Database Retrieval Techniques for Carbon-13 Nuclear Magnetic Resonance Spectrum Simulation

Gary W. Small

Center for Intelligent Chemical Instrumentation, Department of Chemistry, Clippinger Laboratories, Ohio University, Athens, Ohio 45701-2979

Received February 10, 1992

Techniques are described that allow the chemical environments of carbon atoms to be represented in a vector-based format. The database of computed vectors can be used to implement a search of chemical environments for the purpose of retrieving carbon-13 nuclear magnetic resonance chemical shifts. The environment vectors are computed such that they encode both topological and geometrical structural information. The methodology is evaluated by use of a database of 33 polycyclic aromatic compounds and a separate test set of 11 polycyclic aromatics. The shift retrieval method is found to simulate the 11 test spectra to an average deviation of 1.11 ppm between estimated and actual chemical shifts. The spectra simulated by the shift retrieval method are found to be highly similar to spectra simulated by use of a set of previously computed chemical shift models.

INTRODUCTION

Spectrum simulation techniques for carbon-13 nuclear magnetic resonance spectroscopy (^{13}C NMR) allow the chemical shifts of carbon atoms to be estimated directly from chemical structural information. These simulation methods utilize the inherent spectra–structure relationships that make ^{13}C NMR a useful structural investigation tool. Spectrum simulation techniques have potential application in the solution of structure elucidation problems and in the verification of chemical shift assignments.

The two most widely used approaches to predicting ^{13}C NMR chemical shifts are empirical modeling techniques[1-5] and database retrieval methods.[6-11] The modeling techniques attempt to establish empirical relationships between features of the chemical environment and the corresponding chemical shift. Structural features are encoded into numerical parameters for use in the model. Typically, a linear model is computed of the form

$$S = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n \quad (1)$$

where $n+1$ terms are summed to estimate S, the chemical shift of a specific carbon atom whose predicted chemical shift is desired. The X_i terms in eq 1 encode aspects of the chemical environment of the carbon atom in a manner that is linearly related to the chemical shift, whereas the b_i weigh the individual X_i terms. Often, the X_i are computer-generated parameters encoding the steric or electronic characteristics of the molecule. Regression analysis techniques are used to compute the b_i and to select the X_i that are most significant in modeling the chemical shifts.

A database of structures and spectra is used in deriving the models. Once computed, the models can be employed to predict chemical shifts of carbons not included in the actual model development computations. The advantage of the modeling approach is its ability to interpolate chemical shifts. In effect, the computed models are employed to interpolate among the individual carbon atom environments used in the model development.

The database techniques are based on the ability to encode the chemical environment of a carbon atom in a form that can be compared to each member of a library of similarly encoded environments and their associated chemical shifts. In a prediction, each estimated chemical shift is taken as the experimentally observed shift associated with the environment in the library that matches most closely to that of the carbon whose predicted chemical shift is desired. The success of the approach is keyed by having an appropriate database of structures and experimentally observed spectra, as well as an effective means for encoding the chemical environments of carbon atoms in a searchable form. Current schemes for encoding chemical environments focus primarily on the topology of the molecule. Such structural considerations as steric interactions arising from the three-dimensional geometry of the molecule are not considered.

In recent years, work in our laboratory has focused on improving the generality of the chemical shift modeling approach through the design of sophisticated steric and electronic structural parameters.[12-14] This work is motivated by a desire to overcome a drawback of the modeling approach—the presence of an inherent tradeoff between models that produce highly accurate simulated spectra versus models that are applicable to the simulation of a wide variety of structural types. If highly accurate models are desired, those models will tend to be highly specific to the types of carbon atom environments used in the development of the model. In practice, many models are typically required in order to implement a general-purpose spectrum simulation system. Our work in designing new steric and electronic structural parameters is based on a desire to generate models that are more globally applicable while maintaining a high prediction accuracy. While we have been successful in developing improved and more general chemical shift models, it is clear to us that any general-purpose spectrum simulation system based on the modeling approach will require a database of many individual models.

Given the increased availability of large databases of chemical structures and associated ^{13}C NMR spectra, it is apparent that the data are now available to allow large numbers of chemical shift models to be computed. As currently implemented, however, most model development work is performed in a highly interactive manner. Relatively little work has been performed in automating the development of chemical shift models.

Currently, work in our laboratory is focusing on increasing the level of automation used in constructing chemical shift models. A key aspect to automating the model development

0095-2338/92/1632-0279$03.00/0

procedure is the design of methods for selecting appropriate carbon atom environments from a spectra/structure database. The design of methods for the automated selection of a data set for use in computing a given model is the first step in developing an automated model development system.

Automating the selection of a model development data set is identical to retrieving chemical shifts from a database based on comparisons of carbon atom environments. In both cases, a database search is required based on extracting carbon atoms and chemical shifts corresponding to one or more target chemical environments. Unfortunately, the lack of geometrical structural information in most current environment encoding schemes is a major limitation in selecting carbon atoms for use in model development. We have found that geometrical structural information is critical in selecting appropriate carbons for use in developing chemical shift models.

To overcome this limitation, this paper reports a new scheme for encoding both topological and geometrical structural information into a single vector-based representation of the carbon atom environment. Utilizing a set of 44 polycyclic aromatic compounds, this environment encoding scheme is evaluated by implementing a chemical shift retrieval system based on distances between environment vectors. The environment vector approach is evaluated by comparing spectra simulated by use of shift retrieval to the corresponding spectra simulated by use of chemical shift models.

EXPERIMENTAL SECTION

Forty-four polycyclic aromatic compounds and their corresponding ^{13}C NMR spectra were used in this work to evaluate the shift retrieval methodology. The broad-band decoupled ^{13}C NMR chemical shifts were taken from eight literature sources.[15-22] Chemical shifts were referenced to internal or external Me_4Si. This data set was used previously in the development of chemical shift models for polycyclic aromatic compounds.[14] The experimental conditions used in the acquisition of the spectral data are described in detail in the previous work.

The computer software used in this research was written in FORTRAN 77 and implemented on Prime 9955 and Silicon Graphics 4D/440 computers. The chemical shift models used in generating simulated spectra for comparison with the retrieved chemical shifts were computed by use of a set of software tools developed by Small and Jurs.[23] The MM2(87) molecular mechanics software used in computing three-dimensional coordinates for the structures was obtained from the Quantum Chemistry Program Exchange (Department of Chemistry, Indiana University, Bloomington, IN) and was implemented without modification to the force-field parameters.

RESULTS AND DISCUSSION

Overview of Environment Encoding Scheme. The environment encoding scheme used in this work is an extension of a method developed by Small and Jurs.[24] The approach is based on encoding the environment of a carbon atom as

$$\mathbf{e} = (e_0, e_1, e_2, ..., e_n) \tag{2}$$

where e is an $(n+1)$-dimensional vector that describes the structural environment of the carbon in such a way that vectors oriented in similar directions in the $(n+1)$-dimensional space will correspond to atoms having similar ^{13}C NMR chemical shifts. In this scheme, the environments of two carbons can be compared by simply computing the Euclidean distance between the corresponding vector representations of the environments. Analogously, a database of environment vectors can be searched to retrieve the environments that match most closely (i.e., produce smallest Euclidean distances) when compared to a given target environment.

The elements of e are numbers that encode the structural environment radially outward from the carbon atom of interest. For example, e_0 describes the carbon itself, while e_1 describes a group of atoms in the structure located a given radial distance from the carbon. A given element, e_i, is defined as

$$e_i = [\sum_{j=1}^{p_i} Z_j^2]^{1/2}/d^a \tag{3}$$

where Z_j are numerical codes that describe the p_i atoms that have been selected to contribute to e_i. The root sum of squares of the Z_j values is weighted by a distance term, d^a.

The Z_j value for a given atom attempts to encode the influence of that atom in inducing a change in chemical shift. This scheme implies that an atom with a large value of Z_j will have a great influence on the chemical shifts of other atoms in the structure. For example, an oxygen atom would have a larger value of Z_j than an sp^3-hybridized carbon, since the electronegative oxygen atom induces a larger change in the chemical shifts of nearby carbons when introduced into the structure.

For a given atom j, Z_j is defined as

$$Z_j = [\sum_{k=1}^{c_j} P_{\beta,k}^2/c_j]^{1/2} \tag{4}$$

where Z_j is a root mean square average of c_j terms, corresponding to the c_j atoms bonded to atom j. Each $P_{\beta,k}$ is defined as

$$P_{\beta,k} = (M_{\beta,k})(P_{\alpha,j}) + C_{\beta,k} \tag{5}$$

$P_{\alpha,j}$ is a parameter that describes the basic effect of atom j in inducing changes in chemical shifts, while $M_{\beta,k}$ and $C_{\beta,k}$ are multiplicative and additive parameters that adjust $P_{\alpha,j}$ for the effects of atoms bonded to atom j.

Small and Jurs derived the P_α, M_β, and C_β parameters from the experimentally observed ^{13}C NMR chemical shifts of small molecules. For example, P_α for singly-bonded oxygen was derived as

$$P_\alpha = \delta_{\text{methanol}} - \delta_{\text{methane}} = 52.2 \text{ ppm} \tag{6}$$

where δ_{methanol} and δ_{methane} are the actual ^{13}C NMR chemical shifts of the carbon atoms in methanol (49.9 ppm) and methane (−2.3 ppm), respectively. The difference in eq 6 describes the effect on the chemical shift of substituting singly-bonded oxygen for hydrogen, i.e., the influence of singly-bonded oxygen on the carbon chemical shift. Small and Jurs derived carbon, oxygen, nitrogen, fluorine, chlorine, bromine, and iodine parameters for all hybridizations and connectivities. For example, different parameters were derived for 1°, 2°, 3°, and 4° sp^3-hybridized carbons. Thus, the parameters allow different structural units to be distinguished in a manner that is related to the effects on carbon chemical shifts when those structural units are substituted into a molecule.

In the original work by Small and Jurs, atoms were selected for inclusion in eq 3 by their bond distance from the target carbon atom serving as the basis for the environment vector calculation. Thus, e_1 was computed based on the atoms located one bond from the target carbon. The distance parameter, d, in eq 3 was defined as $d = i$ for $i > 0$ and $d = 1$ for $i =$

Table I. Compounds Used for Database Construction and Testing

no.	name	lit ref	backbone no.[a]
	Database Construction		
1	anthracene	15	1
2	phenanthrene	15	2
3	1,4-dimethylphenanthrene	19	2
4	1,4,6-trimethylphenanthrene	19	2
5	3-methylphenanthrene	19	2
6	pyrene	15	3
7	chrysene	15	4
8	1-methylchrysene	15	4
9	2-methylchrysene	15	4
10	3-methylchrysene	15	4
11	benz[*a*]anthracene	15	5
12	7-methylbenz[*a*]anthracene	17	5
13	3,9-dimethylbenz[*a*]anthracene	16	5
14	7,12-dimethylbenz[*a*]anthracene	16	5
15	benzo[*c*]chrysene	15	6
16	picene	16	7
17	dibenz[*a,h*]anthracene	15	8
18	dibenzo[*b,def*]chrysene	16	9
19	benzo[*rst*]pentaphene	16	10
20	coronene	16	11
21	triphenylene	17	12
22	perylene	16	13
23	benzo[*ghi*]perylene	15	14
24	dibenzo[*def,p*]chrysene	16	15
25	benzo[*e*]pyrene	15	16
26	fluoranthene	15	17
27	benzo[*ghi*]fluoranthene	15	18
28	indeno[4,3,2,1]chrysene	16	19
29	benzo[*k*]fluoranthene	15	20
30	benz[*e*]acephenanthrylene	15	21
31	naphth[2,3-*a*]aceanthrylene	16	22
32	indeno[1,2,3-*cd*]fluoranthene	16	23
33	biphenyl	18	24
	Testing		
34	1,8-dimethylanthracene	20	1
35	1-methylphenanthrene	21	2
36	4-methylphenanthrene	21	2
37	6-methylchrysene	15	4
38	benzo[*b*]chrysene	15	25
39	naphtho[2,3-*k*]fluoranthene	16	30
40	benzo[*g*]chrysene	22	27
41	dibenzo[*def,mno*]chrysene	15	28
42	naphtho[1,2,3,4-*def*]chrysene	16	29
43	benzo[*def*]chrysene	15	26
44	dibenz[*a,e*]aceanthrylene	16	31

[a] Refers to backbones displayed in Figure 1.

0. The exponent, *a*, was empirically set at three. The distance term gave more numerical weight to structural features closer to the target carbon. In the original work, *n* in eq 2 was set at five, producing six-dimensional **e** vectors. This selection was based on the fact that structural features located more than five bonds from the target carbon rarely have a significant influence on its chemical shift.

The environment encoding scheme described above has been used extensively in our laboratory to group carbons in similar environments for the purpose of computing chemical shift models.[12-14] These previous applications depend on general clustering of the environment vectors according to structural similarity, but they do not require the same degree of clustering required in a shift retrieval application or in an application in which a database containing thousands of different environments must be interrogated.

Our experience in using the environment vectors has shown that the principal drawback of the scheme described above is a lack of geometrical structural information in the method. The use of bond steps radially outward from the target carbon is an approach based entirely on the topology of the molecule. Shielding effects due to the geometry of the structure are not encoded. For the shift retrieval application described here, it was judged essential to incorporate geometrical structural information into the approach.

Table II. Statistics for Chemical Shift Models

no.	shift range (ppm)	n^a	p^b	R^c	s^d	F^e
1	14.1–27.3	14	2	0.989	0.588	251
2	119.1–135.2	158	11	0.951	0.985	125
3	124.8–136.6	76	6	0.967	0.624	165
4	122.6–135.6	23	4	0.972	1.094	76.6
5	124.5–142.1	66	7	0.937	1.434	59.7
6	127.8–135.1	55	5	0.901	0.815	42.5

[a] Number of chemical shifts used to define model. [b] Number of parameters in computed model. [c] Correlation coefficient. [d] Standard error of estimate in chemical shift units (ppm). [e] *F*-value for significance of the model.

Description of Test Data. To test the environment vector approach in a shift retrieval application, a data set of 44 polycyclic aromatic compounds was employed which had been used previously in a chemical shift modeling study.[14] The 44 compounds encompass 31 different aromatic ring backbones, ranging in size from two to seven individual rings. The compounds were divided into a set of 33 compounds (24 different ring backbones) for use in constructing the database of environments and a set of 11 compounds (seven unique ring backbones) for use in testing the ability of the environment vector approach to retrieve appropriate chemical shifts from the database. Compound names for each of the 44 compounds are listed in Table I along with a corresponding identification number. The structural backbones of these compounds (excluding substituents) are shown in Figure 1.

The structures of the compounds were converted into computer-readable form through the use of a graphical procedure developed by Brügger and Jurs.[25] Molecular mechanics procedures were used to obtain the approximate molecular geometries of the compounds. A force field described by Stuper et al.[26] was used to compute initial atomic coordinates for the compounds, and the MM2(87) force field developed by Allinger et al.[27] was used to refine the coordinates.

Summary of Chemical Shift Modeling Results. This data set represents the most structurally diverse set of compounds for which we have been able to develop a single set of chemical shift models that meet our acceptance criterion of 1 ppm standard error. In effect, the models developed for these compounds represent the current state-of-the-art in chemical shift modeling. The modeling study used the 33 compounds noted above in computing chemical shift models, and the analogous set of 11 compounds to test the computed models. The 392 structurally unique carbons in the 33 compounds were divided into six subsets, and a chemical shift model was developed independently for each subset. The six groups corresponded to (1) methyl carbons, (2) ring atoms two bonds from a ring junction, (3) ring atoms one bond from a ring junction, (4) inner-ring junction atoms shared by more than two rings and atoms within a four-sided bay region, (5) ring junction atoms within a three-sided bay region, and (6) ring junction atoms shared by two rings but not located in a bay region. This final grouping of the atoms was based on the results of several failed attempts to model the data with fewer than six atom groups. The rationale for the atom grouping procedure is explained fully in the paper describing this work.[14]

The number of atoms and range of chemical shifts in each of the six atom groups are listed in Table II, along with the statistics describing the six computed chemical shift models. As noted in the table, each model was characterized by a correlation coefficient greater than 0.9, and the standard errors

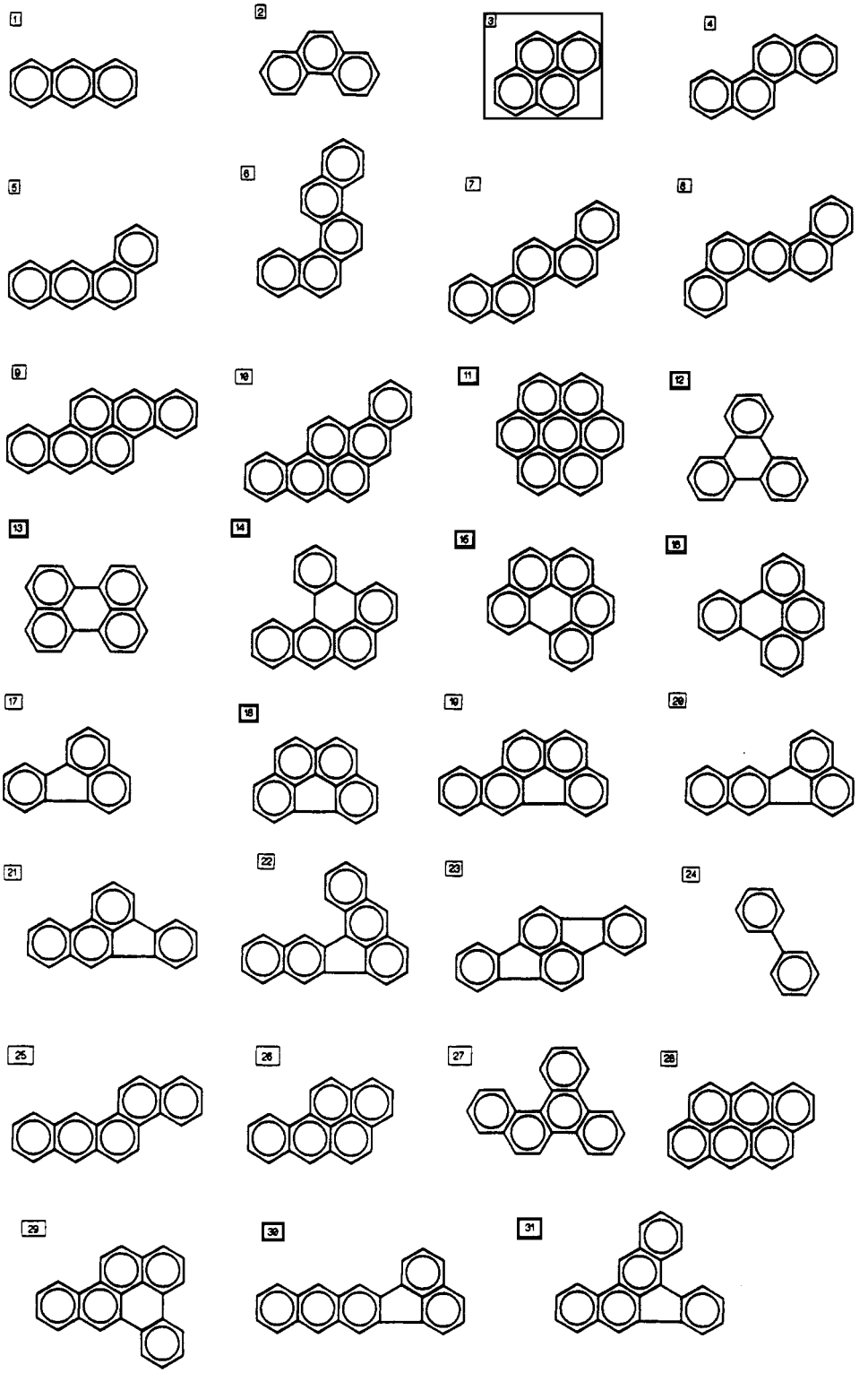

Figure 1. Structural backbones of the 44 compounds listed in Table I.

of estimate were in the range of 1 ppm. The models successfully predicted the spectra of the 11 test compounds to an average deviation of 1.02 ppm between predicted and observed chemical shifts. Based on these standard evaluation methods, the modeling study was judged highly successful.

Inclusion of Geometrical Information in Environment Vectors. Figure 2 illustrates the lack of geometrical information in the basic environment encoding scheme described above. The structures of compounds **1** and **2** are depicted, and one target carbon is selected in each compound (indicated by the solid circle) to serve as the focus for this discussion. The numbers beside each atom indicate the distance in bond steps from that atom to the target carbon. Also plotted are concentric circles in radial steps of 1.5 Å outward from the target carbon. In three dimensions, these circles correspond to concentric spherical shells radiating outward from the target carbon.

In the original environment vector formulation, atoms were selected for inclusion in the e_i calculations based on their bond distances from the target carbon. For these cyclic structures, the bond distance method is inadequate. In both compounds, the second shell contains two atoms located two bonds from the target carbon and one atom located three bonds away. As indicated in eq 3, these atoms would contribute to different

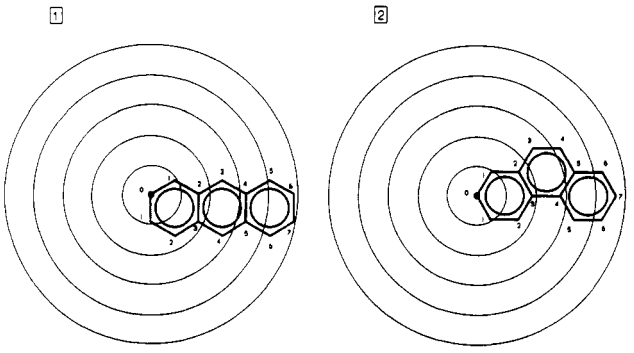

Figure 2. Structures of compounds **1** and **2**, indicating the deficiency in the bond-distance selection scheme for the elements of e_i in eq 3. The numbers indicate the bond distance from each atom to the target carbons denoted by the solid circles. The large circles denote a series of concentric shells spaced 1.5 Å apart, radiating outward from the target carbon. The occurrence in the same shell of atoms with different bond distances emphasizes the deficiency of the bond-step approach.

elements of **e**. Due to the effect of the d^a weighting term in eq 3, the atom three bonds from the target carbon would be assigned significantly less influence than the atoms two bonds away, even though its distance to the target carbon is similar. This is effectively an error in the encoding scheme, given that atoms at equal through-space distances typically influence chemical shifts in a similar manner.

For this reason, it was decided to modify the selection of atoms for inclusion in eq 3. In the new scheme, atom selection would be based on the occurrence of atoms in spherical shells radiating outward from the target carbon. In this scheme, the variables requiring optimization were the number of shells to use and the radial dimension of the shells.

Testing of Environment Vectors for Shift Retrieval. Environment vectors were computed for each of the 640 carbons in compounds **1**–**33** and stored in computer disk files. This set of environment vectors served as the database for use in retrieving chemical shifts for the 11 test compounds. For the test compounds, environment vectors were computed for each of the 196 structurally unique carbon atoms. For each atom, the corresponding environment vector was compared to each of the 640 environment vectors in the database. Based on computed Euclidean distances, the nearest match to each environment was determined. The estimated chemical shifts for the 11 test compounds were taken in each case as the database chemical shift corresponding to the nearest match of the environments. The average deviation between estimated and actual chemical shifts across the 11 compounds was used as an overall estimate of the utility of the environment vector approach.

The above procedure was used multiple times as the individual variables pertaining to the environment vector calculation were varied. Variables studied included (1) a in eq 3, (2) n in eq 2 (i.e., the dimensionality of the vectors), (3) the use of topological vs geometrical criteria in assigning atoms to e_i elements in eq 2, and (4) the radial distance between spherical shells in the implementation of the geometrical-based e_i elements.

The value of a in eq 3 was varied between 1, 2, and 3 during the exploration of the other variables. The values of 2 and 3 produced highly similar results, both producing results superior to a value of 1. A value of 2 was selected finally in order to give the features of the environment remote from the target carbon slightly more weight. It should be stressed, however, that the additional results and conclusions presented below are equally valid for $a = 2$ or $a = 3$.

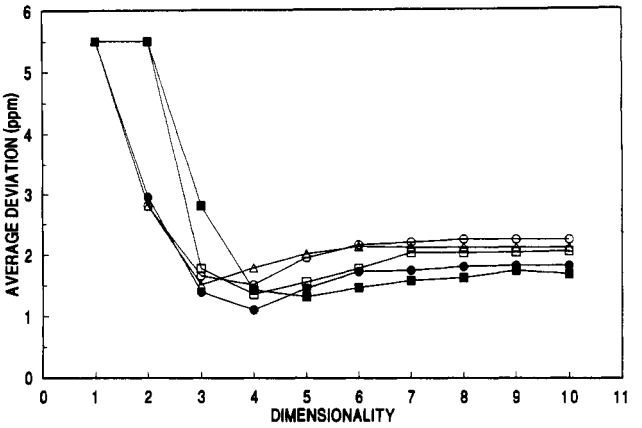

Figure 3. Plot of average deviation between estimated and actual chemical shifts for the set of 11 test compounds as a function of the dimensionality of the geometric-based environment vectors. The plotted curves correspond to spherical shells spaced apart by 1.0 (solid square), 1.3 (open square), 1.5 (solid circle), 1.8 (open circle), and 2.0 Å (open triangle).

The dimensionality of the environment vectors was varied between 1 and 10 ($n = 0$ to $n = 9$) during the comparison of the topological and geometrical criteria for computing e_i. For both topological and geometrical criteria, a dimensionality of four ($n = 3$) produced the lowest overall deviation between estimated and actual chemical shifts for the 11 test compounds. This result was obtained for both $a = 2$ and $a = 3$.

For the topological environment vectors, the dimensionality of 4 differs from the value of six used previously for grouping carbon atoms in the context of the chemical shift modeling work. The dimensionality of six was chosen originally based on the rationale that atoms rarely influence the chemical shifts of other atoms greater than five bonds away. The shift retrieval problem differs from the atom grouping problem, however, in that a correlation is required between intervector Euclidean distances and the chemical shifts of the corresponding carbons. The results obtained here indicate that the inclusion of structural effects four and five bonds from the target carbon interferes with the retrieval of environments with closely matching chemical shifts. This may be due partially to a limitation of the environment vector encoding scheme. As indicated in eq 3, each atom at a particular bond step contributes to a single element of **e**. In large structures, many atoms may be found four or five bonds from the target carbon. Attempting to combine the effects of many atoms into a single vector element may not be effective in reproducing the correlation between chemical structure and chemical shift.

Figure 3 summarizes the investigation of the radial distance between spherical shells for the geometric-based environment vectors as a function of the dimensionality of the vectors. The average deviation between estimated and actual chemical shifts for the 11 test compounds is plotted vs dimensionality for each of five different radial distances. The radial distances tested were 1.0 (solid squares), 1.3 (open squares), 1.5 (solid circles), 1.8 (open circles), and 2.0 Å (open triangles). The best results observed corresponded to a radial distance of 1.5 Å between shells and a dimensionality of four, producing an average deviation of 1.11 ppm between the 196 estimated and actual chemical shifts in the 11 test compounds. This spacing between shells is the same as that displayed in Figure 2.

In each of the curves in Figure 3, the average deviation reaches a minimum, increases, and then levels to a constant value at high dimensionality. The general trend is that larger radial distances produce a minimum at lower dimensionality. As the radial distance increases, more atoms contribute to the

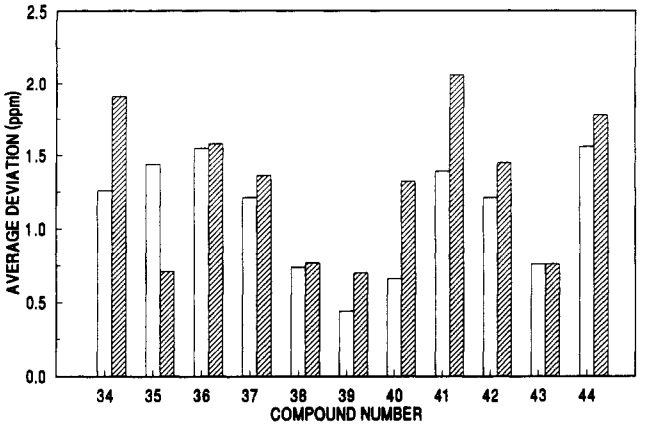

Figure 4. Clustered bar graph comparing the average deviation between estimated and actual chemical shifts for each of the 11 test compounds. The left (open) bar in each cluster indicates the results produced by the four-dimensional geometric-based environment vectors with a shell spacing of 1.5 Å. The right bar (striped) in each cluster indicates the average deviation produced by the four-dimensional topological environment vectors. The geometric-based vectors produce smaller deviations in nine of the 11 cases.

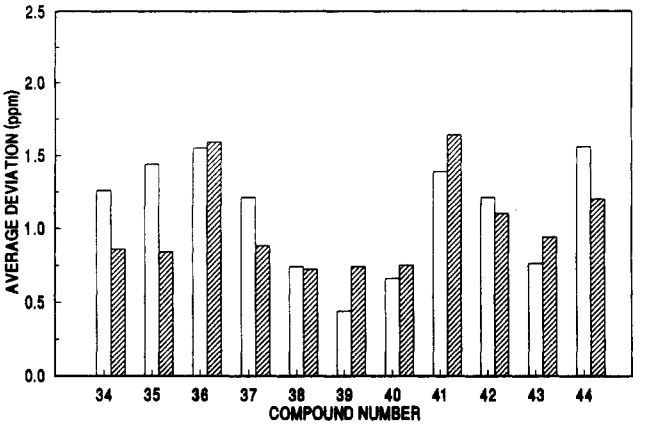

Figure 5. Clustered bar graph comparing the average deviation between estimated and actual chemical shifts for each of the 11 test compounds. The left (open) bar in each cluster indicates the results produced by the four-dimensional geometric-based environment vectors with a shell spacing of 1.5 Å. The right bar (striped) in each cluster indicates the average deviation produced by the application of the previously computed chemical shift models. The chemical shift models produce smaller deviations in six of the 11 cases.

lower dimensional elements of the environment vectors. This supports the conclusion that the inclusion of structural features at large distances from the target carbon serves to interfere with the accuracy of the shift retrieval. This also accounts for the increase in average deviation after the minimum. As noted above, this effect may be partially due to a limitation in the ability of the encoding scheme to combine the effects of many atoms into a single vector element. The curves level at high dimensionality, as few additional atoms are found in the last spherical shells. This produces small values of e_i for the last elements, causing the environment vectors to be very similar.

Figure 4 is a clustered bar graph that compares the best results obtained for the topological and geometrical-based environment vectors. Both calculations used $a = 2$ in eq 3 and a dimensionality of four ($n = 3$ in eq 2). This corresponds to a distance of three bond steps from the target carbon for the topological approach and three spherical shells separated by 1.5 Å for the geometric approach. In the plot, the average deviation between estimated and actual chemical shifts is indicated for each of the 11 test compounds. The open (leftmost) bar in each cluster corresponds to the geometric-based environment vectors, while the striped (rightmost) bar corresponds to the topological environment vectors. For nine of the 11 compounds, the geometric approach produces superior results. For one compound (**43**), the results are equal. Only in the case of compound **35** does the topological approach yield better results. The topological calculation produces an overall deviation of 1.31 ppm, compared to the 1.11 ppm deviation for the geometric approach. For the nine cases in which the geometric-based environment vectors yield better results, an average improvement of 22% in the accuracy of the simulated chemical shifts is observed. The modification of the environment vector calculation to include geometric structural information is clearly an advancement.

Comparison between Shift Retrieval and Chemical Shift Modeling. As a final test, the results of the geometric-based shift retrieval were compared to the prediction results obtained with the computed chemical shift models. In a manner analogous to Figure 4, Figure 5 compares the shift retrieval and modeling results for the 11 test compounds. In Figure 5, the results obtained with the geometric-based environment vectors are presented as the leftmost (open) bar, while the results produced by application of the chemical shift models are indicated by the rightmost (striped) bar.

The models produce the smallest average deviation in chemical shift for six of the 11 compounds, while the shift retrieval approach produces the smallest deviation for the other five compounds. As noted previously, the overall average deviation between estimated and actual chemical shifts was 1.02 ppm for the model-based prediction vs 1.11 ppm for the shift retrieval. In principle, one would expect the modeling approach to produce superior results, given the inherent capability of the models to interpolate among the carbon environments in the original database.

CONCLUSION

The close similarity of the shift retrieval and modeling results in simulating the spectra of the 11 test compounds suggests that the environment vector approach to shift retrieval is highly effective. The incorporation of geometric structural information into the environment vectors is judged an advancement in building a general-purpose method for representing chemical environments.

Clearly, the work reported in this paper represents an initial feasibility study for the shift retrieval method, given the relatively small size of the database and the restriction to polycyclic aromatic compounds. The selection of this test case was motivated by a desire to compare the results of the shift retrieval approach directly to results obtained in a successful chemical shift modeling study. The similarity observed in the shift retrieval and modeling results is very encouraging. Further testing of the method with an expanded database is required, however, before additional conclusions can be drawn regarding the overall utility of the shift retrieval algorithm.

While the work presented here focused on the use of the environment vectors in a spectrum simulation application based on shift retrieval, the method also shows significant promise in helping to automate the chemical shift modeling procedure. In the modeling application, a target carbon environment would be used to search the database of environment vectors for the closest matching chemical environments. The retrieved environments would then be used in constructing a model for use in predicting the chemical shift of the target carbon.

An additional advantage of the environment vector approach is the numerical basis of the representation. Environment representations used in previous shift retrieval applications have often used character strings to encode structural features.[6] Lexical comparisons are then made between the representations. A numerical approach has three advantages. First, character-based encodings of the chemical environment are inherently discrete functions, while a numerical encoding is a continuous function. Comparisons between values of continuous functions often have an advantage in precision. Second, multivariate statistical methods can be applied directly to the numerical data. For example, multivariate techniques such as principal components analysis may be useful in increasing the speed of the environment comparisons. Finally, the numerical-based environment vectors are directly compatible with advances in computer hardware such as vector processing and parallel processing.

ACKNOWLEDGMENT

This paper was presented at the 202nd ACS National Meeting, New York, NY, Aug 26, 1991. Funding for this research was provided by the Shell Development Co., Houston, TX.

REFERENCES AND NOTES

(1) Lindeman, L. P.; Adams, J. Q. Carbon-13 Nuclear Magnetic Resonance Spectrometry: Chemical Shifts for the Paraffins through C_9. *Anal. Chem.* **1971**, *43*, 1245–1252.
(2) Clerc, J. T.; Sommerauer, H. A Minicomputer Program Based on Additivity Rules for the Estimation of ^{13}C-NMR Chemical Shifts. *Anal. Chim. Acta* **1977**, *95*, 33–40.
(3) Ewing, D. F. ^{13}C Substituent Effects in Monosubstituted Benzenes. *Org. Magn. Reson.* **1979**, *43*, 499–524.
(4) Sutton, G. P.; Jurs, P. C. Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Alkyl-Substituted Aromatic Compounds. *Anal. Chem.* **1990**, *62*, 1884–1891.
(5) Ball, J. W.; Anker, L. S.; Jurs, P. C. Automated Model Selection for the Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Cyclopentanones and Cycloheptanones. *Anal. Chem.* **1991**, *63*, 2435–2442.
(6) Bremser, W. HOSE—A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355–365.
(7) Gray, N. A. B.; Crandell, C. W.; Nourse, J. G.; Smith, D. H.; Dageforde, M. L.; Djerassi, C. Computer-Assisted Structural Interpretation of Carbon-13 Spectral Data. *J. Org. Chem.* **1981**, *46*, 703–715.
(8) Bremser, W. Expectation Ranges of ^{13}C NMR Chemical Shifts. *Magn. Reson. Chem.* **1985**, *23*, 271–275.
(9) Crandell, C. W.; Gray, N. A. B.; Smith, D. H. Structure Evaluation Using Predicted ^{13}C Spectra. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 48–57.
(10) Mitchell, T. M.; Schwenzer, G. M. Application of Artificial Intelligence for Chemical Inference. XXV. A Computer Program for Automated Empirical ^{13}C NMR Rule Formation. *Org. Magn. Reson.* **1978**, *11*, 378–384.
(11) Von der Lieth, C. W.; Seil, J.; Köhler, I.; Opferkuch, H. J. ^{13}C NMR Data Bank Techniques as Analytical Tools. *Magn. Reson. Chem.* **1985**, *23*, 1048–1055.
(12) Small, G. W.; McIntyre, M. K. Structure Elucidation Methodology for Disaccharides Based on Carbon-13 Nuclear Magnetic Resonance Spectrum Simulation. *Anal. Chem.* **1989**, *61*, 666–674.
(13) Barber, A. S.; Small, G. W. Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Linear Cyclic Aromatic Compounds. *Anal. Chem.* **1989**, *61*, 2658–2664.
(14) Jensen, K. L.; Barber, A. S.; Small, G. W. Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Polycyclic Aromatic Compounds. *Anal. Chem.* **1991**, *63*, 1082–1090.
(15) *Spectral Atlas of Polycyclic Aromatic Compounds*; Karcher, W., Fordham, R. J., Dubois, J. J., Glaude, P. G. J. M., Lighart, J. A. M., Eds.; D. Reidel: Dordrecht, 1983; Vol. 1.
(16) *Spectral Atlas of Polycyclic Aromatic Compounds*; Karcher, W., Ed.; Kluwer Academic: Dordrecht, 1988; Vol. 2.
(17) Ozubko, R. S.; Buchanan, G. W.; Smith, I. C. P. Carbon-13 Nuclear Magnetic Resonance Spectra of Carcinogenic Polynuclear Hydrocarbons. I. 3-Methylcholanthrene and Related Benzanthracenes. *Can. J. Chem.* **1974**, *52*, 2493–2501.
(18) *The Sadtler Standard Spectra ^{13}C NMR*; Sadtler Research Laboratories: Philadelphia, 1976; Vol. 2, p 274c.
(19) Letcher, R. M. The Anisotropic Effect of 4-Substituents on the Proton NMR Chemical Shift of H-5 in Phenanthrenes. *Org. Magn. Reson.* **1981**, *16*, 220–223.
(20) Wolfenden, W. D.; Grant, D. M. Carbon-13 Magnetic Resonance. V. Conformational Dependence of the Chemical Shifts in the Methylbenzenes. *J. Am. Chem. Soc.* **1966**, *88*, 1496–1502.
(21) Stothers, J. B.; Tan, C. T.; Wilson, N. K. ^{13}C NMR Studies of Some Phenanthrene and Fluorene Derivatives. *Org. Magn. Reson.* **1977**, *9*, 408–413.
(22) Bax, A.; Ferretti, J. A.; Nashed, N.; Jerina, D. M. Complete ^{1}H and ^{13}C NMR Assignment of Complex Polycyclic Aromatic Hydrocarbons. *J. Org. Chem.* **1985**, *50*, 3029–3034.
(23) Small, G. W.; Jurs, P. C. Interactive Computer System for the Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra. *Anal. Chem.* **1983**, *55*, 1121–1127.
(24) Small, G. W.; Jurs, P. C. Determination of Topological Similarity of Carbon Atoms in the Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra. *Anal. Chem.* **1984**, *56*, 1314–1323.
(25) Brügger, W. E.; Jurs, P. C. Molecular Structure Input Program Using a Storage Cathode Ray Tube Terminal. *Anal. Chem.* **1975**, *47*, 781–784.
(26) Stuper, A. J.; Brügger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979; pp 83–90.
(27) Sprague, J. T.; Tai, J. C.; Yuh, Y.; Allinger, N. L. The MMP2 Calculational Method. *J. Comput. Chem.* **1987**, *8*, 581–603.