Computational modeling of biologically active molecules using NMR spectra

Richard D. Beger

Division of Systems Toxicology, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR 72079, USA

The molecular structure and NMR chemical shift information of a compound can be combined to form powerful models of biological activity. NMR spectral data and structure information can be combined on a structural template analogous to 3D-QSAR methodology or orientation independently in spectral space. Surprisingly, quantitative spectrometric data–activity relationship (QSDAR) models built on structure templates are inferior to multi-dimensional QSDAR models built in spectral space. 3D-QSDAR modeling could be useful for estimating chemical toxicity, risk assessment of environmental contaminants and drug lead-compound identifications.

Quantitative spectrometric data–activity relationship (QSDARs) models are based on determining a relationship between the NMR spectra of a set of molecules and their biological activities or physical characteristics. Several papers have recently been published that demonstrate the power of using ^{13}C NMR spectra as input parameters to form models of biological activity [1–10]. QSAR [11–22] and SAR [23–27] are the primary *in silico* modeling techniques used in drug discovery. 3D-QSAR models are typically based on physical fields obtained by superimposing each compound as a whole on a 3D grid. The QSAR and SAR models depend on pattern-recognition techniques for developing a relationship between input physical or structural variables of a molecule and a biological or physical characteristic endpoint. 3D-QSAR models are based partly on electrostatics and partly on molecular geometry [11,12].

NMR spectra reflect quantum mechanical properties that, like QSAR factors, depend on local electrostatics and geometry [28–33]. The ^{13}C NMR spectrum of a compound contains a pattern of frequencies that correspond directly to the quantum mechanical properties of the carbon nuclear magnetic dipole in a magnetic field. The spectral pattern reflects the local electrostatic environment and electron orbital configuration of each carbon atom. The NMR chemical shift tensor is composed of a diamagnetic term and a paramagnetic term [33]. The diamagnetic term is directly related to the electrostatic potential at the nucleus [33], whereas the paramagnetic term is largely dependent on the

orbital configuration [31]. For ^{13}C NMR spectra, the differences between the diamagnetic and paramagnetic terms can be very large, which results in a large spectral range. This means that the resonances from different carbon orbital configurations are generally well-separated from each other, which permits the use of advantageous large-^{13}C NMR spectral diversity to build the QSDAR models.

It is possible to build SDAR and QSDAR models without associating the pattern of NMR spectral features to the molecular nuclei. However, QSDAR modeling based on 1D ^{13}C NMR data should be used on compounds that are primarily carbon-based and for structurally similar compounds. QSDAR modeling using ^{13}C NMR spectra works best when attempted on a set of similar structural motifs or on a set of compounds that have a large proportion of carbon nuclei. Figure 1 shows the QSDAR modeling flow chart, including binning, weighting, and validation steps. To start, the structures of a set of compounds are generated and their ^{13}C NMR spectra are predicted or found in a database. A ^{13}C spectrum is saved as a set of ordered pairs – chemical shift frequencies in parts per million (ppm) and the area under the peak. The area under a specific chemical shift frequency is normalized to an integer value. This normalization is done so that all the spectra have a similar signal-to-noise ratio, and the line width variations (caused by differences in NMR instrumental field strengths, shimming, proton coupling, temperature, pH or solvent) are eliminated. The bins define the number of chemical shift peaks within a ppm range. The bin width used in inputting the ^{13}C NMR spectra

Corresponding author: Beger, R.D. (richard.beger@fda.hhs.gov)
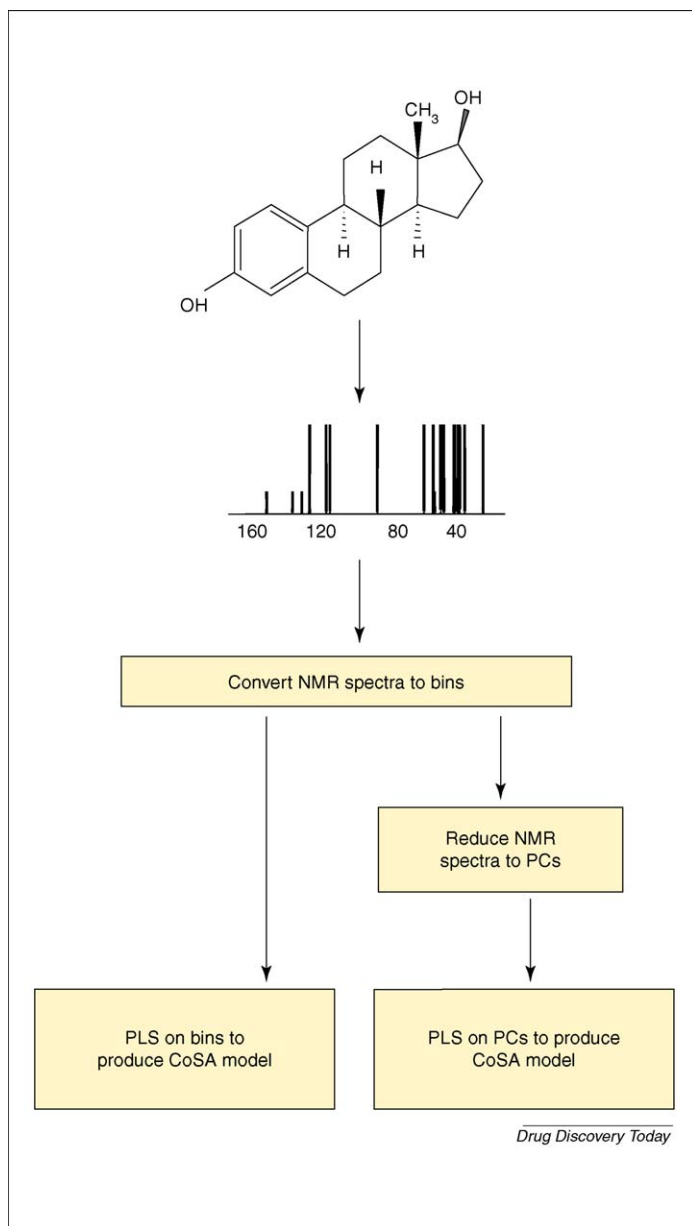
FIGURE 1

The procedural flow chart for QSDAR modeling. First, the structures of a set of compounds are generated and their real or predicted 1D ^{13}C NMR spectra are obtained. The spectra are broken into bins that define the number of chemical shifts peaks with a specific ppm range. The bin width can be allowed to vary and is optimized by testing the model. The bins intensities or principal components (PCs) are used to produce the QSDAR models.

can be optimized by allowing the spectral window width to vary and using the bin width that produces the best cross-validated models.

One limitation of 1D comparative spectra analysis (CoSA) QSDAR models is that they lack direct 3D structural information. The most obvious way to combine structural and spectral information is to establish a specific molecular template as a best, or normal, representative of those causing a particular biological effect. Each carbon atom in this template's backbone is numbered, and all other compounds to be modeled must use the same backbone numbering system. Then, each compound's pattern is

defined by the ordered pair (carbon number, chemical shift) rather than by the previously described system (chemical shift bin number, occupancy number). This means that the chemical shifts have been assigned to the associated carbon atom that produced them. The pattern as defined is correlated with the biological activity of each molecule. The resulting model combines structural information with the assigned simulated ^{13}C NMR chemical shifts. We have named this 1D QSDAR method comparative structurally assigned spectra analysis (CoSASA) to distinguish it from the substantially different structurally unassigned methods previously described as CoSA. The ability to include spatial relationships in SDAR modeling should improve the quality of the results. In fact, when used on the same spectral training set, inferior results by CoSASA were obtained when compared to CoSA [34,35]. In other words, combined structural and spectral information used in a way that is directly encoded in 3D-QSAR was not helpful.

Another way to combine structural and spectral information is to express a molecule's geometrical information in spectral space in a multidimensional matrix that can be built on orientation-independent parameters. In the same way, 2D, 3D and 4D NMR experiments use additional spectral dimensions to reduce spectral overlap [36–42]. The matrix definition does not limit compounds only to those sharing the same backbone template. Compounds in the same model can have dissimilar structures: they can differ in the number and connectivity of carbon atoms, as well as the number and identity of constituents. Following is a description of how a multidimensional QSDAR model is built [43–45].

A molecule's 3D-carbon-connectivity matrix can be built by displaying all the possible carbon-to-carbon connectivities with their assigned carbon NMR chemical shifts on two dimensions and the distance between the two atoms in a third, orthogonal dimension. Figure 2 shows all the carbon-to-carbon connectivities and the 3D-carbon-connectivity matrix for 3β-estradiol. The x-axis shows the chemical shifts of carbon i, the y-axis those of carbon j. The z-axis represents the distance between carbon i and carbon j (r_{ij}). By representing molecules with this matrix, the subjective superposition of molecules on a template is avoided, but we are still restricted to molecules that consist primarily of carbon atoms. Further parameterization needed to develop typical 3D-QSAR models is either avoided or minimized. For flexible molecules, some atom-to-atom distances can vary, so representations of multiple conformations in the 3D-connectivity matrix format can be accommodated in a 4D-connectivity matrix that incorporates the molecule's dynamics in its construction.

The carbon-connectivity matrix shown in Figure 2a is symmetrical around the x-y diagonal. That is, for every connection between atom i and atom j, the identical relationship is represented across the diagonal at the connection between atom j and atom i. Along the x-y diagonal at $r_{ij} = 0$ is the 1D ^{13}C NMR spectrum of 3β-estradiol. The nearest neighbor carbon-to-carbon connections are at $r_{ij} \approx 1.4$ Å with all the other distance related atom-to-atom connections at $r_{ij} > 1.4$ Å.

The information in a 3D-connectivity matrix over-determines the structure of a compound, therefore the information in the matrix that is actually used in a model can be reduced significantly to simplify and accelerate computations. One possible way of reducing the 3D matrix is to cut it into a set of 2D spectral connectivity planes. The first 2D connectivity plane is the plane with
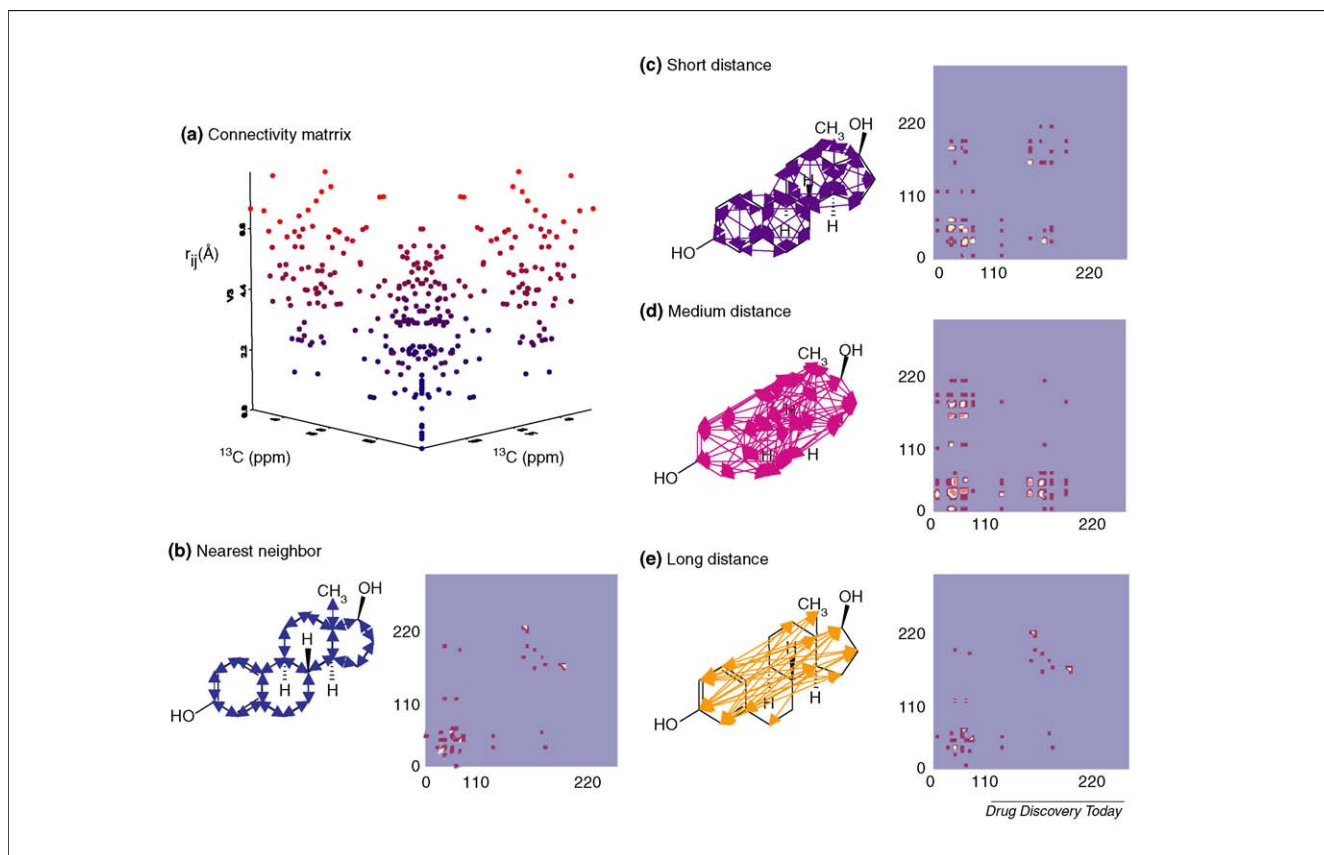
FIGURE 2

Connectivity matrix and connectivity planes. (a) The 3D-connectivity matrix of 3β-estradiol. **(b)** The carbon-to-carbon nearest neighbor structural connectivities and associated projected 2D connectivity plane of the 3D-connectivity matrix for 3β-estradiol. **(c)** The carbon-to-carbon short distance structural connectivities and its associated projected 2D connectivity plane. **(d)** The carbon-to-carbon medium distance structural connectivities and associated projected 2D connectivity plane. **(e)** The carbon-to-carbon long distance structural connectivities and its associated 2D connectivity plane. The x- and y-Axis labels for b–e are ^{13}C ppm.

nearest neighbor through bond-connectivity information. The other 2D planes are constructed by reducing a range of similar distances along the z-axis onto one 2D spectral connectivity plane. This effectively compresses and greatly simplifies distance information along the z-axis. In addition to the bond-connectivity plane, there is a plane for short atom-to-atom distance connections ($2.0 \text{ Å} < r_{ij} < 3.6 \text{ Å}$) of perhaps two bond lengths through space, another for medium atom-to-atom distance connections ($3.6 \text{ Å} < r_{ij} < 6.0 \text{ Å}$), and a fourth for long atom-to-atom distance connections ($r_{ij} > 6.0 \text{ Å}$). The 2D planes can be binned into 2D bins but only half of the 2D bins need to be used in model development. The number of 2D connectivity planes and distance ranges for the 2D connectivity plane can be set by the model developer or by cross validation. Figure 2b–e shows carbon-to-carbon connectivities on the structure of estradiol next to the *in silico* predicted 2D carbon-to-carbon connectivity planes for the nearest neighbor, short-range, medium-range and long-range distances. 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 and used to determine whether the compound is predicted to exhibit the biological activity. This pattern recognition method is called comparative structural connectivity spectra analysis (CoSCoSA), to distinguish it from CoSA and CoSASA.

QSDAR development

Models based on 13C NMR spectra

A ^{13}C NMR CoSASA model of 30 corticosterone steroids with binding affinity gave an explained variance (r^2) of 0.80 and a LOO ('leave-one-out') cross-validated variance (q^2) of 0.73 [34]. The steroid backbone CoSASA model was based on the change in chemical shift frequency of atoms from positions 3, 14 and 20 on the steroid template. Positions 3 and 20 are near the steroid positions, which previous QSAR models have shown are the active regions for corticosterone binding [34,35]. The q^2 of 0.73 for the CoSASA model is slightly better than the QSAR q^2 of 0.68 but the CoSASA is much easier to build and gives direct information as to which atoms are important for the model and biological activity. Without additional effort, more atoms could have been included, and trends in q^2 as a function of number of bins used indicate that further improvements in predictive accuracy of the CoSASA model are possible. A CoSCoSA model of the same 30 steroids binding corticosteroid binding globulin was developed by combining NMR spectral information with structural information to form a 3D-carbon-connectivity matrix. The 3D-carbon-connectivity matrix was built by displaying all possible carbon-to-carbon connections with their assigned carbon NMR chemical shifts and distances between the carbons. The matrix was simplified by compressing data into a 2D [13]C–[13]C correlation spectroscopy (COSY) plane, and

selected theoretical 2D ^{13}C–^{13}C distance connectivity spectral slices to model binding for 30 steroids. Not all the atom-to-atom connectivity information is needed to develop a good CoSCoSA model. It is known for many steroids and corticosteroids that the important binding sites are near the 3 and 17 positions of the molecule. Therefore, effective models can be built using only the nearest-neighbor information and the long-range connections between carbons surrounding positions 3 and 17 that are ∼7.5 Å apart. To include information around positions 3 and 17, through-space carbon-to-carbon connections that were greater than 6.0 Å were included to produce a theoretical 2D ^{13}C–^{13}C distance-connectivity spectrum that contains cross-peaks for atom i and atom j whenever the two carbons were greater than 6.0 Å apart. Principle components (PCs) for the CoSCoSA model were produced by evaluating the 2D ^{13}C–^{13}C COSY and 2D ^{13}C–^{13}C distance-connectivity bins with forward multiple linear regression analysis, and using only the most statistically relevant PCs. The F test for many of the models continued to rise as more PCs were included in modeling. These processes can lead to over-fitting of the model. To avoid over-fitting, the models need to be tested with external test sets or with cross-validation techniques that remove more than 10% of the data.

Modeling using other active nuclei

Often, particularly for chemicals with potentially useful pharmaceutical value or toxicity, compound types to be modeled will contain atoms other than carbon and hydrogen. It that case, NMR spectra from other active nuclei can be used. The most prominent atom that is both biologically important and for which accurate NMR prediction software is available is ^{15}N. Other NMR nuclei that could be used for SDAR or QSDAR modeling are ^{17}O, ^{19}F, and ^{31}P. The optimal nuclei would depend on the molecule training set and biological or chemical end point to be modeled. Figure 3 shows a typical 2D spectral connectivity layout for a 2D ^{13}C–^{15}N heteronuclear CoSCoSA model. In Figure 3, the yellow area is for carbon-to-carbon bonds, the red area is for carbon-to-nitrogen bonds, and the green area is for nitrogen-to-nitrogen bonds. As seen with multi-dimensional ^{13}C–^{13}C models, symmetry-based data duplicates mean that only half of the spectra in the array are necessary to develop a model.

Case study on cephalosporins

Cephalosporins are widely used antibiotics that are similar in structure and mode of action to penicillin. There are two nitrogen atoms in the backbone of cephalosporin molecules. To examine the potential of heteronuclear CoSCoSA methods to model the minimum inhibitory concentrations (MIC) of cephalosporin antibiotics, we developed 2D QSDAR models for 17 compounds using only the through-bond (COSY-type) carbon-to-carbon and through-bond carbon-to-nitrogen connectivities. In producing this 2D CoSCoSA model, we defined the endpoint as log (1 ÷ MIC) MIC) for technical reasons [46–48].

We used bin sizes of 3.0 × 3.0 ppm for carbon-to-carbon COSY connections and 10.0 × 3.0 ppm for the nitrogen-to-carbon direct connections. Nitrogen chemical shifts were predicted from software available on the I-Lab website (ACD Laboratories, www.acdlabs.com/ilab) and carbon shifts were predicted as before. In building the nitrogen-to-carbon connectivity matrix,
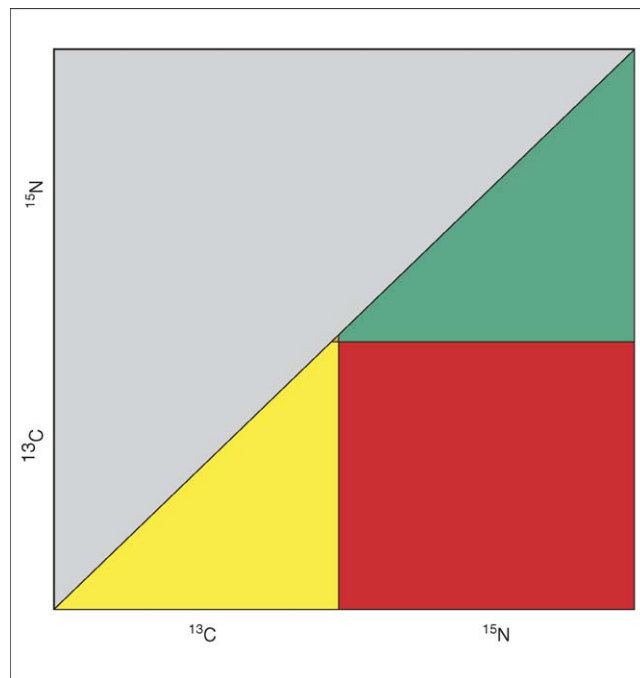

FIGURE 3

Representative set up for a 2D ^{13}C-^{15}N heteronuclear neighbor-to-neighbor bond connectivity matrix. The yellow area is for carbon-to-carbon bonds, the red area is for carbon-to-nitrogen bonds, and the green area is for nitrogen-to-nitrogen bonds. Because of the symmetry of the matrix, only half of the connectivity matrix is used in modeling.

700.0 ppm was added to the predicted nitrogen chemical shifts, so that shifts fell in the range of 300–700 ppm. Thus, a single synthetic spectrum could be defined with carbon-to-carbon connectivity bins occupied from 0 to 240 ppm, and nitrogen-to-nitrogen connectivity from 300 to 700 ppm.

Forward linear regression selected four bins from a total of 101 carbon-to-carbon and 48 carbon-to-nitrogen occupied bins. The selected bins were the following: (−230 nitrogen, 156 carbon); (−280 nitrogen, 162 carbon); (−230 nitrogen, 168 carbon); and (135 carbon, 24 carbon). The cephalosporin CoSCoSA model had a correlation $r^2 = 0.92$, F value = 36.2, $P < 0.000005$, LOO of $q_1^2 = 0.88$, average L4O of $q_2^4 = 0.79$ and SD = 0.03 [48]. These results are indicative of an astonishingly robust model. Figure 4 shows the structurally assigned interpretation of the chemical shifts used to formulate the CoSCoSA model of cephalosporins. The chemical shifts at (−230 nitrogen,168 carbon) identifies the carbon-to-nitrogen bond where acid hydrolysis occurs and allows the cephalosporin to bind to the bacterial enzyme transpeptidase, irreversibly inactivating the enzyme and stopping growth [49]. The chemical shifts connectivity at (−230 nitrogen, 156 carbon) and (−280 nitrogen, 162 carbon) identify two spectral states of another carbon-to-nitrogen bond in the middle of each cephalosporin compound. Because the enzyme transpeptidase binds to dipeptides, it is reasonable to assume that the other carbon-to-nitrogen bond, which looks like part of an amino acid backbone (carbonyl-to-amide-to-α carbon), is involved via a hydrogen bond to the enzyme. The two different chemical shifts for the same carbon-to-nitrogen bond might represent the bond in two different configurations, two different electrostatic potentials or, more likely, a combination of different configurations and associated
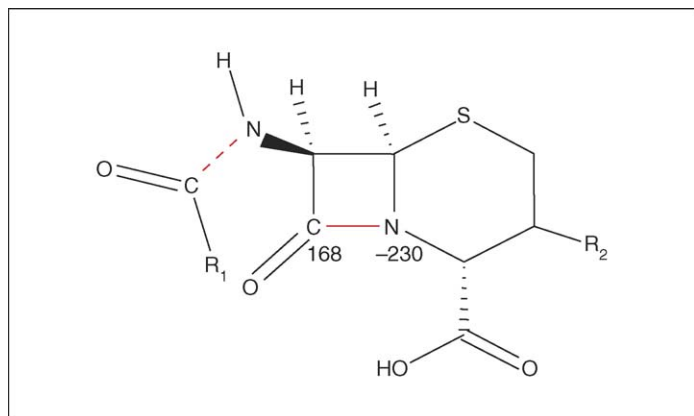
FIGURE 4

Structural explanation of bins selected in the 2D ^{13}C–^{15}N heteronuclear CoSCoSA model of cephalosporin antibacterial activity. Red bond between carbon (168) and nitrogen (−230) corresponds to bond that undergoes acid hydrolysis and then binds to enzyme transpeptidase and irreversibly inactivates it. Red dotted bond between carbon and nitrogen corresponds to bins reflecting carbon (156) nitrogen (−270) and carbon (162) nitrogen (−280) two spectral states and might explain the backbone torsional angle dependence of dipeptides binding affinity to transpeptidase.

electrostatic potentials. It has been shown that peptides interact with the penicillin-binding transpeptidase proteins with binding energy dependent on the backbone torsion angles [50]. This is worth mentioning because most descriptions of cephalosporin activity mechanisms do not discuss the crucial role of the second carbon-to-nitrogen bond. The CoSCoSA model including heteronuclear NMR peaks predicted that both carbon-to-nitrogen bonds contribute significantly to the antibiotic strength of cephalosporin. In this case, ^{15}N and ^{13}C chemical shifts were very important in producing a reliable model of biological activity.

4D QSDAR development

A 4D-connectivity matrix can be defined as the sum of an arbitrary number, say 100, of 3D-connectivity matrices of a molecule's dynamical trajectory. In the simplified version of this 4D-connectivity matrix concept, chemical shifts of atom i and atom j are assumed not to change as the molecule bends or twists, but the distance between atoms i and atom j would differ as a function of the molecular conformation. For any two atoms, molecular dynamics programs can estimate interatomic distance, a value that can change over some range and for which the percentage of time that the distance is within a certain bin will vary, depending on molecular connections, degrees of freedom and so forth. This concept applied to CoSCoSA modeling would treat the distance between atoms as a potential variable rather than as a constant. A score of 100 in a 4D-connectivity matrix will represent unvarying distances between two atoms in terms of bond length and also between more distant atoms if the molecule is very rigid. For two atoms in a flexible molecule, there will a distribution of distance hits along the z-axis varying from 1 to some maximum value (most probable conformation). For all the possible atom pairs, distance distributions will be Gaussian or skewed-Gaussian functions when there is a single maximum distance. When there is more than one maximum, a more complex distribution will be seen. The 4D-connectivity matrix as a way of representing molecular conformation characteristics confers a significant advantage when building models for cases in which the training set includes very diverse compound types. A 4D-connectivity matrix allows for multiple conformations of a molecule to be displayed. The multiple conformations can be put into an 'entropy-like' equation to estimate the effect of configurational entropy [51,52].

Potential applications

Several researchers have demonstrated various QSDAR models capable of predicting the strength of chemical interactions with biological systems. This kind of prediction is useful for estimating pure chemical toxicity, or even risk assessment of environmental contaminants that are composed of complicated mixtures. It is also useful for pharmaceutical lead compound identification and for classifying new drugs with respect to biological binding and identifying the atoms or bonds in a compound that are important for biological activity. QSDAR models that used NMR, MS and IR spectra have been developed when the endpoint is essentially chemical (decomposition) but the process is mediated through the enzymatic capabilities of bacteria [53]. This suggests possible QSDAR applicability in relation to bioremediation efforts.

The type of QSDAR method useful for a particular application depends on the amount of data available for the model training set, the variety of compound types to be modeled, the elemental composition of the compounds in the training set, and the predictive accuracy required. 1D and 2D models are easier to prepare and less computationally intensive to build and validate than 3D and 4D models. As more dimensions are added to a QSDAR model, the 2D or 3D bin sizes must be enlarged to cover the increased dimensional space. This is done so that the models can be built using bins that are populated with hits from a significant portion of the molecules in the training set. This allows the QSDAR model to be developed from stronger statistical generalizations. If the bins are too small the QSDAR model will be built from bins that are populated with an insufficient number of hits, and/or from bins, or information, that is specific to a compound. Using bins with insufficient hits allows a model to memorize the training set, and it cannot generalize the spectral patterns for predictive purposes.

The CoSCoSA modeling system can be applied to receptor-binding systems whose SAR is unknown, a common situation faced by new drug discovery or lead optimization programs in the pharmaceutical industry. Producing QSAR models without detailed structural binding-site information is unreliable and is often based on intuition. By contrast, QSDAR offers a more objective way of building models. The use of spectral-data activity relationships allows an unbiased examination of specific SARs, which can suggest a more meaningful way to discover antibiotics or other pharmaceuticals. This potential was demonstrated in the 2D QSDAR model of cephalosporin activity based on simulated carbon and nitrogen NMR spectra.

Conclusions

Structure and chemical shift information from the 3D-connectivity matrix can be used to produce very accurate models of biological binding activity. The 3D-connectivity matrix uniquely combines quantum mechanical information from the chemical shifts with

internal atom-to-atom distance information for a compound. The combined distance information from nearest neighbor to longer-range distance connectivity information from the 3D-connectivity matrix was used to produce CoSCoSA models that were as accurate as, or more accurate and reliable than, QSAR or E-state [14] models based on separate calculations for electrostatics and steric interactions. The quality of results obtained from QSDAR models based on simulated NMR data should improve as the spectral simulation software improves and prediction errors are reduced. The best modeling approaches for 3D-QSDAR might require fuzzy pattern recognition techniques that have the ability to be built based on the chemical shift distances from bins or features deemed important to the model through training [44]. 3D-QSDAR modeling has widespread application in bioinformatics challenges and is not intended to replace 3D-QSAR and SAR but to provide an alternative technique for computational model developers to use when developing models of biological activity. For example, in the case of cephalosporins binding to the transpeptidase enzyme, the CoSCoSA model can give multiple conformation information about binding mechanisms that could not be obtained from 3D-QSAR or SAR models. The added advantage of this approach is that it can produce these kinds of results very rapidly.

3D-QSDAR models based on NMR spectra are, in a way, a combination of quantum mechanics and topology with conventional SAR and QSAR modeling methods. Instead of using a 3D grid to calculate electrostatic energies, the quantum mechanical energies in the form of chemical shifts are put onto the compound itself in all the possible atom-to-atom connections. The distribution of distances between two atoms is related to molecular topology, shape and configuration entropy. 4D-QSDAR modeling is an *in silico* modeling method that has the potential to evaluate enthalpy and entropy effects at the same time.

Acknowledgements

References

1 Bursi, R. *et al.* (1999) Comparative spectra analysis (CoSA): Spectra as three-dimensional molecular descriptors for the prediction of biological activities. *J. Chem. Inf. Comput. Sci.* 39, 861–867

2 Beger, R.D. and Wilkes, J.G. (2001) 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.* 41, 1322–1329

3 Shade, L. *et al.* (2003) New computerized method for modeling binding affinities to the aryl hydrocarbon receptor using ^{13}C NMR spectra. *Environ. Toxicol. Chem.* 22, 501–509

4 Beger, R.D. *et al.* (2000) Producing ^{13}C NMR, infrared absorption and EI mass spectrometric data models of the monodechlorination of chlorobenzenes, chlorophenols, and chloroanilines. *J. Chem. Inf. Comput. Sci.* 40, 1449–1455

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

6 Beger, R.D. and Wilkes, J.G. (2001) 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.* 41, 1322–1329

7 Jaiswal, M. and Khadikar, P. (2004) QSAR study on (13)C NMR chemical shifts on carbinol carbon atoms. *Bioorg. Med. Chem.* 12, 1793–1798

8 Vanderhoeven, S.J. *et al.* (2004) Nuclear magnetic resonance (NMR) and quantitative structure-activity relationship (QSAR) studies on the transacylation reactivity of model 1beta-O-acyl glucuronides. II: QSAR modeling of the reaction using both computational and experimental NMR parameters. *Xenobiotica* 34, 889–900

9 Khadikar, P.V. *et al.* (2005) Novel use of chemical shift in NMR as molecular descriptor: a first report on modeling carbonic anhydrase inhibitor activity and related parameters. *Bioorg. Med. Chem Lett.* 15, 931–936

10 Latosinska, J.N. (2005) Nuclear quadrupole resonance spectroscopy in studies of biological active molecular systems – a review. *J. Pharm. Biomed. Anal.* 38, 577–587

11 Cramer, R.D. *et al.* (1988) Cross-validation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct.–. Quant. Struct. Act. Relat.* 7, 18–25

12 Cramer, R.D. *et al.* (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* 110, 5959–5967

13 Marshall, G.R. and Cramer, R.D. (1988) Three-dimensional structure-activity relationships. *Trends Pharmacol. Sci.* 9, 285–289

14 Kellogg, et al. (1996) E-state fields: applications to 3D QSAR. *J. Comput. Aided Mol. Des.* 10, 513–520

15 Fang, H. *et al.* (1998) Quantitative comparison of *in vitro* assays for estrogenic assays. *Environ. Health Prospect* 139, 723–729

16 Bradbury, S.P. (1995) Quantitative structure-activity relationship and ecological risk assessment: an overview of predictive aquatic toxicology research. *Toxicol. Lett.* 79, 229–237

17 Pei, J. *et al.* (2005) Improving the quality of 3D-QSAR by using flexible-ligand receptor models. *J. Chem. Inf. Model* 45, 1920–1933

18 De Gregorio, et al. (1988) QSAR modeling with electrotopological state indices: Corticosteroids. *J. Comput. Aided Mol. Des.* 2, 557–561

19 Hopfinger, A.J. *et al.* (1997) Construction of 3D-QSAR models using 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* 119, 10509–10524

20 Senese, C.L. *et al.* (2004) 4D-fingerprints, universal QSAR and QSPR descriptors. *J. Chem. Inf. Comput. Sci.* 44, 1526–1539

21 Tong, W. *et al.* (1997) QSAR models for binding of estrogenic compounds to estrogen receptor α and β subtypes. *Endocrinology* 138, 4022–4025

22 Shi, L.M. *et al.* (2001) QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci.* 41, 186–195

23 Katritzky, A.R. *et al.* (1996) 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.* 100, 10400–10407

24 Fujita, T. *et al.* (1964) A new substituent constant, π, derived from partition coefficient. *J. Am. Chem. Soc.* 86, 5175–5180

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

26 Klopman, G. (1992) MULTICASE1. A hierarchial computer automated structure evaluation program. *Quant. Struct. Act. Rel.* 11, 176–184

27 Good, A.C. *et al.* (1993) Structure-activity relationships from molecular similarity matrices. *J. Med. Chem.* 36, 433–438

28 Bradbury, S.P. (1995) Quantitative structure-activity relationship and ecological risk assessment: an overview of predictive aquatic toxicology research. *Toxicol Letts* 79, 229–237

29 De Dios, A.C. *et al.* (1993) Secondary and tertiary structural effects on protein NMR chemical shifts: an *ab initio* approach. *Science* 260, 1491–1496

30 Cornilescu, G. *et al.* (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* 13, 289–302

31 Beger, R.D. and Bolton, P.H. (1997) 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* 10, 129–142

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

33 Emsley, J.W. *et al.* (1965) Pergamon Press pp. 1–287

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

35 Beger, R.D. *et al.* (2002) Developing comparative structural connectivity spectra analysis (CoSCoSA) models of steroid binding to the corticosteroid binding globulin. *J. Chem. Inf. Comput. Sci.* 42, 1123–1131

36 Bax, A. and Grzesiek, S. (1993) Methodological advances in protein NMR. *Acc. Chem. Res.* 26, 131–138

37 Aue, W.P. *et al.* (1976) Two-dimensional spectroscopy. Application to nuclear magnetic resonance. *J. Chem. Phys.* 64, 2229–2246

38 Bodenhausen, G. and Ruben, D.J. (1980) Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chem. Phys. Lett.* 69, 185–189

39 Bax, A. *et al.* (1983) Sensitivity-enhanced correlation of 15N and 1H chemical shifts in natural-abundance samples via multiple quantum coherence. *J. Am. Chem. Soc.* 105, 7188–7190

40 Bax, A. and Summers, M.F. (1986) 1H and 13C Assignments from sensitivity-enhanced detection of heteronuclear multiple-bond connectivity by 2D multiple quantum NMR. *J. Am. Chem. Soc.* 108, 2093–2094

41 Kumar, A. *et al.* (1980) 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. Commun.* 95, 1–6

42 Clore, G.M. *et al.* (1987) Three-dimensional structure of potato carboxypeptidase inhibitor in solution. A study using nuclear magnetic resonance, distance geometry, and restrained molecular dynamics. *Biochemistry* 26, 8012–8023

43 Beger, R.D. *et al.* (2002) Combining NMR spectral and structural data to form models of polychlorinated dibenzodioxins, dibenzofurans, and biphenyls binding to the AhR. *J. Comput. Aided Mol. Des.* 16, 727–740

44 Beger, R.D. *et al.* (2003) Using simulated 2D ^{13}C-^{13}C NMR spectral data to model a diverse set of estrogens. *Internet Electronic J. Mol. Design* 2, 435–453

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

46 Medical Economics, (2000) Physicians' Desk Reference (54th edn), Medical Economics Company

47 Zinsser, H. (1988) . In *Zinsser Microbiology* (. In *Zinsser Microbiology* 19th edn (Joklik, W.K. and Wilett, H.P., eds), pp. 128–160, Appleton & Lange

48 Beger, R.D. *et al.* (2005) Combining NMR spectral information with associated structural features to form computationally non-intensive, rugged, and objective models of biological activity. In *Drug Discovery Handbook Volume 1: Pharmaceutical Development and Research Handbook* (Shayne, C.Gad, ed.), pp. 227–286, John Wiley & Sons Publisher

49 Grail, B.M. and Payne, J.W. (2002) Conformational analysis of bacterial cell wall peptides indicates how particular conformations have influenced the evolution of penicillin-binding proteins, b-lactam antibiotics and antibiotic resistance mechanisms. *J. Mol. Recognit.* 15, 113–125

50 Compadre, R.L.L. *et al.* (1987) A quantitative structure-activity relationship analysis of some 4-aminophenyl sulfone antibacterial agents using linear free energy and molecular modeling methods. *J. Med. Chem.* 30, 900–906

51 Pickett, S.D. and Sternberg, M.J.E. (1993) Empirical scale of side-chain conformational entropy in protein folding. *J. Mol. Biol.* 231, 825–839

52 Karplus, M. and Kushick, J.N. (1981) Method for estimating the configurational entropy of macromolecules. *Macromolecules* 14, 325–332

53 Beger, R.D. *et al.* (2000) Producing ^{13}C NMR, infrared absorption and EI mass spectrometric data monodechlorination models of chlorobenzenes, chlorophenols, and chloroanilines. *J. Chem. Inf. Comput. Sci.* 40, 1449–1455