Internet Electronic Journal of Molecular Design

July 2003, Volume 2, Number 7, Pages 435–453

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday
Part 1

Guest Editors: Douglas J. Klein and Sonja Nikolić

Using Simulated 2D ^{13}C NMR Nearest Neighbor Connectivity Spectral Data Patterns to Model a Diverse Set of Estrogens

Richard D. Beger, Kathleen J. Holm, Dan A. Buzatu, and Jon G. Wilkes

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

Received: March 21, 2003; Revised: May 2, 2003; Accepted: May 7, 2003; Published: July 31, 2003

Citation of the article:

R. D. Beger, K. J. Holm, D. A. Buzatu, and J. G. Wilkes, Using Simulated 2D ^{13}C NMR Nearest Neighbor Connectivity Spectral Data Patterns to Model a Diverse Set of Estrogens, *Internet Electron. J. Mol. Des.* **2003**, *2*, 435–453, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Using Simulated 2D ^{13}C NMR Nearest Neighbor Connectivity Spectral Data Patterns to Model a Diverse Set of Estrogens[#]

Richard D. Beger,* Kathleen J. Holm, Dan A. Buzatu, and Jon G. Wilkes

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

Received: March 21, 2003; Revised: May 2, 2003; Accepted: May 7, 2003; Published: July 31, 2003

Internet Electron. J. Mol. Des. **2003**, *2* (7), 435–453

Abstract

Motivation. The scope of this investigation is to develop a rapid, objective modeling method that will accurately predict the estrogen receptor binding affinities for a diverse set of compounds.

Method. We have used simulated 2D ^{13}C–^{13}C COSY NMR spectral data to develop a model for 130 diverse organic compounds whose relative binding affinities (RBA) to the estrogen receptor are known. The simulated 2D ^{13}C–^{13}C COSY NMR spectra were generated by using the NMR spectral assignments for predicted carbon chemical shifts to identify nearest neighboring carbon atoms and establish carbon–to–carbon through–bond connectivity spectral patterns of each compound. We call the use of such patterns for model building comparative structural connectivity spectra analysis (CoSCoSA).

Results. For the large number of estrogens, a CoSCoSA multiple linear regression (MLR) model using 16 bins selected from the ^{13}C–^{13}C COSY spectral data had an r^2 of 0.827, a leave–one–out cross–validation q_1^2 of 0.78, and a leave–13–out cross–validation average q_{13}^2 of 0.78. A second CoSCoSA model using 15 bins plus one additional distance–related 3D constraint had an r^2 of 0.833, a q_1^2 of 0.79 and an average q_{13}^2 of 0.78. The predictions for 27 external compounds had q_{pred}^2 of 0.53 for one CoSCoSA model.

Conclusions. The addition of more through–space distance related 3D information, which presently awaits software development to make pattern definition for each compound practical, should improve the predictive accuracy of the CoSCoSA models.

Keywords. Estrogen; quantitative structure–activity relationships; QSAR; ^{13}C NMR.

Abbreviations and notations

2D–QSDAR, Two–dimensional quantitative spectrometric data–activity relationship
3D–QSAR, Three–dimensional quantitative structure–activity relationship
CoSA, Comparative spectral analysis
CoSASA, Comparative structurally assigned spectral analysis
RBA, relative binding affinity
CoSCoSA, Comparative spectral connectivity spectral analysis
CoSY, Correlation spectroscopy
HOSE, hierarchically ordered spherical description of environment
LOO, leave–one–out
MLR, multiple linear regression
NMR, Nuclear Magnetic Resonance

[#] Dedicated to Professor Nenad Trinajstić on the occasion of the 65th birthday.

* Correspondence author; phone: (870) 543–7080; fax: (870) 543–7686; E–mail: rbeger@nctr.fda.gov.

1 INTRODUCTION

The Food Quality Protection Act, passed in 1996, mandates that the United States Environmental Protection Agency will develop screening and testing procedures for endocrine disrupting chemicals [1]. An endocrine disruptor 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 endocrine disruptors to be tested. Over 58,000 compounds are candidates to be screened for their estrogen receptor binding level, and the number of compounds to be screened is growing every day.

This increasing list of substances that can mimic the effects of estrogen in the human body includes environmental estrogens and phytoestrogens (plant–based estrogens). Categories of environmental estrogens include pesticides such as 1,1'–(2,2,2–trichloroethane–1,1–diyl)bis(4–chlorobenzene) (DDT), plasticizers such as bisphenol A, transformer oils such as polycholrinated biphenyls, and many other types of compounds. Genistein and daidzein are two well–known isoflavonoid phytoestrogens that are found in soybeans. There is conflicting evidence regarding the reproductive and cancer risk effects of environmental and plant–based estrogens in humans. Some studies suggest that certain phytoestrogens have beneficial effects on breast and prostate cancers [2–4]; other studies suggest that they have no beneficial or negative effects on cancers [5,6]; and still other studies suggest that environmental estrogens may have negative effects on cancer [7,8]. There is a growing need to develop inexpensive, rapid, and accurate methods to identify compounds with strong estrogenic activity. This would drastically reduce the number of compounds requiring evaluation by biological assays.

In this paper, we developed a couple of two–dimensional quantitative spectrometric data–activity relationship (2D–QSDAR) models for 130 compounds that bind to the estrogen receptor [9,10]. QSDAR and spectrometric data–activity relationship (SDAR) models are based on the spectral–activity leg in the triangular structure–spectra–activity relationship. Three–dimensional quantitative structure–activity relationship (3D–QSAR) [11,12] models are based on the structure–activity leg. Earlier we developed qualitative SDAR models of estrogen receptor binding activity based on a composite of 1D ^{13}C NMR and mass spec [13,14]. We have developed QSDAR models for molecules binding to the corticosterone binding globulin [15], aromatase enzyme [16], and aryl hydrocarbon receptor [17]. A similar method that combined ultraviolet (UV) spectra with 31 other physicochemical parameters was used to build models of polychlorinated biphenyls aryl hydrocarbon receptor binding and cytochrome P4501A induction [18]. Previous comparative spectral analysis (CoSA) models were based on simulated 1D ^{13}C nuclear magnetic resonance (NMR) data [15–17,19]. Advanced Chemistry Development (ACD) Labs [20] now sells software that uses a chemical's structure to predict its ^{13}C NMR spectrum. Other ^{13}C NMR prediction packages include an artificial neural network [21] and NMRscape software from Bio–Rad Laboratories [22]. Because of the ability to predict NMR spectra accurately, the process of building the CoSA models was simple and rapid. We improved the accuracy of CoSA models by displaying all possible carbon–to–carbon connections with their assigned carbon NMR chemical shifts and distances between the carbons in a 3D–connectivity matrix [23–25]. A three–dimensional quantitative spectrometric data–activity relationship (3D–QSDAR) model was developed that is built by combining predicted NMR spectral information with structural information in a 3D–connectivity matrix. The method uses selected 2D ^{13}C–^{13}C COrrelation SpectroscopY (COSY) (through–bond nearest neighbors) and theoretical 2D ^{13}C–^{13}C through–space distance–related connectivity spectral slices from the 3D–connectivity matrix to produce a relationship to biological binding activity. We called this technique comparative structural connectivity spectra analysis (CoSCoSA) modeling [23–25]. The CoSCoSA models for binding to the corticosterone binding globulin and aromatase enzyme were based on principal component regression [23,24]. The CoSCoSA model for binding to the aryl hydrocarbon receptor was based on multiple linear regression of selected bins from the 3D–connectivity matrix [25].

The laws of quantum mechanics govern biological interactions when viewed at the atomic and molecular level. Using NMR data is a way of providing the biological modeling process with quantum mechanical information. An atom's NMR chemical shift is directly related to the quantum mechanically determined state of its nuclear magnetic dipole in a magnetic field [26]. The diamagnetic term of a NMR chemical shift is directly related to the electrostatic potential at its nucleus. While UV, infrared, and NMR spectra are all related to quantum mechanics, one advantage with NMR spectra is that spectral features can be assigned to a specific atom, whereas this simplifying characteristic is not true for UV or IR spectra. Thus, the distribution of a molecule's NMR spectral features directly corresponds to structural elements that, according to structure–activity relationship theory, determine the biological activity of the molecule. Therefore, patterns can be developed from the association of a set of NMR spectra with the known biological characteristics of the compounds in the set. These patterns can then be used for predicting the corresponding activities of any compounds not used in the model building process.

Standard NMR instrumental techniques include 2D ^{1}H–^{1}H COSY [27] experiments in which connectivity relationships through three bonds are found for nearest neighbor protons with off diagonal cross peaks. This experiment is analogous to 2D ^{13}C–^{13}C through–bond nearest neighbor connectivity spectral patterns synthesized for use in the models presented here. In these models a molecule's connectivity spectrum was produced using structurally assigned predicted spectra and adding the nearest neighbor information as off–diagonal cross peaks on a two dimensional plot. As explained below, the spectral/connectivity features were then reduced to bin occupancy levels for the pattern definition used to build the models.

Our CoSCoSA predictive methods bear comparison to several other multi–dimensional NMR experimental techniques. NMR 2D ^{13}C–^{13}C COSY spectral data is similar conceptually to connectivity ideas that are used in the first order $^{1}\chi$ connectivity indices [28], modified adjacency matrix [29], and 3D connectivity indices [30]. A significant difference between these previous topological studies and CoSCoSA modeling is that the topological descriptors are rigorously calculated whereas the CoSCoSA descriptors are simulated from a known database of experimental NMR chemical shifts. Results from our previous comparative structurally assigned spectral analysis (CoSASA) models [15,16,25] suggested that assigned chemical shifts contain modeling information similar to that derived from electrotopological–state indices [31].

In this paper we were limited to using only the 2D ^{13}C–^{13}C COSY spectra. The lack of most three dimensional distance information results from an unanticipated consequence of the available ACD 2D NMR prediction software not using for each molecule's predicted chemical shifts the same atomic numbering system as those found in the "mol" or "pdb" files. To produce long–range carbon–to–carbon distance correlations the numbering system for the three–dimensional structural information must conform to the numbering system of the chemical shifts. It might be theoretically possible by manual methods to renumber the predicted spectra and generate the multitude of long–distance interactions for each molecule in a large, diverse training set. Indeed, we were able to overcome this problem in previous 3D–QSDAR models because all the molecules in those models were both structurally similar and rigid so that we only had to collect distances between atoms once for the entire set of similar molecules [23–25]. This approach was impractical in the current study because there were 130 structurally diverse molecules to model and many of these also exhibited considerable flexibility. We could generate the most probable conformations by molecular dynamics modeling but could not readily associate the resulting inter–atomic distances with the predicted spectral features so as to populate long–distance associated 3D spectral bins.

To investigate potential advantages of long–distance–correlated spectral associations in modeling biological activities for a large, structurally diverse compound training set, we added a single long distance parameter to the otherwise 2D–CoSCoSA model. This long distance parameter was a simple indicator of each molecule's maximum distance possible between any two non–hydrogen atoms.

2 METHODS

2.1 Relative Binding Activity Data

The log relative binding activity (RBA) data for 130 structurally diverse compounds was used to train the CoSCoSA models. The data was produced at National Center for Toxicological Research (NCTR) using a competitive estrogen receptor (ER) binding assay with radiolabeled estradiol

([^{3}H]E_2) in rat uterine cytosol, which was obtained from ovariectomized uteri of Sprague–Dawley rats [9,10,32]. This data set spanned 7 orders in magnitude ranging from a log RBA value of –4 for a weak estrogen, to a log RBA of 2 for a strong estrogen. For a particular molecule, the RBA to the estrogen receptor is defined as one hundred times the ratio of the molar concentrations of 17β–estradiol and the competing compound required to decrease the amount of receptor–bound 17β–estradiol by 50%. Thus 17β–estradiol had an RBA of 100 and a base ten log RBA of 2.0.

2.2 Predicting the ^{13}C NMR Spectral Data

For each of the 130 compounds, the ^{13}C 2D ^{13}C–^{13}C COSY NMR experiment was simulated using the ACD Labs CNMR version 5.0 2D predictor software [20]. The simulated COSY NMR spectra could be saved with two tables per compound. One table showed the assigned carbon chemical shifts and the other indicated through–bond coupling of nearest neighbor carbon atoms. The use of predicted rather than experimentally measured NMR chemical shifts was not necessary for developing the CoSCoSA models, but it saved time and money. Additionally, the use of ^{13}C NMR spectra, each based on the same edition of prediction software (rather than collected from spectral libraries or other sources) eliminated random variability due to the NMR solvent or other experimental factors.

2.3 Producing the ^{13}C COSY Spectral Data

The 2D ^{13}C–^{13}C COSY spectra were predicted for the compounds. We reduced the resolution of all 2D ^{13}C–^{13}C COSY spectra by defining bins, 2.0 parts per million (ppm) wide in both dimensions. The inherent resolution of NMR is much greater than this, but any signal appearing within a 2.0 ppm bin was counted toward the bin population. This choice was made so that many of the bins would be multiply populated, a necessary characteristic for statistical analysis and model validation. The use of such wide bins also reduced the confounding effects on the modeled patterns caused by uncertainties or errors in simulated spectra. An average error of 0.50 ppm in predicted 13C NMR spectra of TAXOL has been reported by ACD Labs CNMR predictor version 6.12 [33]. This is consistent with a 0.53 ppm average uncertainty reported in a previous CoSA study [15]. The specific 2.0 ppm value was chosen because that bin width was used successfully in prior CoSA and CoSCoSA models [17,23–25]. Possible improvements in model efficacy through use of somewhat wider or narrower bins were not investigated. The generally excellent results reported below for this modeling approach probably leave some room for improvement from optimization of this factor alone.

The spectra were saved as two–dimensional bins under the peak within a certain spectral range and normalized to an integer. A single carbon–to–carbon connectivity was assigned an area of 100, two carbon–to–carbon connections in a bin had an area of 200, and so forth. Occupancy of the 120–126 ppm spectral bin represents the same spectral connectivity relationship as that in the 126–120 ppm bin. The data are symmetric across the diagonal of the spectral plane. For this reason, the 240 by 240 ppm 2D spectral plane was consolidated into 7381 of these 2 × 2 ppm bins so that only the bins above and including the diagonal were used. After binning all 130 compounds, only 605 bins from the 7381 bins had "hits" in them. Of the 605 populated bins only 337 bins had more than one "hit". From the remaining 337 multiply populated bins, an increasing number of the mostly highly correlated bins were selected by trial and error and used to construct MLR models until a model was obtained that had an r^2 greater than 0.8 and an F–value greater than 30. We were able to identify 16 bins this way.

The predicted NMR spectra were calculated by a substructure similarity technique called hierarchically ordered spherical description of environment (HOSE) [34]. HOSE uses similar substructural elements from a compound's 2D structure and radial spheres to produce a set of weights that are used to predict the chemical shift. Therefore, the errors produced in the simulated NMR spectra were propagated through the similar structures found in the training set of the QSDAR models. This conveniently reduced the effective error when using a training set to predict unknown sample affinities for compound spectra if they also were predicted using the same HOSE routine referenced to the same version of the predictor software.

2.4 Statistical Analysis

All statistical analysis was performed by Statistica version 6.0 software [35]. For each CoSCoSA model, we used forward multiple linear regression (MLR) on a selected subset of spectral bins until a model had an r^2 greater than 0.82, the F–value was still increasing with the addition of a bin, and the p–value of the bin added was significant ($p < 0.05$). We did not select any bins that had less than 2 "hits". The reason for this is that a bin with one "hit" can inappropriately add to the r^2 of a model but can not improve the leave–one–out cross–validation ($q_1{}^2$) of a model. The use of a large number of very small, singly populated bins is the reason that Bursi had a high r^2 and very low $q_1{}^2$ [19].

2.5 Internal Model Tests

Evaluations of the CoSCoSA models were done using LOO or leave–multiple–out cross–validation procedures in which one or more compounds were systematically excluded from the training set and each developed model (missing any contribution from the excluded compound(s)) was used to predict binding activities of the excluded compounds [36]. The cross–validated r^2 (termed $q_1{}^2$) that results from fitting predictions obtained by cross–validation experiments can be derived from $q_1{}^2 = 1 - \text{PRESS/SSD}$. Here PRESS is the sum of the differences between the actual and predicted activity data for each molecule during LOO cross–validation, and SSD is the sum of the squared deviations between the measured and mean activities of each molecule in the training set. During the LOO cross–validation, each compound was removed from the training and the Beta–coefficients in the MLR equation were recalculated. This new MLR equation was used to recalculate the log RBA of the compound left out. To more rigorously test the validity of the CoSCoSA models and two leave–13–out [10% of the data excluded) cross–validations were performed on each of the models. In these "leave–multiple–samples–out" experiments, the compounds omitted were varied and the results of the two corresponding experiments were averaged.

Table 1. Training Set Predictions. The Table contains the compound number, compound name, experimental log RBA, predicted log RBA from the 16 Bin CoSCoSA model, and predicted log RBA from the 15 Bin +L$_{<7.5Å}$ CoSCoSA model.

No	Compound Name	log RBA	log RBA	CoSCoSA
		Exp	16 Bin	15 Bin + L$_{<7.5Å}$
1	diethylstillbesterol	2.60	1.44	1.43
2	meso–hexestrol	2.48	2.86	2.70
3	ethinyl estradiol	2.28	1.44	1.43
4	4–hydroxyestradiol	2.24	2.45	2.39
5	4–hydroxytamoxifen	2.24	0.58	0.58
6	17β–estradiol	2.00	2.36	1.85
7	α–zearalenol	1.63	0.51	0.51
8	ICI182780	1.57	1.08	1.12
9	dienestrol	1.57	1.44	1.43
10	α–zearalanol	1.48	0.51	0.51
11	2–hydroxyestradiol	1.47	1.26	1.32
12	diethylstilbestrol monomethyl ether	1.31	1.44	1.43
13	3,3'–dihydroxyhestrol	1.19	0.75	0.60
14	droloxifene	1.18	1.63	1.63
15	dimethylstibestrol	1.16	0.04	0.15
16	ICI164384	1.16	1.08	1.12
17	moxestrol	1.14	1.44	1.43
18	17–deoxyestradiol	1.14	0.62	0.79
19	2,6–dimethylhexestrol	1.11	0.64	0.62
20	estriol	0.99	0.62	0.79
21	monomethyl ether hexestrol	0.97	0.51	0.93
22	estrone	0.86	0.62	0.79
23	*p*–meso–phenol	0.6	1.35	1.26
24	17α–estradiol	0.49	1.17	0.79
25	dihydroxymethoxychlorolefin	0.42	–0.10	–0.10
26	mestranol	0.35	1.44	1.43
27	zearalanone	0.32	0.51	0.51
28	tamoxifen citrate	0.21	0.58	0.58
29	toremifene citrate	0.14	0.58	0.58
30	α,α–dimethylbethyl allenolic acid	–0.02	–0.04	–0.02
31	coumestrol	–0.05	0.43	0.05
32	4–ethyl–7–OH–(*p*–meoxyphenol)–dihydro–1–benzopyran–2–one	–0.05	0.11	0.15
33	nafoxidine	–0.14	0.58	0.58
34	clomiphene citrate	–0.14	–0.47	–0.47
35	1,3,5–estratrien–3,6α–17β–triol	–0.15	–0.60	–0.61
36	β–zearalanol	–0.19	0.51	0.51
37	3–OH–estra–1,3,5–trien–16–one	–0.29	–0.08	0.25
38	3–deoxyestradiol	–0.30	–1.47	–1.55

Table 1. (Continued)

No	Compound Name	log RBA	log RBA	CoSCoSA
		Exp	16 Bin	15 Bin + $L_{<7.5Å}$
39	3,6,4'–trihydroxyflavone	–0.35	–0.33	–0.35
40	genistein	–0.36	–2.66	–2.61
41	4,4'–dihroxystilbene	–0.55	–0.56	–0.51
42	dihydroxymethoxychlor (HPTE)	–0.60	–1.47	–2.21
43	monohydroxymethoxychlorolefin	–0.63	–0.10	–0.10
44	2,3,4,5–tetraCl–4'–biphenylol	–0.64	–1.61	–1.56
45	norethynodrel	–0.67	–2.66	–2.61
46	2,2',4,4'–tetrahydroxybenzil	–0.68	–0.80	–0.81
47	β–zearalenol	–0.69	0.51	0.51
48	4,6–dihydroxyflavone	–0.82	–2.07	–1.95
49	equol	–0.82	–0.66	0.05
50	monohydroxymethoxychlor	–0.89	–2.07	–1.95
51	3β–androstanediol	–0.92	–2.66	–2.61
52	bisphenol B	–1.07	–2.66	–2.61
53	phloretin	–1.16	–0.80	–0.81
54	dietheylstilbestrol dimethyl ether	–1.25	–0.66	–0.68
55	2',4,4'–trihydroxychalcone	–1.26	–1.73	–1.71
56	2,5–dichloro–4'–biphenylol	–1.44	–1.61	–1.56
57	4,4'–[1,2–ethanediyl)bisphenol	–1.44	–2.66	–2.61
58	17β–estradiol–16β–OH–16–methyl–3–ether	–1.48	–1.34	–1.34
59	aurin	–1.50	–0.56	–0.51
60	nordihydroguariareticacid	–1.51	–2.66	–2.61
61	4–nonylphenol	–1.53	–1.61	–1.56
62	apigenin	–1.55	–2.07	–1.95
63	kaempferol	–1.61	–2.66	–2.61
64	daidzein	–1.65	–1.61	–1.56
65	3–methylestriol	–1.65	–1.34	–1.34
66	4–dodecylphenol	–1.73	–2.66	–2.61
67	2–ethylhexyl–4–hydroxybenzoate	–1.74	–2.66	–2.61
68	4–tert–octylphenol	–1.82	–2.66	–2.61
69	phenolphthalein	–1.87	–1.47	–1.30
70	kepone	–1.89	–2.66	–2.61
71	heptyl–4–hydroxybenzoate	–2.09	–2.66	–2.61
72	bisphenol A	–2.11	–2.66	–2.61
73	naringenin	–2.13	–2.66	–2.61
74	4–Cl–4'–biphenylol	–2.18	–2.66	–2.61
75	3–deoxyestrone	–2.2	–1.47	–1.55
76	4–octylphenol	–2.31	–2.66	–2.61
77	fisetin	–2.35	–2.14	–2.09
78	3',4',7–trihydroxyisoflavone	–2.35	–2.66	–2.61
79	biochanin A	–2.37	–2.66	–2.61
80	4–OH–chalcone	–2.43	–2.66	–2.61
81	4'–OH–chalcone	–2.43	–2.66	–2.61
82	2,2'–methylenebis[4–chlorophenol)	–2.45	–2.07	–1.95
83	4,4'–dihydroxybenzophenone	–2.46	–2.66	–2.61
84	benzyl–4–hydroxybenzoate	–2.54	–2.66	–2.61
85	2,4–dihyroxybenzophenone	–2.61	–2.66	–2.61

Table 1. (Continued)

No	Compound Name	log RBA	log RBA	CoSCoSA
		Exp	16 Bin	15 Bin + $L_{<7.5Å}$
86	4'–hydroxyflavanone	–2.65	–3.15	–2.96
87	3α–androstanediol	–2.67	–2.66	–2.61
88	4–phenethylphenol	–2.69	–2.66	–2.61
89	prunetin	–2.74	–2.66	–2.61
90	doisynoestrol	–2.74	–2.66	–2.61
91	myricetin	–2.75	–2.66	–2.61
92	2–Cl–4–biphenylol	–2.77	–3.21	–2.61
93	triphenylethylene	–2.78	–2.66	–2.61
94	3'–OH–flavanone	–2.78	–3.43	–3.31
95	chalcone	–2.82	–2.66	–2.61
96	*o*,*p*'–DDT	–2.85	–2.66	–2.61
97	4–heptyloxyphenol	–2.88	–2.66	–2.61
98	dihydrotestosterone	–2.89	–2.66	–2.61
99	formononetin	–2.98	–2.66	–2.61
100	bis–[4–hydroxyphenyl)methane	–3.02	–2.66	–2.61
101	*p*–phenylphenol	–3.04	–2.66	–2.61
102	6–hydroxyflavanone	–3.05	–2.14	–2.09
103	4,4'–sulfonyldiphenol	–3.07	–1.47	–1.30
104	butyl–4–hydroxybenzoate	–3.07	–2.66	–2.61
105	diphenolic acid	–3.13	–2.66	–2.61
106	1,3–diphenyltetramethyldisiloxane	–3.16	–2.66	–2.61
107	propyl–4–hydroxybenzoate	–3.22	–2.66	–2.61
108	ethyl–4–hydrobenzoate	–3.22	–2.66	–2.61
109	phenol red	–3.25	–2.66	–2.61
110	3,3',5,5'–tetraCl–4,4'–biphenyldiol	–3.25	–2.66	–2.61
111	4–tert–amylphenol	–3.26	–2.66	–3.53
112	baicalein	–3.35	–2.66	–2.61
113	morin	–3.35	–2.66	–2.61
114	4–sec–butylphenol	–3.37	–2.07	–1.95
115	4–Cl–3–methylphenol	–3.38	–2.66	–3.53
116	6–hydroxyflavone	–3.41	–2.66	–2.61
117	4–benzyloxyphenol	–3.44	–2.66	–2.61
118	3–phenylphenol	–3.44	–2.14	–2.09
119	methyl–4–hydrobenzoate	–3.44	–2.66	–3.53
120	2–sec–butylphenol	–3.54	–3.20	–3.04
121	2,4'–dichlorobiphenyl	–3.61	–2.66	–2.61
122	4–tert–butylphenol	–3.61	–3.75	–3.53
123	2–Cl–4–methylphenol	–3.66	–2.66	–3.53
124	phenolphthalin	–3.67	–2.66	–2.61
125	4–Cl–2–methylphenol	–3.67	–2.66	–3.53
126	7–hydroxyflavanone	–3.73	–2.66	–2.61
127	3–ethylphenol	–3.87	–2.90	–3.70
128	rutin	–4.09	–2.66	–2.61
129	4–ethylphenol	–4.17	–3.75	–3.53
130	4–methylphenol	–4.50	–3.75	–3.53

2.6 External Model Tests

Additionally, to further test the ruggedness of CoSCoSA models, we predicted the log RBAs of compounds from two published external data sets, namely those of Waller [37] and Kuiper [38]. The log RBAs from those external data sets possessed a greater variability in binding activity. So, by a previously published method, a set of compounds that had their binding activity determined in all three labs (Waller, Kupier, and NCTR) were used to normalize the external data sets to the NCTR data [10,39]. We then built the CoSCoSA models as before and used the resulting MLR equations to predict the log RBA of the compounds in the test set. We used the published normalized log RBA for 27 compounds from Waller and Kuiper data in our external testing of the CoSCoSA models [10]. However, many of the occupied bins for the new compounds from the external data set did not fall into the original 605 occupied bins; the original set of bins comprised only 8.2% of the 2D COSY spectral plane. We inferred that in the different molecular contexts of the external data sets, NMR chemical shift information was expressed in adjacent but non–included bins. NMR chemical shifts exist along a continuum and the process of binning them for this type of pattern recognition inherently compromises the pattern when it barely misses a selected bin.

To account for this source of confusion with the external data, we tried adding various fractions of “near–miss” signals into each compound’s spectrum. With this in mind we used the CoSCoSA model’s MLR equation to predict the normalized log RBA of the compounds in the external test set. However, compounds from the external test set with bins that were one bin away (one of 8 bins surrounding a 2D bin) from the original 605 populated bins were modeled using none, one–quarter, and one–half of that bin's intensity in the nearest neighboring bin used in the original CoSCoSA model.

3 RESULTS AND DISCUSSION

Table 1 shows the compound name, compound number, and two CoSCoSA model predictions of the log RBA for 130 compounds in the training set. Figure 1 shows the CoSCoSA model that was based on the MLR analysis of 16 selected 2D bins from the ^{13}C–^{13}C COSY spectral data. The 16 bin COSY model for the 130 estrogenic compounds had an explained variance r^2 of 0.827, a LOO $q_1{}^2$ of 0.78 and an average leave–13–out cross–validated variance ($q_{13}{}^2$) of 0.78 ± 0.01. The CoSCoSA model was based on COSY bins 28–12 (bin 1), 72–20 (bin 2), 54–28 (bin 3), 50–38 (bin 4), 64–56 (bin 5), 164–104 (bin 6), 152–108 (bin 7), 156–110 (bin 8), 126–112 (bin 9), 140–112 (bin 10), 142–112 (bin 11), 154–112 (bin 12), 154–114 (bin 13), 156–114 (bin 14), 128–116 (bin 15), and 126–120 (bin 16). All 2ppm bins were written using the format *a*–*b*, where *a* and *b* are the ppm values corresponding to the two “connected” atoms. The MLR equation for the 16 bin

CoSCoSA model is:

$$\begin{aligned}\log \text{RBA} = {} & 0.00999\times\text{bin }1 + 0.03173\times\text{bin }2 + 0.0071\times\text{bin }3 + 0.01196\times\text{bin }4 \\ & + 0.02191\times\text{bin }5 + 0.0093\times\text{bin }6 + 0.2329\times\text{bin }7 + 0.01324\times\text{bin }8 \\ & + 0.00737\times\text{bin }9 + 0.02558\times\text{bin }10 + 0.0392\times\text{bin }11 + 0.03094\times\text{bin }12 \\ & - 0.00545\times\text{bin }13 + 0.00526\times\text{bin }14 + 0.00298\times\text{bin }15 - 0.00768\times\text{bin }16\end{aligned}$$

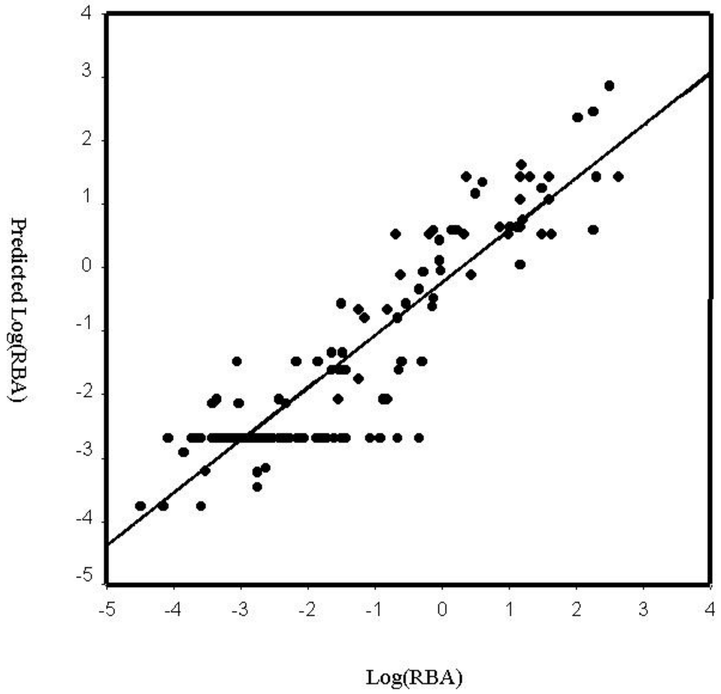

Figure 1. Plot of the predicted log RBA and experimental log RBA based on 16 bins.

All bins had more than three “hits” in the bin except for bins 152–108 and 140–112 that had only two “hits” each. The correlation matrix for the 16 bins were calculated and only two sets of bins had correlation between them that were greater than 0.5. The greatest average correlation between any bin with the other 17 bins was 0.04 and many of the average correlations were much lower than 0.04. The lack of a large correlation among bins suggests that the resulting patterns were based on essentially orthogonal data. The COSY bin 28–12 was most often associated with the CH_3 carbon connected to the CH_2 in the ethyl groups in diethylstilbestrol and hexestrol–like compounds. Twelve of the fourteen compounds with a COSY hit in 28–12 had a log RBA greater than –0.05. Compounds that populated a COSY bin at 154–112 were most often associated with the 3 carbon position connected to the 2 carbon position in the A–ring of 17β–estradiol like compounds. Nine of the ten compounds with a COSY hit in bin 154–112 had a log RBA greater than –0.05. Fourteen compounds had a “hit” in the COSY bin at 128–116. The 128–116 COSY bin was most often associated with the 2 to 3 and 5 to 6 carbon positions in a phenol ring. Twelve of the fourteen compounds with a COSY hit in bin 128–116 had a log RBA less than 0.60. The 24 compounds that had a hit or multiple hits in the COSY bin at 156–114 was most often associated with the hydroxylated carbon of a phenol ring connected to its two nearest neighboring carbons. Only 5 of the 24 compounds with a COSY hit in bin 156–114 had a log RBA less than –1.65. The 6 compounds that had a COSY bin at 64–56 was most often associated with the two carbons between the oxygen ester and the nitrodimethyl of tamoxifen–like compounds. Similar spectral–structural associations could be made for the other COSY bins effectively used for receptor binding prediction in the CoSCoSA models.

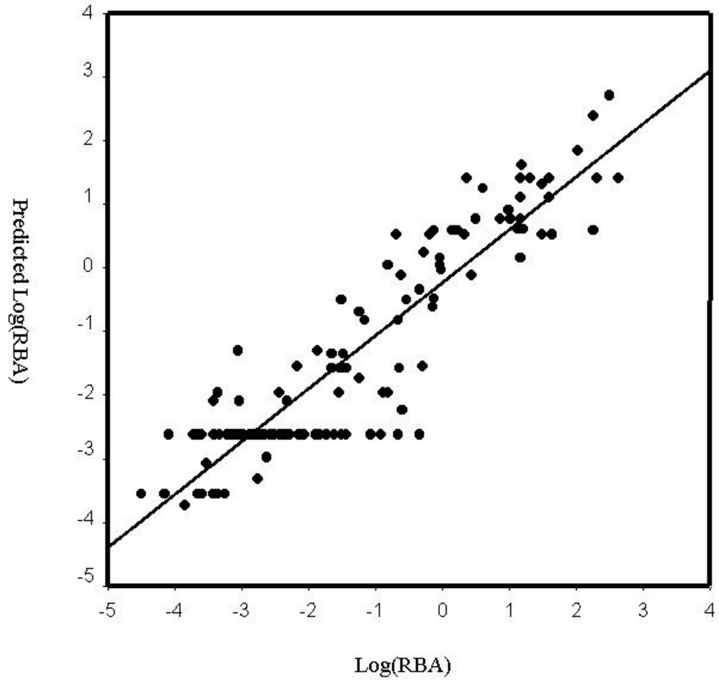

Figure 2. Plot of the predicted log RBA and experimental log RBA based on 15 bins+$L_{<7.5Å}$ variable.

Figure 2 shows results for the CoSCoSA model that was based on the MLR analysis of 15 selected 2D ^{13}C–^{13}C COSY bins plus the one distance variable. The distance variable, $L_{<7.5Å}$, was assigned a value of 1 when the maximum distance between non–hydrogen atoms in a compound was less than 7.5 Å (compact) and a value of zero for all other compounds. The $L_{<7.5Å}$ variable replaced the COSY bin at 154–114 (bin 13) in the previous 16 bin CoSCoSA model. The MLR equation for the 15 bin–with–$L_{<7.5Å}$ CoSCoSA model is

$$\begin{aligned}\log \text{RBA} = {} & 0.00969\times\text{bin } 1 + 0.03122\times\text{bin } 2 + 0.00637\times\text{bin } 3 + 0.01066\times\text{bin } 4 \\ & + 0.02142\times\text{bin } 5 + 0.00902\times\text{bin } 6 + 0.2263\times\text{bin } 7 + 0.01275\times\text{bin } 8 \\ & + 0.00732\times\text{bin } 9 + 0.02507\times\text{bin } 10 + 0.03934\times\text{bin } 11 + 0.02666\times\text{bin } 12 \\ & + 0.00526\times\text{bin } 14 + 0.00329\times\text{bin } 15 - 0.00701\times\text{bin } 16 - 0.91773\times L_{<7.5Å}\end{aligned}$$

This 15 bin–with–$L_{<7.5Å}$ model had an r^2 of 0.83, a $q_1{}^2$ of 0.79 and an average $q_{13}{}^2$ of 0.78 ± 0.01. In this model, the $L_{<7.5Å}$ variable selected 9 compounds all of which had a log RBA lower than –3.26. Smaller, compact molecules tended to bind weakly.

In Figures 1 and 2, the line of compounds predicted to have a log RBA of –2.60 is a set of compounds that did not have a hit in any of the 16 bins used to formulate the two CoSCoSA models. There were 56 compounds in the 16 bin CoSCoSA model that did not have any hit in one of the 16 bins. The 15 bin–with–$L_{<7.5Å}$ model had 52 compounds that did not have a hit in any of the 15 bins or $L_{<7.5Å}$. The removal of the compounds with no hits from the CoSCoSA models did not change the r^2 or q^2 of the model more than 2%. Three of the compounds in the 16 bin CoSCoSA model had residuals greater than 2 standard deviations (3β–androstanediol, genistein, and norethynodrel). In the 16 bin CoSCoSA model only one other compound had a predicted residual greater than 2 standard deviations and it was 4–hydoxy–tamoxifen). The 15 bin–with–$L_{<7.5Å}$ model had 4 compounds with no hits that had residuals greater than 2 standard deviations. They are the same 3 as before and 4,4'–sulfonyldiphenol. There were two compounds in the 15 bin–with–$L_{<7.5Å}$ CoSCoSA model with predicted residuals greater than 2 standard deviations, 4–hydoxy–tamoxifen and dihydroxymethoxychlor (HPTE). Almost all of the compounds with no hits in the 16 bins had experimental log RBA lower than –1.0. The CoSCoSA models did not find a spectral relationship for these weakly binding compounds to the estrogen receptor. Most of the other bins in both CoSCoSA models were used to form a relationship between a spectral bin and binding to the estrogen receptor with a log RBA stronger than –2.60.

Table 2. External Test Set Predictions. The Table contains the compound name, normalized experimental log RBA values to the estrogen receptor for each compound, predicted log RBA from the 16 Bin 2D–CoSCoSA model, predicted log RBA from the 15 Bin +$L_{<7.5Å}$ CoSCoSA model, and predicted log RBA from the CoMFA model [38]. The plus and minus sign reveals the variation seen when using none and one–half of a bin's intensity in a neighboring bins used to formulate the CoSCoSA model.

Name	Normalized log RBA	16–CoSCoSA	15+$L_{<7.5Å}$ –CoSCoSA	CoMFA
2–tert–butylphenol	–4.55	–2.66	–3.53	–3.83
3–tert–butylphenol	–4.82	–1.50 ± 0.64	–2.34 ± 0.66	–3.33
2,4,6,–triCl–4'–biphenylol	–0.16	–1.61	–1.56	–1.60
2–Cl–4,4'–biphenyldiol	–0.61	–1.61	–1.56	–1.49
2,6–dichloro–4'–biphenylol	–1.11	–1.61	–1.56	–2.41
2,3,5,6,tetraCl–4,4'–biphenyldiol	–2.18	–1.61	–1.56	–0.82
2,2',3,3',6,6'–hexaCl–4–biphenylol	–2.74	–2.14	–2.08	–3.06
2,2',3,4',6,6'–hexaCl–4–biphenylol	–2.60	–1.61	–1.56	–2.48
2,2',3,6,6'–pentaCl–4–biphenylol	–1.97	–1.61	–1.56	–3.07
2,2'5,5'–tetraCl–biphenyl	–2.67	–2.66	–2.61	–2.74
2,2',4,4',5,5'–heaxCl–biphenyl	–2.83	–2.66	–2.61	–1.52
2,2',4,4',6,6'–hexaCl–biphenyl	–1.87	–2.66	–2.61	–1.83
2,2',3,3',5,5'–hexaCl–6'–biphenylol	–2.69	–2.36	–2.29	–3.01
4'–deoxyindenestrol	–1.37	2.89 ± 0.63	2.96 ± 0.67	–0.53
4'–deoxyindenestrol	–0.23	2.89 ± 0.63	2.96 ± 0.67	0.111
5'–deoxyindenestrol	–0.59	–0.61	–0.59	–1.00
5'–deoxyindenestrol	0.35	–0.61	–0.59	–0.59
Indenestrol A (R)	1.08	3.95 ± 0.64	4.01 ± 0.67	0.29
Indenestrol A (S)	2.39	3.95 ± 0.64	4.01 ± 0.67	0.62
R 5020	–1.81	–2.45 ± 0.18	–2.48 ± 0.17	–0.70
Zearalenone	0.91	0.51	0.51	–0.12
5–Androstenediol	–0.49	–2.66	–2.61	–0.66
16a–bromoestradiol	1.41	–0.11	0.05	0.33
16–ketoestradiol	–0.38	–0.11	0.05	0.58
17–epi–estriol	0.98	–0.11	0.05	–0.16
2–OH–estrone	–0.19	1.26	1.32	0.36
Raloxifene	1.34	0.17 ± 0.63	0.20 ± 0.66	–0.24

Table 2 shows the predictions for 27 compounds. The first 21 compounds show the predictions for Waller's data set [37] using both the 16 bin and 15 bin–plus–$L_{<7.5Å}$ model of estrogen binding. To make the predictions, we simulated the 2D spectra of the 21 compounds, again using ACD Labs CNMR version 5.0 2D predictor software. The simulated spectra of the test set were binned into the same 605 bins. However, many of the occupied bins for these compounds did not fall into the original 605 occupied bins (that represent only 8.2% of the 2D COSY spectral plane). Therefore, if the simulated spectra did not fall into one of the original 605 populated bins, we put none, one–quarter, and one–half of the bin's intensity into the neighboring bin or bins used in the CoSCoSA model. We then built the CoSCoSA models as before and used its MLR equation to predict the log RBA of the compounds in the test set. Only 6 of the 27 compounds from Waller and Kuiper external data sets had binned COSY chemical shifts that were not in the original 605 bins and that bin that was within one bin of those 16 bins used to formulate a CoSCoSA model. In Table 1, for these 6 compounds we report predicted log RBA using one–quarter intensity in a neighboring bin and plus or minus the deviation seen when predicting the log RBA when using none and one–half intensity in the neighboring bin used for a CoSCoSA model. For Waller's test set and one quarter of a bin's intensity in neighboring bins we achieved a q_{pred}^2 of 0.50 for the 16 bin CoSCoSA model and a q_{pred}^2 0.57 for the 15 bin–plus–$L_{<7.5Å}$ CoSCoSA model. When using one half of a bin's intensity in a neighboring bin we got a q_{pred}^2 of 0.45 for the 16 bin CoSCoSA model and a q_{pred}^2 0.52 for the 15 bin–plus–$L_{<7.5Å}$ CoSCoSA model. Using none of a bin's intensity in a neighboring bin we got a q_{pred}^2 of 0.55 for the 16 bin CoSCoSA model and a q_{pred}^2 of 0.62 for the 15 bin–plus–$L_{<7.5Å}$ CoSCoSA model. A Comparative Mean Field Analysis (CoMFA) model had a q_{pred}^2 of 0.70 for Waller's normalized test set [10,37]. When Indenstrol A (R), Indenestrol A (S), 4'deoxyindenestrol (R), and 4'–deoxyindenestrol (S) are removed from Waller's test set and none of a bin's intensity from neighboring bins a q_{pred}^2 of 0.59 for the 16 bin model and a q_{pred}^2 0.74 for the 15 bin–plus–$L_{<7.5Å}$ model are achieved. Further inspection of the predictions for Indenstrol A (R), Indenestrol A (S), 4'deoxyindenestrol (R), and 4'–deoxyindenestrol (S) revealed that one chemical shift prediction that was highly suspect (142 ppm). When we checked this prediction the structures used in the prediction of this chemical shifts all the compounds had chemical shifts from 134 to 139 ppm and not 142 ppm.

Although both CoSCoSA models had an r^2 of 0.83, the LOO cross–validations of the models were always above 0.78. The cross–validations of both CoSCoSA models remained consistently above 0.78 whether LOO or leave–13–out (10% of training set) calculated them. Compared to CoMFA models formed in three–dimensional space, the ruggedness under cross–validation of the CoSCoSA models is related to the fact that the patterns are representations of a "digital–like" occupancy number of two–dimensional bins. This was true not only for the COSY spectral data, but also for the $L_{<7.5Å}$ variable inputted in a "digital–like" yes or no manner. In contrast, a published CoMFA model of the same 130 compounds based on analogue estimates of electric field spatial distributions at each grid point, had a remarkably good r^2 of 0.91 but much less impressive cross–validation results: q_1^2 of 0.66 and a mean q_{13}^2 of 0.65 [10]. The large falloff in CoMFA model quality under cross–validation indicates the extent to which the model was based on non–linear relationships among the input training data. By using semi–digital data representations and basing our 2D–CoSCoSA models only on multiply populated bins, some of the non–linear relationships are presumably removed.

We made log RBA predictions for 6 compounds from Kuiper's data set that had known experimental log RBA greater than –1.0 shown in Table 1 rows 22 to 27 [38]. We selected all 27 compounds from Kuiper's and Waller's data to make predictions. For 27 compounds, using one–quarter of a bin's intensity in neighboring bins we had a $\mathrm{q}_{\mathrm{pred}}^2$ of 0.42 for the 16 bin model and a q_{pred}^2 of 0.49 for the 15 bin–plus–$\mathrm{L}_{<7.5Å.}$ When using one half of a bin's intensity in a neighboring bin the 16 bin model q_{pred}^2 decreased to 0.38 and the q_{pred}^2 was 0.45 for the 15 bin–plus–$\mathrm{L}_{<7.5Å}$ model. When no intensity was used in the neighboring bins a q_{pred}^2 of 0.46 for the 16 bin C model and q_{pred}^2 of 0.53 for the 15 bin–plus–$\mathrm{L}_{<7.5Å}$ model were obtained. A CoMFA model had a q_{pred}^2 of 0.71 for the 27 compound normalized test set [10,37]. When Indenstrol A (R), Indenestrol A (S), 4'deoxyindenestrol (R), and 4'–deoxyindenestrol (S) are removed the 27 test compounds and one quarter of a bin's intensity in neighboring bins a q_{pred}^2 of 0.53 for the 16 bin model and a q_{pred}^2 0.64 for the 15 bin–plus–$\mathrm{L}_{<7.5Å}$ model are achieved.

When making predictions for Indenestrol A and Indenestrol B a deviation of 1.27 and 1.32 log RBA units due to the different models using none, one–quarter, and one–half of that bin's intensity in the nearest neighboring bin used in the original CoSCoSA models were obtained. This deviation is consistent with the experimental log RBA difference in binding activity between Indenestrol A and Indenestrol B of 1.31 log RBA units. The deviations for the other 3 of the 4 compounds due to different models using none, one–quarter, and one–half of a bin's intensity was about 0.17 to 0.67 log RBA units.

It is possible that the addition of more long–range atom–to–atom bins could make 3D–QSDAR modeling more accurate. When the CoSCoSA modeling becomes much more automated, it will become easier to optimize the variable selection technique and optimum bin size [40]. Previously, we have shown that 2D QSDAR models are very accurate and the addition of other through–space distance related information from a 3D–connectivity matrix made the QSDAR more accurate [23–25]. There is a need for NMR spectral prediction software that will incorporate the same atomic numbering system used in the pdb or mol files generated by cheminformatics programs for the numbering system of the chemical shift predictions. This will make the 3D–CoSCoSA modeling much more feasible, especially for large, structurally diverse data sets that include flexible structures. The 2D–QSDAR models had a q_1^2, and q_{13}^2 that was consistently 0.13 better than corresponding q_1^2 and q_{13}^2 seen in 3D–QSAR models [10]. Both 2D–CoSCoSA models of estrogenic compounds are based on through–bond connectivity 2D bins. Some of the selected bins in CoSCoSA models are associated with chemical shifts of neighboring carbon atoms in phenol rings. Likewise, CoMFA models showed a 0.05 increase in q_1^2 when a phenol indicator was used [10]. The X–ray structure of estradiol–estrogen receptor complex shows a pair of hydrogen bonds between Glu 353 and Arg 394 and the 3 position of estradiol [41]. Similar hydrogen bonds are found in co–crystal structures of diethylstillbesterol, 4–hydroxy–tamoxifen, and raloxifene with the estrogen receptor [40,42]. Without the use of X–ray structural information, the CoSCoSA models selected bins that represented a couple of different quantum states of hydrogen–bonding phenol rings. It has been shown that cross–validation q_1^2 of a CoMFA model is largely dependent on the subjective alignment of a molecule [43], whereas CoSCoSA models are independent of molecular alignment. CoMFA is an “analog” technique that relies on externally calculated three–dimensional physical fields. CoSCoSA modeling is “digital”–like and relies on internal NMR chemical shifts and distance–related atom–to–atom properties. The NMR chemical shifts are directly related to the Hamiltonian representation of the nuclear magnetic moment. The first order approximation for the diamagnetic term of the chemical shielding tensor is proportional to the electrostatic potential at the nucleus [26]. So the difference between two nuclear chemical shifts with similar electron orbital configurations are largely dependent on the difference in electrostatic potential at their respective nuclei. The electrostatic potential term in the Hamiltonian is used for CoMFA field calculations. Therefore, some of the information content the two modeling types are based on is quite similar, but the information is presented in different ways. The predictions based on the averaged CoMFA and CoSCoSA averages were the most accurate, exemplifying one way the two methods can complement each other. The use of a model based on the average of several models is not new, decision tree models that have used the average of several types of models produced better correlated models of internal and external test sets [44].

4 CONCLUSIONS

In the work presented here, the size of the two–dimensional bins used from the reduced 2D ^{13}C–^{13}C COSY matrix was not optimized. In any event, without serious bin size optimization efforts we developed accurate models of estrogenic compounds binding to the estrogen receptor. Since only 8.2% of the available 2D chemical shift “space” had hits in bins (and only 4.6% of the space had multiple hits), we believe that we can use this procedure effectively to build reliable models of this and other specific endpoints for even larger sets of non–congeners. In our 2D–CoSCoSA models, the r^2 and q_1^2 were still increasing with increasing number of bins used in the model. The addition of the one three–dimensional parameter, $L_{<7.5Å}$, increased r^2, q_1^2, q_{13}^2, by 1% and increased the q_{pred}^2 by 6 to 8% over the 16 bin 2D–CoSCoSA model. The CoSCoSA models proposed in this paper have not used information from other 2D planes available in a 3D–connectivity–matrix representation of a compound [23–25] except for $L_{<7.5Å}$, which means all 2D through–space planes for distances greater than 7.5 Å are null planes. Through–space distance 2D planes will provide important information when more than one area in a compound is important for binding activity, as is the case for most steroid receptor binding. The CoSCoSA models shown here were developed with simulated ^{13}C NMR spectral data, which does not involve intensive computation. We believe CoSCoSA models could be used with QSAR models of estrogen binding in phase III of the recently reported integrated "four phase" approach for priority setting of potential estrogenic disruptors [39].

Once 3D information can be obtained, the use of selected structural pharmacophores as "anchors" for defining the through–space distance spectral features in 3D–CoSCoSA models may provide a beneficial approach [23–25]. Overall, CoSCoSA modeling is very fast and can easily be automated. At the very least, 3D–QSDAR modeling presents a new option in modeling techniques that will complement SAR and 3D–QSAR modeling techniques.

Acknowledgment

We thank Dr. Weida Tong of NCTR and Dr. Huixiao Hong from the Northrop Grumman Corporation located at NCTR who provided us with the three–dimensional mol files for all 130 compounds in the training set. These files were used in the ACD CNMR prediction software and to determine the $L_{<7.5Å}$ descriptor used in the second CoSCoSA model.

Supplementary Material

The 130 mol files for the compounds used in training of the two CoSCoSA models in this paper are deposited as supplementary material. The mol files were saved using the ACD ChemSketch software.

5 REFERENCES

[1] U.S. Congress. The Food Quality Protection Act (FQPA) and Safe Drinking Water Act (SDWA), 1996.

[2] R. W. Brueggemeier, X. Gu, J. A. Moblet, S. Hoomprabutra, A. S. Bhat, and J. L. Whetstone, Effects of phytoestrogens and synthetic combinatorial libraries on aromatase, estrogen biosyntheses, and metabolism, *Annals of the New York Academy of Sciences*. **2001**, *948*, 51–66.

[3] A. Cassidy and S. Milligan, How significant are environmental estrogens to women? *Climacteric*. **1998**, *1*, 229–242.

[4] R. S. R. Zand, D. J. Jenkins, T. J. Brown, and E. P. Diamandis, Flavonoids can block PSA production by breast and prostate cancer cell lines, *Clinica Chimica Acta*. **2002**, *317*, 17–26.

[5] C. L. Van Patten, I. A. Olivotto, G. K. Chambers, K. A. Gelmon, T. G. Hislop, E. Templeton, A. Wattie, and J. C. Prior, Effect of soy phytoestrogens on hot flashes in postmenopausal women with breast cancer: a randomized, controlled clinical trial, *Journal of Clinical Oncology*. **2002**, *20*, 1449–1495.

[6] P. Diel, S. Olff, S. Schmidt, and H. Michna, Effects of environmental estrogens bisphenol A, o,p'–DDT, p–tert–octylphenol and coumestrol on apoptosis induction, cell proliferation and the expression of estrogen sensitive molecular parameters in the human breast cancer cell line MCF–7, *Journal of Steroid Biochemistry & Molecular Biology*. **2002**, *80*, 61–70.

[7] W. Washington, L. Hubert, D. Jones, and W. G. Gray, Bisphenol A binds to the low–affinity estrogen binding site, *In Vitro & Molecular Toxicology*. **2001**, *14*, 43–51.

[8] S. C. Nagel, F. S. vom Saal, K. A. Thayer, M. G. Dhar, M. Boechler, and W. V. Welshons, Relative binding affinity–serum modified access (RBA–SMA) assay predicts the relative *in vitro* bioactivity of the xenoestrogens bisphenol A and octylphenol, *Environ Health Perspect*. **1997**, *105*, 70–76.

[9] R. M. Blair, H. Fang, W. S. Branham, B. S. Hass, S. L. Dial, C. L. Moland, W. Tong, L. Shi, R. Perkins, and D. M. Sheehan, The Estrogen receptor relative binding affinities of 188 natural and xenochemicals: Structural diversity of ligands, *Toxicological Sciences*. **2000**, *54*, 138–153.

[10] L. M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R. M. Blair, W. S. Branham, S. L. Dial, C. L. Moland, and D. M. Sheehan, QSAR Models Using a Large Diverse Set of Estrogens, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 186–195.

[11] R. D. Cramer, D. E. Paterson, and J. D. Bunce, 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] P. R. N. Jayatilleke, A. C. Nair, R. Zauhar, and W. J. Welsh, Computational Studies on HIV–1 Protease Inhibitors: Influence of Calculated Inhibitor–Enzyme Binding Affinities on the Statistical Quality of 3D–QSAR CoMFA Models, *J. Med. Chem.* **2000**, *43*, 4446–4451.

[13] R. Beger, J. P. Freeman, J. Lay Jr., J. Wilkes, and D. Miller, ^{13}C NMR and EI mass spectrometric data–activity relationship (SDAR) model of estrogen receptor binding, *Toxicology and Applied Pharmacolgy*. **2000**, *169*, 17–25.

[14] R. D. Beger, J. P. Freeman, J. O. Lay, Jr., J. G. Wilkes, and D. W. Miller, The Use of ^{13}C NMR Spectrometric Data to produce a Predictive Model of Estrogen Receptor Binding Activity, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 219–224.

[15] R. D. Beger and J. G. Wilkes, Developing ^{13}C NMR quantitative spectrometric data–activity relationship (QSDAR) models of steroid binding to the corticosteroid binding globulin, *J. Comput.–Aided Mol. Design*. **2001**, *15*, 659–669.

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

[17] R. D. Beger and J. G. Wilkes, 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**, *41*, 1322–1329.

[18] P. L. Andersson, A. S. A. M. van der Burght, M. van der Berg, and M. Tysklind, Early life–stage mortality in zebrafish (Danio rerio) following maternal exposure to polychlorinated biphenyls and estrogen, *Environ. Toxico. Chem.* **2000**, *19*, 1454–1463.

[19] R. Bursi, T. Dao, T. van Wilk, M. de Gooyer, E. Kellenbach, and P. Verwer, 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.

[20] *ACD Labs* software version 5.0, Toronto, Canada.

[21] J. Meiler, R. Meusinger, and M. Will, Fast Determination of ^{13}C NMR Chemical Shifts Using Artificial Neural Networks, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1169–1176.

[22] Bio–Rad Laboratories, HaveItAll(™) NMR software, Philadelphia, PA.

[23] R. D. Beger, D. A. Buzatu, J. G. Wilkes, and J. O. Lay, Jr., Comparative Structural Connectivity Spectra Analysis (CoSCoSA) Models of Steroid Binding to the Corticosteroid Binding Globulin, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1123–1131.

[24] R. D. Beger and J. G. Wilkes, Comparative structural connectivity spectra analysis (CoSCSA) models of steroids binding to the Aromatase Enzyme, *J. Mol. Recognition*. **2002**, *15*, 154–162.

[25] R. D. Beger, D. A. Buzatu, and J. G. Wilkes, Combining NMR Spectral and Structural Data to Form Models of Polychlorinated Dibenzodioxins, Dibenzofurans, and Biphenyls Binding to the AhR, *J. Comput.–Aided Mol. Design*. **2003**, *16*, 727–740..

[26] J. W. Emsley, J. Feeney, and L. H. Sutcliffe, High Resolution Nuclear Magnetic Resonance. Volume I. Pergamon Press Ltd.: Oxford; **1965** pp 1–287.

[27] W. P. Aue, E. Bartholdi, and R. R. Ernst, Two–dimensional spectroscopy. Application to nuclear magnetic resonance, *J. Chem. Phys.* **1976**, *24,* 2229–2246.

[28] M. Randić, On the characterization of molecular branching, *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.

[29] F. R. A. Burden, Chemically intuitive molecular index based on eigenvalues of a modified adjacency matrix, *Quant. Struct.–Act. Relat.* **1997**, *16*, 309–314.

[30] E. Estada and E. Molina, 3D Connectivity Indices in QSPR/QSAR Studies. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 791–797.

[31] L. B. Kier and L. H. Hall, An electrotopological–state index for atoms in molecules, *Pharm Res.* **1990**, *7*, 801–807.

[32] W. S. Branham, S. L. Dial, C. L. Moland, B. S. Hass, R. M. Blair, H. Fang, L. Shi, W. Tong, R. G. Perkins, and D. M. Sheehan, Phytoestrogens and mycoestrogens bind to the rat uterine estrogen receptor, *J. Nutr.* **2002**, *132*, 658–664.

[33] B. Lefebvre and A. Williams, ^{13}C NMR Chemical Shift Prediction A Comparison of Methods and a Case Study Analysis of Paclitaxel (TAXOL®), 44th ENC poster presentation, **2003**.

[34] W. Bremser, HOSE – a Novel substructure Code. *Anal. Chim. Acta*. **1978**, *103*, 355–36.

[35] Statistica, version 6.0, StatSoft,Tulsa, OK, 2001.

[36] R. D. Cramer, J. D. Bunce, and D. E. Patterson, Cross–validation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies, *Quant. Struct.–Act. Relat.* **1988**, *7*, 18–25.

[37] C. L. Waller, T. I. Opera, K. Chae, H. K. Park, K. S. Korach, S. C. Laws, T. E. Wiese, W. R. Kelce, and L. E. Gray, Jr., Ligand–based identification of environmental estrogens, *Chem. Res. Toxicol.* **1996**, *9*, 1240–1248.

[38] G. G. J. M. Kuiper, B. Carlsson, K. Grandien, E. Enmark, J. Haggblad, S. Nilsson, and J.–A. Gustafsson, Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors α and β, *Endocrinology* **1997**, *138*, 863–870.

[39] H. Fang, W. Tong, R. Perkins, A. Soto, N. Prechtl, and D. M. Sheehan, Quantitative comparison of in vitro assays for estrogenic assays, *Environ*. *Health Prospect*. **1998**, *139*, 723–729.

[40] K. Baumann, H. Albert, and M. von Korff, A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I. Search algorithm, theory, and simulations, *J. Chemometrics* **2002**, *16*, 339–350.

[41] A. M. Brzozowski, A.C. Pike, Z. Dauter, R. E. Hubbard, T. Bonn, O. Engstrom, L. Ohman, G. L. Greene, and J. A. Gustafsson, Carlquist, M. Molecular basis of agonism and antagonism in the oestrogen receptor, *Nature* **1997**, *389*, 753–758.

[42] A. K., Shiau, D. Barstad, P. M. Loria, L. Cheng, P. J. Kushner, D. A. Agard, and G. L. Greene, The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen, *Cell* **1998**, *95*, 927–937.

[43] S. J. Cho and A. Tropsha, Cross–validated R^2–guided region selection for comparative molecular field analysis: A simple method to achieve consistent results, *J. Med. Chem*. **1995**, *38*, 1060–1066.

[44] W. Tong, H. Hong, H. Fang, Q. Xie, and R. Perkins, Decision Forest: Combining the predictions of multiple independent decision tree models, *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 525–531.