Models of Steroid Binding Based on the Minimum Deviation of Structurally Assigned ^{13}C NMR Spectra Analysis (MiDSASA)

Richard D. Beger,*,† Stephen Harris,‡ and Qian Xie‡

Division of Chemistry, National Center for Toxicological Research, Food and Drug Administration, Jefferson, Arkansas 72079-9502, and Northrop Grumman Corporation, Jefferson, Arkansas 72079

Received February 25, 2004

This paper develops a quantitative k-nearest neighbors modeling technique. The technique is used to demonstrate that a compound's biological binding activity to a receptor can be calculated from the minimum of the square root of the sum of squared deviations (SSSD) of a structurally assigned chemical shift on a template between the unknown compound to be predicted and a set of known compounds with known activities. When building models of biological activity, nonlinear relationships are built into the input training data. If a model is developed by selecting only compounds with minimum structurally assigned chemical shift deviations from the unknown compound, some of the nonlinear relationships can be removed. The smaller the total chemical shift deviation between a compound with known activity and another compound with unknown activity, the more likely it will have similar biological, chemical, and physical properties. This means that a model can be produced without rigorous statistics or neural networks. This technique is similar to structure−activity relationship (SAR) modeling, but instead of relying on substructure fragments to produce a model, this new model is based on minimum chemical shift differences on those substructure fragments. We refer to this method as minimum deviation of structurally assigned spectra analysis (MiDSASA) modeling. Modeling by the minimum deviation concept can be applied to other chemoinformatic data analyses such as metabolite concentrations in metabolic pathways for metabolomics research. A MiDSASA template model for 30 steroids binding the corticosterone binding globulin based on the activity factors of the two nearest compounds had a correlation of 0.88. A MiDSASA template model for 50 steroids binding the aromatse enzyme based on the average activity of the four nearest compounds had a correlation of 0.71.

INTRODUCTION

Many different types of models have been developed to predict the binding activity for the steroid-receptor systems.[1−11] These modeling techniques include the standard quantitative structure−activity relationship (QSAR),[2−4] the hybrid electrotopological state (E-state),[5] the self-organizing map (SOM),[6] and the combination QSAR E-state methods.[7] We have developed quantitative models of steroid binding activity based on ^{13}C NMR spectra comparative spectral analysis (CoSA) and comparative structurally assigned spectral analysis (CoSASA).[8,9] Combining NMR spectral information with structural information in a 3D-connectivity matrix led to the development of three-dimensional quantitative spectrometric data-activity relationship (3D-QSDAR) modeling.[10,11] Through-bond nearest neighbors and through-space distance-related connectivity spectral slices from the 3D-connectivity matrix are used to produce a relationship to biological binding activity. We referred to this technique as comparative structural connectivity spectra analysis (CoSCoSA) modeling.[10,11] The CoSA and CoSCoSA models using simulated ^{13}C NMR data yielded higher cross-validated correlations than were seen with comparative molecular field analysis (CoMFA) methods. However, CoSA and CoSCSA modeling used ^{13}C NMR spectrometric data as a set of individual discrete descriptors and not as a continuum. In this paper, we demonstrate how ^{13}C NMR spectrometric data can be treated as a continuum of energy states when applied to a steroid template.

Steroids are necessary for normal development and reproductive development.[12] Corticosteroids are synthesized in the adrenal gland and transported by the corticosterone binding globulin.[1,2,13] It is proposed that the primary purpose of the corticosteroids, including cortisol (hydrocortisone), is to help the body resist infection.[13,14] The male steroid testosterone is catalyzed to the estradiol by the enzyme aromatase. The aromatase enzyme gets its name because it can convert the A-ring in steroids to an aromatic ring.[15,16] Aromatase is a cytochrome P450 complex that converts androgens to estrogens. Estrogen production from aromatase enzyme activity is important in the evolution and development of estrogen-dependent tumors.[17,18] Inhibition of the aromatase enzyme is therapeutically significant because it may control breast cancer.[18]

The ^{13}C NMR spectrum of a compound contains frequencies that correspond directly to the quantum mechanical properties of a nuclear magnetic moment. The diamagnetic quantum mechanical description of the moment depends largely on its electrostatic features and geometry.[19] Ab initio quantum mechanical calculations of ^{13}C chemical shift tensors in proteins reveal that they are dependent on the structural environment.[20] ACD Labs now offers software that will predict ^{13}C NMR one-dimensional and two-dimensional spectra.[21] The frequencies obtained from ^{13}C NMR spectroscopic data correspond directly to the energies obtained when solving the quantum mechanical Schrödinger equation for a

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

† National Center for Toxicological Research, Food and Drug Administration.

‡ Northrop Grumman Corporation.

10.1021/ci049925e CCC: $27.50
Published on Web 05/07/2004

Table 1. Structures of Corticosteroids Used in QSDAR Models of Corticosteroid Binding Globulin Data

no.	structure[a]	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}
1	B	OH	H	H	H	OH	H				
2	E	OH	OH	H							
3	C	=O	H	=O				H	H	H	H
4	B	OH	H	H	H	=O					
5	C	=O	OH	$COCH_2OH$	H			H	H	H	H
6	C	=O	OH	$COCH_2OH$	OH			H	H	H	H
7	C	=O	=O	$COCH_2OH$	OH				H	H	H
8	E	OH	=O								
9	C	=O	H	$COCH_2OH$	H			H	H	H	H
10	C	=O	H	$COCH_2OH$	OH			H	H	H	H
11	B	=O		H	H	OH	H				
12	D	OH	OH	H	H						
13	D	OH	OH	H	OH						
14	D	OH	=O		H						
15	B	H	OH	H	H	=O					
16	E	OH	COMe	H							
17	E	OH	COMe	OH							
18	C	=O	H	COMe	H			H	H	H	H
19	C	=O	H	COMe	OH			H	H	H	H
20	C	=O	H	OH	H			H	H	H	H
21	F	=O	OH	$COCH_2OH$	OH						
22	C	=O	OH	$COCH_2OCOMe$				H	H	H	H
23	C	=O	=O	COMe	H				H	H	H
24	C	=O	H	$COCH_2OH$	H			OH	H	H	H
25[b]	C	=O	H	OH	H			H	H	H	H
26	C	=O	H	COMe	OH			H	OH	H	H
27	C	=O	H	COMe	H			H	Me	H	H
28[b]	C	=O	H	COMe	H			H	H	H	H
29	C	=O	OH	$COCH_2OH$	OH			H	H	Me	H
30	C	=O	OH	$COCH_2OH$	OH			H	H	Me	F

[a] Structures according to refs 1, 22, and 23. [b] H (hydrogen) instead of Me at C_{10} steroid skeleton.

nuclear magnetic moment transition.[19] The NMR chemical shifts (quantum energies) are strongly dependent on the electrostatic potential energy at the carbon nucleus and the type of orbital (wave function) surrounding the carbon nucleus. Both electrostatics and electron orbital wave function have been shown to be important in models of biological activity. The ^{13}C NMR spectrum is similar to the E-state index calculations because they are both largely dependent on electrostatics and valence states.

The chemical shifts in a NMR spectrum are part of an energy continuum. Our previous QSDAR modeling work lost the continuum concept when the chemical shifts in the spectra were binned. This paper demonstrates a method to model biological activity that are calculated from, the set of compounds with the minimum of square root of summed squared deviations (SSSD) between structurally assigned chemical shift differences of a compound with known biological activity and an unknown compound to be predicted by the model. This is similar to structure−activity relationship (SAR) modeling[22,23] but instead of relying on substructure fragments to produce a model, this new model is based on the structures that are most similar based on minimum differences in structurally assigned chemical shifts in the substructure fragments.

PROCEDURES

The 30 steroids specified in Table 1 and Figure 1 have known steroid inhibitor binding affinities to the corticosteroid binding globulin.[1−3,5,6,8,10] The 50 steroids specified in Table 2 and Figure 2 have known steroid inhibitor binding affinities to the aromatase enzyme.[4,9,11] All the ^{13}C NMR spectra were simulated using the ACD Labs CNMR predictor software, version 5.0.[21] The predicted NMR spectra were calculated by

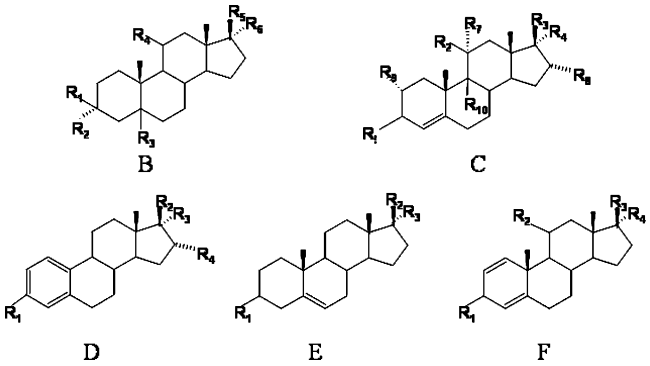

Figure 1. Structures B−F used with Table 1 for the corticosteroid binding globulin steroid series.

a substructure similarity technique called HOSE,[24] which correlates similar structures with similar NMR chemical shifts. We are able to use predicted ^{13}C chemical shifts over experimental chemical shifts because the average uncertainty is very low for ^{13}C NMR prediction. We previously reported the average uncertainty of the predicted ^{13}C NMR data of all the carbon nuclei in the 30 steroid compounds used in the corticosterone binding activity models was 0.53 ppm.[8] The average uncertainty is much lower than the dynamic range of ^{13}C chemical shifts (10−222 ppm) in the study of steroids binding the corticosteroid binding globulin and aromatase enzyme.

One QSDAR model was produced by using the assigned ^{13}C NMR chemical shifts at the 21 positions in the steroid backbone templates, as shown in Figure 3 for corticosterone binding steroids and Figure 4 for aromatase binding steroids. Each molecule requires 21 "bins" in which the corresponding intensity is each carbon's simulated ^{13}C NMR chemical shift. This model combines structural information with the assigned

Table 2. Structures of Steroids Used in Models of Binding to the Aromatase Enzyme

no.	structure	R_1	R_2	R_3	R_4	R_5
1	SA	CH_2OH	=O			
2	SA	CH_2OH	OH	H		
3	SA	CHO	=O			
4	SA	H	=O			
5	SA	Me	OH	H		
6	SB	CH_2OH	=O			
7	SB	CHO	=O			
8	SB	H	=O			
9	SD	CH_2OH	=O		H	
10	SD	CH_2OH	OH	H	H	
11	SD	CHO	=O		H	
12	SD	Me	=O		H	
13	SD	Me	=O		Br	
14	SA	Me	=O			
15	SB	Me	=O			
16	SB	Me	OH	H		
17	SD	Me	OH	H	H	
18	SF	=O				
19	SF	OH	H			
20	SH	H	H	H		
21	SC	Me	=O		H	H
22	SC	CH_2OH	=O		H	H
23	SH	CH_2OH	H	H		
24	SH	Me	=O			
25	SE	=O		=O		CF_2
26	SE	=O		H	H	CH_2
27	SE	OH	H	H	H	CH_2
28	SC	Me	OH	H	H	H
29	SC	CH_2OH	OH	H	H	H
30	SC	$MeC(O)OCH_2$	=O		H	H
31	SC	Me	=O		H	Br
32	SC	Me	=O		Br	H
33	SC	CF_3	=O		H	H
34	SI	Me				
35	SJ	Me				
36	SI	C_2H_5				
37	SJ	C_2H_5				
38	SI	C_3H_7				
39	SJ	C_3H_7				
40	SI	C_nH_9				
41	SJ	C_4H_9				
42	SI	$CH(CH_3)_2$				
43	SJ	$CH(CH_3)_2$				
44	SI	C_6H_5				
45	SJ	C_6H_5				
46	SI	$CH_2C_6H_5$				
47	SJ	$CH_2C_6H_5$				
48	SI	$CH{=}CH_2$				
49	SI	C=CH				
50	SG					

^{13}C NMR chemical shifts. The ^{13}C NMR chemical shifts were saved as a table in EXCEL format for each structurally assigned position. When there was no carbon atom in positions 18, 19, 20, or 21, a zero was entered for the ^{13}C NMR chemical shift. In this particular model, where the lowest chemical shift was 10 ppm, using 0 in template positions where no carbon was available was not problematic. In future studies the penalty for not having an atom in a template position may need to be more severe.

The square root of the sum of the squared differences (SSSD) of the structurally assigned chemical shifts at all the points on the template was calculated for each compound with respect to a test compound. The compounds with minimum deviation in SSSD from an unknown compound were used to produce a model. We refer to this method as minimum deviation of structurally assigned spectra analysis (MiDSASA) modeling. A small subset (usually ~10%) of

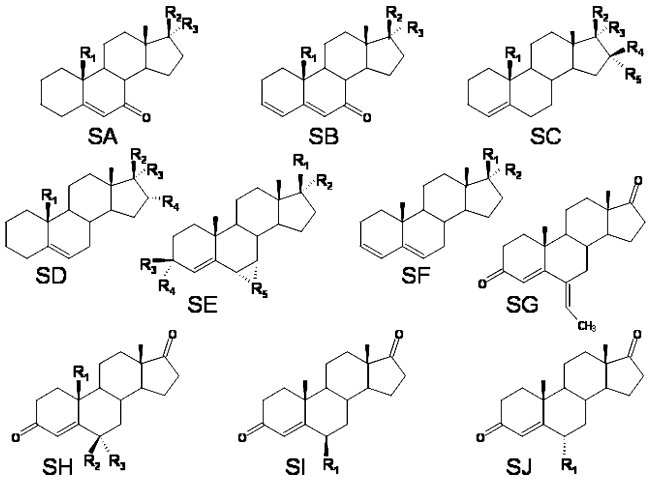

Figure 2. Structures SA−SE used with Table 2 for the aromatase steroid series.

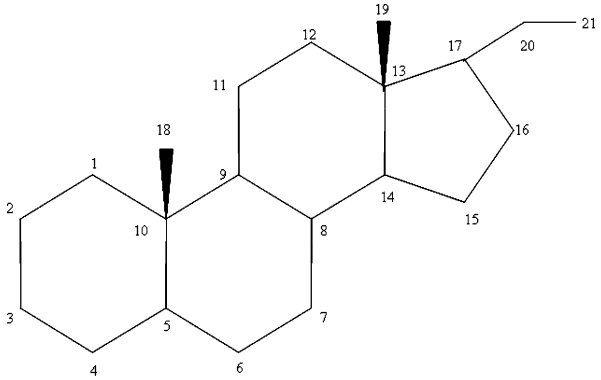

Figure 3. Carbon atom template numbering for corticosterone steroids.

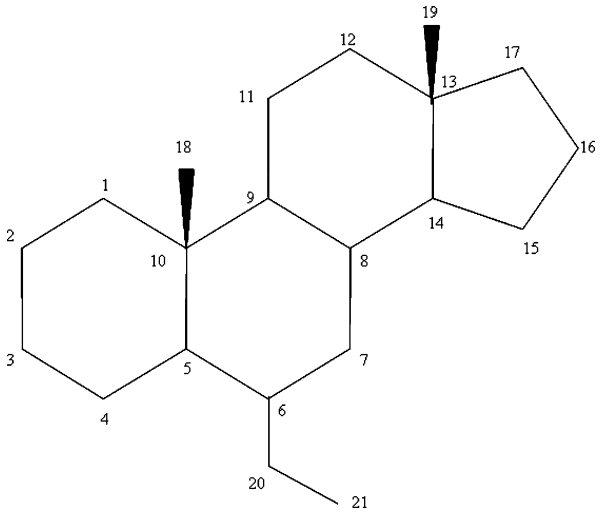

Figure 4. Carbon atom template numbering for aromatase steroids.

compounds having minimum SSSD deviations from the unknown was then extracted from the complete set. Normalized weighted activity factors were then calculated for each compound *i* in the subset using the following formula:

$$\text{activity factor}_i = \frac{\Sigma\text{SSSD (all compounds in subset)} - \text{SSSD}_i}{(N\text{-}1) * \Sigma\text{SSSD (all compounds in subset)}} * \text{activity}_i \quad (1)$$

The subscript *i* denotes the known compounds in the subset of compounds closest in SSSD to the unknown. Activity factors could theoretically be calculated for any number of compounds from the total data set, but we would never use more than 10% of the data set. The factor $1/(N-1)$ makes the sum of all normalized weights equal to one, and in the case for $N = 3$, the normalization factor is 0.5. N is the number of compounds selected for model. The predicted activity of a compound can be expressed as the sum of the activity factors, such that

$$\text{predicted activity} = \text{activity factor}_1 + \text{activity factor}_2 + ... \quad (2)$$

For each compound, the predicted activity and known activity are used in a two-dimensional graph to produce the correlation of the MiDSASA template model.

The concept of breaking the template into substructure pieces can be treated similarly with each of the substructures having a predicted activity factor for each fragment. First the template is divided into a set of substructure fragments. Equation 1 is then applied to each substructure fragment from the compound instead of the whole compound template. The minimum SSSD between chemical shifts on each substructure fragment is then used to produce predicted activity factors for each fragment. Linear regression can be applied to the activity factors for each substructure fragment_i to produce an overall activity modeling equation.

$$\text{predicted activity} = \text{F1} * \text{activity factor}_1 + \text{F2} * \text{activity factor}_2 + ... \quad (3)$$

F1 and F2 are linear regression coefficients that are obtained during partial least squares linear regression. For the corticosterone binding MiDSASA fragment models, we divided the 21-atom template into three fragment parts. Fragment 1 consisted of all positions within 2 bonds of atom 3 (template positions 1−5), fragment 2 was made up of all atoms within 2 bonds of atom 17 (template positions 13−17 and 19−21), and fragment 3 represented the remainder of template (template positions 6−12 and 18).

For the aromatase enzyme binding MiDSASA fragment models, we divided the 21-atom template into four fragment parts. We chose fragment 1 as all positions within 1 bond of atom 3 (template positions 2−4), fragment 2 was made up of all atoms within 2 bonds of atom 17 (template positions 13−17 and 19), fragment 3 as near atom 6 (template positions 5−7, 20, and 21), and fragment 4 as near atom 18 (template positions 1, 10, 9, and 18). The different fragments used in the corticosterone and aromatase models were selected to reflect the structural positions of side chain changes in each individual data set and prevent any fragment overlap.

For comparison purposes only, MiDSASA models were built from activity factors that were averages of the biological activity instead of normalized weighted activity factors of the compounds selected with the minimum SSSD across the template or set of substructure fragments.

$$\text{activity factor} = \Sigma\text{activity compound}_i/N \quad (4)$$

The predicted activity for a compound is then the average of nearest compounds. If the SSSD of compound (i) in eq 1 is zero, the activity factor is calculated by eq 4. The pattern recognition software used for all MiDSASA models was Statistica version 6.1.[25]

When calculating the explained variance (r^2), both the structure and experimental activity of every compound are used in the model to predict activity, which is not the case in MiDSASA modeling. MiDSASA modeling uses the structure of the compound to be predicted by the model but does not use the compound's experimental biological activity. This is why we believe the relationship between experimental biological activity and the activity predicted by MiDSASA modeling is more like a leave-one-out cross-validation (q_1^2) relationship than an explained variance (r^2) relationship. This should not be confused with standard leave-out-cross validations where the compound is left out and the rest of the compounds are used to make a model that will be used to predict the remaining compound.[26] Here the biological activity of the compound is left out, but the structurally assigned chemical shifts on the template are still used in selecting 10% or less of the compounds in the data set. Although MiDSASA calculations are not technically leave-one-cross validations q_1^2, they are not explained correlations (r^2) either.

To evaluate the accuracy of the MiDSASA template models based on the minimal SSSD of assigned chemical shifts on one structural template, we built 500 random models. For each compound in a data set a random model was formed by selecting two random compounds from the corticosterone data set and three compounds from the aromatase data set. The random compound selections were then used in eqs 1 and 2 to produce a random binding activity. This calculation was done 500 times for each compound in a data set. The MiDSASA template model binding activities were then compared to the average of the 500 random MiDSASA calculations to evaluate how much the MiDSASA template models were learning about binding affinity.

RESULTS

Table 3 contains the actual binding activity and the predicted binding activity of the four MiDSASA models for 30 steroids binding to corticosterone binding globulin. In Table 3, the first column represents the compound number, and the second column shows the experimental binding activity. The third column represents the activity predictions from the MiDSASA template model based on the activity factor average of the nearest two compounds using all 21-atom template positions. Since this modeling process does not use the compound it is trying to predict in the model, explained correlation between the experimental binding activity and predicted binding activity is similar to a leave-one-out cross validation. The correlation between experimental binding activity and the predictions made by the 2 compound MiDSASA template model above is 0.88. Figure 5 is a plot of the predicted binding versus experimental binding for the MiDSASA template model of corticosterone binding activity when using only 2 compounds. If the predicted activity of compound 30 is removed from the model, the correlation of the model jumps from 0.88 to 0.91. Problems with the modeling of compound 30 have been reported previously.[3] The fourth column shows the MiDSASA template predictions for the 3 compound corticosterone binding model that has a correlation of 0.79.

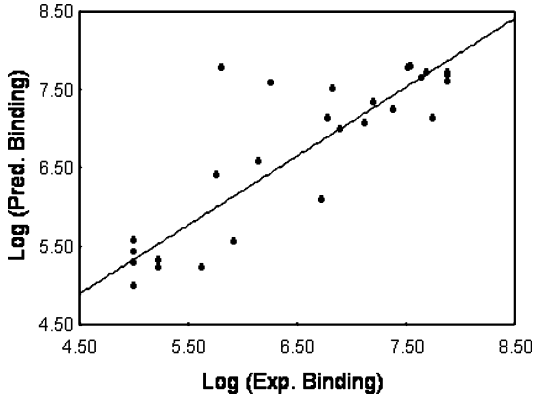

Figure 5. Plot of the predicted binding versus experimental binding for the corticosterone binding models when using only 2 compounds.

Table 3. Real and Predicted Binding Activities (BA) to the Corticosterone Binding Globulin

no.	binding activity	MiDSASA template 2 compd	MiDSASA template 3 compd	MiDSASA 3 fragment 2 compd	MiDSASA 3 fragment 3 compd	av random MiDSASA template 2 compd
1	5.00	5.42	5.28	5.40	5.59	6.37
2	5.00	5.00	5.53	5.43	5.18	6.39
3	5.76	6.43	5.99	6.99	6.66	6.34
4	5.61	5.23	5.12	5.21	5.27	6.31
5	7.88	7.72	7.71	7.48	7.73	6.56
6	7.88	7.61	7.66	7.45	7.75	6.65
7	6.89	7.01	7.35	6.85	7.12	6.49
8	5.00	5.30	5.28	5.18	4.99	6.37
9	7.65	7.65	7.38	7.07	7.30	6.67
10	7.88	7.69	7.60	7.34	7.29	6.62
11	5.92	5.57	5.95	6.30	6.16	6.36
12	5.00	5.00	5.31	5.31	5.24	6.35
13	5.00	5.00	5.31	4.99	5.03	6.39
14	5.00	5.00	5.24	5.10	5.17	6.40
15	5.23	5.23	5.12	5.21	5.27	6.30
16	5.23	5.33	6.13	5.53	5.52	6.41
17	5.00	5.58	6.48	5.68	5.54	6.46
18	7.38	7.24	7.37	6.49	6.44	6.63
19	7.74	7.13	7.36	6.63	6.47	6.57
20	6.72	6.09	5.98	6.85	6.72	6.35
21	7.51	7.78	7.71	7.28	6.86	6.44
22	7.55	7.81	7.81	7.47	7.76	6.63
23	6.78	7.13	7.40	6.80	6.86	6.53
24	7.20	7.36	7.51	7.54	7.43	6.48
25	6.14	6.59	6.25	6.47	6.35	6.39
26	6.25	7.59	7.67	6.86	6.65	6.69
27	7.12	7.08	7.14	5.61	6.06	6.51
28	6.82	7.52	7.57	6.29	6.39	6.70
29	7.69	7.73	7.75	7.37	7.23	6.59
30	5.80	7.78	7.71	7.46	7.63	6.72

The fifth column represents the activity predictions from the MiDSASA fragment corticosterone model based on the activity factor average of the nearest two compounds when using three substructure fragments in this fashion.

Predicted Corticosterone Binding Activity = 0.104 Activity Factor$_1$ + 0.236 Activity Factor$_2$ + 0.595 Activity Factor$_3$. The activity predicted for corticosterone binding has the largest contribution from fragment 3 (template atoms 12−18, 6) and to a smaller degree, fragment 2 (template atoms 13−17, 19−21) and fragment 1 (template atoms 1−5). This model had a correlation of 0.66. The sixth column is the activity prediction from the MiDSASA fragment model based on the average of the nearest three compounds when using three substructure fragments. This model had a correlation of 0.76.

Table 4. Real and Predicted Binding Activities (BA) to the Aromatase Enzyme

no.	binding activity	MiDSASA template 3 compd	MiDSASA template 4 compd	MiDSASA 4 fragment 3 compd	MiDSASA 4 fragment 4 compd	av random MiDSASA template 3 compd
1	−2.92	−2.04	−2.41	−2.14	−2.21	−0.92
2	−3.54	−2.32	−2.37	−2.34	−2.32	−0.83
3	−3.00	−2.38	−2.11	−2.56	−2.42	−0.84
4	−3.26	−2.07	−2.23	−2.26	−2.39	−0.85
5	−2.62	−2.49	−2.48	−2.41	−2.43	−0.85
6	−3.06	−1.85	−2.10	−1.80	−1.98	−0.82
7	−2.14	−2.48	−2.44	−2.35	−2.38	−0.84
8	−2.36	−2.09	−2.23	−1.87	−1.86	−0.82
9	−1.89	−1.43	−1.21	−2.07	−2.09	−0.84
10	−2.88	−1.33	−1.29	−2.31	−2.21	−0.84
11	−2.03	−1.97	−1.74	−2.48	−2.50	−0.84
12	−0.97	−2.04	−1.62	−1.52	−1.58	−0.79
13	−2.93	−1.28	−1.22	−1.93	−1.72	−0.76
14	−1.28	−2.57	−2.54	−2.19	−2.19	−0.96
15	−1.23	−2.68	−2.59	−2.26	−1.98	−0.85
16	−2.61	−2.07	−2.31	−1.98	−1.98	−0.84
17	−2.36	−1.72	−1.44	−2.20	−2.20	−0.87
18	−0.65	0.64	0.22	−1.44	−1.45	−0.74
19	−2.19	−1.46	−1.81	−1.94	−2.03	−0.75
20	−1.03	0.78	0.97	−0.68	−0.65	−0.60
21	0.00	−0.43	−0.22	−0.71	−0.75	−0.75
22	0.46	−0.63	−0.71	−1.22	−1.00	−0.79
23	−0.84	0.79	0.98	−0.37	−0.47	−0.69
24	0.15	0.14	0.12	−0.89	−0.84	−0.81
25	−0.13	0.60	0.68	0.66	0.49	−0.64
26	0.87	−0.61	−0.39	−0.38	−0.58	−0.81
27	−0.51	−1.45	−1.79	−1.55	−1.69	−0.89
28	−1.35	−1.05	−1.49	−1.11	−1.33	−0.81
29	−0.67	−1.48	−1.71	−1.10	−0.99	−0.82
30	−0.89	0.03	−0.27	−0.67	−0.69	−0.77
31	−0.79	−0.39	−0.09	−0.93	−1.02	−0.78
32	−1.09	−0.39	−0.09	−0.87	−1.02	−0.79
33	−1.08	−0.16	−0.30	−0.67	−0.47	−0.78
34	0.56	1.07	0.89	0.68	0.70	−0.60
35	0.87	0.92	0.79	0.61	0.66	−0.55
36	1.56	0.46	0.54	0.69	0.68	−0.59
37	0.94	0.89	0.71	0.79	0.72	−0.61
38	0.94	0.39	0.63	0.58	0.55	−0.64
39	0.78	0.66	0.85	0.64	0.61	−0.62
40	0.65	0.30	0.43	0.43	0.51	−0.57
41	0.53	0.51	0.60	0.52	0.55	−0.57
42	0.21	0.75	0.79	0.51	0.65	−0.58
43	0.04	0.82	0.84	0.53	0.66	−0.60
44	−0.04	0.38	0.21	0.52	0.57	−0.74
45	0.24	0.20	0.09	0.46	0.38	−0.71
46	−0.24	0.20	0.31	0.38	0.48	−0.70
47	0.61	−0.22	0.03	0.20	0.35	−0.69
48	0.91	−0.06	−0.09	0.77	0.72	−0.68
49	−0.32	0.65	0.72	1.09	1.00	−0.60
50	0.96	0.16	0.17	0.52	0.50	−0.68

Predicted Corticosterone Binding Activity = 0.331 Activity Factor$_1$ + 0.323 Activity Factor$_2$ + 0.362 Activity Factor$_3$. Once again the activity predicted for corticosterone binding is predicted mostly by fragment 3, while fragments 1 and 2 have smaller but significant contributions. The seventh column in Table 3 shows the average results of 500 random MiDSASA template calculations using 2 compounds for binding to the corticosterone binding globulin.

Table 4 contains 50 steroids with the actual binding activity and the predicted binding activity from the four MiDSASA models of 50 steroids binding to the aromatase enzyme. In Table 4, the first column is the compound number and the second column shows the experimental binding activity. The third column is the activity predictions from the MiDSASA template model based on the activity factor

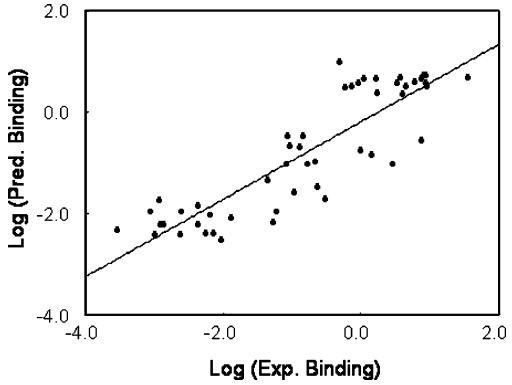

Figure 6. Plot of the predicted binding versus experimental binding for the aromatase enzyme models when using only 3 compounds and 4 substructure activity factors.

average of the nearest three compounds when using all 21-atom template positions. This model had a correlation of 0.68. The fourth column is the activity predictions from the MiDSASA template model based on the activity factor average of the nearest four compounds when using all 21-atom template poisitions. This model has a correlation of 0.71.

The fifth column represents the activity predictions from the MiDSASA fragment model based on the activity factor average of the nearest three compounds when using four substructure fragments. This MiDSASA fragment model had a correlation of 0.76.

Predicted Aromatase Binding Activity = 0.569 Activity Factor$_1$ + 0.274 Activity Factor$_2$ + 0.213 Activity Factor$_3$ − 0.10 Activity Factor$_4$. The activity predicted for aromatase is predicted mostly by fragment 1 (template atoms 2−4) and to a much smaller degree fragment 3 (template atoms 5−7, 20, 21) and fragment 2 (template atoms 13−17, 19−21). Fragment 4 (template atoms 1, 10, 9, 18) has very little effect. Figure 6 is a plot of the predicted binding versus experimental binding for the aromatase enzyme models when using only the 3 compounds and 4 substructure activity factors based on their minimum SSSD of chemical shifts. The sixth column is the activity predictions from the MiDSASA fragment model based on the activity factor average of the nearest four compounds when using four substructure fragments. This MiDSASA fragment model also had a correlation of 0.76.

Predicted Aromatase Binding Activity = 0.525 Activity Factor$_1$ + 0.237 Activity Factor$_2$ + 0.228 Activity Factor$_3$ − 0.03 Activity Factor$_4$. Once again the binding activity predicted for aromatase is predicted mostly by fragment 1 and to a much smaller degree fragment 2 and fragment 3. The seventh column in Table 4 shows the average results of 500 random MiDSASA template calculations using 3 compounds for binding to the aromatase enzyme.

DISCUSSION

Table 5 contains a comparison of the corticosterone binding globulin model performance parameters n, r^2, q^2 and number of components for the QSAR,[2] HE-state/E-state,[5] E-state,[5] SOM,[6] combination QSAR/E-state,[7] CoSA,[8] CoSASA,[8] CoSCSA,[10] and 5 MiDSASA models. The MidSASA models q_1^2 correlations of 0.66−0.88 are favorable when compared to the previously published models of binding to the corticosterone binding globulin. The correlation seen in the MidSASA template model based on two compounds is especially satisfying. The fact that the cross-validations for all other models are true leave-one-out calculations[26] and the correlations seen for the MiDSASA models are much stronger calculations because they leave out more than just the compound that is being calculated indicates that MiDSASA modeling should have good predictability outside the training set. In this case, we have not predicted outside the data set because we are not aware of any other published corticosterone binding globulin data. At the bottom of Table 5 are the average results of 500 random MiDSASA template calculations for corticosterone binding that had a correlation with experimental binding activity of 0.08. The correlation difference between the trained MiDSASA template model and random MiDSASA template model is a significant 0.80.

Table 5. Model Performance Parameters n, r^2, q^2, and Number of Components

model	n	r^2	q_1^2	components
QSAR[2]	31	0.72	0.68[a]	3 (PCs)
HE state/E-state[5]	31	0.98[a]	0.80[a]	3[a](PCs)
E-state[5]	31	0.96[a]	0.79[a]	3[a](PCs)
SOM[6]	31	0.85		3 (PCs)
QSAR/E-state[7]	30	0.82	0.78	3 (atoms)
CoSASA[8]	30	0.80	0.73	3 (atoms)
CoSA[8]	30	0.80	0.78	3 (bins)
CoSCoSA[10]	30	0.84	0.74	3 (PCs)
MiDSASA − template	30		0.88	2 compounds
MiDSASA − template	30		0.79	3 compounds
MiDSASA − 3 fragments	30		0.66	2 compounds
MiDSASA − 3 fragments	30		0.70	3 compounds
500 random MiDSASA − template	30		0.08	2 compounds

[a] 1.0 Å models.

The regression coefficients for each fragment gives an indication of how important that site was within the training set for predicting activity using MiDSASA fragment modeling technique. When the activity factor for each fragment was weak, the binding was always accurately predicted as weak. When the activity factor for each fragment was strong, the binding was always accurately predicted as strong. The linear regression coefficients for each fragment give an indication to how important that fragment is for binding within the data set. There was no preconceived method for selecting the substructure fragments, and it is possible that modeling correlations would change if the fragments were selected differently. The sum of the regression coefficients obtained from linear regression for the predicted binding activity of the corticosterone and aromatase MLR equations when using substructure fragments is near unity. This demonstrates that the quantitative k-nearest neighbor statistics behind substructure fragment MiDSASA fragment modeling is conceptually significant.

Table 6 contains a comparison of the aromatase enzyme model performance parameters n, r^2, q^2, and number of components for the QSAR,[4] molecular quantum similarity measures (MQSM),[27] CoSA,[9] CoSASA,[9] CoSCSA,[11] and 5 MiDSASA models. The MidSASA models leave-out-one correlations of 0.68−0.76 are favorable when compared to the previously published models of binding to the aromatase enzyme. Once again, the cross-validations for all other models are traditional leave-one-out calculations, and the

Table 6. Model Performance Parameters n, r^2, and q^2 for the Aromatase Enzyme

model	n	r^2	q_1^2	components
QSAR[4]	50	0.94	0.72	5 PC
MQSM[27]	50	0.84	0.73	5 PC
CoSA[9]	50	0.82	0.77	5 bins
CoSASA[9]	50	0.75	0.66	5 atoms
CoSCoSA[11]	50	0.77	0.68	5 PC
MiDSASA − template	50		0.68	3 compounds
MiDSASA − template	50		0.71	4 compounds
MiDSASA − 4 fragments	50		0.76	3 compounds
MiDSASA − 4 fragments	50		0.76	4 compounds
500 random MiDSASA − template	50		0.05	3 compounds

correlations for MiDSASA models have much stronger than normal leave-one-out correlations because of the way the model is produced. This is a good indication that the MiDSASA modeling should have decent predictability outside the training set. We have not predicted outside the data set because we are not aware of any other published aromatase enzyme. A MQSM model based on five-factors had an explained variance of 0.84 and a leave-one-out cross-validation of 0.73.[28] The row at the bottom of the Table 6 is the average result of 500 random MiDSASA template calculations using 3 compounds for the aromatase enzyme that had a correlation with experimental binding activity of 0.05. The correlation difference between the trained MiDSASA template model and average random MiDSASA template model of binding to the aromatase enzyme is a robust 0.63.

The problem with the steroid model's performance with compound 30 is due to the uniqueness of that molecule. Compound 30 contains a fluorine atom bound to the number 10 carbon on the steroid backbone, which results in significant shielding of that carbon's nucleus. No other molecule in the steroid data set has this feature. This means that there was no good neighbor (SSSD wise) that could be used to model compound 30 (particularly that feature of compound 30). This may explain why compound 30 is not modeled well by this model and for that matter any of the other published models, and why removing it from the data set dramatically improves the model's performance. If data were available for other steroid molecules with halogens substituted at different positions, then it may be possible to model compound 30 well.

The large decrease in the aromatase enzyme QSAR model cross-validation indicates the extent to which the model was based on nonlinear relationships among the input training data. Selecting only the compounds with the minimum chemical shift deviations in the MiDSASA template and fragment models removes some of the nonlinear relationships. By making a selection based on the minimum summed distances in chemical shift space between a set of known compounds and an unknown compound to be predicted, the model can be produced without rigorous statistics, principal components, or neural networks. This type of modeling only worked because the selection process by minimum deviation was physically significant. The physical significance comes from the fact that the diamagnetic term of a chemical shift tensor depends largely on its electrostatic features and geometry. Another reason MiDSASA models were better than or as good as QSAR models were fewer data points and assumptions were necessary to produce the model. 3D-QSAR modeling requires assumptions about solvent effects, partial charges, dielectrics, and structural conformation used for the calculation of electrostatic potentials. 3D-QSDAR modeling does not rely on any of the previously mentioned assumptions. Introduction of any assumptions and approximations may produce significant error.

We produced models where the activity factor was based on averages of activities rather than on weighted activity factors based on deviations between chemical shift profiles. The model for corticosterone binding based on the nearest two compounds had its correlation drop from 0.88 to 0.75 when activity factor averages were used as opposed to weighted activity factors. When moving from the use of activity factor averages to weighted activity factors, similar decreases in correlation were seen for the compounds that bind the aramatase enzyme. The limit when using the averaged activity factor based on all the compounds in the data set is a zero slope straight line with the y value at the average of all the compounds biological activities. That would result in 0.00 correlations and is the reason that only the nearest compounds were used in the MiDSASA modeling. A flat line correlation is very close to the correlations seen with the average of 500 randomization MiDSASA template model predictions that had 0.08 and 0.05 correlations with log of experimental biological activity. The results of the randomization demonstrated the MiDSASA models were learning significant information about steroids binding to enzyme receptors. We believe that MiDSASA modeling should be based on 10% of a data set or less.

CONCLUSION

The MidSASA modeling approach is a departure from previous attempts where a model is developed based on the comparative analysis of all the data.[1−11] Comparative analysis techniques are often quite replete with nonlinear relationships and that is why the leave-one-out cross-validations of many QSAR models are much lower than their explained variance correlations. The digital-like quality of our previous CoSA and CoSASA models removed some of the nonlinearity associated with a training set, but the relationships developed may not be consistent outside the data set.[8−11] MiDSASA focuses only on the compounds with templates or fragments with the nearest energy profile based on chemical shift energies, so many of the potential nonlinear features are removed. The smaller the total chemical shift deviation between compounds or fragments, the less likely a nonlinear relationship will interfere with the development of the MIDSASA model. The multiple linear regression relationships obtained from using the MiDSASA fragment modeling can aid drug design by revealing which fragments in a molecular template are important for binding and understanding the physical information of the chemical shifts in the important fragments. Since the MiDSASA modeling concept was calculated and validated with only $\leq$10% of the data set, it should be predictive to compounds with the same template outside the data set.

The accuracy and simplicity of the MiDSASA modeling predictions demonstrates that this method can be used in any case where a structural template is part of a data set, and the activity or toxicity of the compounds vary with different side chains attached at various positions of that template.

The MiDSASA modeling method may provide easy and accurate predictions for drug optimization in those kinds of systems. The MiDSASA modeling process does not need to be confined to chemical shifts on a template; E-states or electrostatic potentials could be used. The benefit of using NMR chemical shifts is that they have combined electrostatic and electron orbital information where there is minimal overlap of electron orbital information. The only ^{13}C spectral region where the information about orbitals overlaps is where the sp^3 orbital overlaps with the electropositive sp^2 orbital region. The remedy to this may require the addition of a constant to all sp^3 ^{13}C chemical shifts. Future improvements are planned to the MIDSASA modeling method including the incorporation of normalized ^{15}N chemical shifts in the template. The use of the scalar quantity SSSD in MiDSASA modeling limits the predictions to the range seen in the initial experimental training set. By using the direction of the summed chemical shift deviations in the modeling process, it should be possible to make MiDSASA predictions outside the range of the initial training set. The quantitative k-nearest neighbors modeling concept based on minimum SSSD distances can be applied anywhere a template or series of structural templates are seen in structure−activity relationship modeling. The quantitative k-nearest neighbors modeling can be applied anywhere a network is known like the endogenous metabolite concentrations in metabolic networks.

REFERENCES AND NOTES

(1) Mickelson, K. E.; Forsthoefel, J.; Westphal, U. Steroid-protein interactions. Human corticosteroid binding globulin: Some physiochemical properties and binding specificity. *Biochemistry* **1981**, *20*, 6211−6218.

(2) Good, A. C.; So, S.-S.; Richards, W. G. Structure−activity relationships from molecular similarity matrices. *J. Med. Chem.* **1993**, *36*, 433−438.

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

(4) Oprea T. I.; Garcia A. E. Three-dimensional quantitative structure−activity relationships of steroid aromatase inhibitors. *J. Comput.-Aided. Mol. Des*. **1996**, *10*, 186−200.

(5) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. E-state fields: Applications to 3D QSAR. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 513−520.

(6) Polanski, J. The receptor-like neural network for modeling corticosteroid and testosterone binding globulins. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 553−561.

(7) De Gregorio, C.; Kier, L. B.; Hall, L. H. QSAR modeling with electrotopological state indices: Corticosteroids. *J. Comput.-Aided Mol. Des.* **1988**, *12*, 557−561.

(8) Beger R. D.; Wilkes J. G. Developing ^{13}C NMR quantitative spectrometric data-activity relationship (QSDAR) models of steroid binding to the corticosteroid binding globulin. *J. Comput.-Aided. Mol. Des.* **2001**, *15*, 659−669.

(9) Beger R. D.; Wilkes J. G. ^{13}C NMR quantitative spectrometric data-activity relationship (QSDAR) models to the aromatase enzyme. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1360−1366.

(10) Beger, R. D.; Buzatu, D.; Wilkes, J. G.; Lay, Jr., J. O. Developing comparative structural connectivity spectra analysis (CoSCSA) models of steroid binding to the corticosteroid binding globulin. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1123−1131.

(11) Beger, R. D.; Wilkes, J. G. Comparative structural connectivity spectra analysis (CoSCoSA) models of steroids binding to the aromatase enzyme. *J. Mol. Recognit.* **2002**, *15*, 154−162.

(12) Colborn, T.; von Saal, F. S.; Soto, A. M. Developmental effects of endocrine-disrupting chemicals in wildlife and humans. *Environ. Health Perspect.* **1993**, *101*, 378−384.

(13) Spencer, R. L.; Kalman, B. A.; Dhabhar, F. S. Role of endogenous glucocorticoids in immune system function: regulation as well as counterregulation. In *Handbook of Physiology. Coping with the Environment: Neural and Endocrine Mechanisms*; McEwen, B. S., Ed.; Oxford University Press: Oxford, 2001; Section 7, Vol. IV, pp 381−424.

(14) Pearce D.; Yamamoto K. R. Mineralocorticoid and glucocorticoid receptor activities distinguished by nonreceptor factors at a composite response element. *Science* **1993**, *259*, 1161−1165.

(15) Kellis, J. J.; Vickery, L. E. Purification and characterization of human placental cytochrome P-450. *J. Biol. Chem.* **1987**, *262*, 4413−4420.

(16) Chen, S.; Besman, M. J.; Shively, J. E.; Yanagibashi, K.; Hall, P. F. Human aromatase. *Drug Metab. Rev.* **1989**, *20*, 511−517.

(17) Brodie, A. M. Aromatase inhibitors in the treatment of breast cancer. *J. Steroid Biochem. Mol. Biol.* **1994**, *49*, 281−287.

(18) 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.

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

(20) 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.

(21) CNMR predictor, ACD/Labs Toronto, Canada 2000.

(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) Bremser, W. HOSE − a Novel substructure code. *Anal. Chim. Acta* **1978**, *103*, 355−365.

(25) Statistica version 6.1 Statsoft, Inc., Tulsa, OK, 2003.

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

(27) Baumann, K.; von Korff, M.; Albert, H. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part II. *J. Chemom.* **2002**, *16*, 351−360.

(28) Gironés, X.; Carbó-Dorca, R. Molecular quantum similarity-based QSARs for binding affinities of several steroid sets. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1185−1193.

CI049925E