

Comparative Structural Connectivity Spectra Analysis (CoSCSA) Models of Steroid Binding to the Corticosteroid Binding Globulin

**RICHARD D. BEGER*, DAN A. BUZATU, JON G. WILKES,
and JACKSON. O LAY JR.**

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

*Correspondence: RD Beger, Division of Chemistry, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR 72079-9502, USA

Phone: (870) 543-7080. FAX: (870) 543-7686. E-mail: rbeger@NCTR.FDA.GOV

SUMMARY

A three-dimensional quantitative spectrometric data-activity relationship (3D-QSDAR) model was developed that is built by combining NMR spectral information with structural information in a 3D-connectivity matrix. The 3D-connectivity matrix is built by displaying all possible carbon-to-carbon connections with their assigned carbon NMR chemical shifts and distances between the carbons. Selected 2D ^{13}C - ^{13}C COrrrelation SpectroscopY (COSY) (through-bond nearest neighbors) and selected theoretical 2D ^{13}C - ^{13}C distance connectivity spectral slices from the 3D-connectivity matrix to produce a relationship among the spectral patterns for 30 steroids binding to corticosteroid binding globulin. We call this technique as comparative structural connectivity spectra analysis (CoSCSA) modeling. A CoSCSA principal component linear regression model based on the combination of ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance spectra principal components (PCs) had an r^2 of 0.96 and a leave-one-out (LOO) cross-validation q^2 of 0.92. A CoSCSA parallel distributed artificial neural network (PD-ANN) model based on the combination of ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance spectra had an r^2 of 0.96, a leave-three-out q_3^2 of 0.78, and a leave-ten-out q_{10}^2 of 0.73. CoSCSA modeling attempts to uniquely combine the quantum mechanics information from the NMR chemical shifts with internal molecular atom-to-atom distances into a high powered modeling technique. The CoSCSA modeling technique has the flexibility and accuracy to outperform the cross-validated variance q^2 of previously published quantitative structure-activity relationship (QSAR), quantitative spectral data-activity relationship (QSDAR), self-organizing map (SOM), and electrotopological state (E-state) models.

INTRODUCTION

Many different types of models have been developed to predict the binding activity for the compound-receptor system of the corticosterone binding globulin (1). These corticosteroid binding globulin models include the standard quantitative structure-activity relationship (QSAR) (2), the hybrid electrotopological state (E-state) (3), the self-organizing map (SOM) (4), and the combination QSAR E-state models (5). Previously, we have demonstrated that ^{13}C NMR spectrometric data can be used to produce reliable quantitative spectrometric data-activity relationship (QSDAR) models of binding to the corticosterone binding globulin (6), aromatase enzyme (7), and aryl hydrocarbon receptor (8). These comparative spectral analysis (CoSA) models using simulated ^{13}C NMR data yielded higher cross-validated correlations than were seen with comparative molecular field analysis (CoMFA) methods and comparative structure assigned spectra analysis (CoSASA) (6,7). In this paper, we demonstrate CoSCSA modeling is an improvement over previous CoSA, CoSASA and CoMFA models of binding to the corticosterone binding globulin.

Our previous QSDAR CoSA model for the 30 corticosterone binding globulin steroids was significantly robust with an r^2 of 0.80 and a q^2 of 0.78 (6), but we believed we could do better if we were able to add structural information to the one-dimensional CoSA models. We added structurally assigned chemical shift to the backbone of steroids we obtained a CoSASA model with an r^2 of 0.80 and a q^2 of 0.73 (6). We thought the addition of two-dimensional structural information would drastically improve the model, but the results were approximately the same for the 1D CoSA and 2D CoSASA models.

The power of QSDAR is that it is unnecessary to solve any quantum mechanic calculations or use the structures of molecules for electrostatic calculations as is done in QSAR techniques (2,9-13). QSAR is based on the assumption that there is a relationship between structure and activity of a compound. QSAR modeling results show that receptor binding of a compound can be predicted from a combination of electrostatics potentials and geometrical structural analysis (2,12,13).

^{13}C nuclear magnetic resonance (NMR) chemical shifts have been used to predict and refine chemical structures (14,15). Conversely, the chemical structure of a compound has been used to predict its ^{13}C NMR chemical shifts (16). The ^{13}C NMR spectrum of a compound contains frequencies that correspond directly to the quantum mechanical properties of every ^{13}C nuclear magnetic moment in a chemical structure. The quantum mechanical description of a molecule depends largely on its electrostatic features and three-dimensional geometry (17).

Current QSAR and structure-activity relationship (SAR) models use computer modeling of the molecule or break the molecule into secondary structural pieces (8-12,18-20). Many calculations are used in QSAR, SOM or E-states models. Using a specific molecular structure for computer modeling of each compound dramatically extends the number of calculations required to define the model. Moreover, the selection of the most appropriate 3D structure for each molecule requires a number of assumptions. The necessary simplifying assumptions in some cases give results that are hard to replicate or are inaccurate.

The present research initiative avoids some of the foregoing problems by providing a method for predicting a biological activity of a molecule, by using the ^{13}C NMR spectral data for a test compound and adding the molecules structural connectivity information into a 3D-connectivity matrix. The 3D-connectivity matrix is built by displaying all possible carbon-to-carbon connections and their assigned carbon NMR chemical shifts, so that the x-axis is the chemical shifts of carbon i, the y-axis is the chemical shifts of carbon j, and the z-axis is the distance between carbon i and carbon j (r_{ij}). The information in a 3D-connectivity matrix over determined the compound's structure, so we can reduce the information in the matrix that is used in a model. One way to reduce the information is to reduce the 3D

matrix into a set of 2D planes. We decided to reduce the 3D-matrix into four 2D spectral planes. The first 2D plane was the nearest neighbor through-bond connectivity plane. The three other 2D planes were constructed from compressing distance information on the z-axis, one for all short atom-to-atom connections ($2.0 \text{ \AA} < r_{ij} < 3.6 \text{ \AA}$), one for all medium atom-to-atom connections ($3.6 \text{ \AA} < r_{ij} < 6.9 \text{ \AA}$), and one for all long atom-to-atom connections ($r_{ij} > 6.9 \text{ \AA}$). Similarities between the pattern of 2D spectral data associated with the biological activity of the training set compounds and the spectral data for the test compound are detected by principal component regression to determine whether the compound is predicted to exhibit the biological activity.

Standard NMR instrumental techniques include 2D ^1H - ^1H COSY (21) experiments in which connectivity relationships through three bonds are found for nearest neighbor protons with an off diagonal cross peak. This experiment is similar to 2D ^{13}C - ^{13}C COSY experiments that contain analogous through bond connectivity spectral patterns. For the 3D-QSDAR methods developed in this work, the information contents of such COSY experiments are paralleled in the shortest distance layer of the connectivity matrix. Using the structurally assigned predicted spectra and adding the nearest neighbor information as cross peaks produces this layer.

Our 3D-QSDAR predictive methods also bears comparison to several other multi-dimensional NMR experimental techniques. ^{13}C - ^{13}C COSY experiments have similarities to 2D ^1H - ^{13}C heteronuclear single quantum correlation (HSQC) (22) and ^1H - ^{13}C heteronuclear multiple quantum correlation (HMQC) (23) NMR experiments that show the connectivity for carbons and their attached protons. In practice, 2D ^{13}C - ^{13}C COSY are seldom run because small molecules are rarely fully ^{13}C labeled. Even if the molecules were fully labeled, the ^{13}C through-bond connections usually are obtained directly from other NMR experiments like HCCH or indirectly by combining the information from ^1H - ^1H COSY with ^{13}C - ^1H HMQC and heteronuclear multiple bond correlation (HMBC) (24) NMR experiments. NMR 2D ^{13}C - ^{13}C COSY spectral data is similar conceptually to ideas that are used in the first order $^1\chi$ connectivity indices (25) and the modified adjacency matrix (26).

In 3D-QSDAR, 2D ^{13}C - ^{13}C distance spectra that contain short, medium, and long through-space connectivity spectral patterns that are also produced by using the structurally assigned predicted spectra and selecting a distance range for nucleus to nucleus distance (r). This is analogous to 2D ^1H - ^1H Nuclear Overhauser Effect Spectroscopy (NOESY) NMR experiments where correlations through space are found for neighboring protons that are less than 5 \AA away with an off-diagonal cross-peak (27). The size of the cross peak in the NOESY experiment is dependent on the distance between the protons, the mixing time of the experiment, and the number of different NOE spin diffusion pathways available for dipolar magnetization transfer. ^{13}C - ^{13}C NOESY experiments are for all practical purposes never executed, again because most small compounds are not fully ^{13}C labeled.

The binding of steroids to many receptors including to the corticosteroid binding globulin are strongly linked to the structural environment around position 3 and 17 of the steroids (1-3). In the 3D-connectivity matrix, we can select which 2D spectral-distance range to observe. The distance between position 3 and 17 in a steroid backbone would fall in the long atom-to-atom connections of the 3D-connectivity matrix. We selected a distance where many of the carbon atoms in the A-ring are connected through space to many of the carbon atoms in the D-ring and chains extended off from position 17 in the D-ring. Presently there are no NMR experiments that directly record structural distance information that is greater than 5 \AA apart. The distance related topological indices (28) and the 2D ^{13}C - ^{13}C distance connectivity spectral data are based on similar distance related concepts.

This paper demonstrates that structural information combined with ^{13}C NMR spectra in the form of through-bond and through spatial distance connectivity information can be used to produce a reliable QSDAR model of steroids binding to the corticosteroid binding globulin.

PROCEDURES

Figure 1 and Table 1 contains previously reported steroids binding data used for training these models. Each compound in Table 1 had their ^{13}C NMR spectra simulated using the Advanced Chemistry Developments (ACD) Labs CNMR predictor software, version 4.0 (29). For QSDAR CoSCSA modeling we used the predicted ^{13}C NMR spectra. The use of predicted NMR spectra is not necessary to build the QSDAR models, but it saves time and money. Predicted ^{13}C NMR data allow for the spectra to be independent of the solvent used. The CoSCSA modeling, LOO cross-validation, and prediction processes were completely computerized.

Figure 2 shows the flow chart on the CoSCSA modeling procedure. The structures are used to predict 1D ^{13}C NMR spectra. We reduced the resolution of the 2D spectra to 2.0 and 3.0 ppm in both dimensions to populate many of the NMR bins for statistical analysis and reduce the effects of uncertainties in the simulated spectra. The spectral widths were chosen because of convenience. The 2D ^{13}C - ^{13}C NMR 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. This was done so that all the carbon to carbon connections would have a similar signal-to-noise ratio.

The predicted NMR spectra were calculated by a substructure similarity technique called HOSE (30), which correlates similar structures with similar NMR chemical shifts. 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 the training set to predict unknown sample affinities for compound spectra predicted using the same HOSE routine.

We use the structurally assigned ^{13}C NMR spectra to produce predicted 2D ^{13}C - ^{13}C COSY and theoretical 2D ^{13}C - ^{13}C distance spectra. The arrows in Figure 3A show the through-bond neighboring carbon-to-carbon connections of a steroid backbone molecule without any side chains. These through bond carbon-to-carbon connections were used to simulate a 2D ^{13}C - ^{13}C COSY spectrum of the steroid compounds. The arrows in Figure 3B show the through space carbon to carbon connections that are greater than 6.9 Å apart in a steroid backbone without any side chains. These through space carbon-to-carbon connections and any other through space carbon to carbon distance connections that were greater than 6.9 Å were used to produce a theoretical 2D ^{13}C - ^{13}C distance connectivity spectra that had cross-peaks when two carbons were greater than 6.9 Å apart. The 2D ^{13}C - ^{13}C COSY and 2D ^{13}C - ^{13}C distance connectivity spectra are symmetrical across the diagonal and for modeling purposes only half of each individual spectrum was used. We did not use the 1D ^{13}C NMR spectra in these CoSCSA models because the 1D chemical shifts do not present any new information.

Four CoSCSA models of binding to the corticosteroid binding globulin can be built from the 2D COSY and 2D long range distance spectra. One model can be made from the 2D COSY spectra and one model can be developed using the 2D long-range distance spectra. We can combine the COSY and distance spectra in two different ways, using the combined spectra (3D) or using the combined principal components (PCs) from the COSY PCs and Distance PCs. In Figure 2 arrow A represents the 2D ^{13}C - ^{13}C COSY spectra data that were reduced to PCs. These PCs were then used for multiple linear regression to produce a model for the 2D ^{13}C - ^{13}C COSY data. In Figure 2, arrow B represents the 2D ^{13}C - ^{13}C distance connectivity data that were reduced to PCs. These PCs were then used for multiple linear regression to produce a model for the 2D ^{13}C - ^{13}C distance connectivity data. In figure 3 arrow C represents the procedure that used the combined PCs from the 2D ^{13}C - ^{13}C COSY (arrow A) and the 2D ^{13}C - ^{13}C distance connectivity (arrow B) to produce a combined through bond and through space CoSCSA model. In figure 3 arrow D represents the procedure that used the combined 2D ^{13}C - ^{13}C COSY and the 2D ^{13}C - ^{13}C distance connectivity spectral

data that were then reduced to PCs. These PCs were then used for multiple linear regression to produce a model for the three-dimensional representation of the 2D ^{13}C - ^{13}C COSY and 2D ^{13}C - ^{13}C distance spectra data.

All principal component linear regression statistical analyses were performed by Statsoft Statistica software version 5.5 and 6.0 (31). The CoSCSA QSDAR models were produced in which the connectivity bins were evaluated with forward multiple linear regression analysis using only the most correlated PCs from both the 2D ^{13}C - ^{13}C COSY and 2D ^{13}C - ^{13}C distance connectivity spectra. The F-test for many of the models continued to rise until we reached the number of components in the model equal to the number of compounds in the training set. This is why we limited the number of PC's used in the CoSCSA models to be 3 and 8.

The analysis of each PCR CoSCSA model was done by the LOO cross-validation procedure where each compound is systematically excluded from the training set and its binding activity is predicted by the model (32). The cross-validated r^2 (termed q^2) can be derived from $q^2=1-(\text{PRESS})/\text{SD}$. Where PRESS is the sum of the differences between the actual and predicted activity data for each molecule during LOO cross-validation, and SD is the sum of the squared deviations between the measured and mean activities of each molecule in the training set. We believe that q^2 is a more valid measure than r^2 for assessing the reliability of a mathematical model intended for predictive applications.

Another model for predicting steroid binding activity to the corticosteroid binding globulin was developed using a parallel distributed-artificial neural network (PD-ANN). This artificial neural network consists of an error back-propagating, feed forward code, developed "in house" using JGravity, a Java-based parallel distribution platform (33). We are participating in the beta test phase of JGravity development. The core back-propagation algorithm in the PD-ANN is based on Rumelhart, Hinton, and Williams' generalized delta rule (34).

The PD-ANN receives raw data broken into bins with the population in each bin assigned to its own input node. These nodes are interconnected to a "hidden layer" containing a number of processing elements called hidden nodes. The hidden layer is finally interconnected to the output layer. The training of each neural network in the PD-ANN consists of repeated cycles of presenting examples to the input nodes while simultaneously presenting corresponding output examples to the output node. The back propagation code adjusts weights between the input, hidden, and output layer connections using a gradient descent least squares technique to reduce the error between the net predicted output and the actual output on which it is training. Training stops when the error between the actual output values and the net produced output values drops below a predetermined threshold value.

In this study, spectral data were read into each network through the input nodes and passed through one hidden layer that was connected to an output layer. Thus the neural network architecture consisted of three fully connected layers. A 593 node input layer was connected to a hidden layer, which was then connected to a one node output layer. The 593 nodes of the input layer corresponded to the number of occupied 2D COSY and 2D long-range distance spectral bins. Transfer functions were used within the back-propagation algorithm to calculate the values of the hidden and output nodes. The one node of the output layer corresponded to the binding activity for each molecule.

The parallel distributed network algorithm was used to optimize the neural network configuration. Optimization refers to the search of the multidimensional space created by the adjustable parameters of the back propagation code to find the network configuration that provides the best prediction of binding affinity based on spectral data. The algorithm, which resided on a central computer, distributed neural networks with different configurations in parallel to four personal computers. The impacts of varying permutations and combinations of the number of hidden nodes, the number of (back-propagation) cycles, the transfer functions (sigmoid, hyperbolic tan, or arc tan), and the learning rate were evaluated based on each ones cross-validation results at benchmark cycles during the back-propagation process. Validation or testing of each configuration was performed on each participating computer node, and only the results

were returned to the central computer.

The algorithm on each computer node removed a small number of examples from the training set for testing. Each network then trained to the specified number of cycles on the somewhat abridged set using the parameters assigned by the central computer. When training completed, the removed data was propagated through the trained network and predictions were obtained, the quality of which was subject to validation. On each participating computer node, this process of removing data, training the network, and propagating the removed data was repeated by the algorithm with each network configuration, until all of the training set samples were used to test the network. At that time the master computer assigned a new configuration for the node computer to work on. This assignment and reassignment process was repeated on all the available computer nodes until the master computer finished its task list. The task list is a user-specified list of all the test configurations.

More than 100 varying network configurations were evaluated over a couple of days using four high speed Pentium personal computers. Two types of cross validations were performed. In one case each distributed neural network removed three examples and trained on the rest, rotating through the rest of the data on subsequent training sessions, and thus performing a comprehensive series of leave-three-out cross validations. In another case 10 examples were left out, performing a leave 10 out cross validation.

The configuration consisting of 593 input, 198 hidden, and 1 output nodes yielded the best results when trained for 3400 cycles, with a learning rate of 0.1, and using a sigmoid transfer function. Use of a sigmoid transfer function in the back-propagation algorithm required that the data (input and output) be scaled within a range between 0 and 1. The raw data for this model was actually scaled from 0.1 to 0.9.

RESULTS

Table 2 contains a comparison of the model performance parameters n , r^2 , q^2 , and number of components for the QSAR (2), HE-state/E-state (3), E-state (3), SOM (4), combination QSAR/E-state (5), 2D CoSASA (6), 1D CoSA (6), four CoSCSA, and 2 CoSCSA PD-ANN models. All four CoSCSA models with 8 PCs have a strong correlation (r^2) and cross-validated variance (q^2) and are favorable when compared to the previous published models of binding to the corticosteroid binding globulin.

Figure 4A is a plot of the predicted binding versus experimental binding for the the CoSCSA 2.0 ppm resolution model based on ^{13}C - ^{13}C COSY data. A model based on 8 PC's had an explained correlation (r^2) of this model is 0.93 and the cross-validated variance (q^2) is 0.88, which indicates self-consistency and excellent predictive capability. Figure 4B is a plot of the predicted binding versus experimental binding for the CoSCSA 2.0 ppm resolution model based on ^{13}C - ^{13}C distance greater than 6.9 Å data. Using 8 PCs the explained correlation (r^2) of this model is 0.89 and the cross-validated variance (q^2) is 0.79. Figure 4C is a plot of the predicted binding versus experimental binding for the CoSCSA 2.0 ppm resolution model based on the combined ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance connectivity PCs. The explained correlation (r^2) of this model is 0.96 and the cross-validated variance (q^2) of this model is 0.92, which indicates excellent predictive capability. Figure 4D is a plot of the predicted binding versus experimental binding for the CoSCSA 2.0 ppm resolution model based on the combined ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance connectivity spectral data before principal component extraction. The explained correlation (r^2) of this model is 0.92 and the cross-validated variance (q^2) of this model is 0.81.

All of the spectral bins based on the combination of ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance spectra were used to develop the CoSCSA PD-ANN model. The explained correlation (r^2) of this model is 0.96. Two cross-validated variance

coefficients were calculated for this model. The leave three out (q_3^2) procedure yielded a value of 0.78, and the leave ten out (q_{10}^2) procedure yielded a value of 0.73. These numbers illustrate a high degree of consistency and predictive capability for the PD-ANN model.

DISCUSSION

All four CoSCSA models based on 8 PCs have a q^2 greater than the 0.68 seen for the QSAR model. Three of the four CoSCSA models based on 3 PCs have a q^2 greater than the 0.68 seen for the QSAR model. The only CoSCSA model that did not have a q^2 greater than 0.68 was the ^{13}C - ^{13}C distance connectivity model based on only 3 PCs. The HE-state and E-state models have a greater r^2 than all the QSDAR models but these models are very computational-intensive with many distance formulas used for every point in the grid. Still, all the 2.0 ppm resolution CoSCSA models based on 8 PCs have explained variance (r^2) greater than 0.89 and a cross-validated variance (q^2) greater than of 0.79. All the CoSCSA models with 8 PCs have a predictability that are much better or comparable to the predictability for QSAR, CoSA, CoSASA, HE-state/E-state, and E-state models. The reason we are comparing models based on 8 PCs to models based on 3 or 4 components is that these CoSCSA models are "digital" in nature and the other QSAR, HE-states, and E-states models are in "analog" format. Digital information needs more components to present the same information (10 binary components to represent a number less than 999) as analog electronics (3 variable components to represent a number less than 999) but the resulting information is presented with a higher signal to noise value. Our CoSCSA models see essentially the same thing, although we are displaying same electrostatic information that is used in QSAR or E-states models, our CoSCSA models have a better signal to noise (predictability) than other models when we use more components.

Another explanation for the fact that the cross-validated variance of the CoSCSA model was as good as the other models is that even simulated NMR spectral data are more accurate than the errors introduced by solvent effects, partial charges, dielectrics, and structural conformations used during the calculation of electrostatic potentials. All of the assumptions and approximations are prone to produce significant error. ^{13}C NMR spectral data takes into account all structural conformations and complete solvent effects to produce a "quantum mechanical energy" that represents the average structural environment for every carbon atom in the molecule.

In our earlier CoSA QSDAR model we started with only 256 spectral bins, a number then reduced to 94 spectral bins when all the columns with only zeroes or with only one non-zero entry were removed. In our 2.0 ppm CoSCSA models we started with 6441 two-dimensional bins, a number then reduced to 271 for the ^{13}C - ^{13}C COSY data and 322 for the ^{13}C - ^{13}C distance connectivity data when all the columns with only zeroes were removed. The models used less than 5% of the available 2D connectivity spectral space with this training set and a 2 ppm resolution bin size.

We wanted to determine the effect of combining all bins with only one "hit" in the bin to the nearest bin with a "hit". When multiple bins with "hits" were equally close to the bin with one "hit", we consistently moved the bin with one "hit" to the bin with least number of "hits". When we combined all the bins with one "hit", the 2 ppm ^{13}C - ^{13}C COSY had 93 of the 271 bins removed and the 2ppm ^{13}C - ^{13}C distance connectivity data had 128 of the 322 bins removed. Using the new ^{13}C - ^{13}C COSY data with no bins with only one "hit" in a bin for a model the r^2 increased from 0.93 to 0.94 and q^2 increased from 0.88 to 0.89. With the reduced ^{13}C - ^{13}C distance connectivity data we produced a model where the r^2 increased from 0.89 to 0.91 and q^2 increased from 0.79 to 0.81. Using both the reduced ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance connectivity PCs for a model, the r^2 decreased from 0.96 to 0.95 and q^2 decreased from 0.92 to 0.90. Using both new ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance connectivity combined data before the extraction of PCs the r^2 increased from 0.92 to 0.93 and q^2 increased from 0.81 to 0.84.

We wanted to determine the effect of increasing the bin size to 3.0 ppm. In our 3.0 ppm CoSCSA models we started with 2926 two-dimensional bins, a number then reduced to 199 for the ^{13}C - ^{13}C COSY data and 253 for the ^{13}C - ^{13}C distance connectivity data when all the columns with only zeroes were removed. For the model based on ^{13}C - ^{13}C COSY data, the r^2 decreased from 0.93 to 0.87 and q^2 decreased from 0.88 to 0.79. For the model based on ^{13}C - ^{13}C distance connectivity data, the r^2 increased from 0.89 to 0.90 and q^2 decreased from 0.79 to 0.74. For the model based on the combined ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance connectivity PCs, the r^2 decreased from 0.96 to 0.89 and q^2 decreased from 0.96 to 0.80. For the model based on the combined ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance connectivity data before extraction of principal components, the r^2 was unchanged at 0.92 and q^2 increased from 0.81 to 0.84.

We wanted to determine the effect of using different distance cutoff for the ^{13}C - ^{13}C distance connectivity spectral CoSCSA models. Instead of using all distance connections greater than 6.9 Å we used the same set of atom to atom distance for all the compounds. This meant using the distance connectivity set from the smallest compounds (no chains off of the steroids) for all the compounds. The smallest compounds had 26 distance connectivity interactions greater than 6.9 Å, 13 on each side of the 2D ^{13}C - ^{13}C long-range distance spectra. When using all distance connection interactions greater than 6.9 Å, the number of interactions varied for each compound. When we used only this new ^{13}C - ^{13}C distance connection spectra data with 13 interactions for each compound to build a CoSCSA model, the r^2 of 0.89 did not change and the q^2 decreased from 0.79 to 0.72. When we used the new ^{13}C - ^{13}C distance connectivity PCs based on the same 13 distance connections for each steroid compound with the original ^{13}C - ^{13}C COSY PCs to build a new CoSCSA model, the r^2 decreased from 0.96 to 0.95 and the q^2 decreased from 0.92 to 0.89. Using the original ^{13}C - ^{13}C COSY and the new ^{13}C - ^{13}C distance connectivity combined data before the extraction of PCs the r^2 increased from 0.92 to 0.95 and q^2 increased from 0.81 to 0.93.

It is interesting to note that a strong predictive model was produced by the PD-ANN using the combined COSY and distance ^{13}C spectral data. There was no pre-processing (such as principal components or other types of regression analysis to extract features) performed on the data prior to its use. This model used the raw spectral information. The strength of the model is clearly demonstrated with the q_{10}^2 of 0.73. The total data set consisted of only 30 compounds. Thus each time 10 compounds were left out, the data set use to build the model was depleted by 33%.

The CoSCSA models takes into account the average uncertainty in the predicted ^{13}C NMR data. It therefore reduces the information content of the spectrum by reducing the number of spectral bins and losing the shape of the chemical shift peak. Still, the CoSCSA models retained enough information by increasing the number of chemical shifts in many spectral bins to produce reliable models of binding to the corticosteroid binding globulin.

DISCUSSION

We were able to select combined structure and chemical shift information from the 3D-connectivity matrix to produce very accurate models of steroids binding to the corticosteroid binding globulin. The 3D-connectivity matrix uniquely combines quantum mechanical information from the chemical shifts with nearest neighbor and internal distance connectivity information. The combined information from COSY and long-range distance connectivity information from the 3D-connectivity matrix was able to produce CoSCSA models that are much more accurate and reliable than QSAR or E-state models based on separate calculations for electrostatics and steric interactions. The cross-validated variance of CoSCSA models based on simulated ^{13}C NMR data should improve as the errors introduced by the simulation of the ^{13}C NMR data are further reduced by improved spectral simulation programs.

The 2D ^{13}C - ^{13}C COSY nearest neighbor connectivity spectral data should be important for almost any molecular property or binding affinity. In our CoSCSA models the 2D ^{13}C - ^{13}C COSY had a higher r^2 and q^2 than the ^{13}C - ^{13}C

distance connectivity data. The ^{13}C - ^{13}C distance connectivity data will be important when one or more a distance related structural features are required for molecular binding to a receptor. This is the case for steroids binding to the corticosterone binding globulin because regions around position 3 and 17 of the steroid are important for binding.

The CoSCSA models that combined the ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance connectivity PCs together produced the models with the highest r^2 and q^2 . The combined ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance connectivity PCs models were better than the 3D CoSCSA models because there were twice as many PCs available to build a regression model. In other words, the 3D models have essentially have the half number of PCs as in COSY + distance models but the number of bins contributing to each PC are about double.

In CoSCSA modeling the number and size of bins is very important. Too large a bin size inappropriately lump distinct spectral information into the same category and too small a bin size suffers from false distinctions based on reduced average bin occupancy values that adversely affect the statistics needed to identify and confirm the pattern. If one uses a huge number of bins, the results will be a model with excellent r^2 and pitiful q^2 (35), experimentally without an exhaustive search we have found that 2 ppm and 3 ppm bins seem to work best.

Our investigations showed that the 2 ppm resolution bin seem to work better for the COSY data and the 3 ppm resolution bin seem to work slightly better for the distance connectivity data. This makes sense because there were more COSY nearest neighbor connections than distance connections per molecule and therefore a lower bin size could be used for the COSY data and still produce reliable statistical models. The investigation into moving all the bins with one "hit" in them to the nearest bin with a "hit" had an r^2 and a q^2 that were only slightly improved over the original 2 ppm resolution CoSCSA models. Changing the cutoffs distance from any distance over 6.9 Å to only the same 13 distance connections from the smallest molecules produced only very small changes in r^2 and q^2 that very dependent on the model.

The EVA Infrared spectral (36) descriptors saw a similar fall off for q^2 when the size of the bin became too large. IR descriptors are similar to NMR descriptors in that they are the eigenvalues energies to the quantum mechanics Schrödinger's equation, but NMR descriptors are easily identified with one atomic nucleus whereas IR descriptors are not identified to one atom.

Future improvements to the CoSCSA modeling method include the use of other spectra besides ^{13}C NMR. The most promising NMR spectra is ^{15}N because it is in many organic molecules. Other NMR spectra that could be used are ^1H , ^{17}O , ^{19}F , ^{31}P and others depending on the endpoint and training set. Another major improvement will be the use of multiple structures so flexible compounds can be modeled. A 4D-connectivity matrix can be made as a sum of 100 3D-connectivity matrices. In the 4D-connectivity matrix the chemical shifts of atom i and atom j will not change but the distance between atom i and atom j will fluctuate. A score of 100 in a 4D-connectivity matrix will represent unvarying distances between two atoms as seen in bonds and very rigid molecules. For flexible molecules there will a distribution of distance hits along the z -axis varying from 1 to some maximum. The distributions will be gaussian, skewed-gaussian shaped when there is one maximum distance. When there is more than one maximum other distribution shapes will be seen.

CONCLUSION

In this paper, we have not attempted to optimize the size of the two-dimensional bins or the distance cutoffs used in the ¹³C-¹³C 2D distance spectra. In any event, we have developed very accurate models of steroid compounds binding to corticosteroid binding globulin without having to use an optimization procedure for the bin size and distance connectivity cutoffs. A 3D-connectivity matrix can be reduced into any number of planes. Each distance plane can have any set of distance cutoffs that can be optimized for each individual end point. In this paper, we used only 2 planes (COSY and long range). We did not use short or medium distance ranges because the additional information was not needed to produce reliable CoSCSA models. CoSCSA modeling is a high powered modeling technique because it uniquely combines the quantum mechanics information from the chemical shifts with internal molecular distances. Since we only used 5% of the available 2D chemical shift "space", we believe that we can use this procedure to effectively build reliable models of very large set of non-congeners for a specific endpoint. Optimizing the bin size, the distance cutoffs, multiple structures, multiple spectra and the number of distance connectivity spectra used in a CoSCSA model may be needed in looking at other biological, physical, chemical, ADME, and toxicological endpoints.

#	Structure ^a	Activity	R ₁	R ₂	R ₃	R ₄	R ₅	R ₆	R ₇	R ₈	R ₉	R ₁₀
1	SB	5.00	OH	H	H	H	OH	H				
2	SE	5.00	OH	OH	H							
3	SC	5.76	=O	H	=O				H	H	H	H
4	SB	5.61	H	OH	H	H	=O					
5	SC	7.88	=O	OH	COCH ₂ OH	H			H	H	H	H
6	SC	7.88	=O	OH	COCH ₂ OH	OH			H	H	H	H
7	SC	6.89	=O	=O	COCH ₂ OH	OH				H	H	H
8	SE	5.00	OH	=O								
9	SC	7.65	=O	H	COCH ₂ OH	H			H	H	H	H

10	SC	7.88	=O	H	COCH ₂ OH	OH			H	H	H	H
11	SB	5.92	=O		H	H	OH	H				
12	SD	5.00	OH	OH	H	H						
13	SD	5.00	OH	OH	H	OH						
14	SD	5.00	OH	=O		H						
15	SB	5.23	H	OH	H	H	=O					
16	SE	5.23	OH	COMe	H							
17	SE	5.00	OH	COMe	OH							
18	SC	7.38	=O	H	COMe	H			H	H	H	H
19	SC	7.74	=O	H	COMe	OH			H	H	H	H
20	SC	6.72	=O	H	OH	H			H	H	H	H
21	SF	7.51	=O	OH	COCH ₂ OH	OH						
22	SC	7.55	=O	OH	COCH ₂ OCOMe				H	H	H	H
23	SC	6.78	=O	=O	COMe	H				H	H	H
24	SC	7.20	=O	H	COCH ₂ OH	H			OH	H	H	H
25 ^b	SC	6.14	=O	H	OH	H			H	H	H	H
26	SC	6.25	=O	H	COMe	OH			H	OH	H	H
27	SC	7.12	=O	H	COMe	H			H	Me	H	H
28 ^b	SC	6.82	=O	H	COMe	H			H	H	H	H
29	SC	7.69	=O	OH	COCH ₂ OH	OH			H	H	Me	H
30	SC	5.80	=O	OH	COCH ₂ OH	OH			H	H	Me	F

Table 1. Structures of corticosteroids used in QSDAR models of corticosteroid binding globulin data. ^a Structures according to references (1,22,23), ^b H (hydrogen) instead of Me at C₁₀ steroid skeleton

model	n	r ²	q ²	components
QSAR (2)	31	.72	.68 ^a	3 (PCs)
HE state/ E-state (3)	31	.98 ^a /.96 ^b	.80 ^a /.76 ^b	3 ^a (PCs)/5 ^b (PCs)
E-state (3)	31	.96 ^a /.96 ^b	.79 ^a /.67 ^b	3 ^a (PCs)/4 ^b (PCs)
SOM (4)	31	.85	-	3 (PCs)
QSAR/E-state (5)	30	.82	.78	3 (atoms)
2D CoSASA (6)	30	.80	.73	3 (atoms)
1D CoSA (6)	30	.80	.78	3 (bins)
CoSCSA (COSY)	30	.84/.93	.74/.88	3 (PCs)/8 (PCs)
CoSCSA (distance)	30	.66/.89	.38/.79	3 (PCs)/8 (PCs)
CoSCSA (COSY + distance)	30	.84/.96	.74/.92	3 (PCs)/8 (PCs)
CoSCSA (3D ^c)	30	.78/.92	.68/.81	3 (PCs)/8 (PCs)
CoSCSA-PD-ANN (3D ^c)	30	.96	.78 ^d	593:198:1
CoSCSA-PD-ANN (3D ^c)	30	.96	.73 ^e	593:198:1

Table 2: Model performance parameters n, r², q², and number of components.

^a 1.0 Å models, ^b 2.0 Å models, ^c 3D is the combination of COSY and distance data before PC extraction, ^d is a leave-three-out q₃² and ^e is a leave-ten-out q₁₀².

REFERENCES

1. Mickelson, K. E.; Forsthoefel, J.; and 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.; and Richards, W. G. Structure-activity relationships from molecular similarity matrices. *J. Med. Chem.*, **1993**, 36, 433-438.
3. Kellogg, G. E.; Kier, L. B.; Gaillard, P.; and Hall, L. H. E-state fields: Applications to 3D QSAR. *J. Comput. Aid. Mol. Des.*, **1996**, 10, 513-520.
4. Polanski, J. The receptor-like neural network for modeling corticosteroid and testosterone binding globulins. *J. Chem. Inf. Comput. Sci.*, **1997**, 37, 553-561.
5. De Gregorio, C.; Kier, L. B.; and Hall, L. H. QSAR modeling with electrotopological state indices: Corticosteroids. *J. Comput. Aid. Mol. Des.*, **1988**, 2, 557-561.
6. Beger, R.; and Wilkes, J. Developing ^{13}C NMR quantitative spectrometric data-activity relationship (QSDAR) models of steroid binding to the Corticosteroid binding globulin. *J. Comput. Aid. Mol. Des.*, **2001**, 15: 659-669.
7. Beger, R. D. and 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.
8. Beger, R. D.; and Wilkes, J. G. Models of polychlorinated dibenzodioxins, dibenzofurans, and biphenyls binding affinity to the aryl hydrocarbon receptor developed using ^{13}C NMR data. *J. Chem. Inf. Comput. Sci.*, **2001**, 41:1322-1329.
9. Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; and Lobanov, V. S. Prediction of gas chromatographic retention times and response factors using a general quantitative structure-property relationship. *Anal. Chem.*, **1994**, 66, 1799-1807.
10. Katritzky, A. R.; Mu, L.; Labanov, V. S.; and Karelson, M. Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics. *J. Phys. Chem.*, **1996**, 100, 10400-10407.
11. Fujita, T.; Iwasa, J.; and Hansch, C. A new substituent constant, π , derived from partition coefficient. *J. Am. Chem. Soc.*, **1964**, 86, 5175-5180.

12. Branbury, S. P. Quantitative structure-activity relationship and ecological risk assessment: An overview of predictive aquatic toxicology research. *Toxicology*, **1995**, 25, 67-89.
13. Cramer, R. D.; Paterson, D. E.; and 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.
14. Beger, R. D.; and Bolton, P. H. Protein ϕ and ψ dihedral restraints determined from multidimensional hypersurface correlations of backbone chemical shifts and their use in the determination of protein tertiary structures. *J. Biomol. NMR*, **1997**, 10, 129-142.
15. Wishart, D. S.; and Sykes B. D. Chemical shifts as a tool for structure determination. *Methods in Enzymology*, **1994**, 239, 363-392.
16. Kvasnicka, V. An application of neural networks in chemistry. Prediction of ¹³C NMR chemical shifts. *J. Math. Chem.*, **1991**, 6, 63-76.
17. Emsley, J. W.; Feeney, J.; Sutcliffe, L. H. *High Resolution Nuclear Magnetic Resonance*. Volume I. Pergamon Press Ltd.: Oxford **1965**; Chapter 8, p 287.
18. Hansch, C.; and Leo, A. Exploring QSAR – Fundamentals and applications in chemistry and biology. Washington, D. C. The American Chemical Society, **1995**.
19. 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.
20. Klopman, G. MULTICASE1. A hierarchial computer automated structure evaluation program. *Quant. Struct. Act. Rel.*, **1992**, 11, 176-184.
21. Aue, W. P.; Bartholdi, E.; and Ernst, R. R. Two-dimensional spectroscopy. Application to nuclear magnetic resonance. *J. Chem. Phys.*, **1976**, 64, 2229-2246.
22. Bodenhausen, G; and Ruben, D. J. Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chem. Phys. Lett.*, **1980**, 69, 185-189.
23. Bax, A.; Griffey, R. H.; and Hawkins, B. L. Sensitivity-enhanced correlation of ¹⁵N and ¹H chemical shifts in natural-abundance samples via multiple quantum coherence. *J. Am. Chem. Soc.*, **1983**, 105, 7188-7190.
24. Bax, A.; and Summers, M. F. ¹H and ¹³C Assignments from sensitivity-enhanced detection of heteronuclear multiple-bond connectivity by 2D multiple quantum NMR. *J. Am. Chem. Soc.*, **1986**, 108, 2093-2094.
25. Randic', M. On the characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, 97, 6609-6615.
26. Burden, F. R. A chemically intuitive molecular index based on eigenvalues of a modified adjacency matrix. *Quant. Struct.-Act. Relat.*, **1997**, 16, 309-314.

27. Kumar, A.; Ernst, R. R.; and Wuthrich, K. A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. *Biochem. Biophys. Res. Comms.*, **1980**, 95, 1-6.
28. Balaban, A. T. Highly discriminating distance based topological index. *Chem. Phys. Lett.*, **1982**, 89, 399-404.
29. CNMR predictor, ACD/Labs Toronto, Canada, 2000.
30. Bremser, W. HOSE - a Novel substructure Code. *Anal. Chim. Acta.*, **1978**, 103, 355-365.
31. Statistica, StatSoft, Tulsa, OK, 2001.
32. Cramer, R. D.; Bunce, J. D.; and Patterson, D. E. Cross-validation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct.-Act. Relat.*, **1988**, 7, 18-25.
33. Jgravity, version 1.0 beta-1 build-10, Titan Systems Corporation/LinCom Division, Houston, Texas, 2001.
34. Rumelhart, D.E. and McClelland, T.L., *Parallel Distributed Processing*, Brandford Books/MIT Press: Cambridge, MA (1986).
35. Bursi, R.; Dao, T.; van Wilk, T.; de Gooyer, M.; Kellenbach, E.; and Verwer, P. Comparative spectra analysis (CoSA): Spectra as three-dimensional molecular descriptors for the prediction of biological activities. *J. Chem. Inf. Comput. Sci.*, **1999**, 39, 861-867.
36. Turner, D. B.; Willett, P.; Ferguson, A. M.; and Heritage, T. Evaluation of novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application. *J. Comput. Aid. Mol. Des.*, **1997**, 11, 409-422.
- 37.

FIGURES

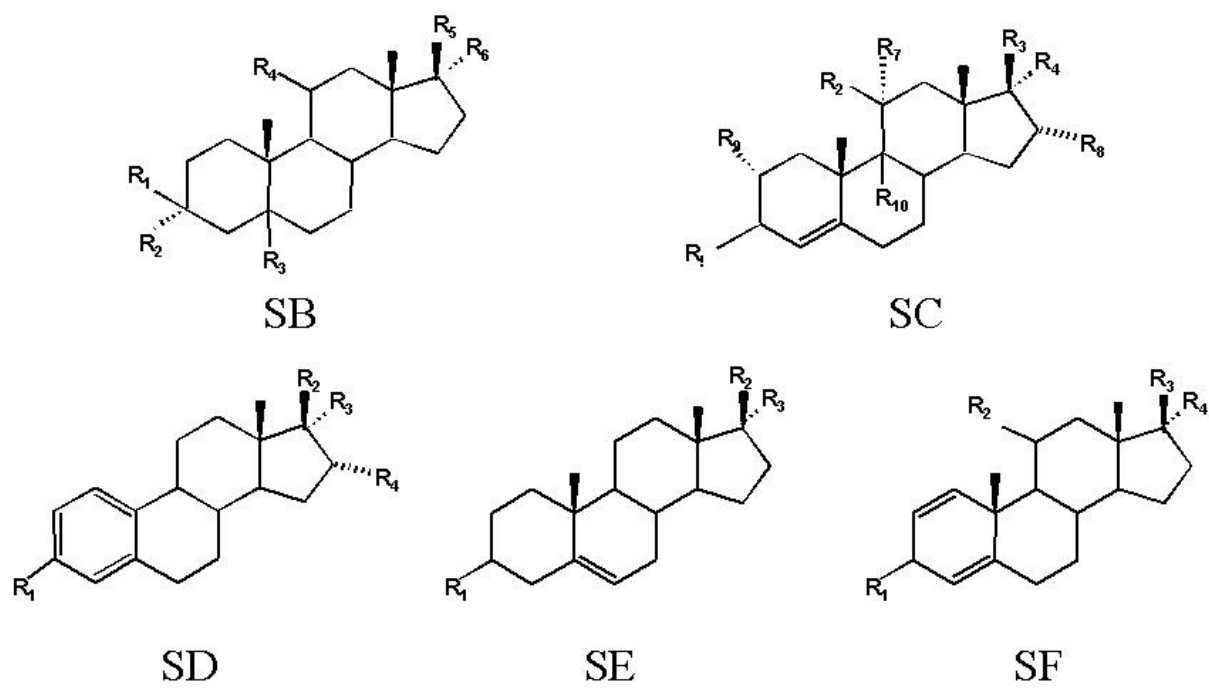


Figure 1. Structures SB -SF used with Table 1 for the corticosteroid binding globulin steroid series.

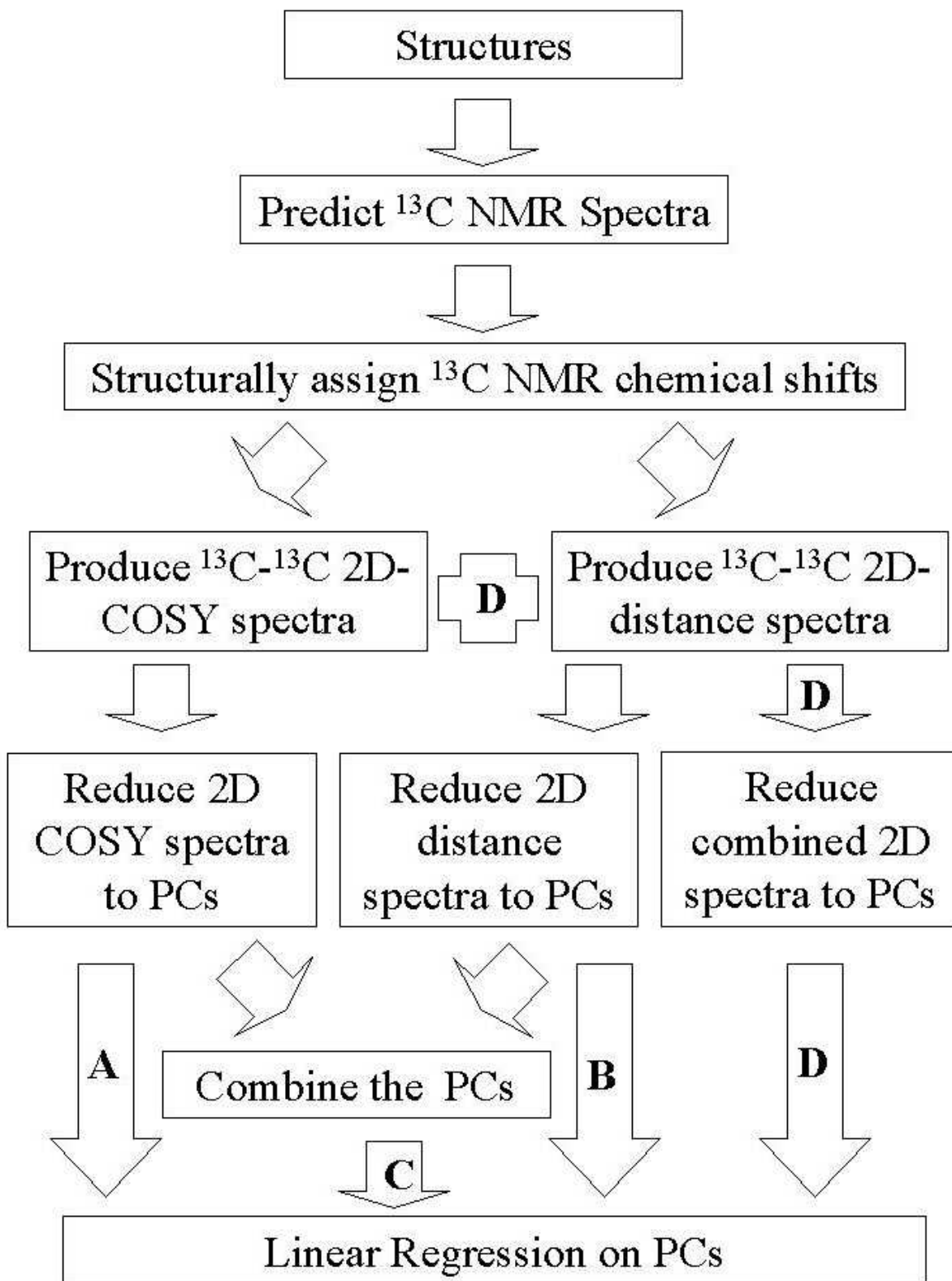


Figure 2. The procedural flow chart for CoSCSA modeling. (A) The ^{13}C - ^{13}C COSY data. (B) The ^{13}C - ^{13}C distance connectivity data greater than 6.9 \AA . (C) The combined ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance connectivity PCs. (D) The combined ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance connectivity data before PCs are extracted from the combined data.

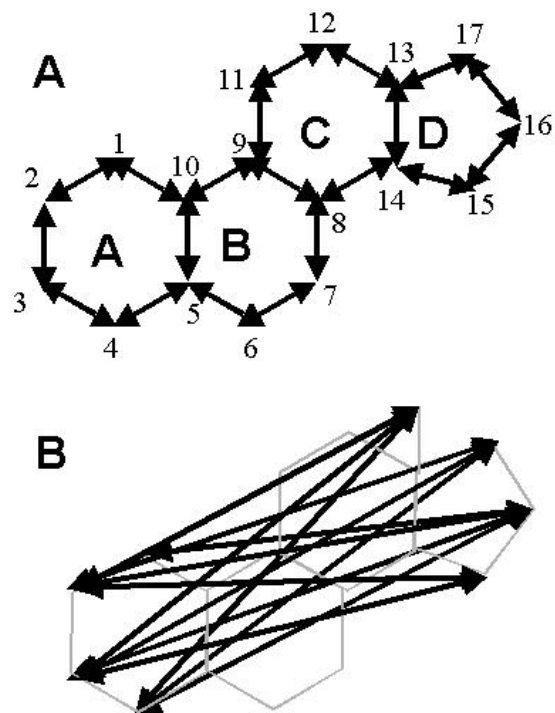


Figure 3. (A) The arrows represent the carbon to carbon through nearest neighbor bond connections for the backbone of a steroid used in the predicted 2D ^{13}C - ^{13}C COSY spectra. (B) The arrows represent the carbon to carbon through spatial distance connections used in the theoretical 2D ^{13}C - ^{13}C distance spectra.

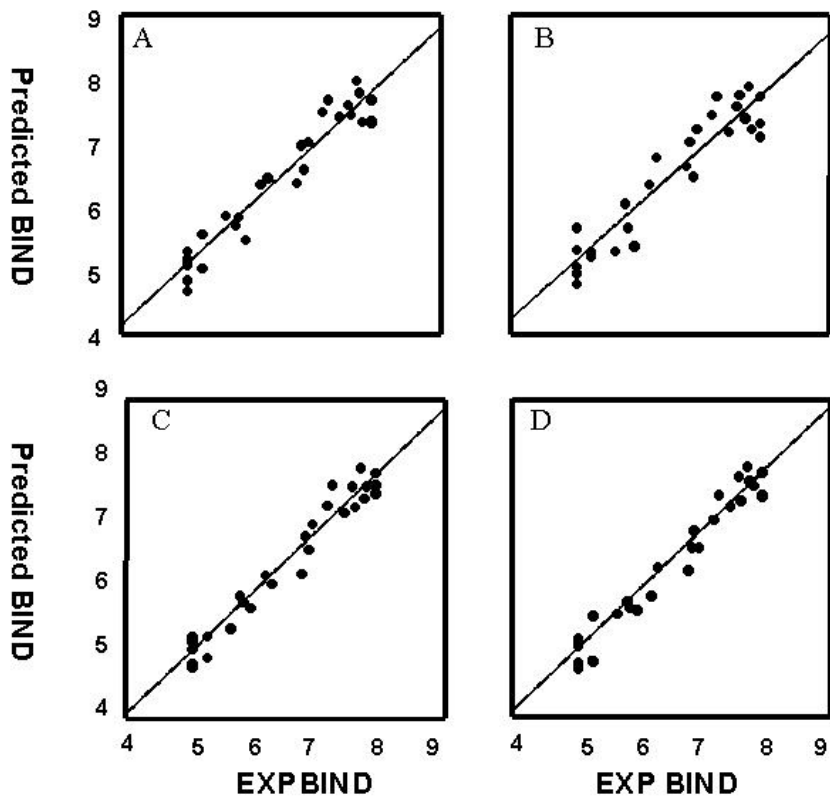


Figure 4. Plot of the predicted binding versus experimental binding. (A) The ^{13}C - ^{13}C COSY data. (B) All the ^{13}C - ^{13}C distance connectivity data greater than 6.9 Å. (C) The combined ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance connectivity PCs. (D) The combined ^{13}C - ^{13}C COSY and ^{13}C - ^{13}C distance connectivity data before PCs are extracted from the combined data.