Use of ^{13}C NMR Spectrometric Data To Produce a Predictive Model of Estrogen Receptor Binding Activity

Richard D. Beger,* 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, Arkansas 72079

Received July 5, 2000

We have developed a spectroscopic data−activity relationship (SDAR) model based on ^{13}C NMR spectral data for 30 estrogenic chemicals whose relative binding affinities (RBA) are available for the alpha (ERα) and beta (ERβ) estrogen receptors. The SDAR models segregated the 30 compounds into strong and medium binding affinities. The SDAR model gave a leave-one-out (LOO) cross-validation of 90%. Two compounds that were classified incorrectly in the SDAR model were in the transition zone between classifications. Real and predicted ^{13}C NMR chemical shifts were used with test compounds to evaluate the predictive behavior of the SDAR model. The ^{13}C NMR SDAR model using predicted ^{13}C NMR data for the test compounds provides a rapid, reliable, and simple way to screen whether a compound binds to the estrogen receptors.

INTRODUCTION

The Food Quality Protection Act, passed in 1996, mandates that the United States Environmental Protection Agency (EPA) develop screening and testing procedures for endocrine disrupting chemicals (EDCs). An 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.

^{13}C nuclear magnetic resonance (NMR) chemical shifts have been used to predict and refine chemical structures.[1,2] Conversely, the chemical structure of a compound has been used to predict its ^{13}C NMR chemical shifts.[3] The ^{13}C NMR spectrum of a compound contains frequencies that correspond directly to the quantum mechanical properties of the nucleus of each carbon atom in a molecule. The quantum mechanical description of a nucleus depends largely on its electrostatic features and three-dimensional geometry.[4] Ab initio quantum mechanical calculations of ^{13}C chemical shift tensors in proteins reveal that they are dependent on the structural geometry and electrostatics.[5] Three-dimensional quantitative structure−activity relationship (3D-QSAR) modeling results show that receptor binding of a compound can be predicted, based in part upon electrostatics and geometrical structure.[6−13] The binding activities of 45 progestagens have been quantitatively modeled with individual molecular ^{1}H NMR, infrared spectra, and mass spectra, as well as simulated infrared (IR) spectra and ^{13}C NMR spectra by comparative spectral analysis (CoSA).[14] The CoSA model produced results that yielded better correlations and predictions than were seen with comparative molecular field analysis (CoMFA),[15−18] but the CoSA quantitative modeling was limited to a set of structurally similar compounds.

Quantum descriptors are believed to enhance the other 3D-QSAR descriptors by providing the atomic scale electromagnetic properties of the compound, but they are difficult to calculate accurately for large compounds. In 3D-QSAR calculations, quantum mechanical descriptors are the most important features because they not only give the electromagnetic properties of the compound but also include topological, geometrical, and electrostatic properties. ^{13}C NMR chemical shifts represent the part of the molecular information that is found in the quantum mechanical description of the molecule. We demonstrate that covariance analysis of patterns in ^{13}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.[19−21] We propose that such modeling based on spectroscopic data−activity relationships be called SDAR.

The frequencies obtained from the ^{13}C NMR spectroscopic data correspond directly to the energies obtained when solving the quantum mechanical Schrodinger equation for the nuclear magnetic moments transitions.[4] The NMR quantum energies are strongly dependent on the electrostatic potential energy of the carbon nucleus and the type of orbital (wave function) surrounding the carbon nucleus. The orbitals in a molecule can be correlated to the LUMO and HOMO quantum states of the molecule. Typically, ^{13}C NMR chemical shifts in the 0−100 ppm range are associated with carbon atoms that have 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 esters). Likewise, ^{13}C NMR chemical shifts in the 100−220 ppm range are associated with carbon atoms that have sp^{2} and sp orbitals, with the more upfield shifts having a positive electrostatic potential

* To whom correspondence should be addressed. Phone: 870 543-7080. FAX: 870 543-7686. E-mail: rbeger@nctr.fda.gov.

10.1021/ci0000878 CCC: $20.00
Published on Web 12/19/2000

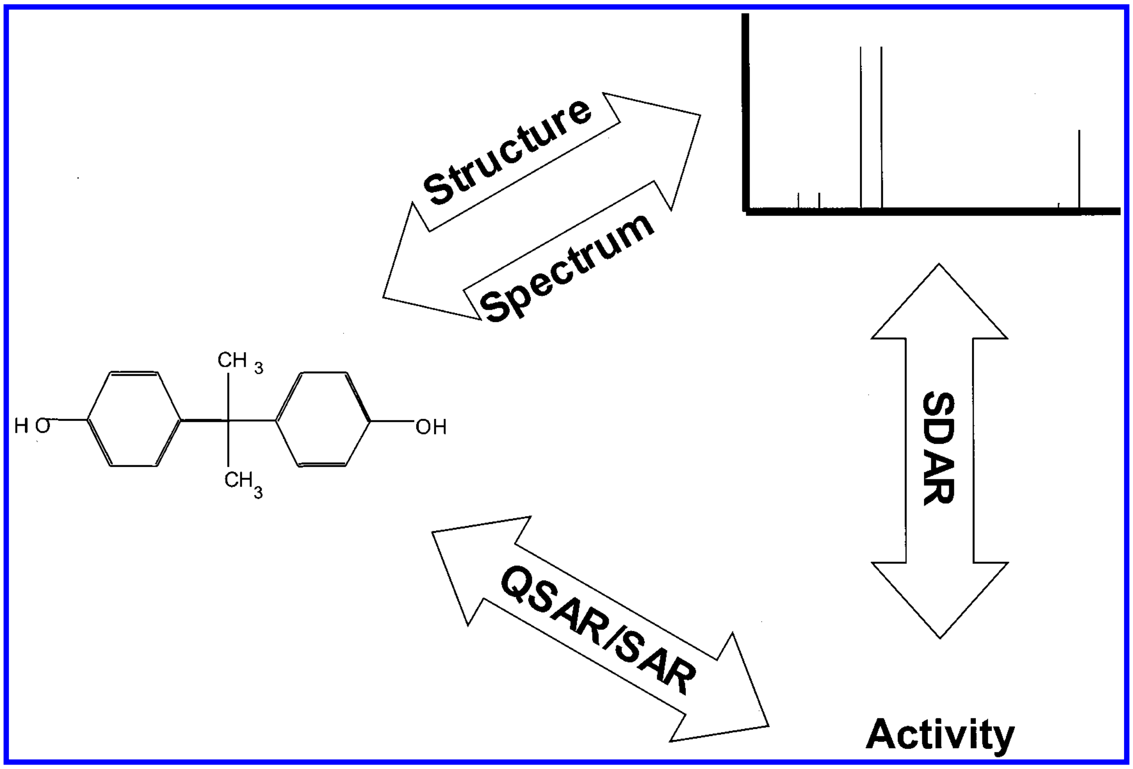

Figure 1. SDAR concept.

(like benzyl groups) and the downfield shifts having a more negative electrostatic potential (like carbonyl groups). The effects of substituents on ^{13}C NMR chemical shifts can be felt from as far as five bonds away or through space directly. The 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 and are given with respect to a reference compound. (When the ^{13}C NMR chemical shifts are given in hertz (Hz), they represent an energy.)

Figure 1 shows how SDAR modeling relates the structure of a compound and other QSAR/SAR modeling techniques.[22,23] NMR spectrometric data are able to distinguish between chemical isomers and are demonstrated by the ^{13}C NMR chemical shift of carbon 17 in 17β-estradiol and 17α-estradiol, which is 3 ppm different. 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}C NMR data.[23] Neural networks have been developed that can accurately identify the presence of 26 structural features present in a compound with the network trained on IR and EI MS data.[24] Like ^{13}C NMR spectral data, the chemical structure of a compound has been used to predict its IR spectrum.[25] SDAR removes the problems associated with structure alignment and structural calculations, but one-dimensional SDAR modeling loses direct three-dimensional structural specific information and that can produce false negatives and positives.

The spectra can be made to function as comparative "*spectrometric digital fingerprints*" for each compound in relation to others and in relation to a biological endpoint, such as estrogen receptor binding. In this paper, the spectral data−activity relationship is "discovered" from the pattern of frequencies of ^{13}C NMR. SDAR modeling bypasses these quantum mechanical calculations and the associated assumptions for dielectric and wave functions needed to solve the quantum mechanical calculations and to predict chemical properties.

METHOD

The estrogenic relative binding affinities (RBAs) of the 30 compounds were taken from previous publications.[9,27] We were able to find most of the ^{13}C NMR spectrometric and EI mass spectrometric data from the Integrated Spectral Data Base System for Organic Compounds web site www.aist.go.jp/RIODB/SDBS/,[28] the *Aldrich Library of ^{13}C and ^{1}H FT NMR Spectra*,[29] and *Spectral Data of Steroids*.[30] We obtained experimental ^{13}C NMR data for five compounds using standard methods (^{13}C NMR spectra were recorded using 99.8% pure $CDCl_3$ or 100.0% pure dimethyl sulfoxide (DMSO) solvents).

The ^{13}C nuclear magnetic resonance spectral analyses of 4-hydroxyestradiol, ICI 164-384, moxestrol, and norethynodrel were performed at 75.46 MHz on a Varian Gemini 300 MHz NMR (Varian Associates, Inc., Palo Alto, CA) spectrometer operating at 301 K. The ^{13}C NMR spectra were run while protons were decoupled during acquistion. Compounds were dissolved in standard $CDCl_3$ or DMSO NMR solvents. The chemical shifts were defined by assigning the $CDCl_3$ peak to 77.0 ppm and the DMSO peak to 39.5 ppm. The spectral width was 21,008 Hz with a 2.6 s delay time between acquisitions. The acquisition time was 0.495 s, and the number of points acquired was 20 800.

The ^{13}C NMR spectra were saved as a set of ordered pairs: chemical shift frequencies in ppm and the area under the peak. The area under a specific chemical shift frequency was first normalized to an integer. A nondegenerate frequency was assigned an area of 25, a doubly degenerate frequency (two ^{13}C NMR chemical shifts at the same frequency) had an area of 50, and so forth. The number 25 was selected arbitrarily in order to normalize the peak intensity in ^{13}C NMR spectroscopic data; it becomes irrelevant later when we autoscale the NMR data. This initial normalization was done so that (1) all the spectra would have a similar signal-to-noise ratio and (2) line width variations due to differences in NMR instrumental field strengths, shimming, coupling to protons, temperature, pH, or solvents would be eliminated. The bin defined the number of significant and distinct chemical shift peaks inside a ppm range. The bin width used in inputting the ^{13}C NMR spectra was optimized by allowing the spectral window width to vary between 0.5, 1.0, 2.0, and 5.0 ppm and determining the best leave-one-out cross-validation. We found that the optimal ppm range for this study was 1.0 ppm. Using a 1 ppm spectral width for the bins, only 141 out of the 221 spectral bins had a ^{13}C NMR chemical shift other than zero for all 30 spectra in them.

The RBAs to the estrogen receptor are defined as the ratio of the molar concentrations of the 17β-estradiol and the competing compound required to decrease the receptor-bound by 50% and multiplied by 100. Thus 17β-estradiol has an RBA = 100 and a log(RBA) = 2.0. Strong binders to the estrogen receptor were classified as those with a log(RBA) over 0.0, and medium binders to the estrogen receptor were classified as those with a log(RBA) $\leq$ 0.0. With a log(RBA) = 0.0 used as a classification division point, there was no change in compound classifications when we used α or β log(RBA) values. Thus, we had 17 strong binders and 13 medium binders in the training set.

The pattern recognition software used was RESolve Version 1.2 (Colorado School of Mines, Boulder, CO). The ^{13}C NMR spectroscopic data and RBA classification for all 30 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.[31]

Autoscaling compares the quantitative response at each NMR spectral bin to all the others in the comparison set. An average value and the standard deviation is calculated for each spectral bin. Then, for each spectrum, the quantitative response at a spectral bin is expressed as the number of standard deviations above or below the average. This data pretreatment step equalizes the distribution of consistent variance for signals with inherently small magnitudes relative to those signals with large magnitudes and corresponding large variance.

Fisher weighting pretreats data in a way that emphasizes those spectral characteristics important in distinguishing defined groups. Fisher weighting for a bin is defined as

$$\mathrm{FW\ (bin)} = \frac{\sum_{i}^{g-1}\sum_{j=1}^{g} (\langle X_i\rangle - \langle X_j\rangle)^2/(\sigma_i^{\,2} + \sigma_j^{\,2})}{g((g-1)/2)}$$

where g is the number of classification categories, $\langle X_i\rangle$ is the average intensity for category I, and σ_i is the variance about the mean of category i. For each NMR spectral bin, the variance between groups is divided by the variance within groups. The resulting dividend becomes a weighting factor that has a magnitude larger than one when a particular spectral bin has an important role in distinguishing groups. Fisher weighting of 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 (such as the ^{13}C NMR signals from carbon atoms not related to the estrogen receptor binding).

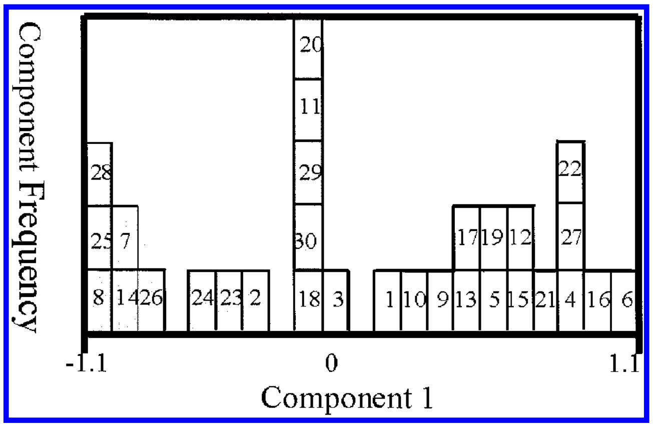

Figure 2. Discriminant function using ^{13}C NMR data in the SDAR model. The X-axis is the first principal component, and the Y-axis is the component frequency. Compounds with a white background have a strong classification predicted by the SDAR model, and compounds with a gray background have a medium classification predicted by the SDAR model.

The number of principal components used in the SDAR model was determined by a systematic search over the number of principal components from 1 to 30 and the corresponding LOO cross-validation. Typically the LOO cross-validation of the SDAR model rises linearly when 1−10 principal components are used and then the LOO cross validation oscillates by 5 to 10% when going from 10 to 20 or 25 principal components used in the SDAR model. When the number of principal components reaches 20−25, a 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.

RESULTS

Based only on ^{13}C NMR spectroscopic data, the statistical pattern recognition program with 8 principal components (PCs) used 81.1% of the total variance and a cross-validation of 90%. The mathematical model significance was 91.9% at one discriminant function, with 28 degrees of freedom. The Wilks discriminant criterion was 99.99%. Figure 2 shows the discriminant function for ^{13}C NMR data. Compounds with a white background have strong RBAs, and compounds with a gray background have medium RBAs. Figure 2 shows a large separation between the 15 strong and the 9 medium binders. In the transition zone between strong and medium RBAs are 4 strong and 2 medium RBAs. Table 1 contains the compound name, an identifying number that is used in Figure 2, the log(RBA) α and log(RBA) β values, the SDAR training input, the SDAR prediction from using ^{13}C NMR, and the SDAR prediction from using ^{13}C NMR data. Table 1 shows that 27 of the 30 compounds are correctly group predicted using only ^{13}C NMR data; 3β-androstanediol,

Table 1[a]

compd	no.	log(RBA) α	log(RBA) *β*	class input	class pred NMR
coumestrol	1	1.97	2.27	S	S
3α-androstanediol	2	−1.15	−0.52	M	M
3*β*-androstanediol	3	0.48	0.85	S	M
4-hydroxyestradiol	4	1.11	0.85	S	S
17α-estradiol	5	1.76	1.04	S	S
17*β*-estradiol	6	2.00	2.00	S	S
bisphenol A	7	−1.30	−0.48	M	M
β-zearanol	8	−1.20	−1.15	M	M
clomifene	9	1.40	1.08	S	S
2-hydroxyestradiol	10	0.85	1.04	S	M
dehydroepiandrosterone	11	−1.40	−1.15	M	M
diethystilbesterol	12	2.67	2.47	S	S
dienestrol	13	2.35	2.61	S	S
dihydrotestosterone	14	−1.30	−0.77	M	M
estriol	15	1.15	1.32	S	S
estrone	16	1.78	1.57	S	S
genistein	17	0.70	1.56	S	S
hexestrol	18	2.48	2.37	S	M
ICI 164-384	19	1.93	2.22	S	S
methoxychlor	20	−2.00	−0.89	M	M
moxestrol	21	1.63	0.70	S	S
nafoxidine	22	1.64	1.20	S	S
nandrolone	23	−2.00	−0.64	M	M
norethindrone	24	−1.15	−2.00	M	M
norethynodrel	25	−0.16	−0.66	M	M
progesterone	26	−3.50	−3.50	M	M
tamoxifen	27	0.85	0.78	S	S
testosterone	28	−2.10	−2.10	M	M
5α-androstanedione	29	<−2.0	<−2.0	M	M
5*β*-androstanedione	30	<−2.0	<−2.0	M	M

[a] In column 2 is the log(RBA) to the estrogen receptor. S stands for a strong Binding classification with a log(RBA) > −0.3, and M stands for a Medium Binding Classification with −0.3 > log(RBA) > −2.7. Column 5 is the input SDAR classification from the log(RBA), and column 6 is the ^{13}C NMR SDAR model LOO predicted classification.

2-hydroxyestradiol, and hexestrol are incorrectly predicted to have a medium RBA using principal components. Only 3*β*-androstanediol and hexestrol are incorrectly predicted to have a medium RBA using the canonical variate function. Strong binding 3*β*-androstanediol is incorrectly predicted, presumably because the compound is similar to the medium RBA 3α-androstanediol and the model does not contain examples of spectra from similar structures that are medium binders.

Figure 3 shows the factor loadings associated with the first canonical variate function for the pattern recognition of the ^{13}C NMR data. The positive peaks in Figure 3 correspond to bins that bias toward a strong RBA for binding to the estrogen receptor, and negative peaks correspond to bins that bias toward a medium RBA. The aliphatic CH_2 bins 30−35 ppm have a bias toward medium RBA. The methyl CH_3 bins 8 and 16 ppm, respectively, have a bias toward strong RBA. Many of the aromatic bins 115−150 ppm have a bias toward strong RBA.

We used the ^{13}C NMR SDAR model to predict a classification of the estrogen binding activity for four compounds not in the training set. First we used the real experimental spectra to correctly predict that 4-androstaindione, corticosterone, bisphenol B,[32] and cholesterol[32] should have a medium classification RBA. Next we used the predicted ^{13}C NMR spectra from ACD Labs[33] to correctly predict that 4-androstaindione and corticosterone should have a medium classification RBA. The use of predicted ^{13}C NMR spectra produced a false positive classification for cholesterol and bisphenol B. Table 2 shows the real and predicted relative binding classifications for 4-androstaindione, corticosterone, bisphenol B, and cholesterol.

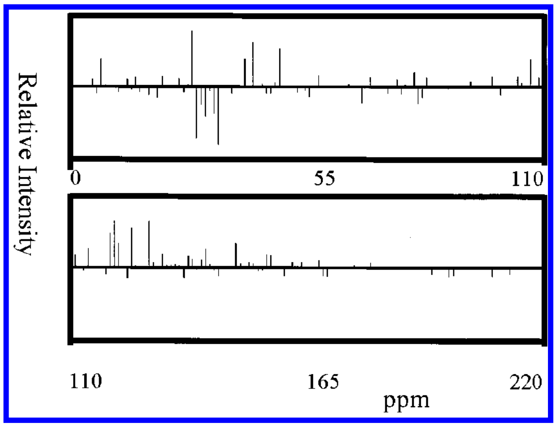

Figure 3. Canonical variate using ^{13}C NMR data in the SDAR model. The *X*-axis is the bin number, and the *Y*-axis is the relative intensity. The bins are numbered from 0 to 220 ppm.

Table 2[a]

				class model	
compd	log(RBA) α	log(RBA) *β*	class	real NMR	pred NMR
4-androstanedione	<−2.0	<−2.0	M	M	M
corticosterone	<−3.0	<−3.0	M	M	M
bisphenol B[30]	<−1.0		M	M	S
choloesterol[30]	<−3.0		M	M	S

[a] In column 2 is the log(RBA) to the estrogen receptor. S stands for a strong binding classification with a log(RBA) > −0.3, and M stands for a medium binding classification with −0.3 > log(RBA) > −2.7. In column 4 is the ^{13}C NMR SDAR model prediction using real ^{13}C NMR data. In column 5 is the ^{13}C NMR and EI MS SDAR model prediction using real ^{13}C NMR and EI MS data. In column 6 is the ^{13}C NMR SDAR model prediction using predicted ^{13}C NMR data.

CONCLUSIONS

Our 8 principal component ^{13}C NMR SDAR model has a leave-one-out cross-validation of 90.0%. Both the α and *β* estrogen receptor binding QSAR models based on 4 principal components with a similar set of compounds and binding data had a 95% correlation between CoMFA calculated fields and the corresponding RBA of both the α and *β* estrogen receptor.[9] The qualitative predictive value of the QSAR was only q^2 > 0.7 and 0.6 for the estrogen receptor α and *β*, respectively, based on LOO cross-validation. A semiqualitative application of the QSAR predictive value based on classifications would be q^2 > 0.97 and 1.0.[9] Both the QSAR and our SDAR reported that 3*β*-androstanediol would have a much lower estrogen binding activity than reported in the literature.[27] 3*β*-Androstanediol has since been reported at a much lower relative binding which would make it fall to a medium classification.[32] The two other false negatives in the SDAR model are 2-hydroxyestradiol and hexestrol. It is curious why these two strong classification estrogen receptor binders are misclassified. Hexestrol is most similar to bisphenol A structurally, which may explain the false negative. Future work with more compounds in the training set and other types of spectra will be used to determine if false negatives can be reduced in SDAR modeling.

Current QSAR models use extensive computer modeling and often break the molecule into secondary structural motifs. Use of molecular structural components requires that the structure of the molecule be determined and disclosed before the model can be used. This causes a number of technical and commercial problems and limits possible applications. In our experimentally based SDAR model, there is no computer modeling of a compound's electrostatic fields or van der Waals surface onto a grid, as is done using QSAR methods.[8] With the SDAR method, there is no need to break the compound into secondary structural motifs, as is done in a CODESSA QSAR and SAR.[22,23] One problem for SDAR modeling is that to obtain a natural abundance ^{13}C NMR spectrum, a lot of highly purified sample or instrument time is needed. Obtaining a lot of highly pure sample can be troublesome, but if the SDAR can be trained using spectra from compounds that have their ^{13}C NMR spectra already available, one should be able to predict the ^{13}C NMR spectra of compounds to be predicted from the SDAR model. As the prediction of the ^{13}C NMR spectra of a compound becomes more reliable the SDAR activity prediction based on predicted ^{13}C NMR spectra will become more reliable.

This SDAR model is fast and may give a straightforward classification for the most active compounds. Moreover, for each compound it requires only experimental data that are readily attainable and often already available. It may be possible for an SDAR model based on spectroscopic data to be used to screen a large number of compounds in the lead discovery process for identifying those few likely by spectral similarity to show estrogenic activity. These predicted "hits" may then have their actual estrogenic relative binding affinities confirmed by expensive in vitro binding assays.

It is interesting to note that the SDAR four predictions using real ^{13}C NMR spectra are correctly classified all four times and the four predictions using predicted ^{13}C NMR spectra are correctly classified only two out of the four times. The predicted spectra are calculated by a substructure technique known as HOSE.[34] HOSE is similar to many SAR techniques in that it breaks a molecule into substructures and searches for similar molecules and substructures. If the ^{13}C NMR did not have any correspondence to the quantum mechanical electronic state of the molecule, the predictions based on the real ^{13}C NMR spectra should have been as bad as the predictions based on the predicted ^{13}C NMR spectra. The fact that the four predictions using real ^{13}C NMR spectra were so much better than the predictions based on predicted ^{13}C NMR spectra may be just a chance occurrence, but it promotes the idea that ^{13}C NMR spectra are associated with the quantum mechanical electronic state of the molecule. A larger and more structurally diverse database is needed to determine whether SDAR modeling can become a rapid and effective screening device, but conceptually the ease, speed, and accuracy of using SDAR modeling is demonstrated in this paper.

Another benefit of the experimental SDAR approach is its flexibility. It 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 compound–receptor system. The ^{13}C NMR spectral data do not change. They can be used as comprehensive chemical descriptors in a new SDAR model of a molecule–receptor or bioactive system. This can be done because the 3D structure of either the receptor or the binding molecule is used in the experimental SDAR model. The only requirement is the availability of known binding values for compounds used to train an SDAR model relative to the new receptor. SDAR is not meant to replace QSAR or SAR but to be used as an alternative to computational drug development when QSAR/SAR modeling is unreliable. Ultimately the combination of SDAR and QSAR/SAR modeling into one model may lead to the most powerful modeling technique.

REFERENCES AND NOTES

(1) Beger, R. D.; Bolton, P. H. Protein ϕ and ψ dihedrals restraints determined from multidimensional hypersurface correlations of backbone chemical shifts and their use in the determination of protein tertiary structures. *J. Biomol. NMR* **1997**, *10*, 129–142.

(2) Wishart D. S.; Sykes B. D. Chemical shifts as a tool for structure determination. *Methods Enzymol.* **1994**, *239*, 363–392.

(3) Kvasnicka, V. An Application of Neural Networks in Chemistry. Prediction of ^{13}C NMR Chemical Shifts. *J. Math. Chem.* **1991**, *6*, 63–76.

(4) Emsley, J. W.; Feeney, J.: Sutcliffe, L. H. *High-Resolution Nuclear Magnetic Resonance;* Pergamon Press Ltd.: Oxford, U.K., 1965; Vol. I, Chapter 8, p 287.

(5) De Dios, A. C.; Pearson, J. G.; Oldfield, E. Secondary and tertiary structural effects on protein NMR chemical shifts: An *ab initio* approach. *Science*. **1993**, *260*, 1491–1496.

(6) Hansch, C.; Leo, A. *Exploring QSAR–Fundamentals and applications in chemistry and biology;* American Chemical Society: Washington, D.C., 1995.

(7) Branbury, S. P. Quantitative structure–activity relationship and ecological risk assessment: An overview of predictive aquatic toxicology research. *Toxicology* **1995**, *25* (1), 67–89.

(8) Tong, W.; Perkins, R.; Strelitz, R.; Collantes, E. R.; Keenan, S.; Welsh, W. J.; Branham, W. S.; Sheehan, D. M. Quantitative Structure–Activity Relationships (QSARs) for Estrogen Binding to the Estrogen Receptor: Predictions across Species. *Environ. Health Perspect*. **1997**, *105*, 1116–1124.

(9) 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.

(10) Katritzky, A. R.; Mu, L.; Labanov, 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.

(11) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S. Prediction of gas chromatographic retention times and response factors using a general quantitative structure–property relationship. *Anal. Chem*. **1994**, *66*, 1799–1807.

(12) Katritzky, A. R.; Rachwal, P.; Law, K. W.; Karelson, M. J. Prediction of polymer glass transition temperatures using a general quantitative structure–property relationship treatment. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 879–884.

(13) Fujita, T.; Iwasa, J.; Hansch, C. A new substituent constant, $\boldsymbol{\pi}$, derived from partition coefficient. *J. Am. Chem. Soc.* **1964**, *86*, 5175–5180.

(14) Bursi, R.; Dao, T.; van Wilk, T.; de Gooyer, M.; Kellenbach, E.; Verwer, P. Comparative Spectra Analysis (CoSA): Spectra as Three-Dimensional Molecualr Descriptors for the Prediction of Biological Activities. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 861–867.

(15) Opera, T. F.; Garcia, A. E. Three-dimensional quantitative structure activity relationships of steroid aromatase inhibitors. *J. Comput.-Aided Mol. Design* **1996**, *10*, 186–200.

(16) Cramer, R. D.; Paterson, 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.

(17) Tong, W.; Collantes, E.; Chen, Y.; Welsh, W. J. A comparative molecular field analysis study of *N*-benzylpiperidines as acetylcholinesterase inhibitors. *J. Med. Chem.* **1995**, *39*, 380–387.

(18) Collantes, E.; Tong, W.; Welsh, W. Use of moment of inertia in comparative molecular field analysis to model chromatographic retention of nonpolar solutes. *J. Anal. Chem.* **1996**, *68*, 2038–2043.

(19) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.

(20) Zhang, J.; Madden, T. L. PowerBLAST: A new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res.* **1997**, *7*, 649−656.

(21) Boguski, M. S.; Schuler, G. D. ESTablishing a human transcript map. *Nature* **1995**, *10*, 369−371.

(22) Klopman, G. Artificial intelligence approach to structure−activity studies. Computer automated structure evaluation of biological activity of organic molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321.

(23) Klopman, G. MULTICASE1. A hierarchial computer automated structure evaluation program. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176−184.

(24) Munk, M. E.; Madison, M. S.; Robb, E. W. The neural network as a tool for multispectral interpretation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 231−238.

(25) Klawun, C.; Wilkins, C. L. Joint Neural Network Interpretation of Infrared and Mass Spectra. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 249−257.

(26) Gasteiger, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Finding the 3D structure of a molecule in its IR spectrum. *J. Anal. Chem.* **1997**, *359*, 50−55.

(27) Kuiper, G. G. J. M.; Carlsson, B.; Grandien, K.; Enmark, E.; Haggblad, J.; Nilsson, S.; Gustafsson, J.-A. Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors α and *β*. *Endocrinology* **1997**, *138*, 863−870.

(28) Integrated Spectral Data Base System for Organic Compounds; U.S.A., web site www.aist.go.jp/RIODB/SDBS/.

(29) *The Aldrich Library of ^{13}C and ^{1}H FT NMR Spectra*, 1st ed.; Pouchert, C. J., Behnke, J., Eds.; Aldrich Chemical Co.: 1993; Vol. 1−3.

(30) *Spectral Data for Steroids*; Frenkel, M., Marsh, K. N., Eds.; Thermodynamics Research Center: College Station, TX, 1994.

(31) Cramer, R. D.; Bunce, J. D.; Patterson, D. E. Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18−25.

(32) Blair, R. M.; Fang, H.; Branham, W. S.; Hass, B. S.; Dial, S. L.; Moland, C. L.; Tong, W.; Shi, L.; Perkins, R.; Sheehan, D. M. The estrogen receptor relative binding affinities of 188 natural and xenochemicals: Structural diversity of ligands. *Toxicol. Sci.* **2000**, *54*, 138−153.

(33) *ACD/Labs CNMR software*, version 4.0; Toronto, Canada, released 1998.

(34) Bremser, W. HOSE−A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355−365.

CI0000878