−2(S)	97%	0.063−0.501
1,1,2-$C_2H_3Br_3$	78−74−0	47,649−8(S)	99%	0.310−2.460
1,1,2,2-$C_2H_2Cl_4$	79−34−5	18,543−4(S)	98+%	1.188−4.750
CH_2Cl_2	75−09−2	27,056−3(S)	99.9%	10.000−100.000
$CHBr_3$	75−25−2	24,103−2(S)	99+%	0.143−4.580
CH_2BrCl	74−97−5	13,526−7(S)	99%	10.000−100.00
1,2-C_2H_4BrCl	107−04−0	23,275−0(S)	98%	0.420−3.364
1,2-$C_2H_4Br_2$	106−93−4	D887(S)	99%	0.070−8.630

[a] F = Fisher Scientific; S = Sigma/Adrich Company.

with the principles stated in the "*Guide for the Care and Use of Laboratory Animals*", National Research Council, 1996, and the Animal Welfare Act of 1966, as amended. Rats were anesthetized with 1 mL/kg of a mixture of ketamine (70 mg/mL) and xylazine (6 mg/mL) prior to undergoing in situ liver perfusion.

Hepatocyte Isolation and Culture. Rat livers were digested by perfusion using the two-step method of Seglen.[13] In the first perfusion step, the liver was perfused via the hepatic portal vein with perfusion buffer (37 °C) consisting of Hank's balanced salt solution (HBSS; pH 7.2) lacking calcium and magnesium and supplemented with 15 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), heparin (2.0 U/mL), and ethylene-bis(oxyethylenenitrilo)-tetraacetic acid (EGTA; 0.5 mM). This first step of the antigrade two-step perfusion method was accomplished by cannulation of the portal vein followed by clamping of the posterior vena cava anterior to the diaphragm. Flow of perfusion buffer was then activated (20 mL/min), immediately followed by the cutting of the posterior vena cava (anterior to the renal vein) to allow drainage of the perfusion buffer.

Following complete removal of blood from the liver, the second step of the liver perfusion was initiated by continuous perfusion with digestion buffer. Digestion buffer (37 °C) consisting of complete HBSS (pH 7.2) supplemented with collagenase (0.26 Wunsche Units/mL) was perfused through the liver to digest interstitial connective tissue. Perfusion was continued until hepatocytes were completely disaggregated (approximately 20 min). Viable primary rat hepatocytes were washed (37 °C) with complete HBSS (pH 7.2) and enriched three times by low speed centrifugation at 50 g for 3 min. Typical viabilities of isolated hepatocytes ranged from 80 to 95% with yields of 250−400 million cells as determined by trypan blue dye exclusion. For cell culture studies, freshly isolated hepatocytes in suspension were adjusted to a cell density of 1.0×10^6 cell/mL in cell attachment medium consisting of CHEEs modified culture medium (CHEE; pH 7.2) supplemented with HEPES (10 mM), insulin/transferrin/sodium selenite solution (final concentrations of 5 μg/mL, 5 μg/mL, and 5 ng/mL, respectively), gentamycin (0.1 mg/mL), and dexamethasone (0.4 mg/mL).

Cells were seeded in 6-well (1.0×10^6 cells/well) culture plates. Plates were previously coated with Type I rat tail collagen (25.0 μg/mL stock) at 2.6 μg/cm^2. After 4 h of incubation in a 95% air/5% CO_2 incubator at 37 °C, the cell attachment medium was removed, and rat hepatocytes were incubated in fresh CHEEs culture medium lacking dexamethasone. Hepatocytes were incubated for an additional 20 h in order to recover from the stress incurred during the isolation procedure. Exposures to halogenated chemicals were initiated at this point.

Cell Culture and Chemical Exposure. All exposures to the test chemicals were accomplished in the VITROBOX exposure system.[14,15] The system was basically a rectangular glass chamber, with interior dimensions of approximately 8 × 25 × 25 cm (∼5 L volume). For each chemical dose, there was a separate VITROBOX chamber and Tedlar dosing bag. Chemical dose ranges are also shown in Table 2.

Prior to exposures (∼30 min) the appropriate chemical concentration was added to the Tedlar dosing bags and placed back into a 37 °C incubator to allow the chemical to volatilize. Immediately prior to cell culture dosing, the chemicals were diluted into the appropriate medium. No other solvents or vehicles (e.g. ethanol) were used in preparation of the dosing solutions. Media in the cell culture plates was replaced with the chemical dosing media, and plates were immediately placed into the chamber without plate lids. The chamber was then closed, and the dosing and capture bags were attached to the chamber. The dosing of the chamber was accomplished by manually compressing the dosing bag, forcing the dosing atmosphere into the chamber, while it exhausted into the capture bag. Following the dosing process, all ports were sealed, and the chamber was placed into a 37 °C incubator. The cells were exposed in the chambers for 4 h and then removed from the chambers for analysis.

Toxicity Measures. Two sets of toxicity measures were used as dependents. One set consisted of the d37 and ARR data reported by Crebelli et al.[5] The values used were part of a larger body of results obtained by measuring the amount

of each halocarbon needed to block mitotic growth of *Aspergillus nidulans*. The second set included six assays which were used to determine cytotoxicity and oxidative stress on hepatocytes. MTT and LDH are indicators of cellular viability, i.e., mitochondrial function and membrane integrity, respectively. Reactive oxygen species (ROS) were measured as indicators of reactive oxidant molecules that are produced in the cell and result in oxidant damage to the cell. Lipid peroxidation (LP) was measured as an indicator of oxidative damage to cellular lipid molecules. Catalase (CAT) activity was determined as a representative antioxidant enzyme. Decreases in catalase activity can contribute to an increased oxidative stress situation in the cell. Total cellular thiols (SH) were determined to assess the antioxidant thiol status.

TWO-DIMENSIONAL ELECTROPHORESIS METHODS

Cell proteins were solubilized directly in the 6-well plates after removal of medium. 400 μL of lysis buffer containing 9 M urea, 4% Igepal CA-630 ([octylphenoxy] polyethoxyethanol), 1% DTT, and 2% ampholytes (pH 8−10.5) were added directly to each well. The culture plates were then placed in a 37 °C incubator for 1 h with intermittent (every 15 min) manual agitation. Following the 1 h solubilization, the entire volume was removed from each well and placed in 2 mL Eppendorf tubes. Each sample was then sonicated with a Fisher Sonic Dismembranator using 3 × 2 s bursts at instrument setting #3. Sonication was carried out every 15 min for 1 h after which the fully solubilized samples were transferred to a cryotube for storage at −80 °C until thawed for analysis.

Proteins were resolved by 2D electrophoresis (2DE) using the ISO-DALT System in which up to 20 first- and second-dimension gels can be run simultaneously.[16] Solubilized protein samples were placed on each of 20 first dimension isoelectric focusing (IEF) tube gels (23.5 cm × 1.5 mm) containing 3.3% acrylamide, 9 M urea, 2% Igepal CA-630, 2% ampholyte (BDH pH 4−8), and isoelectrically focused for 28 000 V−h at room temperature. Each IEF gel was then placed, without equilibration, on a second-dimension slab gel (20 cm × 25 cm × 1.5 mm) containing a linear 11−19% acrylamide gradient, generated by a computer-controlled gradient-maker. Second-dimension slab gels were run for approximately 18 h at 160 V and 8 °C and later stained for 96 h with a colloidal Coomassie brilliant blue stain.[17]

After staining, the 2DE protein patterns were scanned under visible light at 200 μm/pixel resolution using the Fluor-S MAX MultiImager System (Bio-Rad, Hercules CA). Image data were analyzed using PDQuest software (Bio-Rad) running under Windows2000 on a PC-workstation.[18] Background was subtracted, and peaks for the protein spots were located and counted. Because total spot counts and the total optical density are directly related to the total protein concentration, individual protein quantities were thus expressed as parts-per-million (PPM) of the total integrated optical density. A reference pattern was constructed, and gel matching was performed with image analysis software matching each gel in the matchset to the reference gel. A typical 2DE hepatocyte pattern containing the stained protein profile is shown in Figure 1. Numerous proteins that were uniformly expressed in all patterns were used as landmarks to facilitate rapid gel matching. Raw protein expression data corresponding to each matched spot was exported for further statistical analysis.

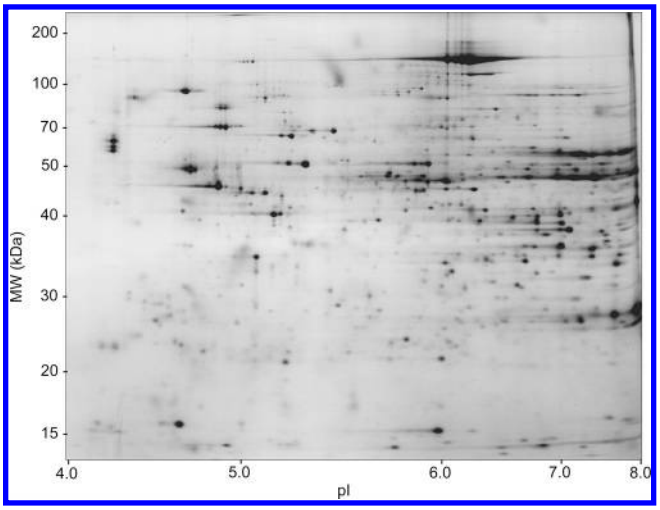

Figure 1. Two-dimensional polyacrylamide gel pattern of primary hepatocyte lysate obtained after monolayer culture. This is a representative of the colloidal Coomassie blue-stained 2D gel pattern from ~100 000 hepatocytes solubilized directly on the culture plate (in situ) after discarding the medium. Protein quantitation and x, y coordinate position is determined using image analysis of each detected spot in the pattern.

CALCULATION OF CHEMODESCRIPTORS

A large set of molecular descriptors (366) was initially calculated for the twenty halocarbons considered in this study; however, many descriptors were removed from the final set. Any descriptor with a value of zero for all twenty compounds was eliminated, and some descriptors that could not be calculated for all twenty compounds were also removed, leaving a final set of 182 molecular descriptors. The software used for calculating these descriptors includes POLLY v2.3,[19] Triplet, [20] Molconn-Z v3.50,[21] and Sybyl v6.4.[22]

POLLY v2.3 and Triplet each calculate approximately 100 topological descriptors including Wiener number,[23] molecular connectivity indices developed by Randić[24] and Kier and Hall,[25] frequency of path lengths of varying size,[25] information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić,[26] Basak et al.,[27] Roy et al.,[28] and Raychaudhury et al.,[29] and parameters defined on the neighborhood complexity of vertices in hydrogen-filled molecular graphs[28] as well as the triplet indices.[20] Molconn-Z v3.50 calculates approximately 150 additional descriptors, including an extended set of molecular connectivity indices, electrotopological state descriptors, general polarity descriptors, hydrogen bonding descriptors, and Kappa shape indices. Finally, Sybyl[22] was used to calculate a relatively small number of geometrical descriptors.

STATISTICAL ANALYSIS

Outliers and Preliminary Analysis. In total, 1401 spots were identified and measured on the gels. In addition to the usual factors that can give rise to missing or outlying readings, electrophoresis gels are susceptible to another

phenomenon—two spots corresponding to different proteins running together and forming one apparent spot. When this is the case, one protein's abundance will be incorrectly recorded as zero, while the total abundance of the two proteins will be assigned to the other protein. This feature creates additional urgency for a robust method able to accommodate individual outlier readings.

We addressed this need using the robust singular value decomposition (rSVD) described in Liu et al.,[30] whose essential features may warrant a brief exposition. Any arbitrary *nxp* data array **X** with elements x_{ij} can be represented as

$$x_{ij} = \sum_{k=1}^{K} r_{ik}c_{jk} + e_{ij}, \text{ or in matrix form } \mathbf{X} = \mathbf{RC^T} + \mathbf{E}$$

where the r_{ik} are 'row markers', the c_{jk} are 'column markers', and the e_{ij} are 'errors'.

In the most familiar form of this representation, the *K* marker vectors defining the matrices **R** and **C** are given by the first *K* eigenvector/value triads of the singular value decomposition. This choice minimizes the sum of squares of the errors e_{ij}—it is a least-squares fit, and in a sense optimal under a Gaussian model of the data.

However, the least squares criterion is a poor choice in settings where the errors e_{ij} are outlier-prone or have some distribution with heavy tails. In a robust singular value decomposition, rather than minimizing the sum of squares of the residuals from the fit, we minimize some other criterion less sensitive to outliers. One such example is the L1, or sum of absolute deviations, criterion. Another is the least trimmed squares (LTS) criterion. In this criterion, we minimize the sum of squares, not of all the fitted residuals, e_{ij}, but of some large fraction of them. This implicitly divides the e_{ij} into two sets—the majority 'covered' residuals, which are incorporated in the sum of squares, and the minority 'uncovered' residuals, which are left out of the sum of squares criterion and so may be arbitrarily large. The latter group is intended to accommodate the outlier elements of the matrix **X**.

The rSVD is conveniently fitted using an alternating robust regression algorithm outlined in Liu et al.,[30] as this method also accommodates missing observations in the array **X**. This feature is not needed in this application, since an abundance is reported for all 1401 spots in all gels, although some of the reported zero abundances are actually the result of proteins merging into some nearby spot.

Even though the original **X** matrix may contain outliers, as a result of the robust fitting the row markers **R** and the column markers **C** should not, in view of the use of an outlier-resistant method to estimate them. We can therefore work with the **R** matrix, in place of the original **X** matrix, in any statistical analyses of the 'cases' in the data set. This is the same as working with principal components instead of the original variables, with two important differences: (i) since the fit is robust, the row markers are not affected by outliers as are the conventional principal components and (ii) many analysts using PCA keep far too few components, leaving relevant 'structure' in the ignored error matrix **E.** Using the rSVD primarily as a way of getting a robust variant of the original data matrix, we will keep a relatively large number (17) of marker vectors.

Compound to Compound Spot Differences. The abundance of a given spot will vary from gel to gel for two reasons—a systematic difference in the expression of that protein as a result of the effect of the halocarbon on the haptocytes and a difference due to random variability. A first step in understanding the relationship between the 1401 proteins and the halocarbon applied is to separate out the systematic differences between one halocarbon to another from the background random variability. In principle, this can be done using a one-way analysis of variance, but in reality the sprinkling of outliers among the spot abundances and the merging of distinct spots make the normality assumptions of the analysis of variance questionable. However, one valuable byproduct of the rSVD is that the product $\mathbf{RC^T}$ gives you a cleaned version of the original data matrix, in which outliers are replaced by values more consistent with the main body of the data. One-way analysis of variance applied to the 'cleaned' abundance of each spot can then test whether that specific spot's abundance varies systematically between the gels produced using different halocarbons.

This analysis gives 1401 *P* values, one for each of the spots. There were in fact numerous spots that differ between the halocarbons. The smallest *P* value was 2×10^{-23}, and there were many highly significant differences, with a frequency breakdown:

$P < 10^{-12}$ 34 spots

$P < 10^{-9}$ 106 spots

$P < 10^{-6}$ 311 spots

These proteins showing highly significant differences between halocarbons constitute a 'shopping list' from which markers of the biological effect of the halocarbon can be expected to be found. Further work on identifying them is ongoing. It is interesting to note that the highly significant differences did not occur in either the most abundant nor the least abundant proteins; rather they were concentrated in the midrange abundances.

Canonical Variates and a Display of the Compounds. The data set comprises 92 gels, and the 1401 spots have been summarized into 17 markers. As there are only 15 compounds tested, the data set provides direct measures of the variability from gel to gel when the same compound is tested. The method of canonical variates is useful in this setting; it provides a system of coordinates in which to plot the compounds, with the property that compounds plotting far apart give gels that are very dissimilar in a multivariate (Mahalanobis distance) sense, while those plotting close together are generally similar in a multivariate sense.

A total of 7 canonical axes were statistically significant in this analysis, carried out using SAS PROC CANVAR. Table 3 shows mean canonical score of each of the compounds' gels. Figure 2 shows the compounds plotted using the two leading canonical variates. The compounds are numbered according to Table 1. This picture shows compound 14 (1,1,2-$C_2H_3Cl_3$) alone and considerably off to the left of the other compounds—even its isomer compound 13 (1,1,1-$C_2H_3Cl_3$). This compound therefore has a quite distinctive proteomic footprint. Conversely, compounds 8 and 11 (1,2-$C_2H_4Br_2$ and 1,2-$C_2H_4Cl_2$) plot close together,

Table 3. Canonical Vector Scores of the Compounds

	compound						
	1	2	3	4	5	6	7
control	1.522	2.180	−1.073	0.145	−0.252	0.555	0.258
C_2Cl_4	3.346	−5.901	−2.939	−3.535	0.543	0.450	0.625
CH_2Br_2	2.689	1.449	4.023	−3.097	2.682	0.726	−3.132
CH_2BrCl	0.239	−4.226	4.386	3.440	0.026	1.575	0.158
CH_2Cl_2	1.515	−4.775	4.463	4.451	−1.558	1.845	1.007
$CHCl_3$	−0.088	−0.977	1.297	−3.321	−5.794	0.750	−2.192
$CHBr_3$	1.980	0.391	0.272	0.822	0.032	−0.964	−0.858
1,2-$C_2H_4Br_2$	−4.776	0.473	−0.351	0.249	0.364	1.111	0.129
1,1,2-$C_2H_3Br_3$	−1.075	−0.585	−0.791	1.296	−0.990	−1.737	0.779
1,2-C_2H_4BrCl	0.097	−1.248	−1.529	2.243	−0.356	−1.145	−0.686
1,2-$C_2H_4Cl_2$	−4.124	1.093	3.286	−3.074	−0.304	−0.787	1.418
C_2HCl_3	2.622	−1.304	−0.448	0.192	0.691	−0.370	−0.333
1,1,1-$C_2H_3Cl_3$	1.704	0.284	1.417	0.354	1.623	−0.016	0.402
1,1,2-$C_2H_3Cl_3$	−10.348	−1.110	−1.570	0.418	0.713	0.227	−0.793
1,1,2,2-$C_2H_2Cl_4$	1.585	−0.315	1.586	−0.125	0.022	−1.117	−0.386

Table 4. q^2 Values for Predicting Log-Scale Toxicity

	toxicity							
descriptors	d37	ARR	EC50MTT	EC50LDH	EC20SH	LECLP	LECROS	LECCAT
bio	0.56	0.60	0.64	0.68	0.66	0.57	0.59	0.67
chemo	0.86	0.86	0.53	0.49	0.67	0.44	0.64	0.71

suggesting that at least in these leading axes the compounds differing just by the exchange of Cl and Br have similar proteomic effects. Compounds 4 and 5 (CH_2BrCl and CH_2Cl_2) are similarly close together, leading to the same conclusion. It is perhaps surprising that such a modest set of compounds (15) could yield statistical significances in so many dimensions (7). Some canonical variates serve primarily to distinguish a single halocarbon from the others—for example variate 5 largely differentiates $CHCl_3$, while variate 7 again loads on $CHCl_3$ but finds a similarity with CH_2Br_2.

Modeling Toxicity. There are two obvious ways to use the proteomics results to model toxicity. One is through the observation that any spot whose abundance might predict toxicity must logically be one that differs significantly from compound to compound. This reasoning then leads to a focus on the spots showing the highest significant differences between the compounds.

That a spot differs between the compounds means that it is a marker for some sort of biological effect of the compounds, but this effect is not necessarily toxicity. Thus the spots that are most significantly different are not necessarily those most predictive of toxicity. And indeed, significant correlations with toxicity do not appear until the spot ordered 8th in overall significance, and the strongest predictors are even further down the list.

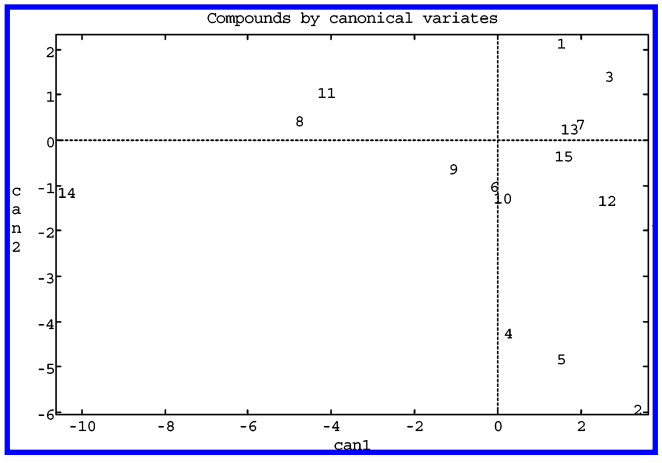

Figure 2. Canonical variates of compounds.

The other method uses the canonical variates. These variates are constructed without any reference to the toxicity measurements, with the less focused objective of maximally differentiating between the halocarbons. It is reasonable to expect that axes which are able to discriminate between the compounds might be useful for predicting the compounds' properties—for example toxicity. We therefore went back to the 92 individual gels, using their canonical scores as predictors of the toxicities recorded in Table 1. The results of ridge regression analysis are summarized by the q^2 values in row 1 of Table 4. The figures are encouraging. We can predict the toxicities of the compounds using the canonical variate scores of a single gel, with a cross-validated q^2 [31] of typically 0.6.

Row 2 of Table 4 also shows, for comparative purposes, the q^2 values obtained by fitting ridge regression models using the 182 molecular descriptors. The figures for the molecular descriptors are not directly comparable with those given by the biodescriptors because they are not based on the same set of compounds. The biodescriptor regressions include control gels in which no halocarbon was applied; clearly no descriptors can be found for this no-compound control, and so it is omitted from the chemodescriptor runs. On the other hand, the chemodescriptor data set included an additional 6 halocarbons for which gels were not made.

Despite these differences in the base, the two sets of q^2 values can be compared qualitatively. This comparison is encouraging. While the chemodescriptors seem better able to predict the two *Aspergillus*-based toxicity measures d37 and ARR, the honors for the other toxicity measures are split.

Using the chemicals that are common to both sets of descriptors, we can carry out a 'test for additional information' to assess whether adding the canonical variates of the biodescriptors to the chemodescriptors gives a significantly better fit. The results of this test are given in Table 5. This

Table 5. *P* Values for Adding Biodescriptors to Chemodescriptors

	toxicity							
	d37	ARR	EC50MTT	EC50LDH	EC20SH	LECLP	LECROS	LECCAT
P value	0.36	0.77	0.004	0.007	0.05	0.13	0.39	0.10

test shows that, for the dependents EC50MTT and EC50LDH, adding the biodescriptors to the chemodescriptors does indeed give a significantly better fit. EC20SH is marginally significant, and on the other toxicity measures there is no perceptible improvement.

DISCUSSION

The spot abundances measured in two-dimensional electrophoresis gels contain conceptually different information than that captured by molecular descriptors. On one hand it can be argued that, since the proteomic map directly measures the result of challenging the cell with the test compound, it should be possible to detect characteristic perturbations in cellular metabolic machinery resulting from the test compound's hepatotoxicity. A contrary argument would be that the molecular descriptors capture the inherent information that leads to both protein expression and toxicity, and so models based on chemodescriptors alone should be able to predict toxicity. In addition, they have the advantage of being computable in silico without experimentation and its consequent costs and potential for adding measurement random variability.

The statistical performance figures do not support any claim of strong superiority of either one of these types of information over the other. Rather they lead to the possibility of combined information sources, using both chemodescriptors and biodescriptors, for the prediction of toxicity.

Ideally, one would like to find the smoking gun in the 2D gel—specific spots in the gel whose presence or absence is associated with hepatotoxicity. There are a number of spots in the gels that differ markedly between halocarbons, and some of these are predictive of toxicity, though none is highly specific. Exploration of the spots associated with toxicity through either over- or underexpression will provide better understanding of the mechanisms of hepatotoxicity.

Summarizing the gel information with canonical variate scores, which are linear combinations of all the spots, provides a compact set of numbers characterizing the biological differences among the compounds. The premise of this work is that the information in the biodescriptors is logically different than that in the chemodescriptors, and hence that using both sources of information may give better results than are obtained with either source alone. The promising results seen so far support this hypothesis.

ACKNOWLEDGMENT

This manuscript is contribution number 389 from the Center for Water and the Environment of the Natural Resources Research Institute. This material is based on research sponsored by the Air Force Research Laboratory, under agreement number F49620-02-1-0138. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government.

REFERENCES AND NOTES

(1) Basak, S. C.; Mills, D.; Gute, B. D. Predicting bioactivity and toxicity of chemicals from mathematical descriptors: A chemical-cum-biochemical approach. In *Advances in Quantum Chemistry: Chemical Graph Theory: wherefrom, wherefor, & whereto*; Klein, D. J., Brandas, E., Eds.; Elsevier-Academic Press: 2005; in press.

(2) Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure−activity relationships (SAR) under TSCA, Section 5. *Environ. Health Perspect.* **1990**, *87*, 183−197.

(3) Hansch, C.; Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(4) Kamlet, M. J.; Abboud, J.-L. M.; Abraham, M. H.; Taft, R. W. Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromatic parameters, *, a and b, and some methods for simplifying the general solvatochromatic equation. *J. Org. Chem.* **1983**, *48*, 2877−2887.

(5) Crebelli R.; Andreoli, C.; Carere, A.; Conti, L.; Crochi, B.; Cotta-Ramusino, M.; Benigni, R. Toxicology of halogenated aliphatic hydrocarbons: structural and molecular determinants for the disturbance of chromosome segregation and the induction of lipid peroxidation. *Chemico-Biological Interactions* **1995**, *98*, 113−129.

(6) Geiss, K. T.; Frazier, J. M. QSAR Modeling of oxidative stress in vitro following hepatocyte exposures to halogenated methanes. *Toxicol. in Vitro* **2001**, *15*, 557−563.

(7) Basak, S. C.; Balasubramanian, K.; Gute, B. D.; Mills, D.; Gorczynska, A.; Roszak, S. Prediction of cellular toxicity of halocarbons from computed chemodescriptors: A hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1103−1109.

(8) Kaufman, J. J.; Koski, W. S.; Roszak, S.; Balasubramanian, K. Correlation between energetics and toxicities of single-carbon halides. *Chem. Phys.* **1996**, *204*, 233−237.

(9) Trohalaki, S.; Pachter, R.; Geiss, K. T.; Frazier, J. M. Halogenated aliphatic toxicity QSARs employing metabolite descriptors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1186−1192.

(10) Randic, M.; Witzmann, F.; Vracko, M.; Basak, S. C. On characterization of proteomics maps and chemically induced changes in proteomes using matrix invariants: Application to peroxisome proliferators. *Med. Chem. Res.* **2001**, *10*, 456−479.

(11) Basak, S. C.; Gute, B. D.; Witzmann, F. A. Information-theoretic biodescriptors for proteomics maps: Development and applications in predictive toxicology. In *Proceedings of the 9th WSEAS International Conference on Computers*, 2005, accepted.

(12) Vracko, M.; Basak, S. C. Similarity study of proteomic maps. *Chemom. Intell. Lab. Syst.* **2004**, *70*, 33−38.

(13) Seglen, P. Preparation of isolated rat liver cells. In *Methods in Cell Biology*; Prescott, D. M., Ed.; Academic Press: New York, 1976; pp 29−83.

(14) Geiss, K. T.; Frazier, J. M. A novel in vitro system for exposures of cell cultures to volatile chemicals. *The Toxicologist* **2000**, *54*, 377.

(15) Geiss, K. T.; Frazier, J. M.; Dodd, D. E. Toxicity screening of halogenated aliphatics using a novel in vitro volatile chemical exposure system, Proceedings of the 12th Halon Options Technical Working Conference, *NIST Special Publication 984*, 2002.

(16) Anderson, N. L. *Two-dimensional electrophoresis: Operation of the ISO-DALT System*; Large Scale Biology Press: Washington, DC, 1991.

(17) Neuhoff, V.; Arold, N.; Taube, D.; Ehrhardt, W. Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie Brilliant Blue G-250 and R-250. *Electrophoresis* **1988**, *9*, 255−262.

(18) Anderson, N. L.; Hofmann, J. P.; Gemmell, A.; Taylor, J. Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. *Clin. Chem.* **1984**, *30*, 2031−2036.

(19) *Polly, v 2.3*; Copyright of the University of Minnesota, 1988.

(20) Filip, P. A.; Balaban, T. S.; Balaban, A. T. A New Approach for Devising Local Graph Invariants: Derived Topological Indices with Low Degeneracy and Good Correlational Ability. *J. Math. Chem.* **1987**, *1*, 61−83.
(21) *Molconn-Z, v 3.50*; Hall Associates Consulting: Quincy, MA, 2000.
(22) *Sybyl, v 6.4*; Tripos Associates, Inc.: St. Louis, MO, 1997.
(23) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(24) Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(25) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Letchworth, Hertfordshire, U.K., 1986.
(26) Bonchev, D.; Trinajstic, N. Information Theory, Distance Matrix and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517−4533.
(27) Basak, S. C.; Roy, A. B.; Ghosh, J. J. Study of the Structure−Function Relationship of Pharmacological and Toxicological Agents using Information Theory. In *Proceedings of the Second International Conference on Mathematical Modelling*; Avula, X. J. R., Bellman, R., Luke, Y. L., Rigler, A. K., Eds.; University of Missouri—Rolla: Rolla, Missouri, 1980.
(28) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications. In *Mathematical Modelling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Lapis, A. I., Rodin, E. Y., Eds.; Pergamon Press: New York, 1984.
(29) Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. Discrimination of Isomeric Structures using Information Theoretic Topological Indices. *J. Comput. Chem.* **1984**, *5*, 581−588.
(30) Liu, L.; Hawkins, D. M.; Ghosh, S.; Young, S. S. Robust Singular Value Decomposition Analysis of Microarray Data. *Proc. Natl. Acad. Sci.* **2003**, *100*, 13167−13172.
(31) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing Model fit by Cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.

CI050252P