

Enhanced Structural Encoding Algorithm for Database Retrievals of Carbon-13 Nuclear Magnetic Resonance Chemical Shifts

Robert C. Schweitzer and Gary W. Small*

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

Received September 30, 1995[®]

An enhanced version of an algorithm is discussed which encodes a description of the chemical environment of carbon atoms in a manner that correlates to carbon-13 nuclear magnetic resonance (¹³C NMR) chemical shifts. The encoding algorithm uses a vector-based approach in which the first dimension of the vector represents the chemical shift of the carbon atom, the second dimension represents the collective influence of atoms one bond away from the carbon on its chemical shift, and each successive dimension represents the influence of the atoms one bond further away. This encoding algorithm is a key component of a ¹³C NMR spectrum simulation procedure in which each of the carbons in a large database of known structures and spectra is represented as a vector. Database search methods based on vector comparisons are used to find the closest matching chemical environments and associated chemical shifts for each of the carbons in a structure input by a user. Enhancements to the original algorithm include an expansion of the number of atom classes treated, the addition of a scheme to treat aromatic systems as a special case, and the use of an expanded vector format to regain some of the information lost by collapsing the molecular structure to a vector representation. To test this algorithm, a database of structures and spectra is split into training and test sets consisting of 16 959 and 4240 structures, respectively. Experiments performed to optimize several parameters associated with the encoding algorithm are followed by comparing the retrieved (i.e., predicted) and actual chemical shifts for the structures in the test set. For the optimal parameter settings found, the median of the mean absolute deviations in chemical shifts for the structures in the test set was 1.30 ppm and was obtained with an expanded vector representation based on 15 dimensions.

INTRODUCTION

Carbon-13 nuclear magnetic resonance spectroscopy (¹³C NMR) is a primary analysis technique for the organic chemist, and the availability of software tools to aid in the interpretation of ¹³C NMR spectra can be of great benefit. One such tool is spectrum simulation, which results in the generation of a predicted spectrum, given a chemical structure. A typical use for spectrum simulation involves the proposal of candidate structures for an unknown, followed by comparison of the experimental ¹³C NMR spectrum with each of the spectra obtained from the spectrum simulation of the candidate structures. This procedure can aid the identification of the structure of the compound under examination.

The simulation of a ¹³C NMR spectrum requires the prediction of the chemical shift for each of the carbons in the corresponding chemical structure. Two practical methods for chemical shift prediction are direct database retrieval methods^{1–6} and empirical modeling methods.^{7–11}

The direct database retrieval method is dependent on the availability of a large database of spectra and structures in which a structural representation of each carbon atom in the database is stored along with its experimentally measured chemical shift. The structural representation requires an algorithm in which the chemical environment of the carbon is encoded in a manner that correlates with its chemical shift. To predict a chemical shift for a carbon atom in a structure

input by the user, the first step is the encoding of the chemical environment of the carbon atom using the same algorithm used to encode and store the chemical environment of each carbon in the large database. The closest chemical environment in the database is found for the carbon atom of interest, and the predicted shift is simply the chemical shift of that closest match. By repeating this procedure for each carbon in the input structure, an entire simulated spectrum can be generated. The direct database retrieval approach has the advantage that chemical shifts can be predicted for carbon atoms in a wide range of chemical environments. A disadvantage of this approach is that the quality of the predicted shift depends on the presence of a carbon atom in the database with a similar chemical environment. On average, chemical shifts predicted by the direct retrieval approach do not have accuracy as high as those predicted by the empirical modeling method.

The empirical model building approach depends on the computation of models that relate chemical structural features to observed chemical shifts. Generally, the models used have the form

$$s_a = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n \quad (1)$$

where s_a is the chemical shift of atom a , the x_i terms are calculated structural descriptors that encode some topological, electronic, or geometrical aspect of the chemical environment of atom a , and the b_i are weighting coefficients. A set of carbons with known chemical shifts is used to build the model, and the b_i terms are calculated by use of regression analysis techniques. To predict a ¹³C NMR spectrum for a

* Author to whom correspondence should be sent. Phone: (614) 593-1748; e-mail: small@helios.phy.ohiou.edu.

[®] Abstract published in *Advance ACS Abstracts*, March 1, 1996.

structure input by a user, one or more models must be built that are applicable to each carbon atom in that structure by use of sets of atoms with known chemical shifts. The models are then used to calculate the chemical shifts of each of the carbons in the input structure, and the resulting shifts are assembled into a spectrum. The empirical modeling method has on average better accuracy of predicted shifts than the direct retrieval approach due to an inherent interpolative capability. The models developed, however, have the disadvantage of being applicable only to a narrow range of chemical environments.

Research in our laboratory is focusing on the ultimate goal of combining these two approaches to gain the relative generality of the database retrieval method and the accuracy of the empirical modeling technique. The first step in this combined approach is the creation of an atom-based database in which each of the carbon atoms is encoded according to its chemical environment and stored along with its experimentally observed chemical shift. The second step is the use of direct retrieval to select a subset of carbon atoms from the atom-based database which have similar chemical environments to the carbon atom in the input structure whose predicted chemical shift is desired (the target carbon). In the third step, an empirical model is built using the subset of carbons obtained from the direct retrieval step, and the chemical shift for the target carbon is calculated by use of this model. By repeating this procedure for each of the carbons in the input structure, a complete simulated spectrum can be assembled.

One of the important steps in this chemical shift prediction methodology is the algorithm used to encode the chemical environment. The initial version of the algorithm used in this work was developed by Small and Jurs in 1984 and has been used in a number of studies.¹² Until this current work, however, all of the studies of this encoding algorithm have involved small collections of data targeted to the prediction of chemical shifts for carbons with a specific and narrow range of chemical environments. In the work presented here, a relatively large database is used, thereby allowing testing of the ability of the algorithm to work with a wide range of chemical environments. In this work, chemical shift predictions are performed by direct retrieval of matching environments from the database. The success of these database retrieval experiments will be evaluated, and improvements which have been made to increase the selectivity of the environmental encoding scheme will be described.

EXPERIMENTAL SECTION

The Sadtler ¹³C NMR database consisting of 29 966 structures and their accompanying chemical shifts was used for this research (Bio-Rad, Inc., Sadtler Division, Philadelphia, PA). To test the utility of including geometrical information in the environmental encoding scheme, the structures were converted from two-dimensional to three-dimensional representations by use of a molecular mechanics procedure reported by Stuper et al.¹³ The coordinates obtained through this procedure were refined by use of MM2 (1987 version), a molecular mechanics package developed by Allinger and co-workers.¹⁴ Of the original 29 966 structures, a set of 21 199 structures was chosen for the current work. The structures chosen were judged to be well modeled, contained only atoms supported by the environmental encoding algorithm used in this work (C, H, O, N, P, S, F, Cl, Br, I), and included a wide range of organic

compounds. The data set was divided randomly into two sets of structures. The larger set contained 16 959 structures and was used to define the database (library set) used subsequently in the retrieval experiments. The smaller set, which contained 4240 structures, was used as a test set.

The computations described in this paper were implemented on a Silicon Graphics 4D/460 computer system running under Irix (version 5.2, Silicon Graphics, Mountain View, CA) and operating in the Center for Intelligent Chemical Instrumentation at Ohio University. The software developed for this work was written in FORTRAN 77.

RESULTS AND DISCUSSION

Overview of Environmental Encoding Scheme. Retrieval of the most structurally similar carbons from a database requires an encoding scheme for the chemical environment of each carbon. The purpose of this scheme is to represent the chemical environment of carbon atoms in a manner that relates to the chemical shift and in a way that facilitates quantitative comparisons between environments. As already mentioned, the basic algorithm used in this work was developed by Small and Jurs and has been used previously to implement chemical shift retrievals.¹² In this scheme, the environment of each carbon atom is represented as an $n+1$ dimensional vector of the form

$$\mathbf{e}_a = (e_0, e_1, \dots, e_n) \quad (2)$$

where the first dimension, e_0 , characterizes atom a , the second dimension, e_1 , describes the influence of atoms one bond away from atom a on its chemical shift, and each successive dimension describes the collective influence of the atoms one bond further from atom a on the chemical shift of that atom. In the work reported here, the vector dimensionality has been varied from two to eight to find an optimal dimensionality.

An element of e_a is defined as

$$e_i = \frac{\sqrt{\sum_{j=1}^{P_i} Z_j^2}}{d^r} \quad (3)$$

where the Z_j are numerical codes that describe the P_i atoms that contribute to e_i . The root sum-of-squares of the Z_j values is weighted by a distance term, d^r , where $d = \max(1, i)$. This ensures that the higher dimensions will have less numerical weight than the lower dimensions and reflects the reality that atoms close to a given carbon atom have a greater effect on its chemical shift than atoms which are far away. For the work reported here, r in eq 3 has been varied from one to five. In the original work, a linear summation was studied in addition to the root sum-of-squares calculation, and, in the work presented here, both approaches will be presented as well.

For a given atom j , Z_j is defined as

$$Z_j = \sqrt{\frac{\sum_{k=1}^{c_j} P_{\beta,k}^2}{c_j}} \quad (4)$$

where Z_j is a root mean square average of c_j terms,

Table 1. Atom Classes Used in Generating Structural Parameters

class	description
1	cyclopropane
2	alkane
3	cyclopropene
4	alkene
5	C=O carbon
6	C=N carbon
7	alkyne
8	cyano carbon
9	aromatic carbon
10	non-furan sp ³ oxygen
11	furan oxygen
12	C=O, N=O, or S=O oxygen
13	non-pyrrole sp ³ nitrogen
14	pyrrole nitrogen
15	nitro group nitrogen
16	pyridine nitrogen
17	non-pyridine, non-nitro sp ² nitrogen
18	cyano group nitrogen
19	fluorine
20	chlorine
21	bromine
22	iodine
23	non-thiophene sp ³ sulfur
24	thiophene sulfur
25	sp ² sulfur
26	sp ³ phosphorus
27	sp ² phosphorus
28	cyclic alkane
29	cyclic alkene
30	cyclic alkyne
31	cyclic non-furan sp ³ oxygen
32	cyclic C=O, N=O, or S=O oxygen where the C,N, or S is in the ring
33	cyclic non-pyrrole sp ³ nitrogen
34	cyclic non-pyridine, non-nitro sp ² nitrogen
35	cyclic non-thiophene sp ³ sulfur
36	cyclic sp ² sulfur
37	cyclic C=O carbon

corresponding to the c_j atoms bonded to atom j . Each $P_{\beta,k}$ is defined as

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

where $P_{\alpha,j}$ is a parameter that describes the basic effect of atom j in inducing changes in chemical shifts, and $P_{\beta,k}$ is an adjusted $P_{\alpha,j}$ that encodes the change in influence of atom j on the target carbon due to the fact that atom k is bonded to atom j . $M_{\beta,k}$ and $C_{\beta,k}$ are multiplicative and additive parameters that implement this adjustment to $P_{\alpha,j}$. The P_{α} , M_{β} , and C_{β} parameters were derived from experimentally observed ¹³C NMR chemical shifts of small molecules and are defined in chemical shift units. These values have been calculated and tabulated for each atom type, hybridization, and connectivity commonly encountered and will be discussed in greater detail below.

Expanded Atom Classes. In the original 1984 work, P_{α} , M_{β} , and C_{β} parameters were calculated for 23 different atom classes for carbon, oxygen, nitrogen, and the halogens, in which each instance of bonding hybridization and connectivity for a given element resulted in a new class. In the work presented here, the number of atom classes has been expanded to 66 and includes new additions for sulfur, phosphorus, and atoms in cyclic environments. These 66 classes can be categorized based on an expanded definition of elemental type, and the resulting 37 categories of classes are listed in Table 1.

A P_{α} value is calculated as the difference in shift of a small reference molecule such as methane and that reference molecule in which one of the hydrogens has been replaced by the atom type of interest. The P_{α} value for secondary sp³ sulfur, for example, is calculated as the difference in shift between methane and a methane molecule in which one of the hydrogens is replaced by methyl sulfide and is computed as

$$\delta\text{CH}_3\text{SCH}_3 - \delta\text{CH}_4 = 18.00 - -2.3 = 20.3 \text{ ppm} \quad (6)$$

In the original work, all of the chemical shifts used to generate P_{α} , M_{β} , and C_{β} parameters were taken from the literature. In the recent parameter development work, most of the chemical shifts used have been taken from the Sadtler database.

The M_{β} and C_{β} parameters are simply the slope and y-intercept, respectively, of a plot of carefully selected P_{β} versus P_{α} values. Table 2 illustrates such a calculation for secondary sp³ sulfur. The first column contains the structures from which the P_{β} values are calculated, and the italicized carbons are the carbons whose chemical shifts are being influenced by the α and β effects being modeled. The chemical shifts of these carbons are listed in column three. The second column describes the atom type of the α atom. The influences of these atoms on the chemical shifts of the underlined atoms are encoded by the P_{α} values in the sixth column. The β atoms are all secondary sp³ sulfurs and serve to modify the effect of the α atoms. The P_{β} values listed in the fifth column of the table are computed as simply the experimentally measured chemical shifts of the italicized carbons in column one minus the chemical shift of the reference carbon (e.g., $\delta\text{CH}_4 = -2.3$ ppm). This is analogous to the procedure used in computing the P_{α} values. The third and fourth structures have the same α atom type and were averaged together to calculate the P_{β} value for that point in the plot.

A linear model was built with the P_{α} values as the independent variable and the P_{β} values as the dependent variable. The resulting slope, M_{β} , was 1.05, and the y-intercept, C_{β} , was 0.64. The correlation coefficient for the regression calculation was 0.977. This approach was repeated to determine the β parameters for each of the atom types for which α parameters were calculated.

Table 3 lists all of the P_{α} , M_{β} , and C_{β} parameters currently used for encoding carbon environments. The first column is a consecutive numbering of each of the 66 atom types in the table. The second column lists the class types as described in Table 1. The third column lists the bond type where 1 = single bond, 2 = double bond, 3 = triple bond, and 4 = aromatic bond. The fourth column gives the connectivity, which is defined as the number of non-hydrogen atoms bonded to the atom of interest, and the P_{α} values are listed in the sixth column.

The fifth column has entries for atom classes in which the P_{α} parameters were used from other atom classes. There were some classes for which it was difficult to choose a good compound for developing P_{α} values. Class 13, for example, is a new class added for carbonyl carbons and raises some problems. The carbon in a carbonyl group has already been used by class 26 for calculating the value for a carbonyl oxygen. Another choice for the carbonyl carbon might be the first carbon in C=C=O, but it is difficult to find

Table 2. Example Calculation of M_β and C_β

structure	α compd	β shift (ppm)	ref shift (ppm)	P_β (ppm)	P_α (ppm)
<chem>CH3CH2SCH3</chem>	sp ³ , secondary, C	14.40	-2.3	16.70	19.50
<chem>CH3CH(CH3)SCH2CH3</chem>	sp ³ , tertiary, C	23.50	-2.3	25.80	26.90
<chem>CH3C(CH3)2SCH2CH2CH3</chem>	sp ³ , quaternary, C	32.10	-2.3	34.40	33.70
<chem>CH3C(CH3)2SC(CH3)3</chem>	sp ³ , quaternary, C	33.20	-2.3	35.50	33.70
<chem>CH3SCCCH2CH2CH3</chem>	sp, secondary, C	92.70	-5.7	87.00	69.20
<chem>C6H5SCH3</chem> ^a	arom, tertiary, C	126.70	-27.5	99.20	101.60
<chem>CH3SSCH3</chem>	sp ³ , secondary, S	22.20	-2.3	24.50	20.30

^a Carbon used is a ring carbon *ortho* to the S substituted carbon.

experimental data for such compounds. Experiments were tried in which a value was arbitrarily chosen for P_α as 10.0 ppm lower than the value for class 26, but it was found that better results were obtained when the parameters for class 11 (C=C) were used. The basic philosophy behind the parameter-based encoding algorithm is that the parameters developed must differentiate between various atom classes as well as possible, and thus the relative values of the parameters are more important than the absolute values. Using values based on experimental data helps to ensure that the relative relationships between classes are maintained, but it was felt that the usage of parameters not strictly based on experimental data would be acceptable for those atom classes for which reasonable experimental data were unavailable. In this instance, and for classes 14–17, values obtained from C=C parameters were used, and these classes in essence can be considered as degenerating to classes 10–12 for their values. The benefit of these extra classes will be seen in the following section which describes the special treatment of aromatic carbons.

The seventh column lists the compound used as a reference compound in computing the P_α values, while columns 8 and 9 list the calculated M_β and C_β values, respectively. Column 10 gives the number of structures used to build the linear models that served as the basis for the computed M_β and C_β values. For some of the atom classes, it was difficult to find many structures for use in building the models. In some cases, such as atom class 10, only one structure (and consequently one value) was found. In this case, the origin was chosen as a second value, and thus the y-intercept was defined as 0.0. In other cases, it was impossible to have substituents on the P_α carbon, and values of 1.0 and 0.0 were used for M_β and C_β , respectively. This was the case for classes 1, 4, and 18, among others.

The 11th column is a subjective rating of the quality of the structures chosen to build the model from which the M_β and C_β values were calculated and judges primarily the range of the independent variable. For this designation, 1 = good, 2 = fair, 3 = poor, and 4 = very poor. For example, in computing M_β and C_β for atom class 31, four structures were used to build the model, but the four structures did not cover the x range very evenly, and the quality of the model is therefore judged to be poor. Four data points were also used to build the model for atom class 36, but these points covered the x range well, and the quality of the model is thus judged to be good.

Column 12, labeled R , gives the linear correlation coefficient describing the regression, and column 13 gives the standard error (in ppm) between the predicted and observed P_β values. Column 14 has entries for atom classes in which the β parameters were used from other atom classes. The

last column lists the literature sources for the data used in generating the α and β parameters.

Aromatic Encoding Scheme. The environmental encoding procedure outlined above utilizes information for the calculation of a Z_j for only those atoms which are directly attached to the atom being considered (eq 4). This approach has been found to give less than optimal results for aromatic carbons. Each carbon atom in an aromatic ring and each of the substituents on the aromatic ring can affect the chemical shifts of all other carbons in the ring. Thus, the method of only incorporating effects of directly attached atoms in the calculation of Z_j in eq 4 is inadequate for aromatic carbons and necessitates the development of a modified encoding scheme for these atoms. This scheme has been limited to aromatic rings where the definition of aromaticity is a six-membered carbon ring composed of alternating single and double bonds. Future improvements could include an extension of this approach to other aromatic systems such as furans, thiophenes, pyrroles, and pyridines.

The aromatic encoding algorithm will be explained through the use of benzoyl chloride as an example (Table 4). Consider the calculation of the Z_j value for atom 1. As described by eq 4, the previous approach would be a root-mean-square average of the P_β values for atoms 2 and 6. The new approach is a root-mean-square average of a P_β value for each of the six atoms in the ring. These P_β values are computed with a new set of aromatic P_α , M_β , and C_β values derived from experimental data for the effect of a given substituent at a certain position (directly attached, *ortho*, *meta*, *para*) from atom 1.

The aromatic P_α parameters have been developed in a manner analogous to the development of the nonaromatic parameters. For example, the P_α for a directly attached Cl is simply the difference in chemical shifts between the carbon in chlorobenzene with the attached Cl and a carbon in cyclohexane (27.5 ppm). Separate parameters are also calculated for a Cl substituted in the *ortho*, *meta*, and *para* positions. For example, the P_α value for a *meta* substituted Cl is computed as the difference in chemical shifts between the carbon in chlorobenzene *meta* to the Cl and a cyclohexane carbon. The complete set of aromatic P_α parameters is listed in Table 5. The table entries are specified by atom class (column 2), bond type (column 3), and connectivity (column 4) in a manner analogous to Table 3. P_α parameters are specified for each case of attachment (directly attached, *ortho*, *meta*, *para*) in columns 5–8. The last column in the table specifies the sequence (id) numbers in the table of degenerate parameters. For example, parameter 37 is degenerate with parameter 4. The structures used to generate the parameters have been taken entirely from the Sadler database. It will be noted that classes 11, 12, 16, 18, 24,

Table 3. Computed P_α , M_β , and C_β Parameters

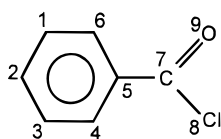
id	class	bond	conn	P_α degn	P_α	ref ^a	M_β	C_β	n	range	R	s	P_β degn	lit.
1	1	1	2		-9.5	2	1	0	0					36
2	1	1	3		-1.4	2	0.965	3.21	6	1	0.996	4.433		36
3	1	1	4		5.6	2	1.07	-7.6	3	2/3	0.999	0.6813		36
4	2	1	1		8.0	1	1	0	0					17
5	2	1	2		19.5	1	1.021	-3.41	9	1	0.999	1.091		15, 17, 18, 21, 23, 27-29
6	2	1	3		26.9	1	1.039	-6.076	10	1	0.999	1.808		15, 17, 20, 23, 29-31
7	2	1	4		33.7	1	1.049	-9.78	9	1	0.998	2.425		15, 17, 20, 23, 28-30
8	3	2	2		109.0	2	1	0	0					36
9	3	2	3		107.3	2	1	0	0					36
10	4	2	1		124.4	1	0.649	0	1	4				18, 32
11	4	2	2		117.0	1	1.048	-3.12	4	1	0.990	6.771		18, 21, 33-35
12	4	2	3		111.0	1	1.028	-5.82	3	1/2	0.999	1.450		20, 34-36
13	5	2	2	11	117.0	1	1.048	-3.12	4	1	0.990	6.771	11	18, 21, 33-35
14	5	2	3	12	111.0	1	1.028	-5.82	3	1/2	0.999	1.450	12	20, 34-36
15	6	2	1	10	124.4	1	0.649	0	1	4			10	18, 32
16	6	2	2	11	114.0	1	1.048	-3.12	4	1	0.990	6.771	11	18, 21, 33-35
17	6	2	3	12	111.0	1	1.028	-5.82	3	1/2	0.999	1.450	12	20, 34-36
18	7	3	1		41.2	1	1	0	0					16
19	7	3	2		69.2	1	1.068	-4.36	4	1	0.999	0.949		19, 22, 30, 32, 36
20	8	3	2	19	69.2	1	1.068	-4.36	4	1	0.999	0.949	19	19, 22, 30, 32, 36
21	9	4	2		101.0	3	1	0	3	3				16, 21
22	9	4	3		101.6	3	1.033	-3.2	11	1	0.992	5.031		21, 31, 35, 36, 38, 39
23	10	1	1		52.2	1	0.813	5.49	4	2	0.999	1.114		16, 21, 37
24	10	1	2		61.5	1	0.733	4.013	6	1/2	0.992	4.823		17, 21, 29, 34, 40
25	11	4	2		116.6	4	0.827	0	2	3				36
26	12	2	1		194.0	2	0.631	-44	2	4				21, 41
27	13	1	1		30.4	1	0.798	9.13	6	1	0.987	4.300		23, 42, 43
28	13	1	2		40.5	1	0.816	3.81	6	1	0.994	3.031		21, 23, 36, 43
29	13	1	3		49.4	1	0.865	-1.23	6	1	0.998	1.854		23, 36, 39, 43
30	13	1	4		59.1	2	0.948	-1.856	3	1/2	0.999	1.432		36
31	14	4	2		91.7	4	0.733	6.44	4	3	0.992	2.767		36
32	14	4	3		95.4	4	0.829	4.979	7	1	0.978	7.76		36
33	15	2	3		65.3	1	0.946	-1.073	6	1	0.994	4.518		36
34	16	4	2		122.3	3	0.946	0	2	3				21, 36
35	17	2	1		167.4	1	0.985	0	2	3				36
36	17	2	2		152.1	5	0.899	0.937	4	1	0.996	3.305		36
37	18	3	1		112.0	2	1	0	0					36
38	19	1	1		72.3	2	0.853	-1.026	2	3				17, 36, 42
39	20	1	1		34.8	2	0.988	2.76	6	1	0.999	1.457		19, 21, 24-26
40	21	1	1		21.7	2	1.013	3.11	5	1	0.999	1.996		19, 21, 44-46
41	22	1	1		-6.5	2	1.069	4.25	5	1	0.999	2.978		19, 21, 46, 47
42	23	1	1		13.4	2	0.965	3.921	4	1	0.999	0.809		36
43	23	1	2		20.3	1	1.048	0.636	6	2	0.977	8.519		36
44	24	4	2		98.8	4	0.984	0	2	3				36
45	25	2	1		251.3	2	0.686	-34	1	4				36
46	25	2	3		39.2	2	1.06	-11.62	5	1	0.999	1.321		36
47	25	2	4		40.5	2	1.227	-19.1	6	1	0.993	6.051		36
48	26	1	3		18.6	1	1.154	-14.56	4	1	0.993	5.429		36
49	27	2	4		18.3	1/2	0.851	0.732	4	1	0.981	3.691		36
50	28	1	2		21.6	2	1.021	-3.41	9	1	0.999	1.091	5	15, 18, 21, 23, 27-29, 36
51	28	1	3		29.8	2	1.039	-6.076	10	1	0.999	1.808	6	15, 20, 23, 29-31, 36
52	28	1	4		34.2	2	1.049	-9.78	9	1	0.998	2.425	7	15, 20, 23, 28-30, 36
53	29	2	2		124.5	2	1.048	-3.12	4	1	0.990	6.771	11	21, 33-36
54	29	2	3		117.0	2	1.028	-5.82	3	1/2	0.999	1.450	12	34, 35, 36
55	30	3	2		88.7	2	1.068	-4.36	4	1	0.999	0.949	19	19, 30, 32, 36
56	31	1	2		64.2	2	0.733	4.013	6	1/2	0.992	4.823	24	21, 29, 34, 36, 40
57	32	1	1		207.5	2	0.631	-44	2	4			26	36, 41
58	33	1	2		42.6	2	0.816	3.81	6	1	0.994	3.031	28	21, 23, 36, 43
59	33	1	3		50.8	2	0.865	-1.23	6	1	0.998	1.854	29	23, 36, 39, 43
60	33	1	4		71.7	2	1	0						36
61	34	2	2		181.3	2	0.899	0.937	4	1	0.996	3.305	36	36
62	34	2	3		128.9	2	1	0						36
63	35	1	2		23.3	2	1.048	0.636	6	2	0.977	8.519	43	36
64	36	2	3		46.0	2	1.06	-11.62	5	1	0.999	1.321	46	36
65	36	2	4		45.9	2	1.227	-19.1	6	1	0.993	6.051	47	36
66	37	2	3	12	111.0	1	1.028	-5.82	3	1/2	0.999	1.450	12	34, 35, 36

^a Reference compound (1 = methane ($\delta = -2.3$), 2 = ethane ($\delta = 5.7$), 3 = cyclohexane ($\delta = 27.5$), 4 = cyclopentane ($\delta = 26.2$), 5 = propane ($\delta = 15.9$)).

and 32 have no P_α parameter since these structural units cannot be substituted directly onto a benzene ring.

It can be seen that the 50 sets of parameters in Table 5 do not contain all of the possible substituents for an aromatic

ring. As an example, consider the parameters with sequence numbers 10 and 11 in Table 5 which correspond to secondary and tertiary carbonyl carbons. These parameters do not distinguish between such entities as aldehydes, ketones,

Table 4. Calculation of Z_j for Atom 1 in Benzoyl Chloride**Old Method: (Only atoms 2 and 6 contribute)**

$P_{\beta,1}$ (atom 6)	P_{α}	M_{β}	C_{β}	P_{β}
	101.0	1.000	0.0	101.0

$P_{\beta,2}$ (atom 2)	P_{α}	M_{β}	C_{β}	P_{β}
	101.0	1.000	0.0	101.0

$$Z_j = ((101.0)^2 + (101.0)^2) / 2)^{1/2} = 101.0$$

New Method: (all 6 atoms contribute)

$P_{\beta,1}$ (atom 5 -- carbonyl carbon meta to atom 1)				
connected atom	P_{α}	M_{β}	C_{β}	P_{β}
8	101.10	1.04	-4.6	100.544
9	101.10	1.00	0.0	101.10
root-mean-square average = 100.822				

$P_{\beta,2} - P_{\beta,6}$ (atoms 1,2,3,4,6)	
no substituents -- default value of 101.0	
root-mean-square average = 101.0	

$$Z_j = ((100.822)^2 + 5 \cdot (101.0)^2) / 6)^{1/2} = 100.97$$

amides, acid chlorides, carboxylic acids, and esters. To obtain the ability to distinguish between these species, the aromatic encoding scheme modifies each P_{α} based on the nonring non-hydrogen atoms directly attached to a given substituent. This modification results in a P_{β} value and is analogous to the procedure used in the nonaromatic algorithm. However, in the nonaromatic algorithm, only one pair of M_{β} and C_{β} values exists for each contributor to Z_j , while in the aromatic scheme, multiple pairs of M_{β} and C_{β} values can exist.

This difference can be seen by studying the structure in Table 4. The calculation of Z_j for atom 1 using the original method includes a contribution for atom 6. There is only one M_{β} and C_{β} pair of values needed for atom 6 since atom 6 is the actual contributor to the Z_j calculation. The calculation of Z_j for atom 1 using the revised method includes a contribution for atom 5. Atom 5, however, is not the actual contributor to the Z_j calculation. The nonring atoms attached to the direct substituent of atom 5 are the actual contributors. The key point is that atom 5 is an aromatic carbon with a carbonyl carbon directly substituted, that atom 5 is in a position *meta* to atom 1 and that the influence of the carbonyl carbon is affected by the atoms attached to it. Since greater than one atom can be bonded to the substituent, greater than one M_{β} and C_{β} pair can be required.

The M_{β} and C_{β} values used in the aromatic approach were calculated by computing regression coefficients for each atom class and relative ring position. For example, if M_{β} and C_{β} values were being computed for Cl attached to a *meta* substituent, substituted benzenes would be assembled with chlorines attached to different *meta* substituents. Table 6 is analogous to Table 2 and illustrates the calculation for this case. The structures labeled β structures contain the sub-

stituents whose effect is being modified by the Cl. The chemical shifts obtained after referencing to a cyclohexane carbon are used as the dependent variable in the regression calculation. The structures labeled α structures do not have chlorine attached to the *meta* substituents, and the resulting chemical shifts are used as the independent variable after referencing to a cyclohexane carbon. The β structures all have a chlorine attached to the *meta* α atom (the *meta* substituent). Column 3 in the table lists the atom type of the α species. Column 4 lists the P_{β} values, and column 5 lists the P_{α} values. A regression analysis of the P_{β} values versus the P_{α} values resulted in a slope, M_{β} , of 1.04, and a y-intercept, C_{β} , of -4.6. The correlation coefficient for the regression calculation was 0.977. The confidence in this regression model is low since the P_{α} values vary so little. This could be expected since substituents in a *meta* position do not have a large influence on the chemical shift. The particular group attached to the α atom also makes a difference in the significance of the influence on the chemical shift.

Table 7 displays the computed M_{β} and C_{β} parameters for all possible atom classes. The atom class, bonding type, and connectivity associated with each set of parameters are listed in the table in a manner analogous to Table 5. As in Table 5, the last column of the table also lists the sequence (id) numbers of any degenerate parameters. It should be noted that there are more parameter sets in Table 7 than in Table 5 since some atom species which could not be directly attached to an aromatic ring, such as a secondary sp^3 cyclopropane carbon or a primary sp^2 carbon, can be attached to a substituent of an aromatic ring.

The use of the M_{β} and C_{β} parameters can be illustrated by returning to the benzoyl chloride example in Table 4. The calculation of Z_j for atom 1 is based on the P_{β} contributions from each atom in the ring. Atom 5 in benzoyl chloride is *meta* to atom 1 and has as its substituent a carbonyl carbon of connectivity 3 (atom class 5). The P_{α} value for a class 5 atom *meta* to a ring carbon is 101.10 (Table 5). The rest of the carbons in the ring have no substituents other than hydrogen, and, as in the original encoding algorithm, the effects of hydrogens are not directly treated. The ring carbons without substituents are assigned the default value of 101.0 for the root-mean-square calculation. The carbonyl carbon substituent on atom 5 has two nonring non-hydrogen atoms attached—chlorine and carbonyl oxygen. The M_{β} and C_{β} values for Cl in a *meta* position are 1.04 and -4.6, respectively. Thus, the P_{β} value for atom 5 for the Cl substituent is

$$P_{\beta} = C_{\beta} + M_{\beta} P_{\alpha} = -4.6 + 1.04 \cdot 101.10 = 100.544 \quad (7)$$

As shown in Table 4, the P_{β} value for atom 5 for the doubly bonded oxygen substituent is 101.10 since M_{β} is 1 and C_{β} is 0 for this case. The root-mean-square average of these two modified P_{β} values is 100.822. The Z_j value for atom 1 can now be calculated using the six P_{β} values, resulting in the value 100.970.

Expanded Vector Representation. A third improvement to the environmental encoding algorithm has been the use of an expanded vector representation. In previous work, either a root sum-of-squares (RSS) summation or a linear summation was used for the calculation of the numerator in

Table 5. Aromatic P_{α} Parameters

id	class	bond	conn	direct	ortho	meta	para	degn
1	1	1	3	116.4	98.2	100.8	97.9	
2	1	1	4	113.5	101.2	100.9	99.2	
3	2	1	1	110.2	102.5	101.7	98.1	
4	2	1	2	116.7	100.4	100.9	98.2	
5	2	1	3	121.3	98.9	100.9	98.3	
6	2	1	4	123.3	97.6	100.5	97.8	
7	3	2	3	113.9	98.1	100.7	99.9	
8	4	2	2	110.2	98.7	101.0	100.3	
9	4	2	3	113.9	98.1	100.7	99.9	
10	5	2	2	109.1	102.2	101.5	106.9	
11	5	2	3	109.8	100.8	101.1	105.5	
12	6	2	2	109.0	100.5	101.0	102.9	
13	6	2	3	113.9	98.1	100.7	99.9	
14	7	3	2	94.9	104.7	100.8	101.2	
15	8	3	2	84.9	104.6	101.7	105.3	
16	9	4	2	106.1	100.3	98.3	98.3	
17	9	4	3	105.2	98.9	97.9	97.9	
18	10	1	1	127.4	87.9	102.2	93.5	
19	10	1	2	132.4	86.6	102.0	93.2	
20	13	1	1	119.2	87.6	101.7	90.7	
21	13	1	2	122.1	84.9	101.7	89.5	
22	13	1	3	123.2	85.2	101.5	89.1	
23	13	1	4	123.2	85.2	101.5	89.1	
24	14	4	3	113.2	92.7	101.9	97.9	
25	15	2	3	120.80	96.0	102.0	107.3	
26	17	2	2	124.6	93.3	101.2	98.3	
27	19	1	1	135.8	88.0	102.7	96.7	
28	20	1	1	106.9	101.1	102.2	98.9	
29	21	1	1	95.0	103.9	102.3	99.2	
30	22	1	1	66.8	109.6	102.3	99.5	
31	23	1	1	103.2	101.8	101.4	97.9	
32	23	1	2	111.1	99.2	101.2	97.4	
33	25	2	3	118.7	95.9	101.6	103.2	
34	25	2	4	113.1	99.7	101.8	106.1	
35	26	1	3	111.1	104.7	100.7	100.9	
36	27	2	4	106.6	103.0	101.5	103.7	
37	28	1	2	116.7	100.4	100.9	98.2	4
38	28	1	3	121.3	98.9	100.9	98.3	5
39	28	1	4	123.3	97.6	100.5	97.8	6
40	29	2	2	110.2	98.7	101.0	100.3	8
41	29	2	3	113.9	98.1	100.7	99.9	9
42	30	3	2	94.9	104.7	100.8	101.2	14
43	31	1	2	132.4	86.6	102.0	93.2	19
44	33	1	2	122.1	84.9	101.7	89.5	21
45	33	1	3	123.2	85.2	101.5	89.1	22
46	34	2	2	124.6	93.3	101.2	98.3	26
47	35	1	2	111.1	99.2	101.2	97.4	32
48	36	2	3	118.7	95.9	101.6	103.2	33
49	36	2	4	113.1	99.7	101.8	106.1	34
50	37	2	3	113.9	98.1	100.7	99.9	

eq 3. Often, there are cases in which two carbon atoms appear to have similar chemical environments based on the RSS calculation, but the linear calculation indicates dissimilarities. The opposite case of a similar linear sum but a different RSS value also occurs. If the vectors resulting from both summations are combined into a single expanded vector, it is much more difficult for these false indications of similarity to arise. The first dimension, e_0 , has only one contributor and therefore is the same for both the RSS and linear calculations and needs only to be used once in the expanded vector. The second through n th dimensions are the vector components 2 through n from the RSS summation, and the remaining dimensions are the vector components 2 through n from the linear summation. The format of the expanded vector is thus $e_0, e_1, \dots, e_{2n+1}$ as defined by eqs 2 and 3 (RSS calculation), with n additional elements appended defined by $e_{1,1}, \dots, e_{1,k}$, where

$$e_{1,i} = \frac{\sum_{j=1}^{P_i} Z_j}{d^r} \quad (8)$$

and P_i , Z_j , and d^r are as defined in eq 3. The dimensionality of the expanded vector is $2n+1$, where n specifies a bond distance removed from the carbon whose environment is being described.

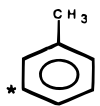
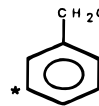
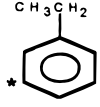
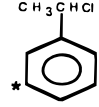
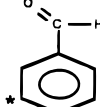
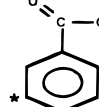
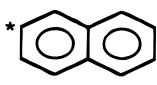
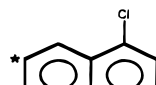
The atoms indicated by the stars in the two structures 4-picoline and 4-cyclopropylpyridine in Figure 1 were determined to be similar by the RSS calculation and somewhat dissimilar by the linear calculation. Each non-hydrogen atom in the structures is labeled with its Z_j value. Table 8 shows the result of combining these Z_j values into e_i values using both the RSS and linear calculations. The differences between the elements of the vectors formed by the RSS and the linear calculations are noted along with the overall Euclidean distance between the two pairs of vectors. As indicated by the Euclidean distance of 0.299, the vectors formed by the RSS calculation are quite similar. The presence of the cyclopropyl carbons does not greatly affect the e_i values computed by the RSS calculations due to the large Z_j values corresponding to the sp^2 hybridized ring carbons. The squaring of the Z_j term in the RSS calculation weights the resulting e_i value toward large values of Z_j . In contrast, large values of Z_j are less heavily weighted in the linear summation. The cyclopropyl carbons thus have greater influence on the e_i values computed with the linear sum. The larger Euclidean distance of 6.827 between the vectors confirms this result. For certain chemical environments, however, the RSS calculation yields a more quantitatively accurate vector representation than the linear sum. For this reason, it was hypothesized that the combined vector approach would form the best overall compromise between the two representations. In the present example, the vector formed by combining the RSS and linear vectors produces a Euclidean distance of 6.833. Results will be presented below which compare the single and combined vectors in terms of their accuracy in retrieving ^{13}C NMR chemical shifts.

Optimization of Environmental Encoding Parameters.

A number of experiments were performed in which the parameters in the environmental encoding algorithm were varied. The first step in each experiment consisted of calculating the environmental vectors for each of the carbon atoms in the library set and test set. Duplicate vectors in each set were removed. This ensured that no one environmental vector was weighted more heavily in the library than any other vector. In the library set, all occurrences of vectors which differed in each dimension by 10^{-6} or less were represented by only one vector. In the test set, a check for duplicate vectors was only made within structures. Thus, every unique vector in every structure in the test set was represented. The number of vectors remaining after removing duplicates depended on the dimensionality of the vector and the exponent of the weighting factor (r in eq 3) and varied from approximately 37 000 to 39 000 for the test sets and 76 000 to 142 000 for the library sets.

The second step involved finding the five closest matching environments in the library set for each vector in the test set. A Euclidean distance calculation was performed for each

Table 6. Example Calculation of M_β and C_β for Aromatic Compounds

α structure	β structure	α atom	P_β	P_α
		sp ³ , secondary, C	101.1	101.7
		sp ³ , tertiary, C	101.1	100.9
		sp ² , tertiary, carbonyl C	101.5	101.5
		tertiary, aromatic C	97.9	98.3

test vector with all of the vectors in the library set. The five library vectors with the smallest Euclidean distances were taken as the five closest matches. The predicted chemical shift for a carbon in the test set was assigned as the chemical shift of the carbon in the library set having the closest matching environmental vector. The difference between the predicted shift and the corresponding actual shift (residual) was recorded for each carbon in the test set, and the mean absolute residual was calculated for each structure in the test set. An overall mean of the mean absolute residuals was computed and used to gauge the relative success of different combinations of parameter settings for the environmental encoding algorithm. The median value of the mean absolute residuals was also computed for comparison.

Figure 2 shows the results of a study in which the exponent of the weighting factor (r in eq 3) was varied from one to five for the RSS calculation for single environmental vectors ranging from two to eight dimensions. The best value obtained for the mean absolute residual was 2.26 ppm for the case in which an exponent of four was used along with an eight-dimensional vector. These conditions were subsequently used to judge the utility of the expanded atom classes and the aromatic encoding scheme.

The optimal parameter values found in this experiment appear to represent the best compromise with respect to including and giving appropriate weight to environmental effects remote from the carbon atom of interest. The fact that only a slight improvement in performance is noted beyond a dimensionality of five (i.e., inclusion of environmental effects four bonds or less from the target carbon) is compatible with the conventional view that environmental effects on chemical shifts are rarely observed over more than four bonds.

By way of comparison, for the version of the algorithm which only used similar atom classes as in the 1984 paper and did not use the aromatic code (version 1), an overall mean of mean absolute residuals of 2.37 was obtained for the eight-dimensional vector with the RSS calculation and an exponent of four for the weighting factor. A second version of the software in which the aromatic code was used, but the cyclic atom classes were not added (version 2), yielded an overall mean of mean absolute residuals of 2.28.

Figure 3 compares the results for the RSS, linear, and combined RSS/linear calculations. A weighting exponent of four was used in each case, and the dimensionality of the single vectors was varied from two to eleven. The corresponding dimensionality of the combined vector was $2n-1$, where n was the dimensionality of the single vector. This produced dimensionalities which ranged from 3 to 21. The best value for the mean of mean absolute residuals was 2.08 ppm and was obtained for the combined vector with a dimensionality of 15.

Figure 4 displays a histogram of the absolute residuals calculated from the 38 568 vectors in the test set for the case in which the combined vector was used with an exponent of four for the weighting factor and a dimensionality of 15. It can be seen that approximately 62% of the carbons have shifts predicted correctly within 1.0 ppm. Similarly, 75.6%, 82.9%, and 87.4% of the shifts are predicted correctly within 2.0, 3.0, and 4.0 ppm, respectively. For this histogram, the count of the number of vectors falling within a given residual range was based solely on the single nearest match to each test vector.

Analysis of the residuals reveals that a relatively small percentage of badly predicted shifts is skewing the overall mean of mean absolute residuals. This is particularly true since all test vectors were used in the calculations, regardless of how bad a match might be. For example, if the nearest matching library vector had a large Euclidean distance, the predicted shift and corresponding residual were still included in the calculation, despite the fact that the large distance indicates the match is poor. Figures 5 and 6 are analogous to Figures 2 and 3, with the exception that they use the median value of the mean absolute residuals rather than the overall mean. The median values are significantly lower than the means as they are more resistant to the influence of large outlying residuals. The best result for the RSS calculation had a median value of 1.37 ppm for a vector dimensionality of eight and a weighting exponent of four. The best result for the combined RSS/linear vector was 1.30 ppm for a vector dimensionality of 15 and a weighting exponent of four. Revisiting the old versions of the software for the median of the mean residuals, a value of 1.57 ppm was obtained for version 1 and a value of 1.42 ppm was obtained for version 2. When compared to the current value of 1.37 ppm, it can

Table 7. Aromatic M_β and C_β Parameters

id	class	bond	conn	direct		ortho		meta		para		degn
				M_β	C_β	M_β	C_β	M_β	C_β	M_β	C_β	
1	1	1	2	1.0	0	1.0	0	1.0	0	1.0	0	
2	1	1	3	1.09	-7.2	0.987	0.007	0.993	0.003	0.995	0.11	
3	1	1	4	1.09	-7.2	0.987	0.007	0.993	0.003	0.995	0.11	
4	2	1	1	1.02	-0.17	0.853	13.2	0.957	3.9	0.850	14.7	
5	2	1	2	1.10	-9.7	0.991	0.04	0.993	-0.04	0.823	17.3	
6	2	1	3	1.02	-0.02	0.985	0.001	0.996	-0.001	0.991	0.07	
7	2	1	4	1.02	-0.02	0.985	0.001	0.996	-0.001	0.991	0.07	
8	3	2	2	1.0	0	1.0	0	1.0	0	1.0	0	
9	3	2	3	1.02	0	0.980	0	0.997	0	1.005	0	
10	4	2	1	1.0	0	1.0	0	1.0	0	1.0	0	
11	4	2	2	1.02	-0.004	0.986	0.001	0.994	0	0.991	0.1	
12	4	2	3	1.02	0	0.980	0	0.997	0	1.01	0	
13	5	2	2	0.712	27.2	1.02	0.03	1.00	-0.029	1.02	0.001	
14	5	2	3	0.990	0.08	0.997	0.005	0.990	-0.019	1.013	0	
15	6	2	1	1.0	0	1.0	0	1.0	0	1.0	0	
16	6	2	2	0.712	27.2	1.02	0.03	1.0	-0.029	1.02	0.001	13
17	6	2	3	0.990	0	0.997	0	0.990	0	1.013	0	
18	7	3	1	1.0	0	1.0	0	1.0	0	1.0	0	
19	7	3	2	0.937	-0.015	0.995	0.011	1.01	0.08	1.03	-0.001	
20	8	3	2	0.937	-0.015	0.995	0.011	1.01	0.08	1.03	-0.001	19
21	9	4	2	1.0	0	1.0	0	1.0	0	1.0	0	
22	9	4	3	0.984	3.8	0.997	0.21	0.996	-0.002	0.746	24.9	
23	10	1	1	1.0	-0.28	0.983	-0.02	0.995	0.003	0.795	21.3	
24	10	1	2	0.946	-0.27	0.988	0	1.01	-1.77	1.0	0.15	
25	11	4	2	0.946	-0.27	0.988	0	1.01	-1.77	1.0	0.15	24
26	12	2	1	1.0	0	1.0	0	1.0	0	1.0	0	
27	13	1	1	0.977	0.04	0.989	0.10	0.983	-0.02	0.996	0.26	
28	13	1	2	1.0	-0.02	0.978	-0.001	0.992	0	0.988	0.18	
29	13	1	3	0.995	-0.2	0.987	0.05	0.991	-0.017	0.986	0.36	
30	13	1	4	0.995	-0.2	0.987	0.05	0.991	-0.017	0.986	0.36	29
31	14	4	2	1.0	-0.02	0.978	-0.001	0.992	0	0.988	0.18	28
32	14	4	3	1.004	0	0.970	0	0.994	0	1.019	0	
33	15	2	3	0.928	-0.04	1.01	0	0.998	-0.06	1.04	0	
34	16	4	2	0.983	0	0.998	0	1.0	0	1.018	0	
35	17	2	1	1.0	0	1.0	0	1.0	0	1.0	0	
36	17	2	2	1.0	0	1.0	0	1.0	0	1.0	0	
37	18	3	1	1.0	0	1.0	0	1.0	0	1.0	0	
38	19	1	1	0.954	-0.15	0.954	-0.05	0.998	0.015	1.031	0	
39	20	1	1	1.08	-10.3	0.995	-0.02	1.04	-4.6	0.823	19.9	
40	21	1	1	1.08	-9.6	1.0	0	0.888	11.1	0.902	11.6	
41	22	1	1	1.01	-0.020	1.04	0.12	1.0	0.024	1.02	0	
42	23	1	1	1.01	0	0.982	0.004	0.997	0.003	0.806	20.3	
43	23	1	2	1.01	0	0.980	-0.001	0.996	0	0.988	-0.11	
44	24	4	2	1.0	0	1.0	0	1.0	0	1.0	0	
45	25	2	1	1.0	0	1.0	0	1.0	0	1.0	0	
46	25	2	3	1.0	0	1.0	0	1.0	0	1.0	0	
47	25	2	4	0.916	0	0.990	0	1.014	0	1.035	0	
48	26	1	3	1.0	0	1.0	0	1.0	0	1.0	0	
49	27	2	4	1.0	0	1.0	0	1.0	0	1.0	0	
50	28	1	2	1.10	-9.7	0.991	0.04	0.993	-0.04	0.823	17.3	
51	28	1	3	1.02	-0.02	0.985	0.001	0.996	-0.001	0.991	0.07	
52	28	1	4	1.02	-0.02	0.985	0.001	0.996	-0.001	0.991	0.07	51
53	29	2	2	1.02	-0.004	0.986	0.001	0.994	0	0.991	0.1	
54	29	2	3	1.02	0	0.980	0	0.997	0	0.991	0	
55	30	3	2	0.937	-0.015	0.995	0.011	1.01	0.08	1.03	-0.001	
56	31	1	2	0.946	-0.27	0.988	0	1.01	-1.77	1.0	0.15	
57	32	1	1	1.0	0	1.0	0	1.0	0	1.0	0	
58	33	1	2	1.0	-0.02	0.978	-0.001	0.992	0	0.988	0.18	
59	33	1	3	0.995	-0.02	0.987	0.05	0.991	-0.017	0.986	0.36	
60	34	2	2	1.0	0	1.0	0	1.0	0	1.0	0	
61	35	1	2	1.01	0	0.980	-0.001	0.996	0	0.998	-0.11	
62	36	2	3	1.0	0	1.0	0	1.0	0	1.0	0	
63	36	2	4	0.916	0	0.990	0	1.014	0	1.035	0	
64	37	2	3	1.02	0	0.98	0	0.997	0	1.01	0	

be seen that the addition of new atom classes and especially the addition of the aromatic code have been useful revisions of the environmental encoding algorithm.

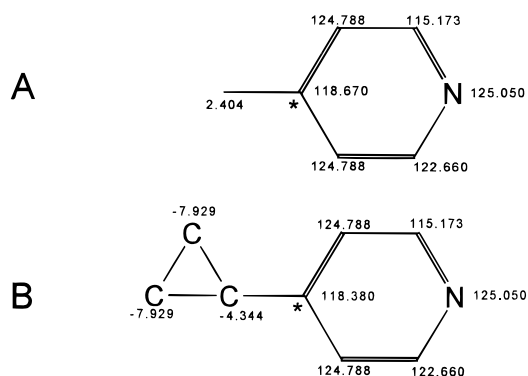
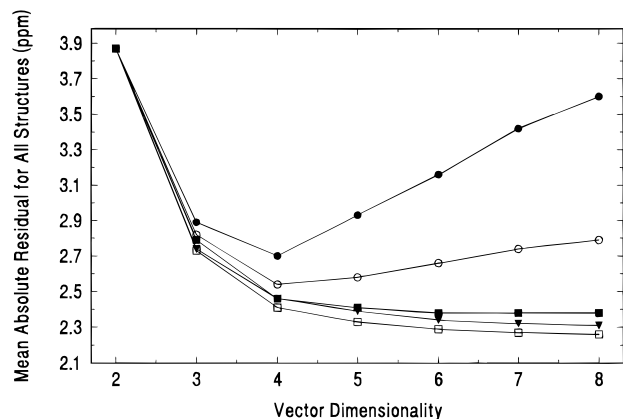
Limitations of the Environmental Encoding Algorithm.

There are several possible reasons for carbons with badly predicted shifts (large residuals). For those carbons whose nearest matching vector had a large Euclidean distance, there

were simply no vectors in the library set with close matches. Large Euclidean distances arise either because there truly is no carbon in the library set with a similar environment, or because of a deficiency in the environmental encoding algorithm that does not encode that environment in a manner that allows the carbon with the similar environment to be discovered.

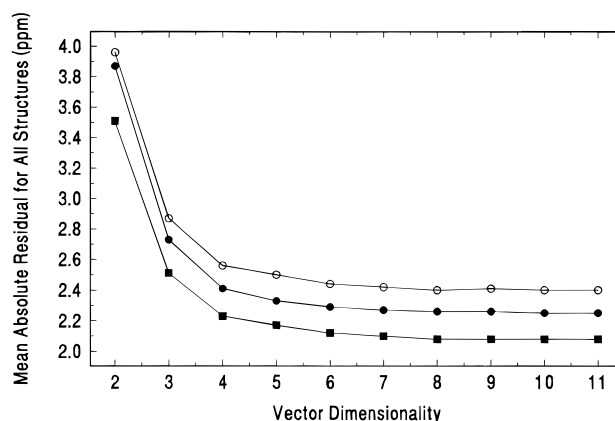
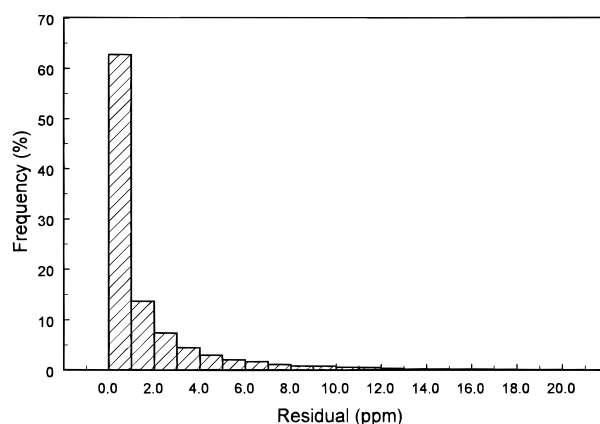
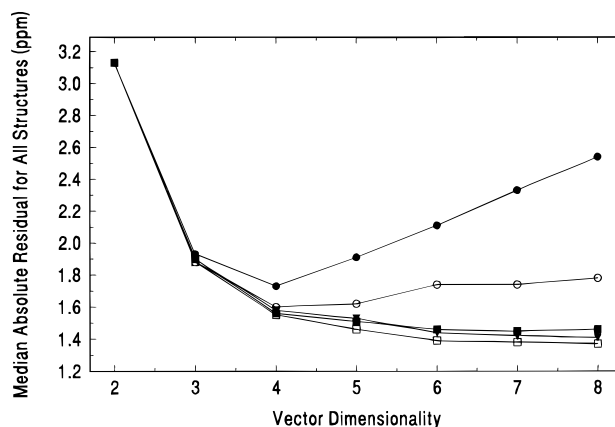
Table 8. Comparison of Environment Vectors

dimension	RSS calculation			linear calculation		
	Figure 1A	Figure 1B	difference	Figure 1A	Figure 1B	difference
0	118.670	118.380	0.290	118.670	118.380	0.290
1	176.49	176.42	0.070	251.980	245.232	6.748
2	10.516	10.493	0.023	14.865	13.874	0.991
3	1.544	1.544	0.0	1.544	1.544	0.0

**Figure 1.** Z_j values are displayed for the atoms in 4-picoline (A) and 4-cyclopropylpyridine (B).**Figure 2.** Plot of mean absolute residuals for the 4240 structures in the test set vs vector dimensionality for each of five different exponents in the weighting factor. Solid circles, open circles, solid squares, open squares, and solid triangles are used for exponents of 1, 2, 3, 4, and 5, respectively.

A second reason for the occurrence of large residuals can be errors or inadequacies in the experimental data or in its usage. A key assumption of the work presented in this paper is that the experimental data in the database are completely accurate. This assumption does not hold in 100% of the cases. A number of carbons with apparently incorrectly assigned shifts have been found and removed from usage. The overall percentage of such errors is small, and it is believed that the majority of such errors giving rise to residuals greater than 10 ppm have been found. A number of other factors such as the solvent used are not taken into consideration in the work presented here and could affect the results obtained.

It has been determined through visual inspection that matches whose Euclidean distances are less than 0.1 generally give very accurate shift predictions. Those matches whose distances range from 0.1 to 1.0 often give good results, and those matches with distances greater than 1.0 typically give poor results. The distribution of the Euclidean distances and residuals is detailed in Table 9 for the same experiment which was described in Figure 4 (a combined vector with a dimensionality of 15 and a weighting exponent of 4).

**Figure 3.** Plot of mean absolute residuals for the 4240 structures in the test set vs vector dimensionality for three different vector types, all of which used a weighting exponent of four. Open circles, solid circles, and solid squares denote linear vectors, RSS vectors, and combined RSS/linear vectors, respectively.**Figure 4.** Histogram of the absolute residuals is displayed for the 38568 vectors in the test set for the case in which a RSS/linear combined vector was used with an exponent of four for the weighting factor and a dimensionality of 15.**Figure 5.** Plot of median absolute residuals for the 4240 structures in the test set vs vector dimensionality for each of five different exponents in the weighting factor. Solid circles, open circles, solid squares, open squares, and solid triangles are used for exponents of 1, 2, 3, 4, and 5, respectively.

Column 2 of Table 9 contains the percentage of vectors from the entire test set which fall within each distance range. Column 3 shows the percentage of vectors from the test set with a residual of greater than 2.0 ppm for the top match and which falls within each distance range. The results in Table 9 confirm the fact that the occurrence of residuals of greater than 2.0 ppm is rare if the vector match is good.

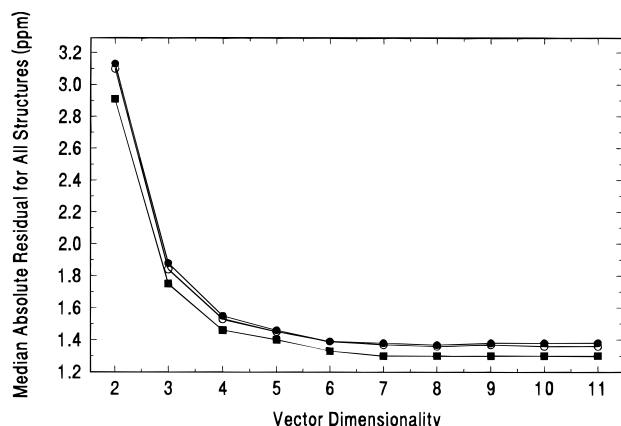


Figure 6. Plot of median absolute residuals for the 4240 structures in the test set vs vector dimensionality for three different vector types, all of which used a weighting exponent of four. Open circles, solid circles, and solid squares are used for linear vectors, RSS vectors, and combined RSS/linear vectors, respectively.

Table 9. Distribution of Euclidean Distances

Euclidean distance	percentage of vectors	
	any residual	>2 ppm residual
<0.001	7170 (18.6%)	561 (6.3%)
<0.01	10 576 (27.4%)	749 (8.4%)
<0.1	20 218 (52.4%)	1737 (19.4%)
<1.0	34 180 (88.6%)	6004 (66.9%)
<10.0	38 544 (99.9%)	8949 (99.8%)
≥0.0	38 568	8969

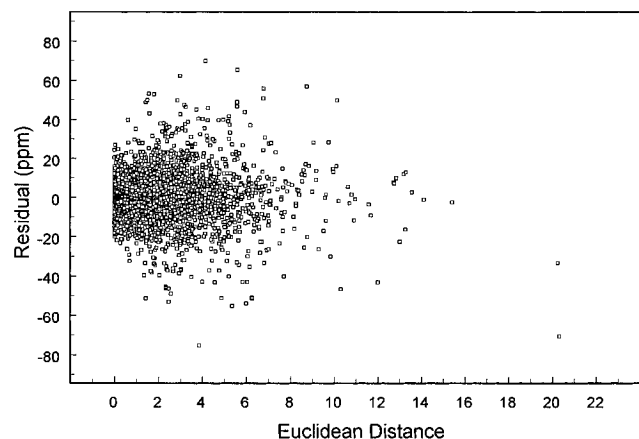
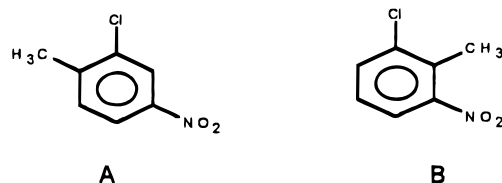


Figure 7. Scatter plot of residual value vs Euclidean distance for the 38 568 vectors in the test set for the case in which a RSS/linear combined vector was used with an exponent of four for the weighting factor and a dimensionality of 15.

Figure 7 displays a scatter plot of the residual values versus Euclidean distances of the vectors associated with the residuals for the experiment described in Table 9. It can be seen that there are some carbons with large residuals, even at small Euclidean distances. There are 55 carbons with residuals greater than 10.0 ppm for a Euclidean distance of 0.1 or less. This corresponds to only 0.27% of the 20 218 carbons in the test set meeting this Euclidean distance cutoff. There are only four carbons with residuals greater than 20.0 ppm for the same Euclidean distance. In contrast, of the 4388 carbons with Euclidean distances greater than 1.0, 860 (19.6%) of them have residuals greater than 10.0 ppm. Of the 55 carbons in Figure 7 which had large residuals and small Euclidean distances, 44 of these are most likely due to incorrectly assigned shifts, and the remaining 11 carbons

indicate possible problems with the encoding algorithm.

No encoding algorithm can perfectly correlate the chemical environment with the chemical shift, and thus further enhancements can always be made. For example, structures A and B below illustrate a problem with the current aromatic encoding scheme. In both compounds, the carbon substituted



by a chlorine has a methyl group in the *ortho* position and a nitro group in the *meta* position. The structures are different, however, and the chemical shifts are affected by the differences. Since the current encoding scheme for aromatic systems does not distinguish between the two positions relative to a carbon which are *ortho* or *meta*, the differences between structures A and B are not encoded.

The current algorithm also does not take *cis/trans* configurations into consideration. This is significant, as substantial shift differences are observed between shifts in a compound with *cis* vs *trans* configurations of a double bond. The need to treat other aromatic systems in addition to six-membered carbon aromatic rings has already been mentioned. Also, conjugation of double bonds has an effect on chemical shifts which is not currently explicitly treated in the environmental encoding algorithm presented in this paper.

A more fundamental problem with the current algorithm is the fact that much of the detailed connection information in the structure is lost by the amalgamation of all information about atoms n bonds from the carbon atom of interest into one scalar value. Essentially a dimensionality reduction has occurred in which the molecular graph is reduced to a series of individual nodes (i.e., the e_i values). The use of the combined vector is one step toward alleviating the ambiguities introduced by the vector-based approach. Further improvements to the multiple vector approach might include some combination of vectors in which the e_i value is the largest Z_j value i bonds from the carbon of interest, or the e_i value is the largest Z_j value attached to the carbon with the largest Z_j value i bonds from the carbon of interest, or the e_i value is the smallest Z_j value attached to the atom with the largest Z_j value i bonds from the carbon of interest, or any number of related vector-based schemes. Some initial work has been performed with various combinations of these new approaches along with RSS or linear combinations in multiple vector implementations, but no results have been obtained which are better than the RSS/linear combination. Further work, however, might prove beneficial.

A geometrical approach has been studied in the past in which the atoms contributing to a particular e_i are those which fall in a particular concentric radial shell centered about the carbon atom of interest.⁴⁸ Three-dimensional coordinates for the structures in the library and test sets were computed by use of molecular mechanics calculations in order to test this approach. However, this method did not offer significant improvements over the topological approach for the more global test set used in this study and was therefore not investigated further.

CONCLUSION

The work presented in this paper has applied the environmental encoding algorithm to a large database and obtained results that predict chemical shifts within 1.30 ppm based on the median value of the mean residuals for the test set of 4240 structures. Approximately 75% of the chemical shifts in the test set of 38 568 carbons were predicted correctly within 2.0 ppm. While it is difficult to compare these results directly to those reported previously by other workers due to differences in the database and test sets used, the level of accuracy in predicted chemical shifts obtained here appears very competitive.

The results presented here have also demonstrated that the Euclidean distance is a good predictor of the quality of the chemical shift prediction. The merits of the vector based approach have been explored, and, although the loss of information resulting from such an approach is a negative side effect, it is felt that the ability to make rapid and quantitative comparisons of chemical environments outweighs this negative factor.

The final use of the methodology must be considered when evaluating the utility of the environmental encoding algorithm. The purpose of the encoding algorithm is to allow the user to select a subset of carbons similar in chemical environment to that of a carbon in a structure input by the user. In our ultimate application, this subset will be used to build an empirical model of the type describe by eq 1 and predict the chemical shift. Thus, the entire burden of shift prediction does not rest on the encoding algorithm.

Finally, it should be noted that the spectrum simulation application described here focuses only on the estimation of chemical shifts. While this information is of greatest importance in solving structure elucidation problems, some applications may require information such as spectral line width and intensity. Simulation of these parameters is beyond the scope of the current methodology, however.

ACKNOWLEDGMENT

This research was supported by the Shell Development Co., Houston, TX. The Sadtler Division of Bio-Rad Laboratories, Inc. is acknowledged for providing the spectral and chemical structural data used in this work.

REFERENCES AND NOTES

- Bremser, W. HOSE—A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355–365.
- Bremser, W. Expectation Ranges of ^{13}C NMR Chemical Shifts. *Magn. Reson. Chem.* **1985**, *23*, 271–275.
- Chen, L.; Robien, W. The CSEARCH-NMR Data Base Approach to Solve Frequent Questions Concerning Substituent Effects on Carbon-13 NMR Chemical Shifts. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 217–223.
- 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.
- Cheng, H. N.; Kasehagen, L. J. Integrated Approach for ^{13}C Nuclear Magnetic Resonance Shift Prediction, Spectral Simulation and Library Search. *Anal. Chim. Acta* **1994**, *285*, 223–235.
- Von der Lieth, C. W.; Seil, J.; Köhler, Opferkuch, H. J. ^{13}C NMR Data Bank Techniques as Analytical Tools. *Magn. Reson. Chem.* **1985**, *23*, 1048–1055.
- Chen, L.; Robien, W. OPSI: A Universal Method for Prediction of Carbon-13 NMR Spectra Based on Optimized Additivity Models. *Anal. Chem.* **1993**, *65*, 2282–2287.
- 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.
- 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.
- Clouser, D. L.; Jurs, P. C. Simulation of ^{13}C Nuclear Magnetic Resonance Spectra of Tetrahydropyrans Using Regression Analysis and Neural Networks. *Anal. Chim. Acta* **1994**, *295*, 221–231.
- Jurs, P. C.; Ball, J. W.; Anker, L. S.; Friedman, T. L. Carbon-13 Nuclear Magnetic Resonance Spectrum Simulation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 272–278.
- 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.
- 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.
- Sprague, J. T.; Tai, J. C.; Yuh, Y.; Allinger, N. L. The MMP2 Computational Method. *J. Comput. Chem.* **1987**, *8*, 581–603.
- 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.
- Levy, G. C.; Lichter, R. L.; Nelson, G. L. Carbon-13 Nuclear Magnetic Resonance Spectroscopy, 2nd ed.; Wiley-Interscience: New York, 1980; p 30.
- Spiesecke, H.; Schneider, W. G. Effect of Electronegativity and Magnetic Anisotropy of Substituents on C^{13} and H^1 Chemical Shifts in CH_3X and $\text{CH}_3\text{CH}_2\text{X}$ Compounds. *J. Chem. Phys.* **1961**, *35*, 722–730.
- Savitsky, G. B.; Ellis, P.; Namikawa, K.; Maciel, G. E. Effect of *cis-trans* Isomerism on the Carbon-13 Chemical Shifts of Some Unsymmetrically Substituted Ethylenes. *J. Chem. Phys.* **1968**, *49*, 2395–2404.
- Maciel, G. E.; Simeral, L.; Elliott, R. L.; Kaufman, B.; Cribley, K. Additivity in the Carbon-13 Chemical Shifts of 1,2-Disubstituted Ethanes. *J. Phys. Chem.* **1972**, *76*, 1466–1469.
- Couperus, P. A.; Clague, A. D. H.; an Dongen, J. P. C. M. Carbon-13 Chemical Shifts of some Model Olefins. *Org. Magn. Reson.* **1976**, *8*, 426–431.
- Johnson, L. F.; Jankowski, W. C. Carbon-13 NMR Spectra; Wiley-Interscience: New York, 1972.
- Strong, A. B.; Ikenberry, D.; Grant, D. M. Calculation of Carbon-13 Chemical Shifts in Simple Hydrocarbons. *J. Magn. Reson.* **1973**, *9*, 145–165.
- Sarneski, J. E.; Surprenant, H. L.; Molen, F. K.; Reilley, C. N. Chemical Shifts and Protonation Shifts in Carbon-13 Nuclear Magnetic Resonance Studies of Aqueous Amines. *Anal. Chem.* **1975**, *47*, 2116–2124.
- Ejchart, A. Substituent Effects on Carbon-13 NMR. Part 2. Chemical Shifts in the Saturated Framework of Secondary Aliphatic Derivatives. *Org. Magn. Reson.* **1981**, *15*, 22–24.
- Miyajima, G.; Takahashi, K. Carbon-13 Nuclear Magnetic Resonance Spectroscopy. III. Chloro-substituted Ethanes and Ethylenes. *J. Phys. Chem.* **1971**, *75*, 331–334.
- Chukovskaya, E. C.; Dostovalova, V. I.; Vasil'eva, T. T.; Freidlina, R. Kh. Carbon-13 NMR Spectra of some Polychloroalkenes. *Org. Magn. Reson.* **1976**, *8*, 229–232.
- Paul, E. G.; Grant, D. M. Additivity Relationships in Carbon-13 Chemical Shift Data for the Linear Alkanes. *J. Am. Chem. Soc.* **1963**, *85*, 1701–1702.
- Friedel, R. A.; Retcofsky, H. L. Carbon-13 Nuclear Magnetic Resonance Spectra of Olefins and other Hydrocarbons. *J. Am. Chem. Soc.* **1963**, *85*, 1300–1306.
- Christl, M.; Reich, H. J.; Roberts, J. D. Nuclear Magnetic Resonance Spectroscopy. Carbon-13 Chemical Shifts of Methylcyclopentanes, Cyclopentanols, and Cyclopentyl Acetates. *J. Am. Chem. Soc.* **1971**, *93*, 3463–3468.
- NIH/EPA Chemical Information System; Fein-Marquart Associates: Baltimore, MD, 1980.
- Lauer, D.; Motell, E. L.; Traficante, D. D.; Maciel, G. E. Carbon-13 Chemical Shifts in Monoalkyl Benzenes and Some Deuterio Analogs. *J. Am. Chem. Soc.* **1972**, *94*, 5335–5338.
- van Dongen, J. P. C. M.; de Bie, M. J. A.; Steur, R. ^{13}C -NMR and CNDO/2 Calculations of Compounds Containing SP-Hybridised Carbon. Acetylenes and Cumulenes. *Tetrahedron Lett.* **1973**, 1371–1374.
- Dorman, D. E.; Jautelat, M.; Roberts, J. D. Carbon-13 Nuclear Magnetic Resonance Spectroscopy. Quantitative Correlations of the Carbon Chemical Shifts of Acyclic Alkenes. *J. Org. Chem.* **1971**, *36*, 2757–2766.
- Rojas, A. C.; Crandall, J. K. Carbon-13 Nuclear Magnetic Resonance Spectroscopy. Substituted Vinyl Ethers and Acetates. *J. Org. Chem.* **1975**, *40*, 2225–2229.

- (35) Dharni, K. S.; Stothers, J. B. ^{13}C NMR Studies Part V. Carbon-13 Spectra of some Substituted Styrenes *Can. J. Chem.* **1965**, *43*, 510–520.
- (36) Sadtler Standard C-13NMR Spectra; Sadtler Research Laboratories: Philadelphia, PA, 1976.
- (37) Roberts, J. D.; Weigert, F. J.; Kroschwitz, J. I.; Reich, H. J. Nuclear Magnetic Resonance Spectroscopy. Carbon-13 Chemical Shifts in Acyclic and Alicyclic Alcohols *J. Am. Chem. Soc.* **1970**, *92*, 1338–1347.
- (38) Lauterbur, P. C. C^{13} Nuclear Magnetic Resonance Spectroscopy. I. Aromatic Hydrocarbons *J. Am. Chem. Soc.* **1961**, *83*, 1838–1846.
- (39) Nash, C. P.; Maciel, G. E. Carbon-13 Nuclear Magnetic Resonance Spectra of some Aromatic Amines and Imines *J. Phys. Chem.* **1964**, *68*, 832–836.
- (40) Hatada, K.; Nagata, K.; Yuki, H. Carbon-13 NMR Spectra of Alkyl Vinyl Ethers, and their Structures and Reactivities *Bull. Chem. Soc. Jpn.* **1970**, *43*, 3195–3198.
- (41) Firl, J.; Runge, W. ^{13}C -NMR Spectrum of Ketene *Angew. Chem.* **1973**, *12*, 668–669.
- (42) Spiess, H.; Schneider, W. G. Substituent Effects on the C^{13} and H^1 Chemical Shifts in Mono-substituted Benzenes *J. Chem. Phys.* **1961**, *35*, 731–738.
- (43) Lichter, R. L.; Roberts, J. D. Nitrogen-15 Magnetic Resonance Spectroscopy. XV. Natural-Abundance Spectra. Chemical Shifts of Hydrazines *J. Am. Chem. Soc.* **1972**, *94*, 4904–4906.
- (44) Lauterbur, P. C. Some Applications of Carbon-13 Nuclear Magnetic Resonance Spectra to Organic Chemistry *Ann. N. Y. Acad. Sci.* **1958**, *70*, 841–857.
- (45) Pehk, T.; Lippmaa, E. Carbon-13 Chemical Shifts of Monosubstituted Cyclohexanes *Org. Magn. Reson.* **1971**, *3*, 679–687.
- (46) Maciel, G. E. Carbon-13 Chemical Shifts of Vinyl Carbons *J. Phys. Chem.* **1965**, *69*, 1947–1951.
- (47) Marker, A.; Doddrell, D.; Riggs, N. V. Deshielding of Carbon-13 Nuclei by Attached Iodine Atoms. Solvent Effects on Carbon-13 Chemical Shifts in Alkyl Iodides *J. Chem. Soc., Chem. Commun.* **1972**, 724–725.
- (48) Small, G. W. Database Retrieval Techniques for Carbon-13 Nuclear Magnetic Resonance Spectrum Simulation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 279–285.

CI950142Q