^{13}C NMR Quantitative Spectrometric Data-Activity Relationship (QSDAR) Models of Steroids Binding the Aromatase Enzyme

Richard D. Beger,* Dan A. Buzatu, Jon G. Wilkes, and Jackson O. Lay, Jr.

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

Received January 26, 2001

Five quantitative spectroscopic data-activity relationships (QSDAR) models for 50 steroidal inhibitors binding to aromatase enzyme have been developed based on simulated ^{13}C nuclear magnetic resonance (NMR) data. Three of the models were based on comparative spectral analysis (CoSA), and the two other models were based on comparative structurally assigned spectral analysis (CoSASA). A CoSA QSDAR model based on five principal components had an explained variance (r^2) of 0.78 and a leave-one-out (LOO) cross-validated variance (q^2) of 0.71. A CoSASA model that used the assigned ^{13}C NMR chemical shifts from a steroidal backbone at five selected positions gave an r^2 of 0.75 and a q^2 of 0.66. The ^{13}C NMR chemical shifts from atoms in the steroid template position 9, 6, 3, and 7 each had correlations greater than 0.6 with the relative binding activity to the aromatase enzyme. All five QSDAR models had explained and cross-validated variances that were better than the explained and cross-validated variances from a five structural parameter quantitative structure−activity relationship (QSAR) model of the same compounds. QSAR modeling suffers from errors introduced by the assumptions and approximations used in partial charges, dielectric constants, and the molecular alignment process of one structural conformation. One postulated reason that the variances of QSDAR models are better than the QSAR models is that ^{13}C NMR spectral data, based on quantum mechanical principles, are more reflective of binding than the QSAR model's calculated electrostatic potentials and molecular alignment process. The QSDAR models provide a rapid, simple way to model the steroid inhibitor activity in relation to the aromatase enzyme.

INTRODUCTION

The aromatase enzyme catalyzes the conversion of testosterone to estradiol by the aromatization of the A-ring in steroids.[1,2] Estrogen production from aromatase enzyme activity is important in the evolution and development of estrogen-dependent tumors.[3,4] Inhibition of the aromatase enzyme, a cytochrome P450 complex that converts androgens to estrogens, is therapeutically significant because it may control breast cancer.[4]

Standard three-dimensional quantitative structure−activity relationship (3D-QSAR) models have been produced for 50 steroid inhibitors of the aromatase enzyme.[5] ^{13}C NMR data has been used to produce a reliable classification for spectrometric data-activity relationship (SDAR) models of the estrogen receptor system[6] and monodechlorination rates.[7] We have developed a quantitative relationship between spectra and certain properties or activities for binding to the corticosterone binding globulin[8] and aryl hydrocarbon receptor.[9] Quantitative spectrometric data-activity relationships (QSDAR) is based on the spectral-activity leg in the triangular structure-spectrum-activity relationship. The binding activity of 45 progestagen steroids to a steroid receptor have been quantitatively modeled with simulated ^{13}C NMR spectra by comparative spectral analysis (CoSA).[10] These CoSA models, using simulated ^{13}C NMR data, yielded better correlations and predictions than were seen with comparative molecular field analysis (CoMFA) methods. This paper demonstrates that CoSA of simulated ^{13}C NMR spectral data can be used to produce a reliable quantitative spectrometric data-activity relationship (QSDAR) model of steroids binding to the aromatase enzyme.

QSAR is based on the assumption that there is a relationship between the structure and activity of a compound. QSAR modeling results have been able to show that receptor binding of a compound can be predicted from a combination of electrostatic potentials and geometrical structural analysis.[11−13] However, using a specific molecular structure for computer modeling of each compound dramatically extends the number of calculations required to define the model. Moreover, the selection of the most appropriate 3D structure for each molecule requires a number of assumptions. The necessary simplifying assumptions in some cases give results that are hard to replicate or are inaccurate. An advantage of QSDAR is that it is not necessary to solve any quantum mechanical calculations or use the structures of molecules for electrostatic calculations, as is done in QSAR techniques.[5,14−17]

Using QSAR modeling results, receptor binding of a compound can be predicted based, in part, upon electrostatics and geometrical structure. The ^{13}C NMR spectrum of a compound contains frequencies that correspond directly to the quantum mechanical properties of the ^{13}C nuclear magnetic moment. The quantum mechanical description of magnetic moment, in turn, depends largely on its electrostatic features and geometry.[18] Therefore, we postulated that we could use ^{13}C NMR data in much the same way that QSAR uses comparative molecular field analysis (CoMFA) of

* Corresponding author phone: (870)543-7080; fax: (870)543-7686; e-mail: rbeger@nctr.fda.gov.

10.1021/ci010285e CCC: $20.00
Published on Web 08/03/2001

Table 1. Structures of Steroids Used in QSDAR Models of Binding to the Aromatase Enzyme[a]

no.	binding activity	structure	R_1	R_2	R_3	R_4	R_5
1	−2.92	SA	CH_2OH	=O			
2	−3.54	SA	CH_2OH	OH	H		
3	−3.00	SA	CHO	=O			
4	−3.26	SA	H	=O			
5	−2.62	SA	Me	OH	H		
6	−3.06	SB	CH_2OH	=O			
7	−2.14	SB	CHO	=O			
8	−2.36	SB	H	=O			
9	−1.89	SD	CH_2OH	=O		H	
10	−2.88	SD	CH_2OH	OH	H	H	
11	−2.03	SD	CHO	=O		H	
12	−0.97	SD	Me	=O		H	
13	−2.93	SD	Me	=O		Br	
14	−1.28	SA	Me	=O			
15	−1.23	SB	Me	=O			
16	−2.61	SB	Me	OH	H		
17	−2.36	SD	Me	OH	H	H	
18	−0.65	SF	=O				
19	−2.19	SF	OH	H			
20	−1.03	SH	H	H	H		
21	0.00	SC	Me	=O		H	H
22	0.46	SC	CH_2OH	=O		H	H
23	−0.84	SH	CH_2OH	H	H		
24	0.15	SH	Me	=O			
25	−0.13	SE	=O		=O		CF_2
26	0.87	SIE	=O		H	H	CH_2
27	−0.51	SIE	OH	H	H	H	CH_2
28	−1.35	SC	Me	OH	H	H	H
29	−0.67	SC	CH_2OH	OH	H	H	H
30	−0.89	SC	$MeC(O)OCH_2$	=O		H	H
31	−0.79	SC	Me	=O		H	Br
32	−1.09	SC	Me	=O		H	H
33	−1.08	SC	CF_3	=O		H	H
34	0.56	SI	Me				
35	0.87	SJ	Me				
36	1.56	SI	C_2H_5				
37	0.94	SJ	C_2H_5				
38	0.94	SI	C_3H_7				
39	0.78	SJ	C_3H_7				
40	0.65	SI	C_nH_9				
41	0.53	SJ	C_4H_9				
42	0.21	SI	$CH(CH_3)_2$				
43	0.04	SJ	$CH(CH_3)_2$				
44	−0.04	SI	C_6H_5				
45	0.24	SJ	C_6H_5				
46	−0.24	SI	$CH_2C_6H_5$				
47	0.61	SJ	$CH_2C_6H_5$				
48	0.91	SI	$CH{=}CH_2$				
49	−0.32	SI	C≡CH				
50	0.96	SG					

[a] The structure column refers to Figure 1.

constitutional, topological, geometrical, electrostatic, and quantum descriptors to model receptor binding of a compound.[14−17] By combining the ^{13}C NMR data into a composite set of descriptors for CoSA and putting them into statistical software programs for comparative spectral analysis, it is possible to produce a QSDAR model of the inhibitor compounds binding to the enzyme. In QSDAR, each NMR chemical shift functions as a quantum mechanical identifier of a four- to eight-atom structural moiety.[6−9] These quantum mechanical identifiers are used in a manner similar to that in current QSAR models that break the molecule into secondary structural motifs.

PROCEDURES

The 50 compounds specified in Table 1 and Figure 1 have known steroid inhibitor binding affinities to the enzyme.[5,19−23] Simulated ^{13}C NMR spectra were determined using the ACD Labs CNMR Predictor software, version 4.0.[24] In this software, the predicted NMR spectra are calculated by a substructure similarity technique called HOSE,[25] which correlates similar substructures with similar NMR chemical shifts. Since some of the compounds are in the predictor software spectral database, the resultant predicted spectrum of these compounds is their true experimental NMR spectrum. Simulated ^{13}C NMR spectra were used to make the QSDAR model completely computer driven. No experimental NMR spectra were acquired.

Figure 2 shows a flowchart of the procedures that we used to make the CoSA and CoSASA models. We used the unassigned, simulated ^{13}C NMR data points for CoSA. Unassigned 1D ^{13}C NMR chemical shifts were segregated into bins over a 0−256 ppm range. The CoSA QSDAR models were produced with a spectral bin width of 1.0 ppm. The 1.0 ppm spectral width was used because of convenience and because that spectral width was used in prior QSDAR and SDAR models based on ^{13}C NMR spectral data.[6−9] It was also chosen here to take into account some of the uncertainties produced by use of simulated NMR data. Finally, this width allowed many of the bins to become populated with a significant number of "hits" so meaningful statistical correlations could be obtained.

For CoSA, the ^{13}C NMR chemical shifts were defined as the area under the peak within a certain spectral range and normalized to an integer. A single chemical shift frequency in the 1.0 ppm spectral bin was assigned an area of 100, two chemical shifts in the 1.0 ppm spectral bin had an area of 200, and so forth. This was done so that all the chemical shifts would have a similar signal-to-noise ratio and to eliminate line width variations due to shimming, temperature, and predicted line shapes. In the CoSA model, there were 256 bins, each of which was populated or not depending on the pattern of simulated chemical shifts. This approach did not require an identification of the shift with the carbon that produced it. Therefore, its predictions were not limited to compounds having a steroid backbone. Two CoSA QSDAR models were developed: one model was based on the individual unassigned spectral bins and the other was based on the principal components (PCs) of variation calculated using contributions from each of the 1.0 ppm spectral bins. Both CoSA QSDAR models were evaluated with PLS multiple regression analysis using only the most correlated individual spectral bins or PCs.

Another QSDAR model was produced by using the assigned ^{13}C NMR chemical shifts at the 17 positions in the steroid backbone template, as shown in Figure 3. This requires 17 "bins" in which the corresponding intensity is each carbon's simulated ^{13}C NMR chemical shift. This model combines structural information with the assigned simulated ^{13}C NMR chemical shifts. We call this procedure comparative structurally assigned spectra analysis (CoSASA) because the way in which spectra are defined differs substantially from CoSA methods disclosed above. Two CoSASA QSDAR models were developed: one model was based on the individual assigned carbon atom chemical shifts and the other model was based on the PCs built from all the assigned carbon atom chemical shifts in the steroid backbone.

The errors produced in simulating NMR spectra are propagated through similar structures found in the training

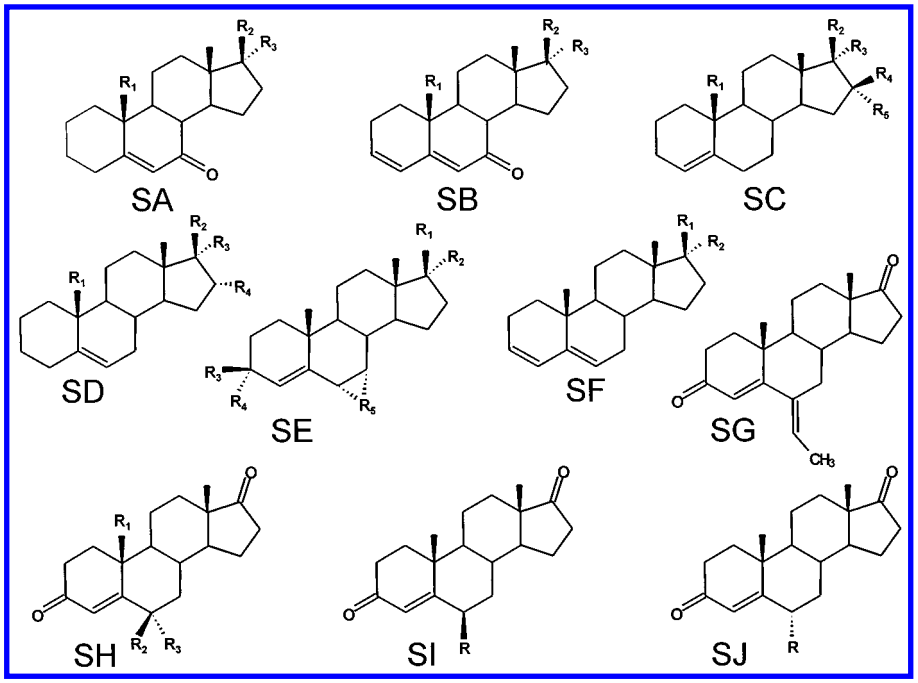

Figure 1. Structures SA−SJ used with Table 1 for the enzyme aromatase steroid series.

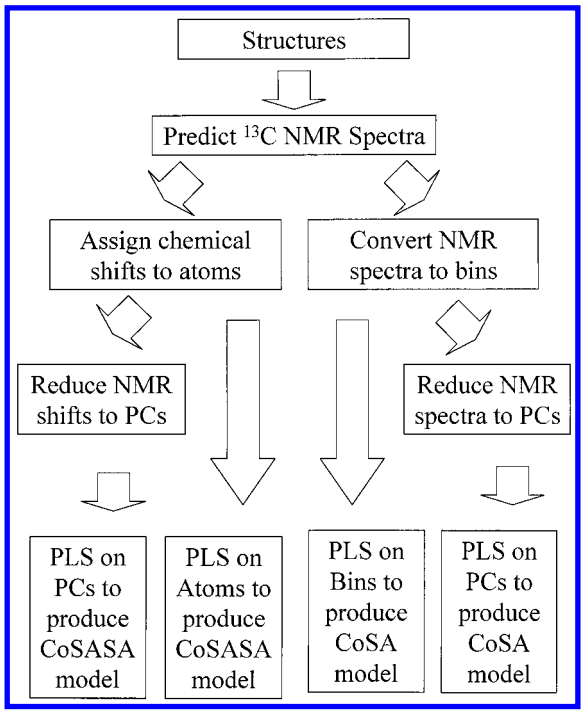

Figure 2. The procedural flowchart for CoSA and CoSASA modeling.

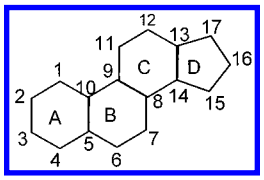

Figure 3. Steroid backbone template carbon atom numbering scheme.

set of the CoSASA and CoSA QSDAR models. This conveniently reduces the effective error when using the training set to predict unknown sample affinities using the predicted spectra of the unknowns. Before PLS multiple regression analysis was done on the simulated ^{13}C NMR data, the spectral bin columns with only zero values and spectral bin columns with only one nonzero number were removed from the data set.

To compare QSDAR to more common QSAR techniques a quantitative structure−activity relationship (QSAR) model for all 50 compounds was built from descriptors that were generated using Cerius2 version4.5.[27] We generated 256 descriptors categorized into constitutional, topological, geometrical, electrostatic, thermodynamic, and E-state descriptors. From these 256 descriptors, PLS multiple linear regression was performed, and the top five structural descriptors were used to produce and cross-validate the QSAR model.

The pattern recognition software used was Statistica version 5.5.[26] The simulated ^{13}C NMR spectroscopic data for all 50 steroids from Table 1 were input into the software. The analysis of all multiple regression QSDAR and one QSAR model were done by the leave-one-out (LOO) cross-validation procedure in which each compound is systematically excluded from the training set, and its inhibitor binding activity is predicted by the model.[26] The cross-validated r^2 (termed q^2) can be derived from $q^2 = 1 - (\text{PRESS})/\text{SD}$. Where PRESS is the sum of the differences between the actual and predicted activity data for each molecule during LOO cross-validation, and SD is the sum of the squared deviations between the measured and mean activities of each molecule in the training set.

A model for the aromatase binding activity (BA) was developed using an artificial neural network (ANN) program, NETS 3.0.[29] This is an error back-propagating neural network that is based on Rummelhart, Hinton, and Williams' generalized delta rule as explained in Chapter 8 of *Parallel Distributed Processing*.[30] Spectral data was put into the network through a series of input nodes and passed through a hidden layer that was connected to an output layer. Thus, the neural network architecture that was used consisted of three layers. An 87 node input layer was connected to a 29 node hidden layer which was then connected to a one node

Table 2. Model Performance Parameters r^2, q^2, and Number and Type of Components Used in the Model

model	r^2	q^2	components used
CoMFA	0.94	0.72	five PCs
QSAR	0.73	0.66	five parameters
CoSASA	0.75	0.66	five atoms
CoSASA	0.75	0.67	five PCs
CoSA	0.82	0.77	five bins
CoSA	0.78	0.71	five PCs
CoSA (neural network)	1.0	0.75	all 87 bins

output layer. The 87 nodes of the input layer corresponded to the number of multiply occupied C^{13} NMR bins for each molecule. The 29 nodes of the hidden layer corresponded to approximately 1/3 times the number of input nodes. The one node of the output layer corresponded to the binding activity for each molecule. This was a fully connected network.

A series of tests were performed on the neural network to arrive at this configuration. The number of hidden nodes was varied during these tests as well as the number of training (back-propagation) cycles. The previously mentioned configuration (87 input, 29 hidden, 1 output) yielded the best results. The optimal results were obtained with 45 000 training cycles. NETS 3.0 uses a sigmoid transfer function, which requires that all of data (input and output) be scaled to a range between 0 and 1. The data for this model was actually scaled over the range 0.1−0.9.

Once the optimal configuration for the network was established, the network's ability to learn the relationship between the NMR spectra and binding affinities (BAs) was tested using a LOO cross-validation. This was implemented by removing the spectrum and BA data for each molecule and training the neural network for 45 000 cycles on the remaining data. After each retraining period, the excluded NMR data was propagated through the network, and the network produced a predicted BA for that molecule. This procedure was repeated until the neural network had predicted BAs for each of the 50 molecules.

RESULTS

We used five components in all the PLS models because that was the maximum shown in the CoMFA analysis.[5] Table 2 contains a comparison of the model performance parameters r^2, q^2, and number of components for the previously published CoMFA model,[5] the QSAR model, the two CoSASA models, and the two CoSA models of steroid inhibition to the enzyme. All PLS QSDAR models have an explained variance (r^2) that is lower than the previous published model of binding to the enzyme but have a LOO cross-validated variance (q^2) that is better or close to the previous published QSAR model of binding to the enzyme.[5] The statistical results were further tested and validated by randomizing the binding activity data. As expected, the statistical correlation using incorrect binding data gave low r^2 values.

Figure 4A is a plot of the predicted binding versus experimental binding for the 1.0 ppm resolution CoSA models when using only the five most correlated spectral bins. The explained correlation (r^2) of this model is 0.82 and LOO cross-validated variance (q^2) is 0.77. These five spectral bins corresponded to the chemical shift (cs) frequencies between 34 ppm <= cs < 35 ppm, 19 ppm <= cs < 20 ppm, 35 ppm <= cs < 36 ppm, 129 ppm <= cs < 130, and 18 ppm <= cs < 19 ppm. The 34 ppm spectral bin was primarily associated with positions 6 and 7 on the steroid template. Since the spectra were not assigned, this identification was not used in the correlation analysis. Similarly, the 19 and 18 ppm spectral bins were associated with the methyl position 19 above position 16 on the steroid template. Figure 4B is a plot of the predicted binding versus experimental binding for the 1.0 ppm resolution CoSA models when using five principal components, based on 87 spectral bins. The explained correlation (r^2) of this model is 0.78, and LOO cross-validated variance (q^2) is 0.71.

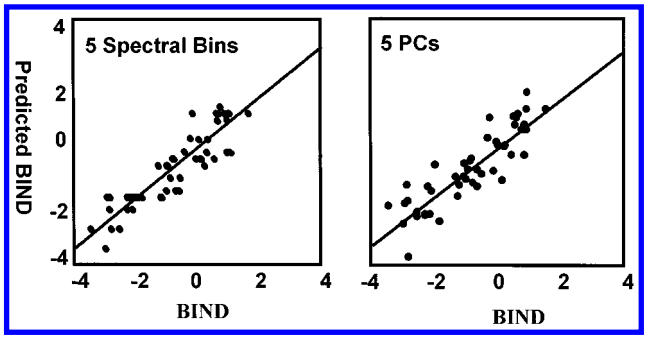

Figure 4. Plot of the predicted binding versus experimental binding for the CoSA QSDAR models: (A) five spectral bins and (B) five principal components.

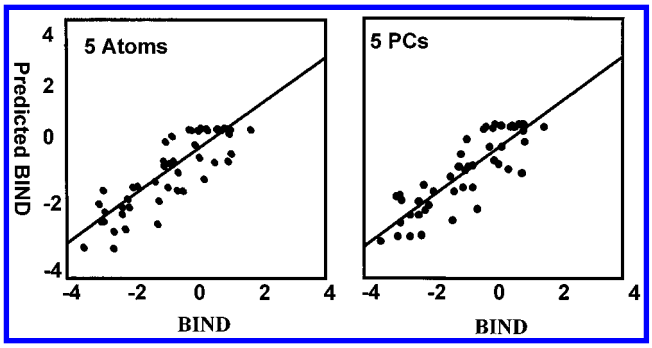

Figure 5. Plot of the predicted binding versus experimental binding for CoSA QSDAR models: (A) five atoms and (B) five principal components.

Figure 5A is a plot of the predicted versus experimental binding for the steroid backbone CoSASA model when using only the five most correlated assigned ^{13}C NMR chemical shifts. The five ^{13}C NMR chemical shifts used in this model were the steroid template positions 3, 9, 6, 7, and 12. Individual ^{13}C NMR chemical shifts from atoms in the steroid template position 3, 9, 6, and 7 all had correlations greater than 0.6 with the relative binding activity to the enzyme. The explained correlation to the binding activity data (r^2) of this model is 0.75, and the LOO cross-validated variance (q^2) is 0.66, which indicates self-consistency and good predictive capability, respectively. Figure 5B is a similar plot for the CoSASA model using the five most correlated principal components, each including contributions from all 17 carbon atoms' assigned chemical shifts. The explained correlation (r^2) of this model is 0.75, and the LOO cross-validated variance (q^2) is 0.67, also indicating self-consistency and good predictive capability.

In the 50 compound training set the steroid backbone template position 3 showed the highest correlation to inhibit aromatase. The polar negative keto group at position 3

(compounds **20, 23−25, 34−50**) was easily detected and correlated to enzyme inhibition when the chemical shift of position 3 was between 198 and 200 ppm. Similarily, a correlation between a polar negative region around atom positions 2−4 and inhibitor activity to the enzyme was found in the previous QSAR model.[5] The keto group at position 7 (compounds **1−8, 14−16**) was indentified with chemical shifts between 200 and 201 ppm and correlated to weak inhibitor activity. Similarily, a correlation between the chemical shift from the position 7 keto group and weak inhibitor activity to aromatase was found in the previous QSAR model.[5] The weak inhibitor activity of the androst-5-ene compounds was identified by chemical shifts between 117 and 127 ppm from position 6. Again, a similar correlation was found in QSAR models between all the androst-5-ene compounds and lower inhibitor activity to the enzyme.[5] In contrast to these comparisons, the strong correlation of the chemical shift from position 9 with aromatase inhibitor activity was unexpected. The chemical shifts between 55 and 58 ppm in position 9 are for steroids with 6-akyl substitutions. These are stronger inhibitors to the enzyme. It is significant that a correlation around position 9 and inhibitor activity to the enzyme was not reported in the previous QSAR model, whereas the CoSASA QSDAR model based on chemical shifts from five atoms correctly showed a high correlation of position 9 to inhibitor activity.

The CoSASA model was based on the chemical shift frequency change of atoms from carbon atoms at the 3, 9, 6, 7, and 12 positions on the template. The CoSA model with chemical shifts not assigned to particular carbons showed the strongest correlations when substituents caused the chemical shifts of atoms in the 6, 7, 9, and 13 positions to appear in those bins. By CoSASA, the chemical shift from position 9 in the steroid backbone had the strongest correlation to inhibitor activity. Position 9 is the middle of the steroid molecule (four carbon bonds to terminal positions of 3 and 17), and it was not directly involved in most of the side chain changes for the 50 steroids of the training set. However, it was able to monitor, and respond to, the side chain changes anywhere in the whole molecule. The reason the CoSASA and CoSA models are not based on exactly the same atoms is that the chemical shift frequencies from the atoms in the COSASA model may fall into chemical shift frequency ranges from the CoSA model that do not have a high correlation to the binding activity. In terms of statistical measures of quality, all four PLS QSDAR models using five components had a standard error (SE) between 0.5 and 0.7 and $p < 0.00001$.

Figure 6 (A and B) plots the explained variance (r^2) and cross-validated variance (q^2) results for the CoSA models as a function of the number of spectral bins or principal components. Figure 6 (C and D) plots the r^2 and q^2 results for the CoSASA models versus the number of atoms or principal components. All plots in Figure 6 show that the r^2 and q^2 consistently increase as a function of the number atoms, spectral bins, or principal components used in any of the QSDAR models. Figure 6 shows that the PLS QSDAR models based on bins and atoms have r^2 and q^2 values that were still rising at five components. Because of this we believe that we could have used more spectral bins or atoms in the QSDAR models to get higher r^2 and q^2 correlations, but we stopped at five components so we could compare

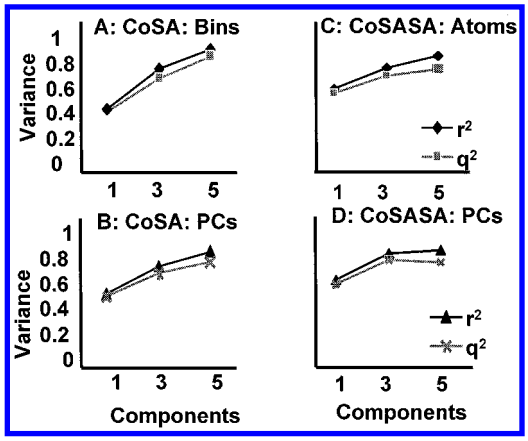

Figure 6. Plot of CoSA ((A) five bins and (B) five PCs) and CoSASA ((C) five atoms and (D) PCs) QSDAR model parameters r^2 and q^2 versus number of atoms or number of principal components used to produce the model.

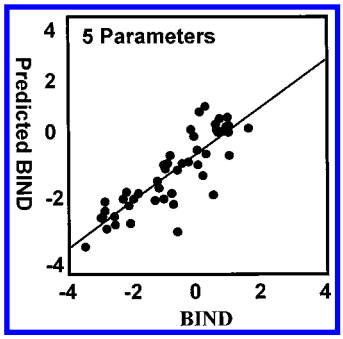

Figure 7. Plot of the predicted binding versus experimental binding for the QSAR model based on five structural parameters.

our results with the previously published QSAR CoMFA results.[5] Figure 6 also shows that the QSDAR models based on principal components have a lower r^2 and q^2 than corresponding QSDAR models based on individual atoms or spectral bins. This is because the principal components include chemical shift contributions that do not have a high correlation with inhibitor activity. The "irrelevant" information dilutes the ability of the model to make a good quality prediction of binding affinity.

Figure 7 is a plot of the predicted binding versus experimental binding for the five structural parameter QSAR model when using, as in the QSDAR models, the five most highly correlated structural parameters. The r^2 of this model is 0.73 and the q^2 is 0.66, which indicates self-consistency and good predictive capability. However, the quality of the fit for the QSAR structural parameter model is generally inferior to that of the QSDAR models, especially the QSDAR CoSA model based on the five most correlated bins.

DISCUSSION AND CONCLUSIONS

A possible explanation for the fact that the explained variance of all the QSDAR models was consistently lower than the explained variance of the CoMFA QSAR model is that ^{13}C NMR based on substructure predictions are only a two-dimensional representation of a three-dimensional molecule. Important 3-D information may be lost in the conversion. One possible explanation for the observation that the q^2 of QSDAR models are just as good for CoSA as the QSAR model is that even the predicted ^{13}C NMR spectra are more representative of molecular activity than QSAR

models based on calculating electrostatic fields and requiring a molecular alignment process. Another possible reason that the q^2 of QSDAR models are just as good as or better than QSAR is that the ^{13}C NMR bins are more unrelated to each other (orthogonal), whereas the variables for QSAR modeling are highly interrelated (nonorthogonal). A possible explanation for the fact that the q^2 of the CoSASA models are lower than those of the CoSA models is that some of the side chain chemical shift information (not used by the CoSASA models) was needed for a better cross-validated model.

The predicted activities in the CoSASA model seem to be restricted to between −3.25 and 0.65, while the experimental data lay between −3.54 and 1.56. One reason is the training compounds for this model have experimental binding data not weighted evenly across the whole range of binding activity. The maximum and minimum extreme experimental binding data points are singular points that are over 0.5 log units away from the next nearest experimental binding data and therefore have very little weight in any of the models. The five structural parameter QSAR does worse in predicting compound 3 (the maximum extreme) of −3.10 as opposed to the −3.25 predicted by the CoSASA model. The five-parameter QSAR does better in predicting a minimum extreme of 1.36 but for a molecule that has an experimental binding of 0.21 that ranks 36 out of 50 in binding activity. The five structural parameter QSAR prediction for compound 36 (the experimental minimum) is 0.46 which is about the same as our CoSASA prediction of 0.48 for compound 36. The CoSA model based on bins was able to correctly rank as extremes both compound 3 (with a predicted binding activity of −3.5) and compound 36 (predicted with the second lowest binding activity of 1.33). The CoMFA model produced the largest deviation between actual and experimental binding activity for compound 36.[5]

Our QSAR model was built from 257 constitutional, topological, geometrical, electrostatic, thermodynamic, and E-state descriptors of which the top five were actually incorporated into PLS analysis. This model had an r^2 and a q^2 that were very similar to both CoSASA models. The r^2 and q^2 values of the five structural parameter QSAR model were lower than both CoSA models. It is interesting to note that three of the five structural parameters it selected from the 257 parameters were for the number of carbon atom types in the molecule. Essentially, our 1D ^{13}C NMR CoSA modeling is accomplishing the same thing but in a much more simple and sensitive way.

Bursi et al.[10] used 8192 spectral bins and chemical shift peak shaping for their CoSA model. In contrast, the two CoSA models started with only 256 spectral bins that were reduced to 87 spectral bins when all the columns with zeros or with only one nonzero entry were removed and the two CoSASA models were based on 17 bins. CoSA models condense the information content of the spectrum by reducing the number of spectral bins and losing the shape of the chemical shift peak. The NMR chemical shift frequency has information about atom environment, but the peak shape is greatly affected by shimming and temperature. From these results, we conclude that the two CoSA models retained enough information to produce reliable models of inhibitor binding to the enzyme. Further, use of simulated rather than experimental ^{13}C NMR data provides a facile and rapid way to model binding of structurally similar compounds for which real spectra are not available. Because the simulated ^{13}C NMR spectra is assigned during the prediction, it is easy to implement CoSASA models. The CoSASA modeling attempts to combine two-dimensional QSAR modeling approaches[31,32] with our nonassigned QSDAR CoSA modeling technique. The CoSASA models have results that are poorer than CoSA modeling results, because the CoSA has the added side chain information that is not present in CoSASA models.[8] This was unexpected but significant because QSDAR CoSA is easier to implement in an objective way.

The Artificial Neural Network CoSA (ANN CoSA) model is different from the other CoSA models in that all of the ^{13}C NMR data was used to build the model. All of the 87 ^{13}C NMR bins were used as input for the neural network. The neural network model was able to reproduce the training data as perfect as possible with a r^2 value of 1.00. The model was evaluated as already mentioned for predictive capability with a LOO cross-validation. This analysis resulted in a q^2 value of 0.75. This is slightly less precise in predicting unknowns than the five bin CoSA model, but again emphasis must be made on the fact that all of the NMR data was used in the development of this model. Unlike the five bin CoSA model, no decisions were made regarding the exclusion of input data. Neural networks are known for their ability to handle noisy data and using predicted NMR spectrum presents no difficulty on that score. A drawback of the neural network approach is the time required for the development of the model. This includes the time to develop the optimal network architecture and learning parameters as well as the time that it takes to retrain the network to perform a LOO cross-validation for evaluation purposes.

The accuracy of the QSDAR model predictions shows that simulated ^{13}C NMR spectra can be used in PLS regression analysis and with ANNs to model binding of structurally similar compounds to a receptor. Like electrotopological-states calculations, simulated ^{13}C NMR spectral data have information about the electronic structure and topological environment of an atom[33,34] to produce a model that has a better LOO cross-validated variance that was seen in QSAR and SAR models based on calculations for electrostatics and steric interactions. We took precautions by reducing the resolution of the chemical shift peak to minimize the effect of uncertainties in the simulation of ^{13}C NMR data that could lead to systematic errors. We were unable to obtain all of the compounds and therefore are unable to produce all of the experimental NMR spectra and determine the exact degree of systematic error from the predicted spectra. Nevertheless, the cross-validated variance of QSDAR models based on simulated ^{13}C NMR data may improve as the errors introduced by simulation of the ^{13}C NMR data are further reduced using later editions of the spectral simulation programs.

A major benefit of the QSDAR model approach is that simulated spectra can be saved and used for other compound-receptor systems by simply exchanging the relative binding affinities. The ^{13}C NMR spectral data do not change and can be used as comprehensive descriptors in new QSDAR models for many other biological endpoints.

ACKNOWLEDGMENT

We thank Christie McKenzie a teacher from the STRIVE (Science Teachers Research Involvement for Vital Education)

program at NCTR who performed data management and some data analysis in this work. We thank Dr. Weida Tong the senior computational chemist from the ROW Sciences at NCTR who helped us perform QSAR modeling based on structural parameters.

REFERENCES AND NOTES

(1) Kellis, J. J.; Vickery, L. E. Purification and characterization of human placental cytochrome P-450. *J. Biol. Chem*. **1987**, *262*, 4413−4420.
(2) Chen, S.; Besman, M. J.; Shively, J. E.; Yanagibashi, K.; Hall, P. F. Human Aromatase. *Drug Metab. Rev.* **1989**, *20*, 511−517.
(3) Brodie, A. M. Aromatase inhibitors in the treatment of breast cancer. *J. Steroid Biochem. Mol. Biol.* **1994**, *49*, 281−287.
(4) Brodie, A. M.; Santen, R. J. Aromatase and its inhibitors in breast cancer treatment−overview and perspective. *Breast Cancer Res. Treat.* **1994**, *30*, 1−6.
(5) Oprea, T. I.; Garcia, A. E. Three-dimensional quantitative structure−activity relationships of steroid aromatase inhibitors. *J. Comput.-Aided Mol. Design*. **1996**, *10*, 186−200.
(6) Beger, R.; Freeman, J.; Lay, J., Jr.; Wilkes, J.; Miller, D. ^{13}C NMR and EI Mass Spectrometric Data-Activity Relationship (SDAR) Model of Estrogen Receptor Binding. *Toxicol., Appl. Pharmacol.* **2000**, *169*, 17−25.
(7) Beger, R.; Freeman, J.; Lay, J., Jr.; Wilkes, J.; Miller, D. Producing ^{13}C NMR, Infrared Absorption and EI Mass Spectrometric Data Models of the Monodechlorination of Chlorobenzenes, Chlorophenols, and Chloroanilines. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1449−1455.
(8) Beger, R. D.; Wilkes, J. G. Developing ^{13}C NMR quantitative Spectrometric data-activity relationship (QSDAR) Models to the Corticosteroid Binding Globulin. *J. Comput.-Aided Mol. Design* **2001**, in press.
(9) Beger, R. D.; Wilkes, J. G. Models of Polychlorinated Dibenzodioxins, Dibenzofurans, and Biphenyls Binding Affinity to the Aryl Hydrocarbon Receptor Developed Using ^{13}C NMR Data. *J. Chem. Inf. Comput. Sci*. **2001**, in press.
(10) Bursi, R.; Dao, T.; van Wilk, T.; de Gooyer, M.; Kellenbach, E.; Verwer, P. Comparative spectra analysis (CoSA): Spectra as three-dimensional molecular descriptors for the prediction of biological activities. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 861−867.
(11) 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.
(12) 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.
(13) Hansch, C.; Leo, A. *Exploring QSAR − Fundamentals and applications in chemistry and biology*; The American Chemical Society: Washington, DC, 1995.
(14) 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.
(15) 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.
(16) Fujita, T.; Iwasa, J.; Hansch, C. A new substituent constant, π, derived from partition coefficient. *J. Am. Chem. Soc.* **1964**, *86*, 5175−5180.
(17) Branbury, S. P. Quantitative structure−activity relationship and ecological risk assessment: an overview of predictive aquatic toxicology research. *Toxicology* **1995**, *25*, 67−89.
(18) Emsley, J. W.; Feeney, J.; Sutcliffe, L. H. *High-Resolution Nuclear Magnetic Resonance*; Pergamon Press Ltd.: Oxford, 1965; Vol. I, pp 1−287.
(19) Numazawa, M.; Mutsumi, A.; Hoshi, K.; Koikie, R. 19-hydroxy-4-androsten-17-one: Potential competitive inhibitor of estrogen biosynthesis. *Biochem. Biophys. Res. Comm.* **1989**, *160*, 1009−1014.
(20) Numazawa, M.; Mutsumi, A. 6α,7α-cyclopropane derivatives of androst-4-ene: A novel class of competitive aromatase inhibitors. *Biochem. Biophys. Res. Comm.* **1991**, *177*, 401−406.
(21) Numazawa, M.; Mutsumi, A.; Hoshi, K.; Oshibe, M.; Ishikawa, E.; Kigawa, H. Synthesis and biochemical studies of 16- or 19-substituted androst-4-enes as aromatase inhibitors. *J. Med. Chem.* **1991**, *34*, 2496−2504.
(22) Numazawa, M.; Oshibe, M. 6-Akyl- and 6-arylandrost-4-ene-3, 17-diones as aromatase inhibitors. Synthesis and structure−activity relationships. *J. Med. Chem.* **1994**, *37*, 1312−1319.
(23) Numazawa, M.; Mutsumi, A.; Tachibana, M.; Hoshi, K. Synthesis of androst-5-en-7-ones and androst-3,5-diene-7-ones and their related 7-deosy analogues as conformational and catalytic probes for the active site of aromatase. *J. Med. Chem.* **1994**, *37*, 2198−2205.
(24) ACD/Labs CNMR software version 4.0; Toronto, Canada.
(25) Bremser, W. HOSE − a Novel substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355−365.
(26) StatSoft Statistica software version 5.5; Tulsa, OK.
(27) Accelrys Cerius2 version 4.5; San Diego, CA.
(28) 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.
(29) Baffes, P. T.; Shelton, R. O.; Phillips, T. A. *NETS Version 3.0*; Technology Branch, Lyndon B. Johnson Space Center: 1991.
(30) Rumelhart, D. E.; McClelland, T. L. *Parallel Distributed Processing*; Brandford Books/MIT Press: Cambridge, MA, 1986.
(31) 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.
(32) Klopman, G. MULTICASE1. A hierarchial computer automated structure evaluation program. *Quant. Struct. Act. Relat.* **1992**, *11*, 176−184.
(33) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. E-state fields: Applications to 3D QSAR. *J. Comput.-Aided Mol. Design* **1996**, *10*, 513−520.
(34) De Gregorio, C.; Kier, L. B.; Hall, L. H. QSAR modeling with electrotopological state indices: Corticosteroids. *J. Comput.-Aided Mol. Design* **1998**, *12*, 557−561.

CI010285E