

# $^{13}\text{C}$ NMR and Electron Ionization Mass Spectrometric Data-Activity Relationship Model of Estrogen Receptor Binding

Richard D. Beger,<sup>1</sup> James P. Freeman, Jackson O. Lay, Jr., Jon G. Wilkes, and Dwight W. Miller

Division of Chemistry, National Center for Toxicological Research, Food and Drug Administration, Jefferson, Arizona 72079

Received June 21, 2000; accepted August 17, 2000

**$^{13}\text{C}$  NMR and Electron Ionization Mass Spectrometric Data-Activity Relationship Model of Estrogen Receptor Binding. Beger, R. D., Freeman, J. P., Lay, Jr., J. O., Wilkes, J. G., and Miller, D. W. (2000). *Toxicol. Appl. Pharmacol.* 169, 17–25.**

Two Spectroscopic Data-Activity Relationship (SDAR) models based on  $^{13}\text{C}$  nuclear magnetic resonance (NMR) and electron ionization mass spectra (EI MS) data were developed for 108 compounds whose relative binding affinities (RBA) to the estrogen receptor are known. The  $^{13}\text{C}$  NMR and EI MS data were used as spectrometric digital fingerprints to reflect the electronic and structural characteristics of the compounds. Both SDAR models segregated the 108 compounds into 20 strong, 15 medium, and 73 weak relative binding classifications. The first SDAR model, based on  $^{13}\text{C}$  NMR data alone, gave a leave-one-out (LOO) cross-validation of 75.0%. The second SDAR model, based on a composite of  $^{13}\text{C}$  NMR and EI MS data, gave a LOO cross-validation of 82.4%. Many of the misidentifications from the cross-validations were between medium and weak classifications, where there were fewer specific spectrometric characteristics to identify the relationship of spectra to estrogen receptor binding. Real and predicted  $^{13}\text{C}$  NMR chemical shifts were used to test the predictive behavior of both SDAR models. The ease of use and speed of SDAR modeling may facilitate their use with other toxicological endpoints.

$^{13}\text{C}$  NMR chemical shifts have been used to predict and refine chemical structures (Beger and Bolton, 1997; Wishart and Sykes, 1994). Conversely, the chemical structure of a compound has been used to predict its  $^{13}\text{C}$  NMR chemical shifts (Kvasnicka, 1991). The  $^{13}\text{C}$  NMR spectrum of a compound contains a pattern of frequencies that correspond directly to the quantum mechanical properties of the carbon nuclei in the molecule and which reflect the proximity and connectivity of nearby atoms. The quantum mechanical description of a molecule depends largely on its electrostatic features and three-dimensional geometry (Emsley *et al.*, 1965). *Ab initio* quantum mechanical calculations of  $^{13}\text{C}$  chemical shift tensors in proteins reveal that they are dependent on the

structural environment and electrostatic potential (De Dios *et al.*, 1993).

QSAR (quantitative structure-activity relationship) modeling results show that receptor binding of a compound can be predicted, based in part upon electrostatics and geometrical structure (Hansch and Leo, 1995; Branbury, 1995; Tong *et al.*, 1997). The binding activity of 45 progestagens has been quantitatively modeled with individual molecular  $^1\text{H}$ -NMR, infrared, and mass spectra, as well as with simulated infrared (IR) spectra and  $^{13}\text{C}$  NMR spectra by comparative spectral analysis (CoSA) (Bursi *et al.*, 1999). The CoSA model produced results that yielded better correlations and predictions than were seen with comparative molecular field analysis (CoMFA), but the CoSA quantitative modeling was limited to a set of structurally similar compounds.

SDAR (spectrometric data-activity relationship) models use  $^{13}\text{C}$  NMR chemical shifts and electron ionization mass spectra (EI MS) patterns in much the same way that QSAR uses constitutional, topological, geometric, electrostatic, and quantum descriptors (Katritzky *et al.*, 1996, 1994; Fujita *et al.*, 1964; Cramer *et al.*, 1988b; Tong *et al.*, 1995; Collantes *et al.*, 1996) to model receptor binding of a compound. SDAR models eliminate the need to calculate electrostatics or quantum mechanical descriptors. The molecular alignment limitation of QSAR is avoided in SDAR modeling because the spectrometric data are independent of assumptions about spatial orientation of the free molecules and contain the structural and quantum mechanical information about the compound (Bursi *et al.*, 1999). The spectra can function as comparative “spectrometric digital fingerprints” for each compound, both in relation to the others and in relation to a biological endpoint, such as estrogen-receptor binding. We will demonstrate that covariance analysis of patterns in  $^{13}\text{C}$  NMR spectroscopic data can be used to predict the intensity of a compound’s interaction with a binding site. Covariance analysis is routinely done on large genetic databases to determine sequence and genetic similarity (Altschul *et al.*, 1990; Zhang and Madden, 1997; Boguski and Schuler, 1995; Julier *et al.*, 1994). We propose that such modeling based on spectroscopic data-activity relationships be called SDAR.

The frequencies obtained from the  $^{13}\text{C}$  NMR spectroscopic

<sup>1</sup> To whom correspondence should be addressed. Fax: (870) 543-7686. E-mail: rbeger@NCTR.FDA.GOV.

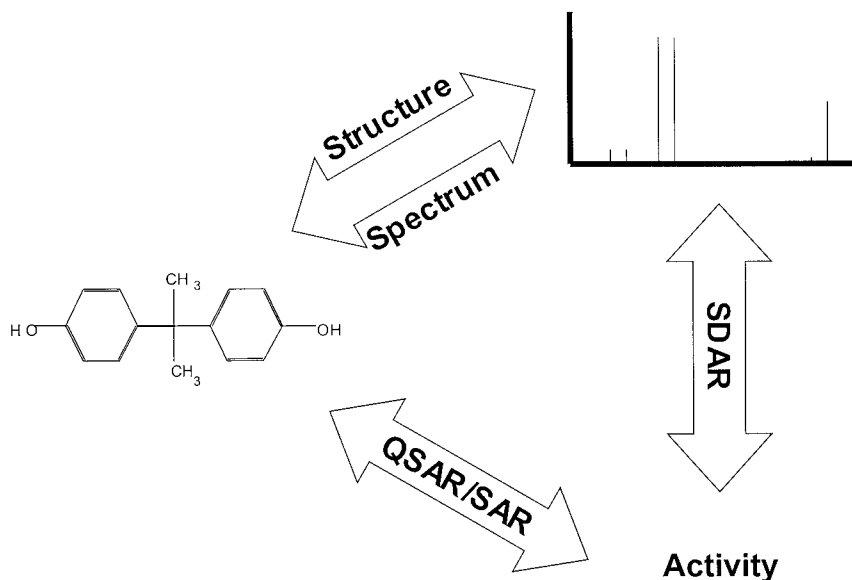


FIG. 1. The SDAR concept.

data correspond directly to the energies obtained when solving the quantum mechanical Schrödinger equation for a nuclear magnetic moment transition (Emsley *et al.*, 1965). The NMR quantum energies are strongly dependent on the electrostatic potential energy of the carbon nucleus and the type of orbital (wavefunction) surrounding the carbon nucleus. The wavefunction surrounding the atoms in a molecule can be correlated to the lowest unoccupied molecular orbital (LUMO) and highest occupied molecular orbital (HOMO) quantum states (Kollman *et al.*, 1973). Typically,  $^{13}\text{C}$  NMR chemical shifts in the 0 to 100 ppm range are associated with carbon atoms that have  $\text{sp}^3$  orbitals with the more upfield shifts having a positive electrostatic potential (like methyl groups) and the downfield shifts having a more negative electrostatic potential (like ester bonds). Likewise,  $^{13}\text{C}$  NMR chemical shifts in the 100 to 220 ppm range are associated with carbon atoms that have  $\text{sp}^2$  and  $\text{sp}$  orbitals with the more upfield shifts having a positive electrostatic potential (like benzyl groups) and the downfield shifts having a more negative electrostatic potential (like carbonyl groups). The effects of substituents on  $^{13}\text{C}$  NMR chemical shifts can be felt from as far as five bonds away or through space directly. The absolute energies in NMR spectra are not used in SDAR because the NMR spectra are given as parts per million (ppm) chemical shifts that are dimensionless numbers defined with respect to a reference compound. (When  $^{13}\text{C}$  NMR chemical shifts are given in hertz (Hz), they can represent energies.)

EI MS data represent a mass-size description of molecular substructures and often the whole molecule. By combining the  $^{13}\text{C}$  NMR and EI mass spectral data into a composite set of molecular descriptors and putting them into pattern recognition programs, it should be possible to produce predictive SDAR

models in a way conceptually analogous to the process used in many QSAR models.

Figure 1 shows the relationship of structure, SDAR, and QSAR/SAR modeling techniques. The top arrow shows how a structure is characterized through NMR, IR, and MS experiments. Spectrometric data experiments can then be used to refine and predict a structure. NMR spectrometric data are able to distinguish between stereo-isomers, as demonstrated by the 3 ppm deviation observed in  $^{13}\text{C}$  NMR chemical shifts of carbon 17 between  $17\beta$ -estradiol and  $17\alpha$ -estradiol. The NMR chemical shifts of the six carbons in a phenol or benzyl compound all change when a substitution is made in the 2, 3, or 4 position. The direction and magnitude of the six NMR chemical shift changes are dependent on the substitution in the benzyl or phenol ring. Neural networks have been developed that can accurately identify the presence of a broad range of structural features present in a compound with the network trained on IR and  $^{13}\text{C}$  NMR data (Munk *et al.*, 1996). Neural networks have also been developed that can accurately identify 26 structural features with the network trained on IR and EI MS data (Klawun and Wilkins, 1996). Similar to  $^{13}\text{C}$  NMR spectral data, the chemical structure of a compound has been used to predict its IR spectrum (Gasteiger *et al.*, 1997). The bottom arrow shows the lock and key relation between structure and activity, the relationship exploited in QSAR and SAR (Klopman, 1984, 1992) modeling techniques. The arrow between the spectrum and activity is what we refer to as SDAR. SDAR removes the problems associated with structure alignment and structural calculations. However, the one-dimensional data used in SDAR modeling loses direct three-dimensional structure-specific information and this information lost can lead to false negatives and positives predictions.

This SDAR model can be rapidly derived and instantly consulted on an ordinary personal computer. It gives a straightforward classification for the most active compounds with estrogen binding activity. Moreover, it requires only experimental data (<sup>13</sup>C NMR and EI mass spectra) that are readily obtainable. The SDAR model based on experimental spectroscopic data can be used to screen a large number of compounds to identify those few likely by spectral similarity to show estrogenic activity. These “hits” would have their estrogenic relative binding affinities determined by expensive *in vitro* binding assays. Once an SDAR model has been developed using training with real data, it should be possible to predict whether or not other compounds bind strongly to the estrogen receptor as quickly as their <sup>13</sup>C NMR and EI mass spectra can be obtained and compared in the SDAR model.

Since <sup>13</sup>C NMR spectra require either a substantial amount of purified compound or a long acquisition time, we hoped to demonstrate that a compound’s predicted <sup>13</sup>C NMR spectrum could be used in lieu of its instrumentally determined <sup>13</sup>C NMR spectrum without compromising the accuracy of the SDAR model. EI MS data can be obtained fairly quickly with small samples or else obtained from many online computer sources (SDBS, 2000; NIST, 1998).

An endocrine disrupting chemical (EDC) is defined as “an exogenous agent that interferes with the production, release, transport, metabolism, binding, action, or elimination of natural hormones in the body responsible for the maintenance of homeostasis and the regulation of developmental processes.” Estrogenic compounds represent a significant subset of the EDCs to be tested. Over 86,000 compounds are candidates to be screened for their estrogen receptor binding level, and the number of compounds to be screened is growing every day. There is a growing need to develop inexpensive and rapid methods to screen these compounds. The development of SDAR modeling gives a fast and straightforward classification for the most active compounds. The use of spectrometric data would allow SDAR modeling approaches to be used in predictive toxicology.

## METHODS

The estrogenic relative binding affinities (RBAs) of 108 compounds (Table 1) were derived from previous publications (Kuiper *et al.*, 1997; Blair *et al.*, 2000; Zava and Duwe, 1997; Hopert *et al.*, 1998). Most of the <sup>13</sup>C NMR spectrometric and EI mass spectrometric data are in the Integrated Spectral Data Base System for Organic Compounds web site [www.aist.go.jp/RIODB/SDBS/](http://www.aist.go.jp/RIODB/SDBS/) (SDBS, 2000), the NIST MS database software version 1.6 (NIST, 1998), the *Aldrich Library of <sup>13</sup>C and <sup>1</sup>H FT NMR Spectra* (Pouchert and Behnke, 1993), *Spectral Data of Steroids* (Frenkel and Marsh, 1994), and ACD/Labs CNMR software version 4.0 (ACD/Labs, 2000).

The <sup>13</sup>C-NMR spectral analyses of 4-hydroxy-estradiol, ICI 164-384, moxestrol, norethynodrel, clomiphene, coumestrol, daidzein, nafoxidine, naringenin, and genistein were performed at 75.46 MHz on a Varian Gemini 300 MHz NMR (Varian Associates, Palo Alto, CA) spectrometer operating at 301 K. Compounds were dissolved in CDCl<sub>3</sub> or DMSO. The chemical shifts were defined by assigning the CDCl<sub>3</sub> peak to 77.0 ppm and the DMSO peak to 39.5 ppm, depending on which solvent is actually used.

The samples were also analyzed by direct exposure probe (DEP) mass spectrometry (MS). The mass spectrometers were operated in the electron ionization (EI) mode, with 70-V electron energy and the ion source temperature at 150°C. Samples, in solution, were applied to the rhenium wire of the DEP and the solvent was allowed to evaporate before the analysis was begun. MS data were collected until the current of the DEP exceeded 500 mA. Norethynodrel, ICI 164-384, 4-hydroxytamoxifen, 4-hydroxy-estradiol, clomiphene, daidzein, and genistein were analyzed on a Finnigan TSQ 700 and moxestrol was analyzed on a Finnigan 4500 (Finnigan Corp., San Jose, CA) mass spectrometer.

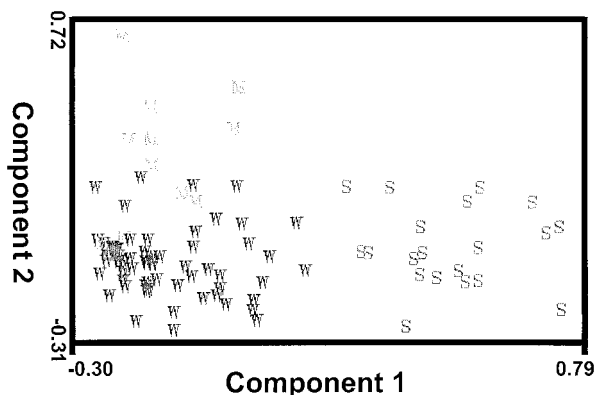
In the SDAR model EI MS data from *m/z* 100 to 549 were used and anything below *m/z* 100 was omitted. Factor analysis of preliminary SDAR models that included the *m/z* 50 to 100 range revealed that the *m/z* 50 to 100 did not contain much information important to this model of estrogen receptor binding. Unassigned 1D <sup>13</sup>C NMR chemical shifts were segregated into bins over a 0 to 222 ppm range. The <sup>13</sup>C NMR frequencies were shifted to bins 550 to 770, so bin 550 was the <sup>13</sup>C NMR spectrum for frequencies inside 0 to 1 ppm and bin 770 was the <sup>13</sup>C NMR spectrum for frequencies inside 220 to 221 ppm. The <sup>13</sup>C NMR spectra were saved as the area under the peak. The area under a specific chemical shift frequency was first normalized to an integer. A non-degenerate frequency was assigned an area of 25, a doubly degenerate frequency (two <sup>13</sup>C NMR chemical shifts at the same frequency) had an area of 50, etc. This was done so that all the spectra would have similar signal-to-noise ratios and to eliminate linewidth variations due to differences in NMR instrumental field strengths, shimming, temperature, pH, or solvents. The width of the <sup>13</sup>C NMR spectral bin was set to 1.0 ppm.

The relative binding affinities (RBAs) of a molecule to the estrogen receptor is defined as the ratio of the molar concentrations of 17β-estradiol and the competing compound required to decrease the receptor-bound compound by 50% then multiplied by 100. Thus 17β-estradiol had an RBA of 100 and a log RBA of 2.0. Strong binders to the estrogen receptor were classified as those with a log RBA over -0.30; weak binders were classified as those with a log RBA less than or equal to -2.70, and medium binders were everything between -2.70 and -0.30. The classification boundaries were set by trial and error and cluster analysis. There were 20 strong binders, 15 medium binders, and 73 weak binders in the training set.

The analysis of the SDAR model was done by the leave-one-out (LOO) cross-validation procedure, in which each compound is systematically excluded from the training set and its relative binding activity class predicted by an SDAR model generated without its contribution (Cramer *et al.*, 1988a). The pattern recognition software used was Resolve Version 1.2 (Colorado School of Mines, Golden, CO). The <sup>13</sup>C-NMR and EI MS spectroscopic data for all 108 compounds were input into the software. The spectroscopic data were then autoscaled and Fisher-weighted prior to principal component analysis. The discriminant analysis was based on the canonical variate vector and LOO cross-validation was used for each compound to maximize the size of the training set.

Autoscaling compares the quantitative response at each mass spectral *m/z* or NMR chemical shift bin to all the others in the comparison set. An average value with standard deviation is calculated for each *m/z* or bin. Then, for each spectrum, the quantitative response at an *m/z* or bin is expressed as the number of standard deviations above or below the average. This data pretreatment step equalizes the weight of consistent variance from signals with inherently small magnitudes (25 units for NMR bin 558 representing a single methyl carbon) to those signals with large magnitudes (130,000 area counts at *m/z* bin 91, probably arising from a tropylium fragment ion). Autoscaling, which automatically compensates for gross magnitude variations, is particularly important for this application in which two completely different types of analytical spectra are formed into a composite surrogate representative of important molecular characteristics affecting estrogen binding.

Fisher weighting treats data in a statistical way that emphasizes those spectral characteristics important in distinguishing defined groups. In this SDAR model, the three groups are strong, medium, or weak estrogen receptor binders. For each mass spectral *m/z* or NMR bin, the variance between groups



**FIG. 2.** The discriminant function using  $^{13}\text{C}$  NMR data in the SDAR model. The X-axis is the first principal component and the Y-axis is the second principal component. Compounds shown with an S have a strong classification; compounds shown with an M have a medium classification, and compounds shown with a W have a weak classification.

is divided by the variance within groups. The dividend is a weighting factor with a magnitude larger than one when a particular  $m/z$  or NMR bin has an important role in distinguishing groups. Fisher weighting all of the raw spectra before pattern recognition increases the power of discriminant analysis to classify spectra correctly. It is particularly important in this application because it deemphasizes the relative importance of irrelevant spectral information.

The number of principal components used in the SDAR model was determined by a systematic search over the number of principal components from one to 30 and the corresponding LOO cross-validation. Typically the LOO cross-validation of the SDAR model rises linearly when 1 to 15 principal components are used and then the LOO cross validation oscillates by 5 to 10% when going from 15 to 22 or 25 principal components are used in the SDAR model. When the number of principal components reaches 25 to 30, a slow steady decline is seen in the LOO cross-validation of the model. We selected the number of principal components that gave the best LOO cross-validation for the SDAR model.

The ACD CNMR predictor 4.0 program was used to predict the  $^{13}\text{C}$  NMR spectra of 2-ethylphenol, 3-deoxyestradiol, 3-methylestriol, dimethylstilbestrol, and 4,4'-dihydroxystilbene (ACD/Labs, 2000). The predicted NMR spectra are calculated by the substructure technique HOSE (Bremser, 1978).

## RESULTS

Based only on  $^{13}\text{C}$  NMR spectroscopic data, the statistical pattern recognition program with 22 principal components (PCs) used 89.8% of the total variance and had a cross-validation of 75.0%. The first principal component had 32.5% of the cumulative variance. Figure 2 is a two-dimensional display of the discriminant function from the 22-dimensional  $^{13}\text{C}$  NMR data SDAR model. Compounds shown with an S had a  $\log(\text{RBA}) > -0.30$ ; compounds shown with a W had a  $\log(\text{RBA}) < -2.70$ , and compounds shown with an M had intermediate values. The strong, medium, and weak classification cutoff positions were optimized by trial and error. Figure 2 shows that the 20 strong binders that are estrogens and synthetic estrogens had a positive component 1 and were separated from the 88 medium and weak binders. Most of the medium binders are phytoestrogens and androgens that clustered to-

gether with a positive canonical variate component 2 at the top of Fig. 2. Table 1 contains the compound name, the  $\log(\text{RBA})$  input (where N stands for "not determined" because its RBA to the estrogen receptor was weak or did not bind to the estrogen receptor), the SDAR training input, the SDAR prediction using  $^{13}\text{C}$  NMR data, and the SDAR prediction using  $^{13}\text{C}$  NMR and EI MS data.

Figure 3 shows the factor loadings associated with the first canonical variate function for the pattern recognition of the  $^{13}\text{C}$  NMR data. The positive peaks in Fig. 3 correspond to bins that bias toward a strong classification for binding to the estrogen receptor and negative peaks correspond to bins that bias toward a weak and medium classification. The aliphatic  $\text{CH}_2$  bins, 30 to 35 ppm, showed a bias toward weak classification. The methyl  $\text{CH}_3$  bins 8 and 16 ppm, respectively, showed a bias toward strong classification. Many of the aromatic bins, 115 to 150 ppm, showed a bias toward strong classification.

Based on the composite  $^{13}\text{C}$  NMR and EI MS data, the statistical pattern recognition program with 21 principal components (PCs) included 80.4% of the total variance and had a cross validation of 82.4%. The first principal component had 27.2% of the cumulative variance. Figure 4 is a two-dimensional display of the discriminant function from the 21-dimensional  $^{13}\text{C}$  NMR and EI MS SDAR model. Figure 4 shows that the 20 strong binders (estrogens and synthetic estrogens) that had a positive component 1 were well separated from the 88 medium and weak binders. Most of the medium binders are phytoestrogens and androstrogens that clustered together with a positive canonical variate component 2 at the top of Fig. 4.

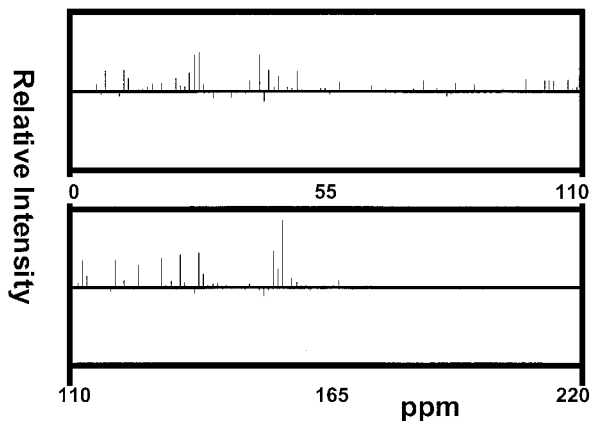
Figure 5 shows the factor loadings for the first canonical variate for the SDAR model based on  $^{13}\text{C}$  NMR (Fig. 5a) and EI MS (Fig. 5b) spectra. The positive peaks in Figs. 5a and 5b correspond to bins that bias toward a strong classification and negative peaks correspond to bins that bias toward a medium or weak classification. Many of the canonical variate  $^{13}\text{C}$  NMR bins that showed a strong classification bias in Fig. 3 are present in Fig. 5a. In Fig. 5b, the mass spectral portion of the canonical variate is split evenly into bins that bias a strong classification and a not-strong classification.

Figure 6 shows the discriminant function of the  $^{13}\text{C}$  NMR SDAR model shown in Fig. 2 with the predicted estrogen receptor binding of 2-ethylphenol displayed with an **A** when using the real experimental  $^{13}\text{C}$  NMR data and **a** when using the predicted  $^{13}\text{C}$  NMR spectra. Similarly, the predicted estrogen receptor binding of 3-deoxyestradiol is shown at **B** when using the real  $^{13}\text{C}$  NMR data and at **b** when using the predicted  $^{13}\text{C}$  NMR spectra. The predicted estrogen receptor binding of 3-methylestriol is shown at **D** when using the real experimental  $^{13}\text{C}$  NMR data and at **d** when using predicted  $^{13}\text{C}$  NMR spectra. The predicted estrogen receptor binding of dimethylstilbestrol is shown at **E** when using the real experimental  $^{13}\text{C}$  NMR data and at **e** when using predicted  $^{13}\text{C}$  NMR spectra. The predicted estrogen receptor binding of 4,4'-dihydroxystilbene is shown at **F** when using the real experimental  $^{13}\text{C}$  NMR data and at **f**

**TABLE 1**  
**Model Training Compounds**

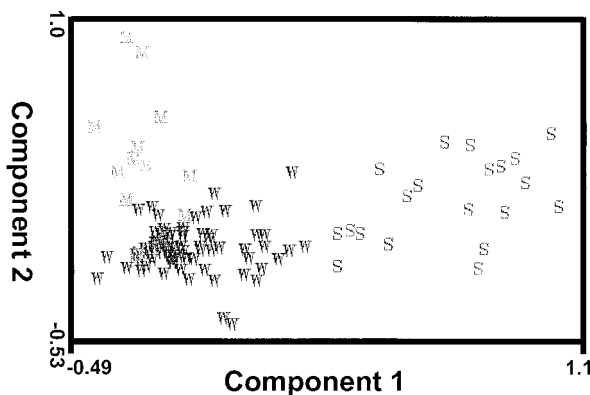
Compound	log (RBA)	Class	Class predicted NMR	Class predicted NMR/MS	Compound	log (RBA)	Class	Class predicted NMR	Class predicted NMR/MS
1,4-Diphenyl-1,3-butadiene	N	W	W	W	DDD-o,p'	N	W	W	W
1,6-Dimethylnaphthalene	N	W	W	W	DDD-p,p'	N	W	W	W
1,8-Octanediol	N	W	M	M	DDE-p,p'	N	W	S	S
2,2'-Dihydroxy-4-methoxybenzophene	N	W	W	W	DDT-o,p'	-2.85	W	W	W
2,2'-Methylenebis(4-chlorophenol)	-2.45	M	M	M	DDT-p,p'	N	W	W	W
2,4,5-T	N	W	W	W	Diethylstilbesterol	2.60	S	S	S
2,4-D	N	W	W	W	Daidzein (28)	-1.65	M	M	M
2,4-Dihydroxybenzophenone	-2.61	M	W	M	Dexamethasone	N	W	W	W
2-Chloro-4-methylphenol	-3.66	W	W	W	Dibenzo-18-crown-6	N	W	W	W
2-Chlorophenol	-3.67	W	W	W	Dibutyl phthalate	N	W	W	W
2-Furaldehyde	N	W	W	W	Dieldrin	N	W	S	S
2-Hydroxy-estradiol (25)	1.04	S	S	S	Dienestrol	1.57	S	M	S
2-Phenylphenol	N	W	W	W	Diethyl phthalate	N	W	W	W
2-sec-Butylphenol	-3.54	W	M	W	Dihydrotestosterone	N	W	M	W
3-Phenylphenol	-3.44	W	M	W	Diisobutyl phthalate	N	W	W	W
4,4'-Dihydroxybenzophenone	-2.46	M	W	W	Dimethyl phthalate	N	W	W	W
4,4'-Methylenebis(N,N-dimethyl)	N	W	W	S	Diphenolic acid	-3.13	W	S	S
4,4'-Methylenedianiline	N	W	W	W	Dopamine	N	W	W	W
4,4'-Methylene(2,6-ditertbutylphenol)	N	W	W	W	Estra-1,3,5(10)-trien-3-ol	1.14	S	S	W
4,4'-Sulfonylphenol	-3.07	W	W	W	Estra-1,3,5(10)-trien-3,6 $\alpha$ ,17 $\beta$ -triol	-0.15	S	S	S
4-(Benzyloxy)phenol	-3.44	W	W	W	Estriol	0.99	S	S	S
4-Chloro-2-methylphenol	-3.67	W	W	W	Estrone	0.86	S	S	S
4-Chloro-3-methylphenol	-3.38	W	M	W	Ethylcinnamate	N	W	W	W
4-Ethylphenol	-4.17	W	W	W	Ethynyl estradiol	2.28	S	S	S
4-Nonylphenol	-1.45	M	M	W	Etiocholan-17 $\beta$ -ol-3-one	N	W	W	W
4-Hydroxyestradiol (25)	0.85	S	S	S	Eugenol	N	W	W	W
4-Hydroxytamoxifen	2.24	S	S	W	Genistein (27, 28)	-0.35	M	M	M
4-Phenylphenol	-3.04	W	W	W	Heptanal	N	W	W	W
4-Stilbenol	N	W	S	S	Hesperetin (27)	N	W	M	M
4-tert-Amylphenol	-3.26	W	W	W	Hexachlorobenzene	N	W	W	W
4-tert-Butylphenol	-3.61	W	W	W	Hexestrol	0.56	S	W	S
4-tert-Octylphenol	-1.82	M	M	W	Hexyl alcohol	N	W	W	W
5 $\alpha$ -Androstane-3 $\alpha$ ,17 $\beta$ -diol	-2.67	M	W	M	ICI-164,384	1.16	S	W	S
5 $\alpha$ -Androstane-3 $\beta$ ,17 $\beta$ -diol	-0.92	M	W	M	Isoeugenol	N	W	W	W
Aldrin	N	W	W	M	Kaempferol (27)	-1.55	M	M	M
Aurin	-1.49	M	M	M	Lindane	N	W	W	S
Benzyl alcohol	N	W	W	W	Melatonin	N	W	M	M
Benzylbutyl phthalate	N	W	W	W	Mestranol	0.35	S	S	S
Bis(2-ethylhexyl) phthalate	N	W	W	S	Methoxychlor	-3.2	W	W	W
Bis(2-hydroxyphenyl)-methane	N	W	W	W	Moxestrol	1.14	S	S	S
Bis(4-hydroxyphenyl)-methane	-3.02	W	W	W	Nafoxidine	-0.14	S	M	S
Bisphenol A	-2.11	M	M	M	Norethynodrel	-0.65	M	W	M
Bisphenol B	-1.07	M	S	M	Phenolphthalein	-1.87	M	W	W
Butyl-4-aminobenzoate	N	W	W	W	Phenol red	-3.25	W	W	W
n-Butylbenzene	N	W	W	M	Progesterone	N	W	W	W
Caffeine	N	W	S	W	Quercetin (27)	N	W	W	M
Cholesterol	N	W	W	M	Suberic acid	N	W	W	W
Chrysene	N	W	W	W	Tamoxifen	0.21	S	S	S
Chrysin	N	W	M	M	Testosterone	N	W	W	W
Cineole	N	W	W	W	Triphenylethylene	-2.78	W	W	W
Cinnamic acid	N	W	W	W	Triphenylphosphate	N	W	W	W
Clomiphene	-0.14	S	M	W	Vanillin	N	W	W	W
Corticosterone	N	W	W	S	17 $\alpha$ -Estradiol	0.49	S	S	S
Coumestrol (28)	-0.05	S	S	S	17 $\beta$ -Estradiol	2.0	S	S	S

*Note.* In column two is the log (RBA) to the estrogen receptor. N, not determined due to weak or nonbinding. S, strong-binding classification with a log (RBA) > -0.3; M, medium-binding classification with -0.3 > log (RBA) > -2.7; W, weak-binding classification with a log (RBA) < -2.7. Column three is the input SDAR classification from the log (RBA); column four is the <sup>13</sup>C NMR SDAR model LOO predicted classification, and column 5 is the <sup>13</sup>C NMR and EI MS SDAR model LOO predicted classification.

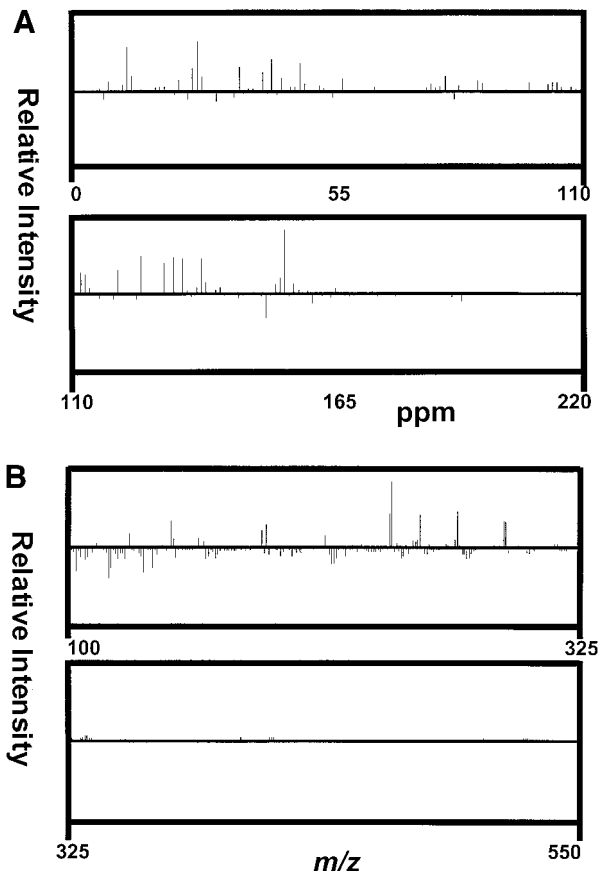


**FIG. 3.** The canonical variate using  $^{13}\text{C}$  NMR data in the SDAR model. The X-axis is the bin number and the Y-axis is the relative intensity. The bins are numbered from 0 to 110 and 110 to 220 ppm.

when using predicted  $^{13}\text{C}$  NMR spectra. 2-Ethylphenol was correctly predicted to be a weak binder by the  $^{13}\text{C}$  NMR SDAR model when using both real experimental and predicted  $^{13}\text{C}$  NMR spectra. 3-Deoxyestradiol was correctly predicted to be a strong binder by the  $^{13}\text{C}$  NMR SDAR model when using both real experimental and predicted  $^{13}\text{C}$  NMR data. 3-Methylestriol was incorrectly predicted to be a strong binder in the  $^{13}\text{C}$  NMR SDAR model when using real  $^{13}\text{C}$  NMR data, but correctly predicted to be a medium binder in the SDAR model when using predicted  $^{13}\text{C}$  NMR data. Dimethylstilbesterol was correctly predicted to be a strong estrogen binder by the  $^{13}\text{C}$  NMR SDAR model when using the real experimental and predicted  $^{13}\text{C}$  NMR data. 4,4'-Dihydroxystilbene was incorrectly predicted to be a strong binder instead of a medium binder in the  $^{13}\text{C}$  NMR SDAR model when using real  $^{13}\text{C}$  NMR data and incorrectly predicted to be a weak binder in the SDAR model when using predicted  $^{13}\text{C}$  NMR data.

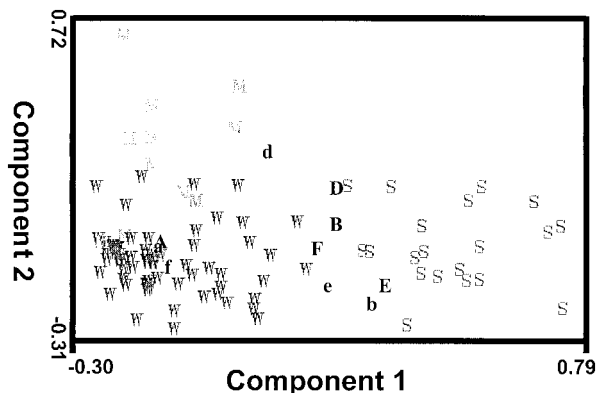


**FIG. 4.** The discriminant function using  $^{13}\text{C}$  NMR and EI mass spectral data in the SDAR model. The X-axis is the first principal component and the Y-axis is the second principal component. Compounds shown with an S have a strong classification; compounds shown with an M have a medium classification, and compounds shown with a W have a weak classification.

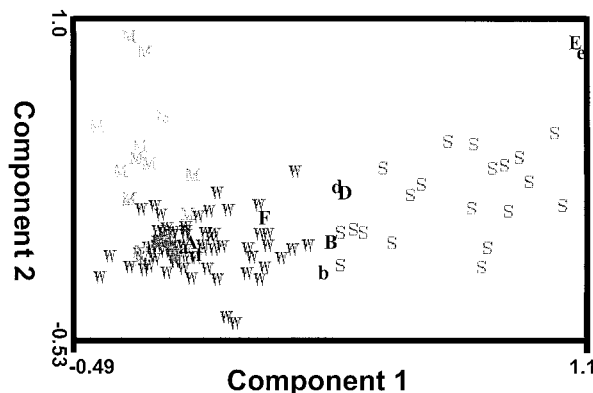


**FIG. 5.** The canonical variate using  $^{13}\text{C}$  NMR and EI mass spectral data in the SDAR model. The X-axis is the bin number and the Y-axis is the relative intensity. A.  $^{13}\text{C}$  NMR data with the bins numbering from 0 to 110 and 110 to 220 ppm. B. EI mass spectral data with the bins numbering from  $m/z$  100 to 325 and  $m/z$  325 to 550.

Figure 7 shows the discriminant function of the  $^{13}\text{C}$  NMR and EI MS SDAR model shown in Fig. 4. The notation for the predictions of 2-ethylphenol, 3-deoxyestradiol, 3-methylestriol, dimethylstilbesterol, and 4,4'-dihydroxystilbene are the same as used in Fig. 6. The actual positions of **E** and **e** are off the Fig. 7 display region and are at the same component 2 on the y-axis but the component 1 (x-axis) is equal to 3.1, not 1.1 as displayed in the figure. 2-Ethylphenol was correctly predicted to be a weak binder by the  $^{13}\text{C}$  NMR and EI MS SDAR model when using both real experimental and predicted  $^{13}\text{C}$  NMR spectra. 3-Deoxyestradiol was correctly predicted to be a strong binder by the  $^{13}\text{C}$  NMR and EI MS SDAR model when using both real experimental and predicted  $^{13}\text{C}$  NMR data. 3-Methylestriol was incorrectly predicted to be a strong binder instead of a medium binder in the  $^{13}\text{C}$  NMR and EI MS data SDAR model when using real and predicted  $^{13}\text{C}$  NMR data. Dimethylstilbesterol was correctly predicted to be a strong binder by the  $^{13}\text{C}$  NMR EI MS SDAR model when using both real experimental and predicted  $^{13}\text{C}$  NMR data. 4,4'-Dihydroxystilbene was incorrectly predicted to be a weak binder by the



**FIG. 6.** The discriminant function for the <sup>13</sup>C NMR SDAR model. The position of predicted estrogen receptor binding of 2-ethylphenol with **A** when using the real <sup>13</sup>C NMR spectral data and with **a** when using predicted <sup>13</sup>C NMR data. The position of predicted estrogen receptor binding of 3-deoxyestradiol with **B** when using the real <sup>13</sup>C NMR spectral data and with **b** when using predicted <sup>13</sup>C NMR data. The position of predicted estrogen receptor binding of 3-methylestriol with **D** when using the real <sup>13</sup>C NMR spectral data and with **d** when using predicted <sup>13</sup>C NMR data. The position of predicted estrogen receptor binding of dimethylstilbesterol with **E** when using the real <sup>13</sup>C NMR spectral data and with **e** when using predicted <sup>13</sup>C-NMR data. The position of predicted estrogen receptor binding of 4,4'-dihydroxystilbene with **F** when using the real <sup>13</sup>C NMR spectral data and with **f** when using predicted <sup>13</sup>C MR data.



**FIG. 7.** The discriminant function for the <sup>13</sup>C NMR and EI MS SDAR model. The positions of predicted estrogen receptor binding are displayed with the same notation as used in Fig. 6.

<sup>13</sup>C NMR EI MS SDAR model when using both real experimental and predicted <sup>13</sup>C NMR data.

In all four SDAR predictions of 2-ethylphenol, it was correctly classified as a weak binder. The removal of the hydroxyl group from the 3 position in estradiol and estriol lowers their binding constant to the estrogen receptor by more than a factor of 100. All four SDAR model predictions of 3-deoxyestradiol correctly see this loss in binding activity by predicting a strong classification. Three of four SDAR model predictions of 3-methylestriol incorrectly predicted a strong binding classification. Only the prediction using predicted <sup>13</sup>C NMR with the <sup>13</sup>C NMR SDAR model correctly classified 3-methylestriol in the medium classification. All four SDAR model predictions of

dimethylstilbesterol correctly see a loss in binding activity with respect to diethylstilbesterol (DES) by predicting them outside but near the strong classification zone. All four 4,4'-dihydroxystilbene predictions were incorrect and three of the four predictions were weak classifications instead of its medium classification.

### DISCUSSION AND CONCLUSIONS

There were five false negatives from the <sup>13</sup>C NMR SDAR model and four false negatives from the composite <sup>13</sup>C NMR and EI MS SDAR model. There were five false positives from the <sup>13</sup>C NMR SDAR model and seven false positives from the composite <sup>13</sup>C NMR and EI MS SDAR model. A majority of compounds that fail the LOO cross-validation test do so because of the confusion between medium and weak classification. This is consistent with the fact that the SDAR model “learns” that binding strongly to the estrogen receptor is a well-defined relationship, whereas there are many ways for a molecule to bind weakly to the estrogen receptor.

The SDAR model that uses all the <sup>13</sup>C NMR and EI MS

**TABLE 2**  
**Model Test Compounds**

Compound	log (RBA)	Class	Class model real NMR	Class model real NMR + MS	Class model predicted NMR	Class model predicted NMR + MS
2-Ethylphenol	< -4.0	W	W	W	W	W
3-Deoxyestradiol	-0.30	S	S	S	S	S
3-Methylestriol	-1.65	M	S	S	M	S
Dimethylstilbesterol	1.19	S	S	S	S	S
4,4'-Dihydroxystilbene	-0.55	M	S	W	W	W

*Note.* In column two is the log (RBA) to the estrogen receptor. S, strong-binding classification with a log (RBA) > -0.3; M, medium-binding classification with -0.3 > log (RBA) > -2.7; W, weak-binding classification with a log (RBA) < -2.7. In column 3 is the <sup>13</sup>C NMR SDAR model prediction using real <sup>13</sup>C NMR data. In column 4 is the <sup>13</sup>C NMR and EI MS SDAR model prediction using real <sup>13</sup>C NMR and EI MS data. In column 5 is the <sup>13</sup>C NMR SDAR model prediction using predicted <sup>13</sup>C NMR data. In column 6 is the <sup>13</sup>C NMR and EI MS SDAR model prediction using predicted <sup>13</sup>C NMR data and real EI MS data.

spectrometric data had the best LOO cross-validation. When a compound had both classification predictions from the two SDAR models, there was only one false negative (clomiphene) and there were two false positives (dieldrin and p,p'-(dichloro diphenyl)-2,2,dichloroethylene). The false negative compounds have very few compounds that are structurally similar in the SDAR models. False positive compounds are not a major concern since they would be tested experimentally for estrogen-receptor binding. Clomiphene is a false negative with a log (RBA) of  $-0.14$ , which is close to the medium and strong estrogen-receptor classification divisions. Compounds with a log (RBA) near a classification division are harder to predict. It appears that for a SDAR model to correctly predict a compound's binding activity, it must have structurally similar compound as a basis for the prediction.

The addition of infrared absorption (IR) spectrometric data has been used to increase the accuracy of SDAR models of monodechlorination of chlorobenzenes, chlorophenols, and chloroanilines (Beger *et al.*, 2000) and IR data may be able to increase the accuracy of the SDAR models estrogen-receptor binding. IR absorption spectra are based on quantum mechanical principle and, similar to  $^{13}\text{C}$  NMR chemical shifts, have been used to predict molecular structure (Hemmer *et al.*, 1999). Conversely, similar to  $^{13}\text{C}$  NMR spectra, chemical structures have been used to predict IR spectra (Gasteiger *et al.*, 1997). A drawback of SDAR modeling is the loss of three-dimensional site-specific information, which can lead to false negatives. One way to increase the accuracy of SDAR modeling is to use all SDAR models that predict a compound as a strong binder to the estrogen receptor be tested experimentally for estrogen receptor binding. This way the six false negative compounds that were predicted correctly in one SDAR model and incorrectly predicted in the other SDAR model would still be tested experimentally. Each spectrometric data set (NMR, IR, and EI MS) is a two-dimensional representation of the molecule and each spectrometric data set contains unique information about the molecule. Using the combined results of multiple SDAR models with different spectrometric data is another reason to add IR spectral data to SDAR modeling of estrogen receptor binding.

Another objective of building SDAR models is to use them in high-throughput virtual drug design, working in real time with combinatorial chemistry production. An SDAR model can be built as fast as the spectrometric digital fingerprints can be acquired and predicted. There is no need to align molecular compounds or even to know the structure of the compound before an SDAR can be built. Another benefit of SDAR modeling is that there is no need to calculate quantum mechanical properties where inaccuracies can be built into the QSAR model by assumptions used to solve Schrödinger's equation. In drug discovery, the activity of a lead compound can be predicted for a particular endpoint activity once a model using its  $^{13}\text{C}$  NMR and real EI MS spectrum has been defined for

comparison with spectra for at least 30 compounds whose activity in this respect is known.

The accuracy of the SDAR model predictions based on real  $^{13}\text{C}$  NMR and on predicted  $^{13}\text{C}$  NMR spectral data were almost the same. The predictions for 2-ethylphenol, 3-deoxyestradiol, and dimethylstilbesterol were correct for all four predictions. The predictions for 3-methylestriol were incorrectly classified as strong in three of four predictions. The only SDAR model prediction that was correct for 3-methylestriol was that based on predicted  $^{13}\text{C}$  NMR data using the  $^{13}\text{C}$  NMR data model. Predicting the effects of point changes from a compound that is a strong binder will be the toughest test for SDAR modeling. The predictions for 4,4'-dihydroxystilbene were wrong all four times. The prediction of compounds in the medium binding classification is hard because of nonspecific binding; the electronics or geometry can be altered in many different ways to lower the binding of a compound a little. The variance analysis used in SDAR models are optimized to find the set of NMR frequencies that are best associated with strong binding. The accuracy of the SDAR model predictions proves that  $^{13}\text{C}$  NMR spectra can be used in covariant analysis to discriminate between molecules that may bind to the estrogen receptor. SDAR modeling is not meant to replace QSAR or SAR modeling, but can be used as an alternative when QSAR/SAR modeling is unreliable. Ultimately the combination of SDAR and QSAR/SAR into one model may lead to the more powerful modeling technique.

The SDAR model approach can be used for other compound-receptor systems by simply exchanging the relative binding affinities of the estrogen-receptor system with those appropriate for the alternative system to be modeled and training a new SDAR for that endpoint. The  $^{13}\text{C}$  NMR, EI MS, and other spectral data do not change. They can be used as comprehensive descriptors in a new SDAR model of many different biological end points. The only requirement is the availability of a suitable training set for which the strength in relation to that endpoint has already been determined. Thirty to 45 similar compounds may represent a good start for a training set, especially if one uses it only to predict the activity of other structurally similar compounds. However, for effective toxicological prediction involving a variety of potential structural types, a large, diverse training set such as that used in this work will be important.

## REFERENCES

- ACD/Labs CNMR software version 4.0, Toronto, Canada.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
- Beger, R., Freeman, J., Lay Jr., J., Wilkes, J., and Miller, D. (2000). Producing  $^{13}\text{C}$  NMR, Infrared Absorption and EI Mass Spectrometric Data Models of the Monodechlorination of Chlorobenzenes, Chlorophenols, and Chloroanilines. *J. Chem. Inf. Comput. Sci.* [in press].
- Beger, R. D., and Bolton, P. H. (1997). Protein  $\phi$  and  $\psi$  dihedrals restraints determined from multidimensional hypersurface correlations of backbone

- chemical shifts and their use in the determination of protein tertiary structures. *J. Biomol. NMR* **10**, 129–142.
- Blair, R. M., Fang, H., Branham, W. S., Hass, B. S., Dial, S. L., Moland, C. L., Tong, W., Shi, L., Perkins, R., and Sheehan, D. M. (2000). The estrogen receptor relative binding affinities of 188 natural and xenochemicals: Structural diversity of ligands. *Toxicol. Sci.* **54**, 138–153.
- Boguski, M. S., and Schuler, G. D. (1995). Establishing a human transcript map. *Nature* **10**, 369–371.
- Branbury, S. P. (1995). Quantitative structure-activity relationship and ecological risk assessment: An overview of predictive aquatic toxicology research. *Toxicology* **25**, 67–89.
- Bremser, W. (1978). HOSE—A novel substructure code. *Anal. Chim. Acta* **103**, 355–365.
- Bursi, R., Dao, T., van Wilk, T., de Gooyer, M., Kellenbach, E., and Verwer, P. (1999). Comparative spectra analysis (CoSA): Spectra as three-dimensional molecular descriptors for the prediction of biological activities. *J. Chem. Inf. Comput. Sci.* **39**, 861–867.
- Collantes, E., Tong, W., and Welsh, W. (1996). Use of moment of inertia in comparative molecular field analysis to model chromatographic retention of nonpolar solutes. *J. Anal. Chem.* **68**, 2038–2043.
- Cramer, R. D., Bunce, J. D., and Patterson, D. E. (1988a). Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct. Act. Relat.* **7**, 18–25.
- Cramer, R. D., Paterson, D. E., and Bunce, J. D. (1988b). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**, 5959–5967.
- De Dios, A. C., Pearson, J. G., and Oldfield, E. (1993). Secondary and tertiary structural effects on protein NMR chemical shifts: An *ab initio* approach. *Science* **260**, 1491–1496.
- Emsley, J. W., Feeney, J., and Sutcliffe, L. H. (1965). In *High Resolution Nuclear Magnetic Resonance*. Vol. I, pp. 1–287. Pergamon Press, Oxford.
- Frenkel, M., and Marsh, K. N. (1994). *Spectral Data for Steroids*. Thermodynamics Research Center: College Station.
- Fujita, T., Iwasa, J., and Hansch, C. (1964). A new substituent constant,  $\pi$ , derived from partition coefficient. *J. Am. Chem. Soc.* **86**, 5175–5180.
- Gasteiger, J., Schuur, J., Selzer, P., Steinhauer, L., and Steinhauer, V. (1997). Finding the 3D structure of a molecule in its IR spectrum. *J. Anal. Chem.* **359**, 50–55.
- Hansch, C., and Leo, A. (1995). *Exploring QSAR—Fundamentals and applications in chemistry and biology*. American Chemical Society, Washington, DC.
- Hemmer, M. C., Steinhauer, V., and Gasteiger, J. (1999). The prediction of the 3D structure of organic molecules from their infrared spectra. *Vibrat. Spectrosc.* **19**, 151–164.
- Hopert, A.-C., Beyer, A., Frank, K., Strunk, E., Wunsche, W., and Vollmer, G. (1998). Characterization of estrogenicity of phytoestrogens in an endometrial-derived experimental model. *Environ. Health Perspect.* **106**, 581–586.
- Integrated Spectral Data Base System for Organic Compounds web site. (2000). [www.aist.go.jp/RIODB/SDBS/](http://www.aist.go.jp/RIODB/SDBS/).
- Julier, C., Lucassen, A., Villedieu, P., Delepine, M., Levy-Marchal, C., and Danze, P. M. (1994). Multiple DNA variant association analysis: Application to the insulin gene region in type I diabetes. *Am. J. Hum. Genet.* **55**, 1247–1254.
- Katritzky, A. R., Ignatchenko, E. S., Barcock, R. A., and Lobanov, V. S. (1994). Prediction of gas chromatographic retention times and response factors using a general quantitative structure-property relationship. *Anal. Chem.* **66**, 1799–1807.
- Katritzky, A. R., Mu, L., Labanov, V. S., and Karelson, M. (1996). 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.* **100**, 10400–10407.
- Klawun, C., and Wilkins, C. L. (1996). Joint neural network interpretation of infrared and mass spectra. *J. Chem. Inf. Comput. Sci.* **36**, 249–257.
- Klopman, G. (1984). Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. *J. Am. Chem. Soc.* **106**, 7315–7321.
- Klopman, G. (1992). MULTICASE1. A hierarchical computer automated structure evaluation program. *Quant. Struct. Act. Rel.* **11**, 176–184.
- Kollman, P. A., Giannini, D. D., Duax, W. L., Rothenberg, S., and Wolff, M. E. (1973). Quantitations of long range effects in steroids by molecular orbital calculations. *J. Amer. Chem. Soc.* **95**, 2869–2873.
- Kuiper, G. G. J. M., Carlsson, B., Grandien, K., Enmark, E., Haggblad, J., Nilsson, S., and Gustafsson, J.-A. (1997). Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors  $\alpha$  and  $\beta$ . *Endocrinology* **138**, 863–870.
- Kvasnicka, V. (1991). An application of neural networks in chemistry. Prediction of <sup>13</sup>C NMR chemical shifts. *J. Math. Chem.* **6**, 63–76.
- Munk, M. E., Madison, M. S., and Robb, E. W. (1996). The neural network as a tool for multispectral interpretation. *J. Chem. Inf. Comput. Sci.* **36**, 231–238.
- NIST 1998 Mass Spectral Library*, version 1.6. (2000). U.S. Department of Commerce, National Institute of Standards and Technology, Standard Reference Data Program, Gaithersburg, MD.
- Pouchert, C. J., and Behnke, J. (1993). *The Aldrich Library of <sup>13</sup>C and <sup>1</sup>H FT NMR Spectra*, Vols. 1–3, 1st ed., Aldrich Chemical Co.
- Tong, W., Collantes, E., Chen, Y., and Welsh, W. J. (1995). A comparative molecular field analysis study of N-benzylpiperidines as acetylcholinesterase inhibitors. *J. Med. Chem.* **39**, 380–387.
- Tong, W., Perkins, R., Strelitz, R., Collantes, E. R., Keenan, S., Welsh, W. J., Branham, W. S., and Sheehan, D. M. (1997). Quantitative structure-activity relationships (QSARs) for estrogen binding to the estrogen receptor: Predictions across species. *Environ. Health Perspect.* **105**, 1116–1124.
- Wishart, D. S., and Sykes, B. D. (1994). Chemical shifts as a tool for structure determination. *Meth. Enzymol.* **239**, 363–392.
- Zava, D. T., and Duwe, G. (1997). Estrogenic and antiproliferative properties of genistein and other flavonoids in human breast cancer cells *in vitro*. *Nutr. Canc.* **27**, 31–40.
- Zhang, J., and Madden, T. L. (1997). PowerBLAST: A new network BLAST application for interactive or automated sequence analysis and annotation. *Genom. Res.* **7**, 649–656.