Classification of Environmental Estrogens by Physicochemical Properties Using Principal Component Analysis and Hierarchical Cluster Analysis

Takahiro Suzuki,*[,†] Kunihito Ide,[†] Masaru Ishida,[†] and S. Shapiro[‡]

Chemical Resources Laboratory, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku,
Yokohama 226-8503, Japan, and Institut für orale Mikrobiologie und allgemeine Immunologie,
Zentrum für Zahn-, Mund- und Kieferheilkunde der Universität Zürich, Plattenstrasse 11, Postfach,
CH-8028 Zürich 7, Switzerland

Received July 12, 2000

A structurally diverse assortment of 60 environmental estrogens was divided into two main clusters (“A”, “B”) and a pair of subclusters (“C1”, “C2”) by applying principal component analysis to selected 1D and 2D molecular descriptors and subjecting the PCs to hierarchical cluster analysis. Although clustering was predicated solely on physicochemical properties, the dependence on particular physicochemical parameters of xenoestrogen binding affinities (pK_i) to murine uterine cytosolic estrogen receptor (ER) proved greater for compounds within (sub)clusters than for compounds between (sub)clusters. Quantitative structure-binding affinity relationships derived using molecular descriptors and PCs suggested differences in the driving forces for xenoestrogen-ER binding for different (sub)clusters. The modeling power for xenoestrogen-ER binding affinities of a combination of TLSER and WHIM 3D indices was much greater than that of combinations of 1D and 2D molecular descriptors or the PCs derived therefrom. The clusterings obtained using PCs also proved applicable to the 3D-QSARs.

INTRODUCTION

There is widespread scientific and public concern about the effects of environmental chemicals that mimic or obstruct the action of estrogens in animals and humans.[1] Environmental estrogens (“xenoestrogens”), for instance, are suspected of promoting breast cancer.[2,3] Some 70 chemicals (including many nonsteroids) are regarded as possible xenoestrogens,[4,5] though a common structural feature responsible for estrogenic activity has yet to be identified.

Quantitative structure−activity relationships (QSARs) for binding of xenoestrogens to estrogen receptors (ERs) have been obtained using classical regression techniques, comparative molecular field analysis, and hologram QSAR.[4,6−11] However, these efforts have not shed much light on the nature of xenoestrogen-ER interactions. We sought to explore the interactions between murine uterine cytosolic ERs (mainly ER-α) and a set of environmental estrogens by generating QSARs for binding affinity (expressed as pK_i, the negative logarithm of the xenoestrogen-ER binding constant K_i in μmol) using a collection of molecular descriptors and principal components (PCs) derived therefrom. Hierarchical cluster analysis (HCA) was applied to the PCs to determine how the xenoestrogens may be grouped. It was anticipated that clustering based solely on structural information would reveal differences between the mechanisms by which these diverse environmental chemicals exert their endocrinological effects.

METHODS

Data Sets. Sixty estrogenic molecules[4,5,12] covering a broad range of chemical types (6 phenolics, 3 phthalates, 16 PCBs, 6 DDTs, 15 other pesticides, 3 DESs, 8 steroids, and 3 phytoestrogens; see Figure 1) were examined. Experimental pK_i values for 49 of 60 compounds were taken from Waller et al.[4]

Descriptors for Classification/Clustering. Many hundreds of molecular descriptors are available for construction of QSARs;[13,14] 15 one- and two-dimensional descriptors which have proven especially useful for characterizing xenoestrogen-ER interactions[4,6,10] were selected for PCA: molecular weight (MW, g/mol), van der Waals volume (V_w, cm^3/mol), melting point (T_m, °C), dipole moment (μ, D), standard enthalpy of formation ($\Delta H_f°$, kJ/mol), HOMO energy (ϵ_{HOMO}, eV), LUMO energy (ϵ_{LUMO}, eV), the most negative partial atomic charge in a molecule (q^-, au), the most positive partial atomic charge in a molecule (q^+, au), the sum of all negative partial atomic charges in a molecule (Q_T^-, au), absolute (Mulliken) electronegativity [$EN = -(\epsilon_{HOMO} + \epsilon_{LUMO})/2$, eV], absolute hardness [$HD = -(\epsilon_{HOMO} - \epsilon_{LUMO})/2$, eV], the zero- and first-order simple molecular connectivity indices ($^0\chi$ and $^1\chi$, respectively),[15] and the *n*-octanol/water partition coefficient (log *P*). Although devoid of explicit three-dimensional (3D) information, the calculated values of most of these indices are nonetheless sensitive to molecular conformation. Moreover, Ajay et al.[16] and Brown and Martin[17] reported that 2D descriptors were more effective at separating active from inactive compounds than some 3D descriptors.

Molecular geometries were optimized initially using PC-Model v5.13 and GMMX v1.6 (Serena Software, Bloom-

[*] Corresponding author phone and fax: (+81 45) 924-5255; e-mail: tsuzuki@res.titech.ac.jp.
[†] Tokyo Institute of Technology.
[‡] Zentrum für Zahn-, Mund- und Kieferheilkunde der Universität Zürich.

10.1021/ci000333f CCC: $20.00
Published on Web 04/05/2001

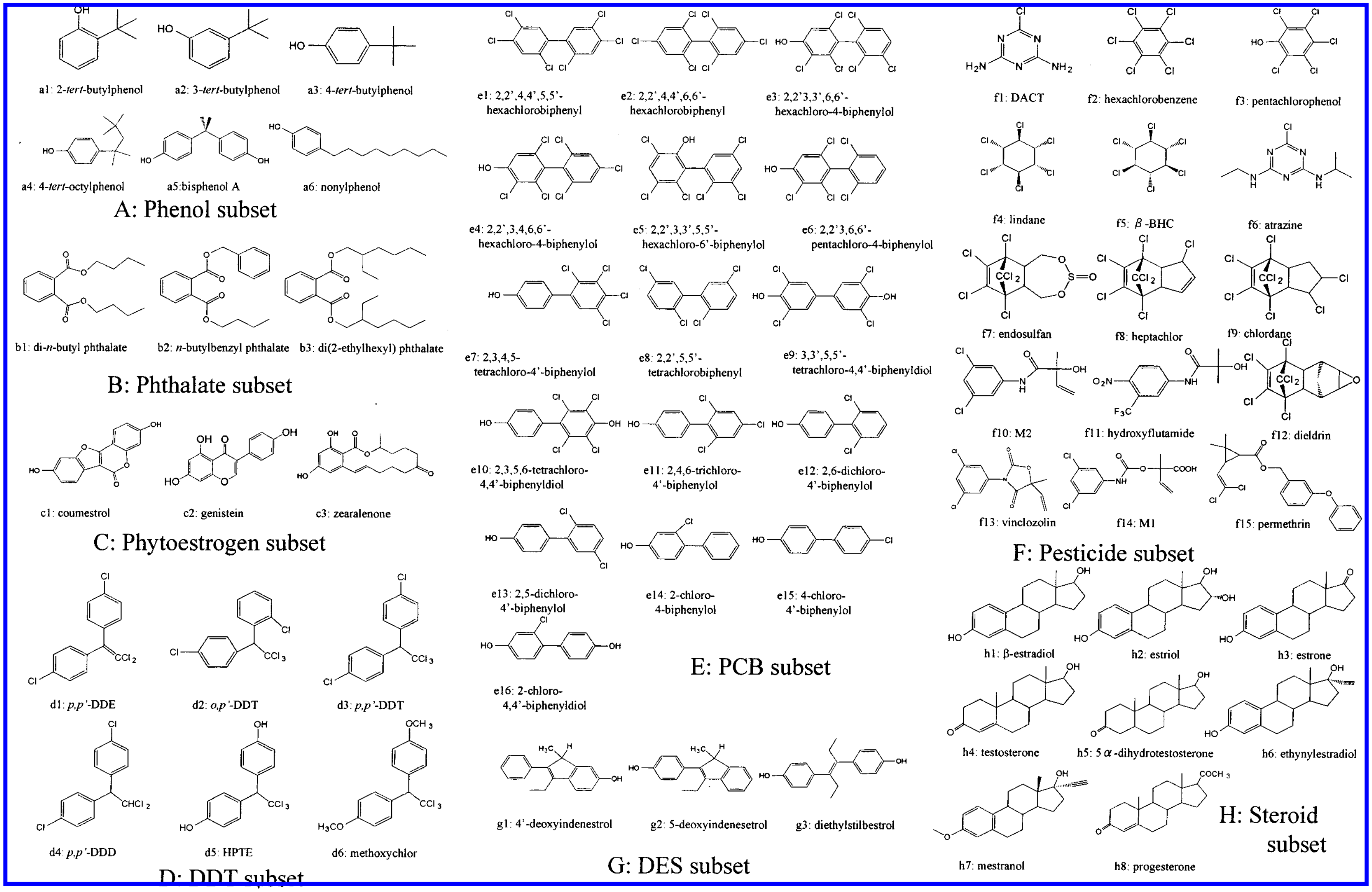

Figure 1. Structures of the 60 xenoestrogens used for HCA. d5: 2,2-bis(*p*-hydroxyphenyl)-1,1,1-trichloroethane, f1: diaminochlorotriazine, f5: *β*-hexachlorocyclohexane, f10: 3′,5′-dichloro-2-hydroxy-2-methylbut-3-enanilide, f14: 2-[[(3,5dichlorophenyl)carbamoyl]oxy]-2-methyl-3-butenoic acid.

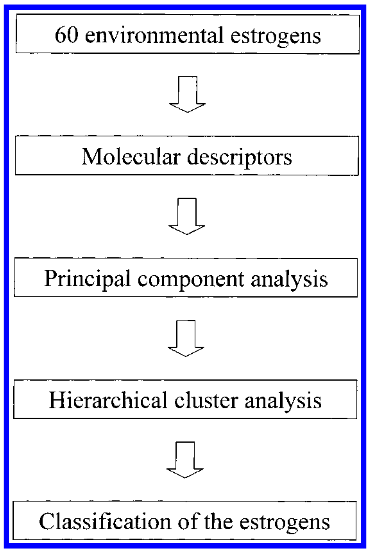
60 environmental estrogens
Molecular descriptors
Principal component analysis
Hierarchical cluster analysis
Classification of the estrogens

Figure 2. Flow diagram for procedures used in classifying environmental estrogens.

ington, IN) and refined using the AM1 Hamiltonian[18] and Eigenvector Following routine[19] as implemented in MOPAC 6.0 and VAMP v4.40; quantum mechanical descriptors (including TLSER descriptors; vide infra) were computed using these same programs. Molecular connectivity indices were obtained from SMILES string notations using MOLCONN-X v2.0 (Hall Associates Consulting, Quincy, MA) with the standard control option. Experimental log P values were taken from the LOGKOW databank,[20] where experimental values were lacking they were calculated using CHEMICALC-2.[21] Melting points for which experimental values could not be found in the literature were predicted using the group contribution methods of Yalkowsky and coworkers[22−24] or of Joback and Reid.[25]

Clustering. Clustering of environmental estrogens was accomplished as outlined in Figure 2. Since the descriptors were of nonuniform dimensionality PCA was performed using a variance-covariance matrix of data normalized to a mean of 0 and a variance of 1. HCA of environmental estrogens was achieved in the space defined by values of the first six PCs; Euclidean distances between combinations of PC scores were determined using Ward's method.[26] PCA and HCA were performed using Excel Multivariate Analysis v3.0 (Esumi Co., Ltd., Tokyo, Japan) on a microcomputer running Windows 95 as its operating system. Additional statistical analyses were done using MINITAB v11.21 (Minitab, Inc., State College, PA).

Descriptors for 3D-QSARs. The different classes of environmental estrogens examined in this study (Figure 1) contained no common structural elements, precluding application of orientation-sensitive (e.g. STERIMOL)[27,28] or alignment-sensitive (e.g. CoMFA,[29] SEAL[30]) measures of three-dimensionality. Instead, 3D rototranslational invariant WHIM[31,32] and TLSER[33] descriptors were employed. WHIM indices were calculated from Sybyl.mol2 files of geometry-optimized structures using DRAGON v1.1 (Chemometrics and QSAR Research Group, Dipartimento di Scienze dell'Ambiente e del Territorio, Università degli Studi di Milano-Bicocca, Italy; available from http://www.disat.unimib.it/chm/Dragon.html); the TLSER V_{mc} (reduced molecular volume) index was calculated from MOPAC.arc files for these same geometry-optimized structures using PC-Model. In contrast to many other 3D-QSAR methods currently in vogue the 3D descriptors used in this work were not dependent on structural alignment or identification of putative bioactive conformers, thus eliminating a substantial measure of conjecture. The correlation matrix for the 6 TLSER descriptors + 66 directional WHIM descriptors + 33 nondirectional WHIM descriptors for the complete xenoestrogen data set ($n = 60$) was analyzed; when a descriptor pair was identified for which $r > 0.90$[34] the descriptor possessing a clearer intuitive physicochemical meaning and/or which was easier to calculate was retained and the other deleted. The descriptor pool was further refined by eliminating indices for which correlations with pK_i ($n = 49$) < 0.1.[35]

Nonlinear Modeling. The theory and general practice of artificial neural networks (ANNs) and their applications in chemistry have been reviewed in depth.[36] A fully connected three-layer neural network was used to relate xenoestrogen-ER binding to a selected set of 3D molecular descriptors used for multiple linear regression (MLR) (vide infra). The hidden layer contained variable nodes, and the input and hidden variables each had a bias neuron. A sigmoid transfer function was used for each neuron, and connection weights were adjusted iteratively by back-propagation using the generalized delta rule to minimize mean square errors between desired and actual outputs. Input and output data were normalized between 0.05 and 0.95, and models were evaluated on the basis of correlation coefficient (r) and root-mean-square error (RMSE).[37]

RESULTS AND DISCUSSION

Classification of Xenoestrogens. High correlations ($r \geq 0.8$) were found for only five of the 105 nonredundant pairwise combinations of 1D and 2D descriptors: q^-, q^+ (0.854); EN, ϵ_{HOMO} (0.842); V_w, ${}^1\chi$ (0.923); V_w, ${}^0\chi$ (0.928); ${}^0\chi$, ${}^1\chi$ (0.975). The variability of the 15 1D and 2D descriptors were redistributed across a set of orthogonal indices by PCA; the first six principal components were weighted according to their eigenvalues as follows: PC_1, 32.0%; PC_2, 20.6%; PC_3, 14.8%; PC_4, 12.6%; PC_5, 7.0%; PC_6, 4.9%. Loading plots for the first three PCs (Figure 3a) showed PC_1 dominated by ${}^0\chi$, ${}^1\chi$, and V_w, which describe molecular size or bulk; PC_2 by ϵ_{LUMO} and HD, reflecting molecular polarity; and PC_3 mainly by log P, with lesser contributions by $V_{w,}$ ${}^1\chi$, and ϵ_{HOMO}. Loading plots of PC_4, PC_5, and PC_6 (Figure 3b) revealed PC_4 (dominated by ϵ_{HOMO} and HD) and PC_6 (dominated by μ) to reflect different aspects of molecular polarity; whereas PC_5, highly correlated with T_m, was a complex function of intrinsic properties such as bulk, polarity, and cohesiveness.[38]

Classification of all environmental estrogens ($n = 60$) by HCA of the first six PC values is summarized in Figure 4. Xenoestrogens fell into two clusters ("A", "B") and two subclusters ("C1", "C2"). Cluster A contains 15 relatively small mono- and bicyclic compounds: the 3 *tert*-butyl phenols (a1-a3), the 7 PCBs with ≤ 4 chlorine and/or hydroxy substituents (e8,e11-e16), and 5 smaller pesticides (f1-f5). Cluster B contains all of the DDTs (d1-d6), the remaining 9 PCBs (e1-e7,e9-e10), 3 pesticides (f8,f9,f12), and 2 DES-like molecules (g1,g2). Except for g1 and g2 all members of cluster B contain ≥ 4 chlorine and/or hydroxy and/or methoxy moieties. Subcluster C1 consists of the 3

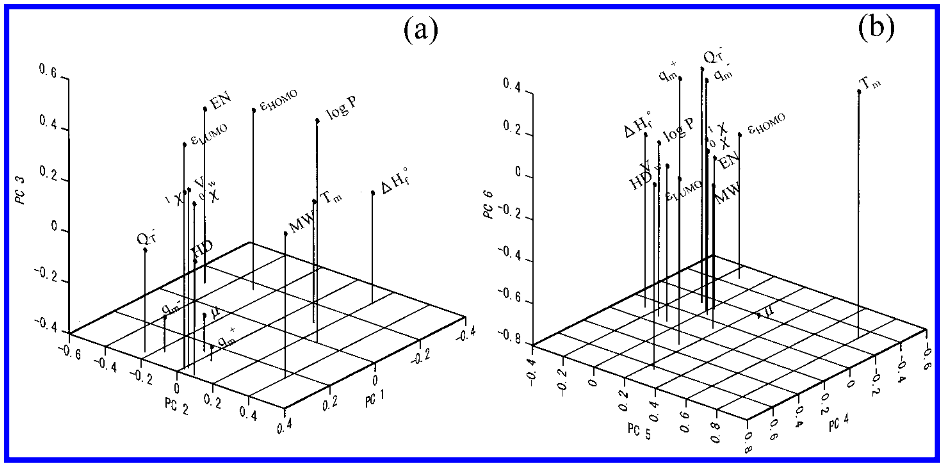

Figure 3. Loading plots for (a) PC$_1$, PC$_2$, and PC$_3$ and (b) PC$_4$, PC$_5$, and PC$_6$.

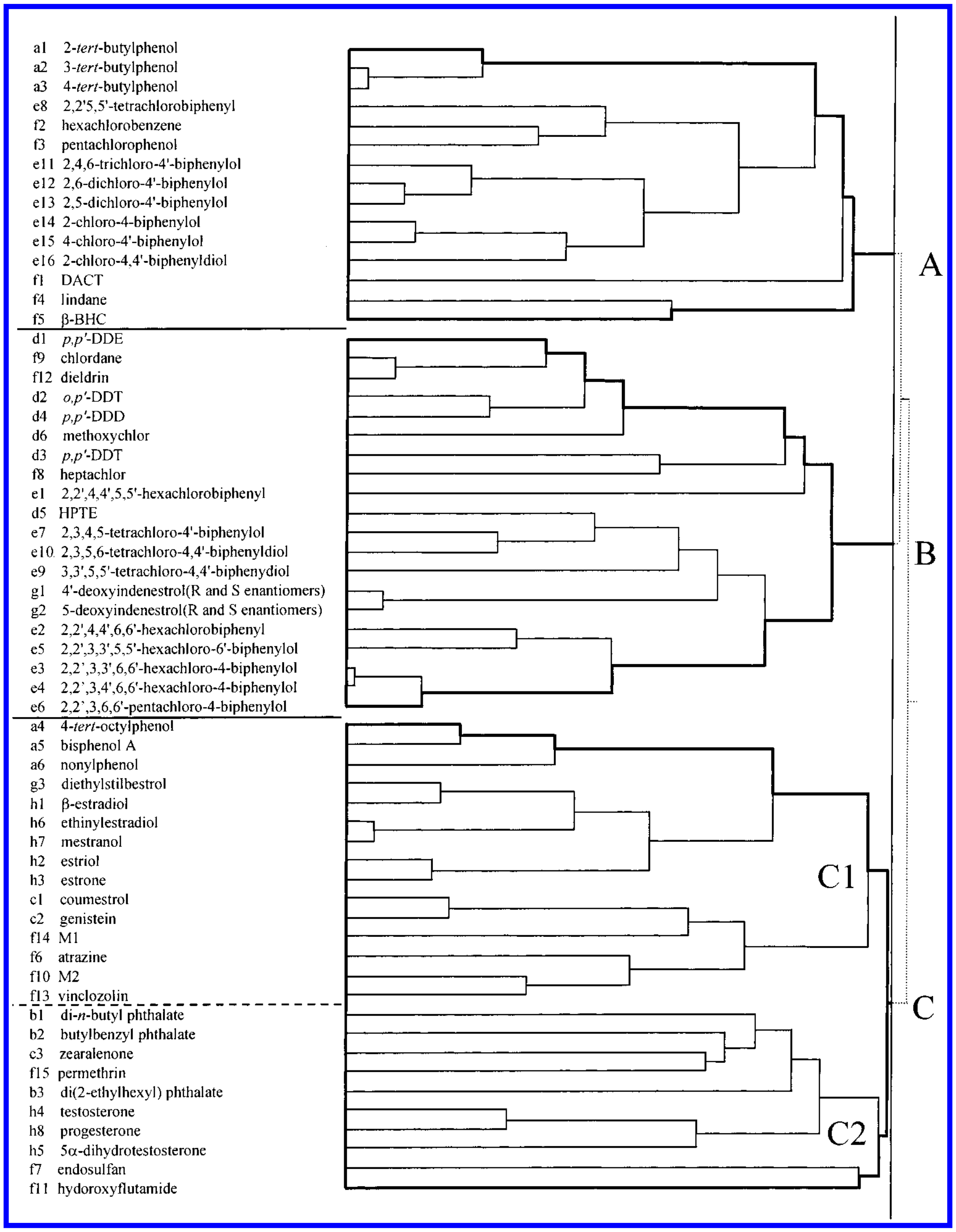
a1 2-tert-butylphenol
a2 3-tert-butylphenol
a3 4-tert-butylphenol
e8 2,2'5,5'-tetrachlorobiphenyl
f2 hexachlorobenzene
f3 pentachlorophenol
e11 2,4,6-trichloro-4'-biphenylol
e12 2,6-dichloro-4'-biphenylol
e13 2,5-dichloro-4'-biphenylol
e14 2-chloro-4-biphenylol
e15 4-chloro-4'-biphenylol
e16 2-chloro-4,4'-biphenyldiol
f1 DACT
f4 lindane
f5 β-BHC
d1 p,p'-DDE
f9 chlordane
f12 dieldrin
d2 o,p'-DDT
d4 p,p'-DDD
d6 methoxychlor
d3 p,p'-DDT
f8 heptachlor
e1 2,2',4,4',5,5'-hexachlorobiphenyl
d5 HPTE
e7 2,3,4,5-tetrachloro-4'-biphenylol
e10 2,3,5,6-tetrachloro-4,4'-biphenyldiol
e9 3,3',5,5'-tetrachloro-4,4'-biphenydiol
g1 4'-deoxyindenestrol(R and S enantiomers)
g2 5-deoxyindenestrol(R and S enantiomers)
e2 2,2',4,4',6,6'-hexachlorobiphenyl
e5 2,2',3,3',5,5'-hexachloro-6'-biphenylol
e3 2,2',3,3',6,6'-hexachloro-4-biphenylol
e4 2,2',3,4',6,6'-hexachloro-4-biphenylol
e6 2,2',3,6,6'-pentachloro-4-biphenylol
a4 4-tert-octylphenol
a5 bisphenol A
a6 nonylphenol
g3 diethylstilbestrol
h1 β-estradiol
h6 ethinylestradiol
h7 mestranol
h2 estriol
h3 estrone
c1 coumestrol
c2 genistein
f14 M1
f6 atrazine
f10 M2
f13 vinclozolin
b1 di-n-butyl phthalate
b2 butylbenzyl phthalate
c3 zearalenone
f15 permethrin
b3 di(2-ethylhexyl) phthalate
h4 testosterone
h8 progesterone
h5 5α-dihydrotestosterone
f7 endosulfan
f11 hydoroxyflutamide
A
B
C
C1
C2

Figure 4. Dendrogram of relationships between environmental estrogens obtainedby HCA of 15 1D and 2D molecular descriptors.

larger phenols (a4-a6), 4 pesticides (f6,f10,f13-f14), *trans*-diethylstilbestrol (g3), 5 steroids (h1-h3,h6,h7), and 2 phytoestrogens (c1,c2). The molecules in this subcluster structurally resemble archetypal estrogenic species such as

β-estradiol (h1) and *trans*-diethylstilbestrol (g3). The compound positioned closest to β-estradiol in the dendrogram is *not* another steroid but rather *trans*-diethylstilbestrol. Subcluster C2 includes all the phthalates (b1-b3), 3 larger pesticides (f7,f11,f15), 3 steroids (h4,h5,h8), and a phytoestrogen (c3). The average molecular volume of the compounds in subcluster C2 (400.6 Å^3 ± 69.5 Å^3, $n = 10$) is larger than that of subcluster C1 (320.7 Å^3 ± 41.9 Å^3, $n = 15$). Testosterone (h4), 5α-dihydrotestosterone (h5), and progesterone (h8) occurred together in subcluster C2, whereas the other steroids (all of which have an aromatized ring A) occurred in subcluster C1. At 30 °C the aromatic ring contributes ca. −1.5 kcal/mol to the free energy of binding of β-estradiol to human uterine ER, comparable to the contribution of the 3-hydroxy group.[39]

PCR and MLR QSARs for (Sub)clusters. Although clustering of environmental estrogens (Figure 4) was accomplished in the absence of biological data, it was reasoned that the dependence of pK_i on particular combinations of physicochemical parameters would be greater for compounds within (sub)clusters than for compounds between (sub)clusters.[40] To test this hypothesis QSARs formulated from linear combinations of PCs

$$pK_i = \sum_{i=1}^{n} a_i PC_i + c \qquad (1)$$

were obtained at the 0.05 level of significance for (sub)clusters by principal component regression (PCR) and compared to QSARs for (sub)clusters obtained by "best subsets" MLR using the 15 1D and 2D descriptors.

For cluster A the PCR QSAR was

$$pK_i = \underset{(\pm 0.142)}{0.877\ PC_3} - \underset{(\pm 0.210)}{0.923\ PC_4} + \underset{(\pm 0.415)}{1.473\ PC_5} + \underset{(\pm 0.298)}{0.477} \qquad (2)$$

$$|t\text{-statistic}|:\quad 6.17,\quad 4.39,\quad 3.55\quad (\text{for } PC_3,\ PC_4,\ PC_5)$$

$$n = 12,\quad r^2 = 0.866,\quad q^2 = 0.404,\quad \sigma = 0.681,\quad F = 17.3$$

where n = sample size, r^2 = index of determination, q^2 = cross-validated (L−O−O) index of determination, σ = standard deviation, and F = Fisher distribution statistic. The MLR counterpart to eq 2 is

$$pK_i = \underset{(\pm 0.099)}{0.579 \log P} + \underset{(\pm 0.002)}{0.011\ T_m} + \underset{(\pm 0.255)}{0.822\ \epsilon_{HOMO}} + \underset{(\pm 2.489)}{2.741} \qquad (3)$$

$$\text{VIF}:\quad 1.1,\quad 1.0,\quad 1.0\quad (\text{for } \log P,\ T_m,\ \epsilon_{HOMO})$$

$$|t\text{-statistic}|:\quad 5.82,\quad 5.44,\quad 3.22$$

$$n = 12,\quad r^2 = 0.896,\quad q^2 = 0.453,\quad \sigma = 0.600,\quad F = 23.0$$

which has a slightly improved statistical significance compared to eq 2 (VIF = variance inflation factor).

For cluster B the PCR QSAR is

$$pK_i = \underset{(\pm 0.607)}{-0.848\ PC_3} + \underset{(\pm 0.413)}{0.546\ PC_5} + \underset{(\pm 0.616)}{0.964\ PC_6} \underset{(\pm 0.848)}{-0.033} \qquad (4)$$

$$|t\text{-statistic}|:\quad 1.40,\quad 1.32,\quad 1.56$$

$$n = 15,\quad r^2 = 0.531,\quad q^2 < 0,\quad \sigma = 1.052,\quad F = 4.1$$

and the corresponding MLR QSAR

$$pK_i = \underset{(\pm 0.002)}{0.009\ T_m} + \underset{(\pm 2.379)}{3.252\ HD} - \underset{(\pm 0.269)}{0.704 \log P} - \underset{(\pm 9.941)}{13.203} \qquad (5)$$

$$\text{VIF}:\quad 1.7,\quad 2.0,\quad 1.7$$

$$|t\text{-statistic}|:\quad 3.97,\quad 1.37,\quad 2.61$$

$$n = 15,\quad r^2 = 0.642,\quad q^2 = 0.289,\quad \sigma = 0.919,\quad F = 6.6.$$

Although PC_6 is a partial function of μ, the latter contributed negligibly to the MLR QSAR for cluster B xenoestrogens (eq 5). The best MLR QSAR for cluster B xenoestrogens had a higher statistical quality than the best PCR QSAR (eq 4); with $q^2 < 0$ this PCR QSAR was of no predictive value.

For subcluster C1 the PCR and MLR QSARs are

$$pK_i = \underset{(\pm 0.221)}{1.612\ PC_3} + \underset{(\pm 0.404)}{1.266\ PC_5} + \underset{(\pm 0.307)}{0.327} \qquad (6)$$

$$|t\text{-statistic}|:\quad 7.29,\quad 3.14$$

$$n = 14,\quad r^2 = 0.835,\quad q^2 = 0.702,\quad \sigma = 1.061,\quad F = 27.7$$

and

$$pK_i = \underset{(\pm 0.966)}{4.716\ \epsilon_{HOMO}} + \underset{(\pm 0.016)}{0.055\ V_w} + \underset{(\pm 9.815)}{34.661} \qquad (7)$$

$$\text{VIF}:\quad 1.2,\quad 1.2$$

$$|t\text{-statistic}|:\quad 4.88,\quad 3.51$$

$$n = 14,\quad r^2 = 0.846,\quad q^2 = 0.751,\quad \sigma = 1.023,\quad F = 30.3$$

respectively. PC_3 is governed mainly by $\log P$, but the MLR QSAR for subcluster C1 did not contain this descriptor.

Experimental and calculated pK_i values using both PCR and MLR QSARs (eqs 2−7) for clusters A and B and subcluster C1 are listed in Table 1. Since binding data are available for only eight of the 10 xenoestrogens in subcluster C2 (b1,b2,c3, f7,f11,h4, h5,h8), the QSARs derivable for this subcluster will be of questionable robustness; moreover, with the exception of zearalenone (c3) there is little variation in pK_i value among the members of subcluster C2 (the relatively high pK_i of c3, 2.222, has been attributable to a

Table 1. Observed and Predicted pK_i Values

		calcd pK_i (calibration)			
compound	obsd pK_i	PCR models (eq 2)	MLR models (eq 3)	eq 10a	ANN
		Cluster A			
2-*tert*-butylphenol	−2.367	−3.105	−2.787	−2.345	−2.999
3-*tert*-butylphenol	−2.597	−2.608	−2.289	−2.433	−3.007
4-*tert*-butylphenol	−2.207	−1.845	−1.600	−1.790	−3.015
2,2′5,5′-tetrachlorobiphenyl	−0.792	0.127	0.432	−1.282	−0.742
2,4,6-trichloro-4′-biphenylol	1.316	0.732	0.845	1.083	2.198
2,6-dichloro-4′-biphenylol	0.519	0.049	0.089	1.827	0.453
2,5-dichloro-4′-biphenylol	0.446	0.281	0.177	0.279	0.542
2-chloro-4-biphenylol	−0.509	−0.477	−0.618	0.778	−0.383
4-chloro-4′-biphenylol	−0.746	0.181	−0.929	1.202	−0.816
2-chloro-4,4′-biphenyldiol	0.939	0.184	0.744	−0.187	2.017
DACT	−3.000	−2.913	−2.733	−1.787	−2.984
lindane	−3.000	−2.599	−3.329	−2.743	−2.894

		calcd pK_i (calibration)			
compound	obsd pK_i	PCR models (eq 4)	MLR models (eq 5)	eq 10a	ANN
		Cluster B			
p,p′-DDE	−3.000	−1.746	−2.758	−3.067	−2.993
o,p′-DDT	−0.462	−1.885	1.347	−1.189	−1.558
p,p′-DDD	−3.000	−1.962	−1.589	−2.271	−2.908
methoxychlor	−1.839	−2.306	−1.657	0.148	−0.174
p,p′-DDT	−3.000	−2.089	−2.333	−1.988	−2.869
2,2′,4,4′,5,5′-hexachlorobiphenyl	−0.934	−1.263	−1.456	−1.219	−1.061
HPTE	1.301	0.424	1.289	0.266	−0.390
2,3,4,5-tetrachloro-4′-biphenylol	1.345	−0.266	−0.533	−0.841	0.625
2,3,5,6-tetrachloro-4,4′-biphenyldiol	−0.380	0.259	−0.276	1.206	−0.127
3,3′,5,5′-tetrachloro-4,4′-biphenyldiol	−0.290	0.503	0.991	0.062	−0.738
2,2′,4,4′6,6′-hexachlorobiphenyl	−0.117	−1.325	−0.830	−0.822	−0.742
2,2′,3,3′,5,5′-hexachloro-6′-biphenylol	−0.812	−0.717	−0.790	−1.035	−0.652
2,2′,3,3′,6,6′-hexachloro-4-biphenylol	−0.847	−0.443	−0.805	−1.399	−1.108
2,2′,3,4′6,6′-hexachloro-4-biphenylol	−0.732	−0.400	−0.801	−1.353	−0.673
2,2′,3,6,6′-pentachloro-4-biphenylol	−0.203	0.247	−0.074	−0.419	−0.389

		calcd pK_i (calibration)			
compound	obsd pK_i	PCR models (eq 6)	MLR models (eq 7)	eq 10a	ANN
		Cluster C1			
4-*tert*-octylphenol	−0.121	0.557	0.460	−0.312	0.032
bisphenol A	−0.164	0.595	0.358	0.488	−0.082
nonylphenol	0.080	−0.056	0.720	−1.170	0.193
trans-diethylstilbestrol	3.155	2.554	2.265	3.697	3.172
β-estradiol	2.585	1.985	2.061	0.562	2.321
ethinylestradiol	3.523	2.606	3.466	0.393	2.277
estriol	1.854	3.350	2.050	0.573	2.185
estrone	2.357	1.884	1.296	0.549	2.247
coumestrol	1.032	0.326	0.864	−0.105	0.907
genistein	0.409	−1.033	−0.308	1.330	−0.028
M1	−3.000	−1.425	−1.581	−2.802	−2.787
atrazine	−3.000	−4.046	−3.691	−1.760	−2.828
M2	−3.000	−2.023	−1.053	−3.135	−2.907
vinclozolin	−3.000	−2.565	−4.196	−2.210	−2.859

		calcd pK_i (calibration)			
compound	obsd pK_i	PCR models	MLR models	eq 10a	ANN
		Cluster C2			
di-*n*-butyl phathalate	−2.002			−1.794	−2.638
n-butylbenzyl phthalate	−1.883			−1.128	−1.170
zearalenone	2.222			−0.482	0.944
testosterone	−1.462			−1.462	−1.706
progesterone	−3.000			−0.654	−1.445
5α-dihydrotestosterone	−1.000			−1.033	−1.508
endosulfan	−2.778			−3.033	−2.610
hydroxyflutamide	−3.000			−1.104	−2.735

common hydrogen bonding site for its two phenolic hydroxy groups)[39].

For the 49 compounds for which experimental pK_i values are available the composite PCR QSAR was

$$pK_i = \underset{(\pm 0.118)}{-0.446\ \mathrm{PC}_2} + \underset{(\pm 0.125)}{0.559\ \mathrm{PC}_3} - \underset{(\pm 0.182)}{0.565\ \mathrm{PC}_4} + \underset{(\pm 0.201)}{0.293\ \mathrm{PC}_5} - \underset{(\pm 0.202)}{0.765} \qquad (8)$$

$$|t\text{-statistic}|:\ 3.77,\ 4.46,\ 3.10,\ 1.46$$

$$n = 49,\quad r^2 = 0.490,\quad q^2 = 0.330,\quad \sigma = 1.397,\quad F = 10.6$$

The equivalent 3-variable MLR QSAR was the following equation:

$$pK_i = \underset{(\pm 0.0079)}{0.0243\ V_w} + \underset{(\pm 0.00161)}{0.00410\ T_m} + \underset{(\pm 0.395)}{2.08\ \epsilon_{\mathrm{HOMO}}} + \underset{(\pm 3.684)}{14.530} \qquad (9)$$

$$\mathrm{VIF}:\ 1.0,\ 1.0,\ 1.1$$

$$|t\text{-statistic}|:\ 3.08,\ 2.54,\ 5.28$$

$$n = 49,\quad r^2 = 0.473,\quad q^2 = 0.316,\quad \sigma = 1.405,\quad F = 13.4$$

Notably, log *P* and HD are not involved in this equation.

The PCR and MLR QSARs for cluster A and subcluster C1 reliably predicted pK_i values for components of their learning sets, whereas the qualities of learning set predictions for cluster B and for the composite PCR QSAR (eq 4) were much lower. Only the QSARs for subcluster C1 (eqs 6 and 7) had acceptable predictivities ($q^2 \geq 0.5$)[41] for compounds outside the learning set. Overall, the QSARs formulated either from 1D and 2D descriptors or PCs derived therefrom will not be useful for projecting pK_i values for suspected xenoestrogens. However, it is noteworthy that the r^2 values for PCR QSARs for cluster A (eq 2) and subcluster C1 (eq 6) were significantly higher than that of the composite PCR QSAR (eq 8), implying that at least in these two instances HCA predicated on physicochemical properties can enhance QSAR quality for individual xenoestrogen (sub)clusters relative to that of an unordered composite of xenoestrogens.

Ordering the structurally diverse environmental estrogens by HCA afforded insights into the properties of ligands important for ER binding. PC_1 and PC_2 had no bearing on pK_i for the (sub)clusters. The *t*-statistics for eqs 2 and 6 indicated that PC_3 was the dominant variable for these QSARs, whereas $PC_{3,}$ PC_5, and PC_6 contributed about equally to eq 4. Certain 1D and 2D descriptors (T_m, log *P*, HD, ϵ_{HOMO}, V_w) occurred in some but not all MLR QSARs. These differences in QSAR formulation are attributable to variations in the driving forces for ligand-ER complexation among the different (sub)clusters identified using HCA. For example, T_m (which occurs in the MLR QSARs for clusters A and B) is also an important descriptor for the stability of complexes formed by organic molecules with cyclodextrins,[42] including β-cyclodextrin complexes with heterocyclic (O, S)[43] and steroidal[44−46] compounds. Our results suggest similarities between the driving forces for guest molecule-cyclodextrin complexation and xenoestrogen-ER binding.

QSAR Modeling Using 3D Descriptors. Two composite xenoestrogen 3D-QSARs were derived. The best 5-variable composite QSAR was

$$pK_i = \underset{(\pm 11.98)}{-21.64\ \pi_I} - \underset{(\pm 2.35)}{6.63\ \eta_{1p}} + \underset{(\pm 0.90)}{2.86\ \gamma_{1s}} - \underset{(\pm 0.72)}{3.38\ \gamma_{3s}} - \underset{(\pm 4.52)}{15.24\ D_v} + \underset{(\pm 1.98)}{10.85} \qquad (10a)$$

$$\mathrm{VIF}:\ 1.2\ \ 1.4,\ 1.6,\ 1.5,\ 2.2$$

$$|t\text{-statistic}|:\ 1.81,\ 2.80,\ 3.17,\ 4.70,\ 3.37$$

$$n = 49,\quad r^2 = 0.587,\quad q^2 = 0.510,\quad \sigma = 1.271,\quad F = 12.23$$

where π_I (polarizability index) is a TLSER descriptor and η_{1p}, γ_{1s}, γ_{3s}, and D_v are WHIM descriptors. The best 6-variable composite QSAR was

$$pK_i = \underset{(\pm 11.53)}{-37.71\ \pi_I} + \underset{(\pm 13.05)}{29.81\ \epsilon_B} - \underset{(\pm 4.42)}{13.29\ \eta_{1e}} - \underset{(\pm 2.24)}{5.70\ \eta_{1p}} - \underset{(\pm 2.12)}{7.77\ \theta_{2s}} - \underset{(\pm 0.57)}{1.33\ \gamma_{3s}} + \underset{(\pm 3.99)}{11.47} \qquad (10b)$$

$$\mathrm{VIF}:\ 1.2,\ 1.1,\ 1.5,\ 1.5,\ 1.2,\ 1.1$$

$$|t\text{-statistic}|:\ 3.27,\ 2.29,\ 3.01,\ 2.54,\ 3.67,\ 2.34$$

$$n = 49,\quad r^2 = 0.654,\quad q^2 = 0.463,\quad \sigma = 1.178,\quad F = 13.2$$

where ϵ_B is a TLSER descriptor and η_{1e}, η_{1p}, γ_{3s}, and θ_{2s} are WHIM descriptors. The pK_i values calculated from eq 10a are listed in Table 1.

Regressing TLSER and WHIM descriptors for each (sub)-cluster produced the following QSARs: Cluster A:

$$pK_i = \underset{(\pm 0.53)}{2.61\ V_{mc}} + \underset{(\pm 15.38)}{61.72\ \epsilon_B} + \underset{(\pm 2.06)}{10.24\ \eta_{1p}} - \underset{(\pm 0.37)}{2.55\ \gamma_{3s}} - \underset{(\pm 2.40)}{8.35\ D_v} - \underset{(\pm 3.41)}{17.48} \qquad (11)$$

$$\mathrm{VIF}:\ 4.1,\ 1.4,\ 1.7,\ 2.2,\ 2.2$$

$$|t\text{-statistic}|:\ 4.97,\ 4.01,\ 4.97,\ 6.90,\ 3.49$$

$$n = 12,\quad r^2 = 0.979,\quad q^2 = 0.906,\quad \sigma = 0.315,\quad F = 54.9$$

Cluster B:

$$pK_i = \underset{(\pm 9.51)}{37.14\ q^+} + \underset{(\pm 59.41)}{192.80\ \pi_I} + \underset{(\pm 8.92)}{42.65\ \epsilon_B} + \underset{(\pm 0.82)}{3.36\ \gamma_{1s}} - \underset{(\pm 2.49)}{4.68\ \eta_{1p}} - \underset{(\pm 11.28)}{41.89} \qquad (12)$$

$$\mathrm{VIF}:\ 4.8,\ 5.1,\ 1.5,\ 2.6,\ 2.0$$

$$|t\text{-statistic}|:\ 3.90,\ 3.25,\ 4.78,\ 4.10,\ 1.88$$

$$n = 15,\quad r^2 = 0.855,\quad q^2 = 0.656,\quad \sigma = 0.646,\quad F = 10.7$$

Subcluster C1:

$$pK_i = \underset{(\pm 0.40)}{3.70\ V_{mc}} - \underset{(\pm 7.97)}{23.56\ \pi_I} + \underset{(\pm 59.90)}{173.75\ \epsilon_B} - \underset{(\pm 1.44)}{8.17\ \eta_{1p}} - \underset{(\pm 10.44)}{33.06} \qquad (13)$$

$$\mathrm{VIF}:\ 1.6,\ 1.5,\ 2.8,\ 2.3$$

$$|t\text{-statistic}|:\ 9.14,\ 2.96,\ 2.90,\ 5.68$$

$$n = 14,\quad r^2 = 0.979,\quad q^2 = 0.956,\quad \sigma = 0.418,\quad F = 104.7$$

Table 2. Optimum Parameters (Weights) of the Neural Network Model of (5+1):(3+1):1

	hidden-layer neuron			
	1	2	3	bias
Input Layer				
π_I	21.4286	9.6451	6.9299	
η_{1p}	15.8129	1.9330	12.4913	
γ_{1s}	−5.4530	3.8136	−1.8107	
γ_{3s}	7.1666	−7.9333	1.7574	
D_v	3.7111	−5.8863	−3.5100	
bias	−18.3281	−4.5190	−4.3802	
Output Layer				
pK_i	−4.2710	4.7479	2.2116	−0.8654

Although HCA was accomplished using PCs derived independent of TLSER and WHIM descriptors, the 3D-QSARs formulated for the (sub)clusters using these 3D descriptors were of very much better statistical quality than those obtained using 1D and 2D indices, with remarkably high q^2 values for (sub)clusters A (eq 11) and C1 (eq 13) and with acceptable[41] q^2 scores for cluster B (eq 12) and the composite 5-variable QSAR (eq 10a). Of the 105 TLSER and WHIM descriptors available, only a handful were needed to formulate these high-quality 3D-QSARs for xenoestrogen binding to ER. The covalent basicity ϵ_B, representing the hydrogen bond accepting ability of ligands, occurred with a positive coefficient in each (sub)cluster 3D-QSAR and in the 6-variable composite QSAR (eq 10b), though not in the 5-variable composite QSAR (eq 10a). V_{mc} occurred with a positive coefficient in 2/3 of the (sub)cluster 3D-QSARs; for those QSARs in which V_{mc} occurs molecular bulk promotes interaction with ER. The highest partial positive atomic charge q^+ enhanced interaction of cluster B xenoestrogens with ER (eq 12). The polarizability index π_I occurred with a positive coefficient for cluster B but with a negative coefficient for subcluster C1 (eq 13) and for both composite QSARs (eqs 10a,b). Five WHIM descriptors occurred in the 3D-QSARs: η_{1p}, a directional descriptor representing inverse atomic density distribution weighted by polarizability projected along the first principal axis; η_{1e}, a directional descriptor representing inverse atomic density distribution weighted by Mulliken atomic electronegativity projected along the first principal axis; γ_{1s} and γ_{3s}, directional descriptors representing symmetry weighted by electrotopology projected along the first and third principal axes, respectively; θ_{2s}, a directional descriptor representing molecular shape weighted by electrotopology projected along the second principal axis; and D_v, a nondirectional descriptor representing total atomic density weighted by van der Waals volume. While the physicochemical meanings of these WHIM descriptors are vague, they appear to be information-rich indices which contribute significantly to the high statistical quality of the 3D-QSARs for xenoestrogen binding.

Nonlinear Modeling. Since there may be a strong nonlinear component to the relationship between xenoestrogen structure and ER binding affinity QSAR modeling was performed using ANNs. The architecture of the ANN was (5 inputs +1 bias):(3 hidden-layer nodes + 1 bias):(1 output). Connection weights between the input and hidden-layers and between the hidden and output-layers and bias elements for the hidden neurons and the output neuron are listed in Table 2. The pK_i values calculated by this model are listed in Table 1. The relevant statistical parameters of the best ANN-derived model corresponding to the composite 3D-QSAR (eq 10a), derived using the same molecular descriptors (π_I, η_{1p}, γ_{1s}, γ_{3s}, D_v), were as follows: $r^2 = 0.888$, $q^2 = 0.561$, RMSE = 0.620. The best nonlinear model had a much-improved value of r^2 compared to eq 10a, but the q^2 score, while still exceeding the threshold for acceptability,[41] was only marginally better than that achieved using linear regression.

Comparison with Other Classical QSARs for Estrogen Binding. Gao et al.[11] formulated QSARs for estradiol derivatives and nonsteroidal compounds using MLR. They obtained separate models for training sets comprised of congeneric molecules (16α-substituted estradiols, 11β-substituted estradiols, metahexestrol derivatives, etc.) using a variety of 1D and 2D descriptors, plus some steric or bulk indices such as Taft's E_s parameter, molar volume calculated by the method of McGowan,[47] and STERIMOL substituent parameters L, B_1, and B_5. In contrast this, we examined a much more diverse assortment of estrogenic species using a much more diverse assortment of descriptors, including quantum mechanical descriptors and rototranslational invariant 3D parameters. The resulting 3D-QSARs were of good to exceptional quality and should be of great value for assessing the estrogenic potential of chemicals.

CONCLUSIONS

Coherent groupings of structurally diverse environmental estrogens were obtained by hierarchical cluster analysis in conjunction with principal component analysis. Variations in QSAR formulation between (sub)clusters indicated that ligand−receptor interactions responsible for xenoestrogen-ER binding are not uniform throughout the data set. Xenoestrogen groupings (Figure 4) should prove useful for designing experiments to evaluate combinations of environmental estrogens,[48] since for a set of 100 suspected environmental estrogens there are 4950 unique dyadic combinations. The observed effect of pairs of agents (additivity, synergy, antagonism) is expected to reflect to some extent the mechanism of binding of each component.[49] The approach employed in this work provides important insights into the similarities and differences of ligand-ER interactions for estrogenic substances.

ACKNOWLEDGMENT

This work was supported by the Nippon Life Insurance Foundation, Osaka, Japan.

Supporting Information Available: Tables of values for molecular descriptors for the 60 xenoestrogens (Figure 1) and correlation coefficients for pairs of descriptors. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Colborn, T. Environmental estrogens: health implications for humans and wildlife. *Environ. Health Perspect.* **1995**, *103*(*Suppl. 7*), 135−136.

(2) Sonnenschein, C.; Soto, A. M. An updated review of environmental estrogen and androgen mimics and antagonists. *J. Steroid Biochem. Mol. Biol.* **1998**, *65*, 143−150.

(3) Wiseman, H. Importance of oestrogen, xenoestrogen and phytoestrogen metabolism in breast cancer risk. *Biochem. Soc. Trans.* **1999**, *27*, 299−304.

(4) Waller, C. L.; Oprea, T. I.; Chae, K.; Park, H.-K.; Korach, K. S.; Laws, S. C.; Wiese, T. E.; Kelce, W. R.; Gray, L. E., Jr. Ligand-

based identification of environmental estrogens. *Chem. Res. Toxicol.* **1996**, *9*, 1240−1248.

(5) Keith, L. H. *Environmental Endocrine Disruptors: A Handbook of Property Data;* John Wiley & Sons: New York, 1997.

(6) Tong, W.; Perkins, R.; Xing, L.; Welsh, W. J.; Sheehan, D. M. QSAR models for binding of estrogenic compounds to estrogen receptor α and β subtypes. *Endocrinology* **1997**, *138*, 4022−4025.

(7) Bradbury, S. P.; Mekenyan, O. G.; Ankley, G. T. Quantitative structure−activity relationships for polychlorinated hydroxybiphenyl estrogen receptor binding affinity: an assessment of conformer flexibility? *Environ. Toxicol. Chem.* **1996**, *15*, 1945−1954.

(8) Gantchev, T. G.; Ali, H.; van Lier, J. E. Quantitative structure−activity relationships/ comparative molecular field analysis (QSAR/CoMFA) for receptor-binding properties of halogenated estradiol derivatives. *J. Med. Chem.* **1994**, *37*, 4164−4176.

(9) Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Yu; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Sheehan, D. M. Evaluation of quantitative structure−activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 669−677.

(10) Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary quantitative structure−activity relationship (QSAR) analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164−168.

(11) Gao, H.; Katzenellenbogen, J. A.; Garg, R.; Hansch, C. Comparative QSAR analysis of estrogen receptor ligands. *Chem. Rev.* **1999**, *99*, 723−744.

(12) Strategic Program on Environmental Endocrine Disruptors '98, Japan Environment Agency; May 1998; available at http://www.eic.or.jp/eanet/e/end/sp98.html.

(13) Karelson, M. *Molecular Descriptors in QSAR/QSPR;* Wiley-Interscience: New York, 2000.

(14) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(15) Kier, L. B., Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Letchworth, Herts, U.K., 1986.

(16) Ajay, W.; Walters, P.; Murcko, M. A. Can we learn to distinguish between "drug-like"and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(17) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(18) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(19) Baker, J. An algorithm for the location of transition states. *J. Comput. Chem.* **1986**, *7*, 385−395.

(20) Sangster, J. *LOGKOW − A Databank of Evaluated Octanol−Water Partition Coefficients;* Sangster Research Laboratories: Montréal, Canada, 1993.

(21) Suzuki, T. Development of an automatic estimation system for both the partition coefficient and aqueous solubility. *J. Comput.-Aid. Mol. Des.* **1991**, *5*, 149−166.

(22) Simamora, P.; Miller, A. H.; Yalkowsky, S. H. Melting point and normal boiling point correlations: applications to rigid aromatic compounds. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 437−440.

(23) Simamora, P.; Yalkowsky, S. H. Group contribution methods for predicting the melting point and boiling point of aromatic compounds. *Ind. Eng. Chem. Res.* **1994**, *33*, 1405−1409.

(24) Krzyzaniak, J. F.; Myrdal, P. B.; Simamora, P.; Yalkowsky, S. H. Boiling point and melting point prediction for aliphatic, non-hydrogen bonding compounds. *Ind. Eng. Chem. Res.* **1995**, *34*, 2530−2535.

(25) Joback, K. G.; Reid, R. C. Estimation of pure-component properties from group contributions. *Chem. Eng. Comm.* **1987**, *57*, 233−243.

(26) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Statist. Assoc.* **1963**, *58*, 236−244.

(27) Verloop A.; Hoogenstraaten, W.; Tipker, J. Development and application of new steric substituent parameters in drug design. *Drug Design* **1976**, *7*, 165−207.

(28) Shapiro, S.; Guggenheim, B. Inhibition of oral bacteria by phenolic compounds. Part 2. Correlations with molecular descriptors. *Quant. Struct.-Act. Relat.* **1998**, *17*, 338−347.

(29) Cramer, R., III.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(30) Kearsley, S. K.; Smith, G. M. An alternative method for the alignment of molecular structures: maximizing electrostatic and steric overlap. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615−633.

(31) Todeschini, R.; Gramatica, P. The WHIM theory: new 3D-molecular descriptors for QSAR in environmental modelling. *SAR QSAR Environ. Res.* **1997**, *7*, 85−115.

(32) Todeschini, R.; Gramatica, P. New 3D molecular descriptors: the WHIM theory and QSAR applications. *Perspect. Drug Discov. Des.* **1998**, *9−11*, 355−380.

(33) Famini, G. R.; Wilson, L. Y. Using theoretical descriptors in linear solvation energy relationships. *Theor. Comput. Chem.* **1994**, *1*, 213−241.

(34) The adopted cutoff $r > 0.9$ corresponds to the upper limit of the recommended VIF range ($1 < \mathrm{VIF} < 5$) as a measure of cross correlation.[33]

(35) Katrizky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.

(36) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design: An Introduction;* 2nd ed., Wiley-VCH: Weinheim, 1999.

(37) Suzuki, T.; Ebert, R.-U.; Schüürmann, G. Development of both linear and nonlinear methods to predict the liquid viscosity at 20° C of organic compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1122−1128.

(38) Cramer III, R. D. BC(DEF) Parameters. 1. The intrinsic dimensionality of intermolecularinteractions in the liquid state. *J. Am. Chem. Soc.* **1980**, *102*, 1837−1849.

(39) Anstead, G. M.; Carlson, K. E.; Katzenellenbogen, A. The estradiol pharmacophore: ligand structure-estrogen receptor binding affinity relationships and a model for the receptor binding site. *Steroids* **1997**, *62*, 268−303.

(40) *Concepts and Applications of Molecular Similarity*; Johnson, M., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.

(41) Oprea, T. I.; Ho, C. M. W.; Marshall, G. R. *De novo* design. Ligand construction and prediction of affinity. In *Computer-Aided Molecular Design. Applications in Agrochemicals, Materials, and Pharmaceuticals;* Reyonolds, C. H., Holloway, M. K., Cox, H. K., Eds.; ACS Symposium Series 589; American Chemical Society: Washington, DC, 1995; pp 64−81.

(42) Szejtli, J. *Cyclodextrin Technology*; Kluwer Academic Publishers: Dordrecht, NL, 1988.

(43) Carpignano, R.; Marzona, M.; Cattaneo, E.; Quaranta, S. QSAR study of inclusion complexes of heterocyclic compounds with β-cyclodextrin. *Anal. Chim. Acta* **1997**, *348*, 489−493.

(44) Hamasaki, K.; Ueno, A.; Toda, F.; Suzuki, I.; Osa, T. Molecular recognition indicators of modified cyclodextrins using twisted intramolecular charge-transfer fluorescence. *Bull. Chem. Soc. Jpn.* **1994**, *67*, 516−523.

(45) Krismundsdóttir, T.; Loftsson, T.; Holbrook, W. P. Formulation and clinical evaluation ofa hydrocortisone solution for the treatment of oral disease. *Int. J. Pharm.* **1996**, *139*, 63−68.

(46) Cserháti, T.;, Forgács, E. Inclusion complex formation of steroidal drugs withhydroxypropyl-β-cyclodextrin studied by charge-transfer chromatography. *J. Pharm. Biomed. Anal.* **1998**, *18*, 179−185.

(47) Abraham, M. H.; McGowan, J. C. The use of characteristic volumes to measure cavity terms in reversed-phase liquid chromatography. *Chromatographia* **1987**, *23*, 243−246.

(48) Suzuki, T.; Ide, K.; Ishida, M. Evaluation of synergistic effects of environmental oestrogens. *J. Pharm. Pharmacol.* **1999**, *51*(*Suppl.*), 142.

(49) Berenbaum, M. C. What is synergy? *Pharmacol. Rev.* **1989**, *41*, 93−141.

CI000333F