Molecules **1997**, *2*, 114-128

molecules

ISSN 1420-3049

^{13}C NMR Spectral Prediction by Means of Generalized Atom Center Fragment Method

Jun Xu*

BIO-RAD Laboratories, Sadtler Division, 3316 Spring Garden Street, Philadelphia, PA 19104, USA
*Present address: Oxford Molecular Group, Inc. 810 Gleneageles Court, Suite 300 Baltimore, MA 21286, USA
Tel. 410-821-5980 x-339, Fax 410-296-0712, E-mail: jxu@oxmol.com

Received: 21 October 1996 / Accepted: 11 April 1997 / Published: 20 August 1997

Abstract: Knowledge-based NMR spectral prediction relies on the correlations between substructures and sub-spectra. To extract the correlations, a systematic substructure measurement has been developed to classify substructures according to their chemical shift values. Historically, the atom center fragment (ACF) concept has been used as a means to systematically measure substructures for NMR spectral prediction. The assumption behind this concept is that the chemical shift value of an atom is influenced by its chemical environment. Based upon the study of the ACF-type approaches, a generalized atom center fragment (GACF) approach is proposed in this paper. In the GACF approach, a substructure consists of a center atom, core layer, and external layers. The center atom and the core layer, are identified as the super center atom. The external layers are the chemical environment. A number of algorithms have been developed to measure GACF substructures from a structure database, and create the NMR knowledge base for NMR spectral prediction.

Keywords: ^{13}C NMR, NMR spectral prediction, atom center fragment (ACF), .generalized atom center fragment (GACF).

Introduction

Empirical NMR spectral prediction approaches correlate substructures and sub-spectra by means of sub-structural encoding. The oldest encoding method is the additivity model, which consists of a set of frame structures, substituents, and calculation rules [1,2]. A more general additivity model was reported by the Small and Jurs [3], and enhanced by Schweitzer and Small [4] later. Recently, a number of neural network approaches have been employed for the encoding [5-8]. Bremser proposed HOSE code to systematically encode substructures for ^{13}C NMR knowledge extraction [9]. Robien adopted HOSE method [10], and built up a direct knowledge base retrieval method for ^{13}C NMR spectral prediction [11]. His newer publications can be found from *Anal. Chim. Acta* , **1990,** *229*, 17 and *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 291.

The critical part of NMR knowledge generation is the systematic substructure measurement. The chemical shift value of a carbon atom is influenced by the chemical environment of the atom. HOSE code uses layer (or level) to define the chemical environment (Figure 1). The first layer is defined as all the atoms being one-bond-away from the central atom (or focus atom); the second layer has the atoms being two-bond-away, etc. This idea can be represented in atom center fragment (ACF) concept, which has been addressed by many authors in different ways [12-14].

The effects of the environmental atoms on the central atom are not completely determined by the topological distance between the center atom and the environmental

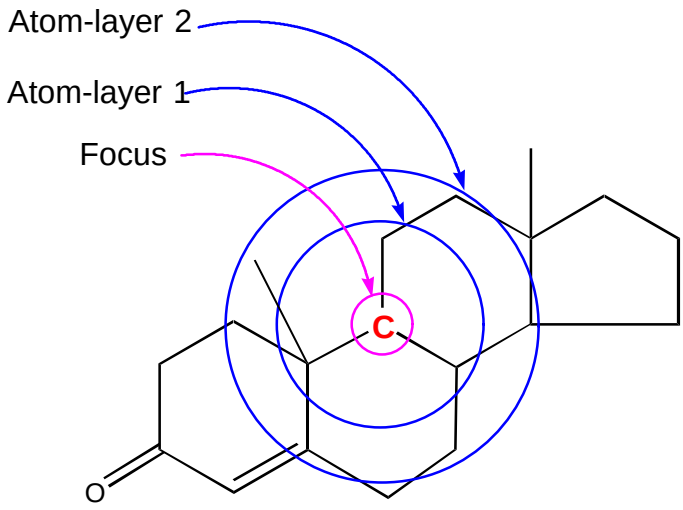


Figure 1. The center atom (focus atom) and atom layers in HOSE code approach.

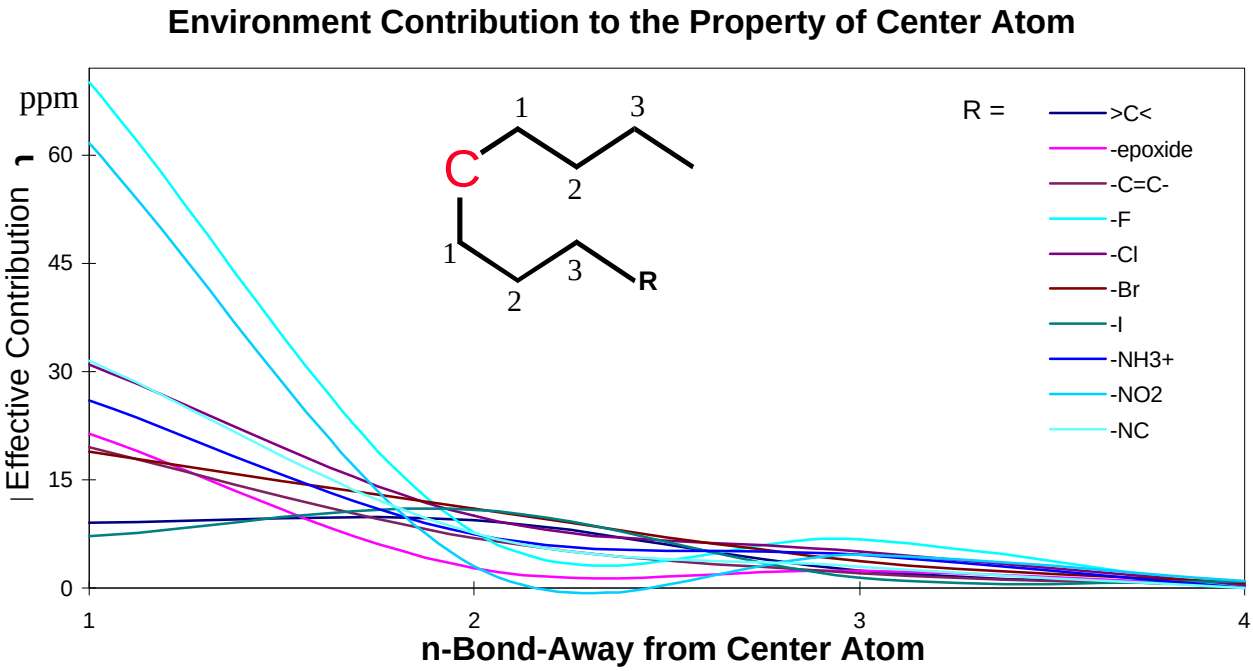


Figure 2. Chemical environment effective contribution to a central carbon atom in an aliphatic structure.

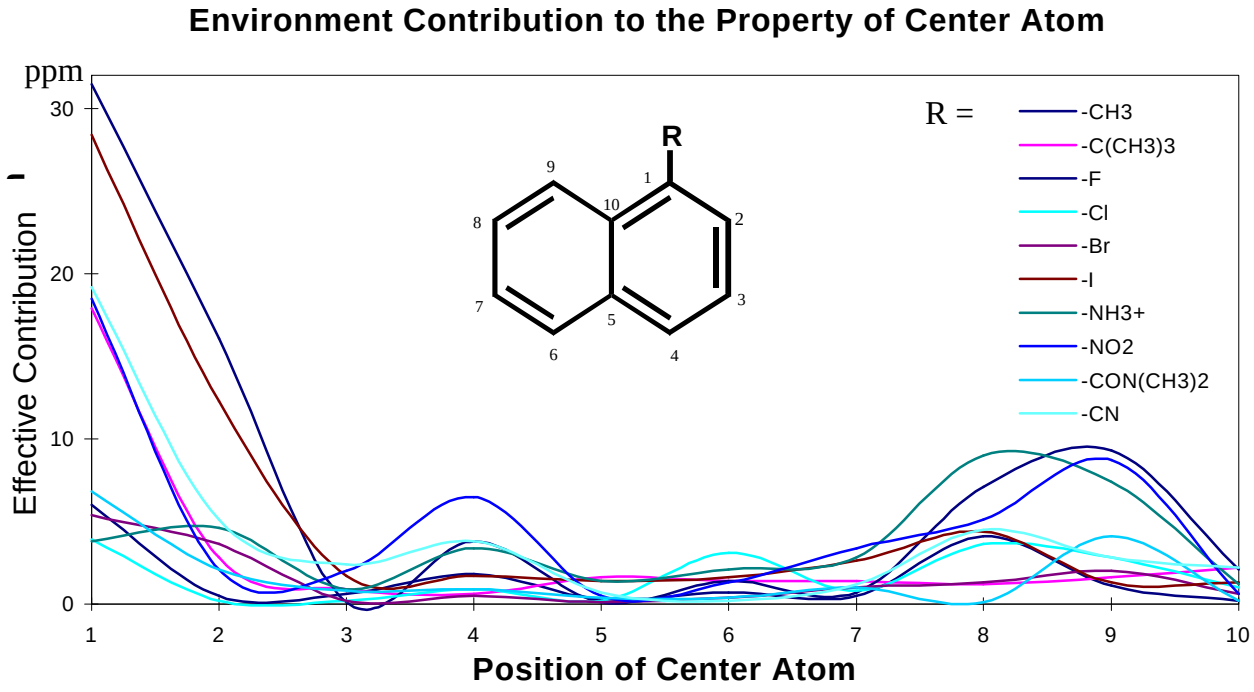


Figure 3. Chemical environment and its effect on central carbon atoms in aromatic system.

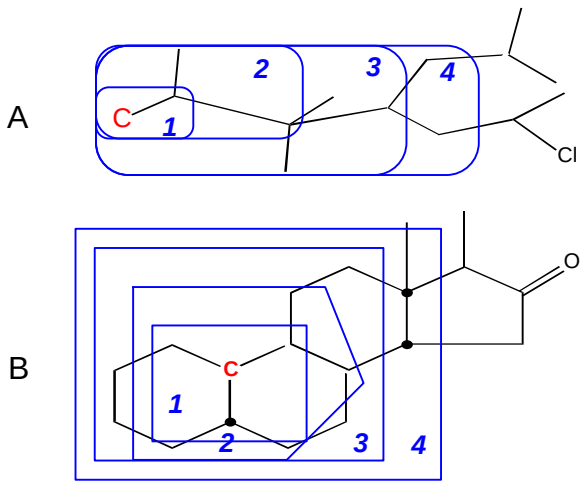


Figure 4. Atom layers in different systems. A: Layer 4 is too far for this center atom (in red color). B: Layer 4 is still not far enough for this center atom (in red color). The numbers represent atom layers.

atoms. The effects are influenced by the topological distance and the bonding types. These effective contributions of environmental atoms/groups to the central atom are shown in Figures 2 and 3.

The data for these Figures are adopted from reference 1; the plots use the chemical shift absolute values to emphasize the contributions. Figures 2 and 3 are just showing the tendency, the continuous curves do not mean there is any effective contribution between 1-bond-away and 2-bond-away, etc.

From these Figures, it can be seen that if a center atom is in an acyclic aliphatic system, the effect of an environmental substituent to the center atom decreases along the increasing number of the atom-layer. If the substituent is four bonds away from the center atom, it has almost no effective contribution to change the center atom's chemical shift. In Figure 3, however, even R-group is at the position 4 (four-bonds-away), its effect on the center atom's chemical shifts is still significant. It is because the center atom is in an aromatic system.

The conclusion from these Figures is that the simple ACF measurement works only for acyclic and non-conjugated systems. Another problem in the simple ACF approach is that the number of atom layers should be included in an ACF substructure. For example, if an ACF includes four atom layers, then it may be too far for the center atom in an aliphatic system, but not far enough for the one in a ring or conjugated system (see Figure 4).

To have an objective substructure measurement, it is necessary to extend the simple ACF concept to generalized atom center fragment concept.

Generalized atom center fragment (GACF)

In the additivity model, the chemical environment of a center atom has two parts (see Figure 5), i.e., frame (core) structure and substituents (environmental substructures).

The GACF approach takes this into account in the substructure measurement. A GACF substructure consists of a center atom, core structure and environmental layers (substituents). The core structure characterizes the different bonding system, which is also responsible for the special chemical behavior. According to topological and chemical properties, core structures are classified into the following nine classes:

1. Independent non-aromatic single ring system
2. Fused non-aromatic ring system
3. Bridged ring system
4. Spiro ring system
5. Independent aromatic single ring system
6. Fused aromatic ring system
7. Conjugated system
8. Cumulene system
9. Acyclic system

The structural features of these systems are listed in Table 1.

A core structure can be classified into more than one ring class simultaneously. In this case, the assigned class is chosen by applying priority:

spiro ring > bridged ring > fused ring

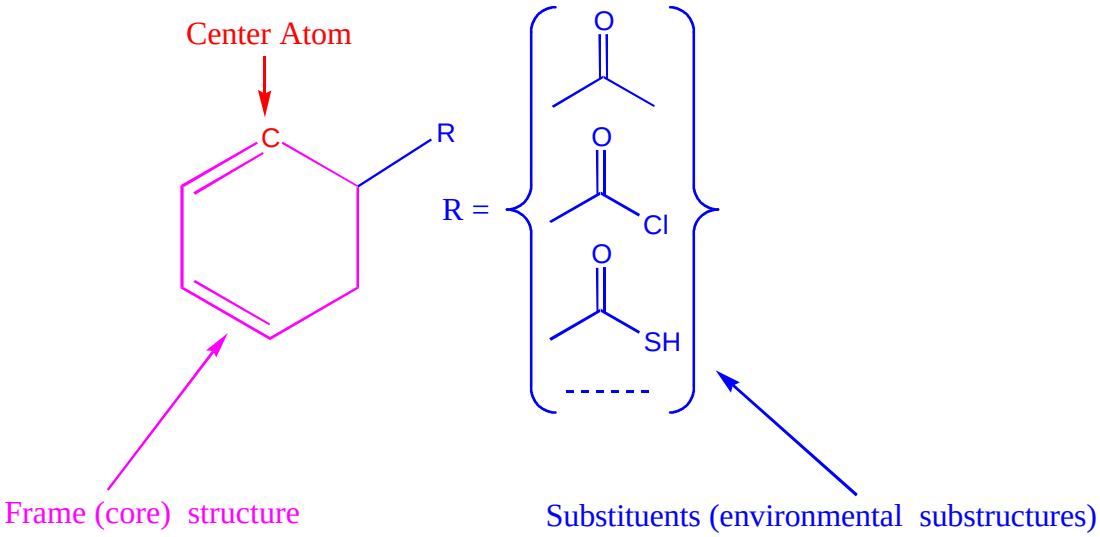
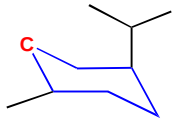
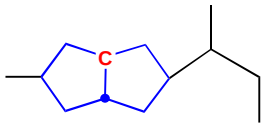
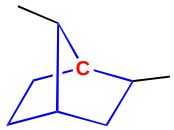
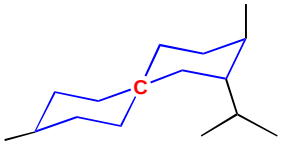
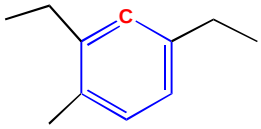
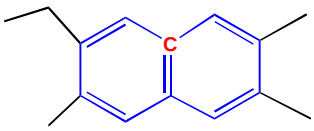
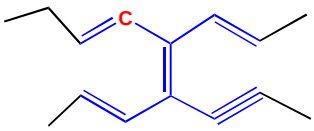
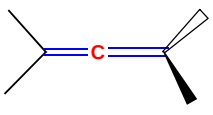
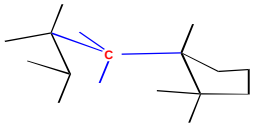


Figure 5. Center atom, core structure and environmental substructures.

Table 1. Core structure classification*

Chemical Environment Classification	Example
Class 1: Independent non-aromatic single ring system	
Class 2: Fused non-aromatic ring system	
Class 3: Bridged ring system	
Class 4: Spiro ring system	
Class 5: Independent aromatic single ring system	
Class 6: Fused aromatic ring system	
Class 7: Conjugated system	
Class 8: Cumulene system	
Class 9: Acyclic system	

*Red: center atom. Blue: core structure. Black: chemical environment.

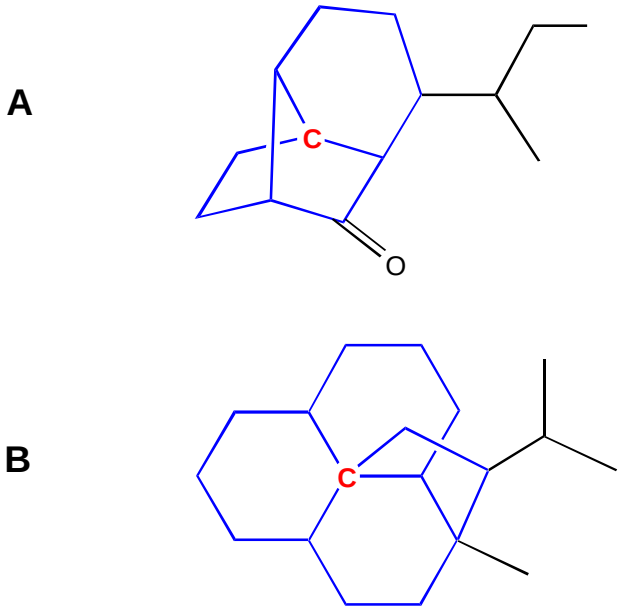


Figure 6. Choosing a ring class for an ambiguous center atom. A: the center atom is in bridged ring and fused ring systems; it is assigned to bridged ring system. B: the center atom is in fused ring, bridged ring and spiro ring systems; it is assigned to spiro ring system.

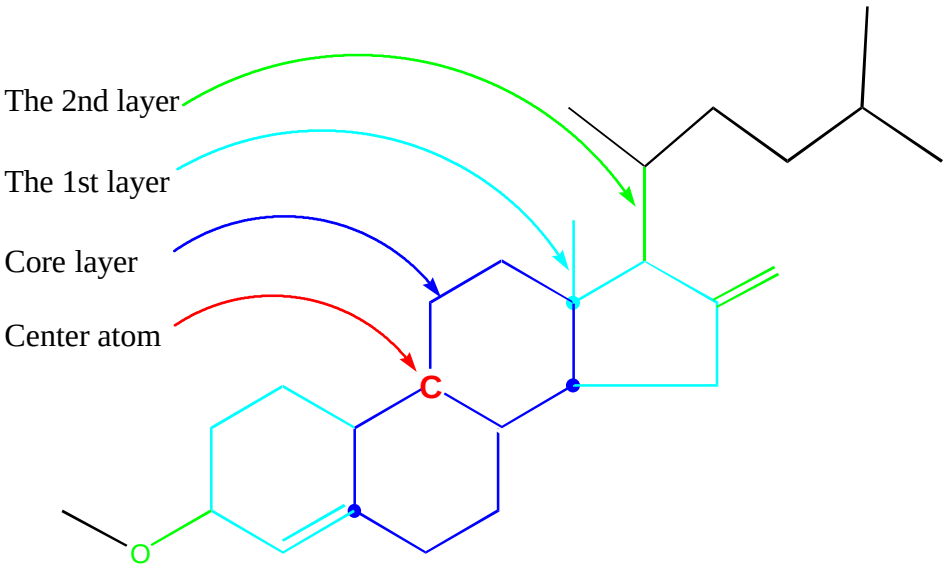


Figure 7. Example of a GACF.

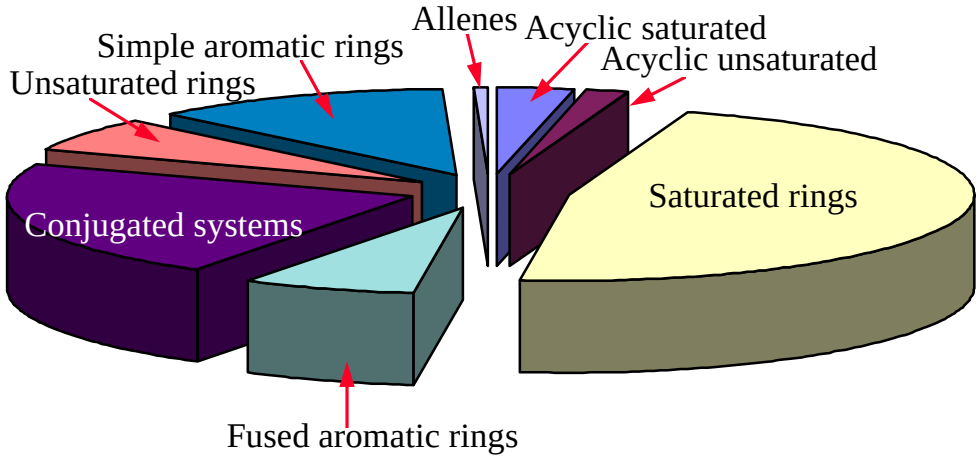


Figure 8. The result shows that our ^{13}C NMR database has well diversified substructural information, which is good enough for general ^{13}C NMR prediction. Cumulane system will not have very good ^{13}C NMR prediction because of insufficient information in the database.

The examples are shown in Figure 6.

Therefore, a GACF substructure is measured in the following steps:

1. select a center atom
2. get a class for the center atom
3. capture a core structure for the center atom (the core structure measurement is shown in blue in Table 1 in the right column)
4. capture environmental substituents (which are part of the GACF) for the GACF according to the number of the layers

An GACF example is illustrated in Figure 7.

^{13}C NMR knowledge extraction and chemical shift prediction

50,000 structures with ^{13}C chemical shift assignments have been selected as input data set for ^{13}C NMR knowledge extraction. A number of graph theory algorithms have been developed to measure GACF substructures. The main algorithms are independent ring, fused ring, bridged ring, spiro ring perception algorithms, and conjugated system perception algorithm, etc. These algorithms are all based upon the GMA algorithm reported in our previous work [15].

The molecular diversity of the input data has been analyzed by means of our in-house algorithms, and shown in Figure 8.

Total 64,307 GACF (General Atom Center Fragment, 1-GACF means the first layer of GACF, so and so forth) substructures (up to 2 layers) extracted from 565,513 assigned chemical shifts. The GACF class distribution is shown in Figure 9. Different GACF reflects a carbon atom with a different core structure and chemical environment; therefore Figure 9 shows the atomic diversity.

It is known that larger numbers of atom-layers will increase the size of a substructure, the size of the knowledge body, and the accuracy of the NMR spectral prediction. But, too large a knowledge body will reduce the search performance. The larger size of the GACF has less chance to be matched; that is, the knowledge will not be used very often. Figure 10 shows that 2-GACF (each GACF substructure has 2-atom-layer chemical environment) knowledge body has significantly more knowledge entries in classes 9, 5 and 6, and less entries in classes 1, 3, 4, and 7. We generate 0~2 GACF substructures for ^{13}C NMR knowledge base, where 0-GACF substructures have only core atoms, no environmental atoms.

^{13}C NMR knowledge body is the correlation table of GACF substructures and ^{13}C chemical shifts. Its format is

Table 2. The Format of GACF-Chemical Shift Correlation Table

GACF	class	layer	shift	median	maximum	minimum	σ	*fn*
R—C≡N	9	1	116.87	117.80	122.81	108.1	3.40	263

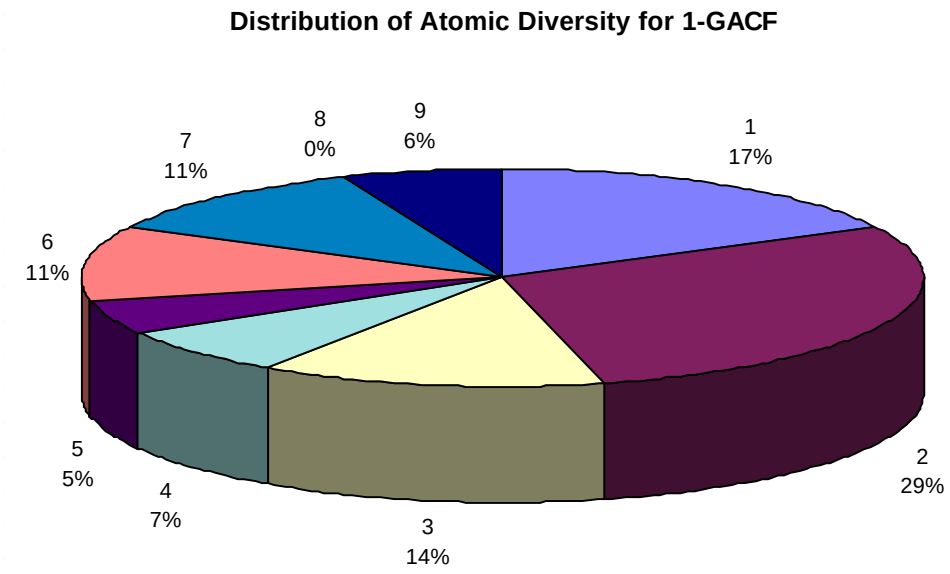


Figure 9. The GACF class distribution. The number over the percentage figure is the code of a GACF class (refer to Table 1.). The largest portion of this knowledge body is regarding fused non-aromatic carbon atom chemical shifts. 0% actually means <1% (due to the poor resolution of the graphic display).

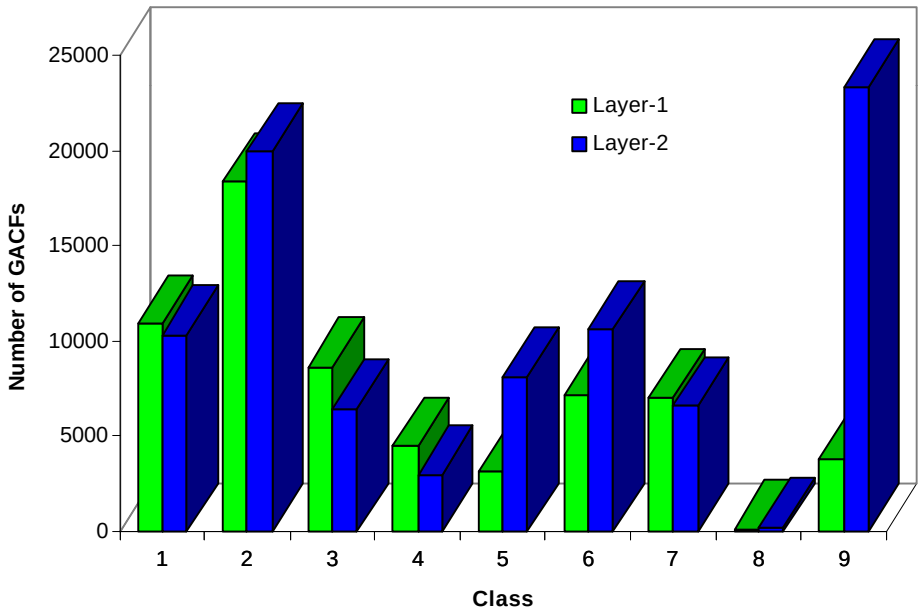


Figure 10. Comparison of the sizes of 1-GACF knowledge body and 2-GACF knowledge body.

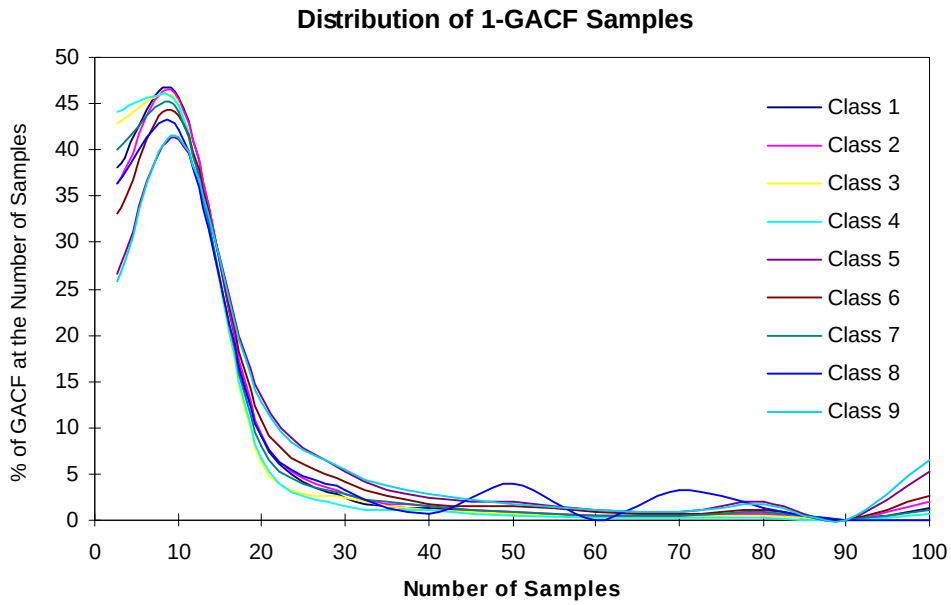


Figure 11. Distribution of 1-GACF *fn*.

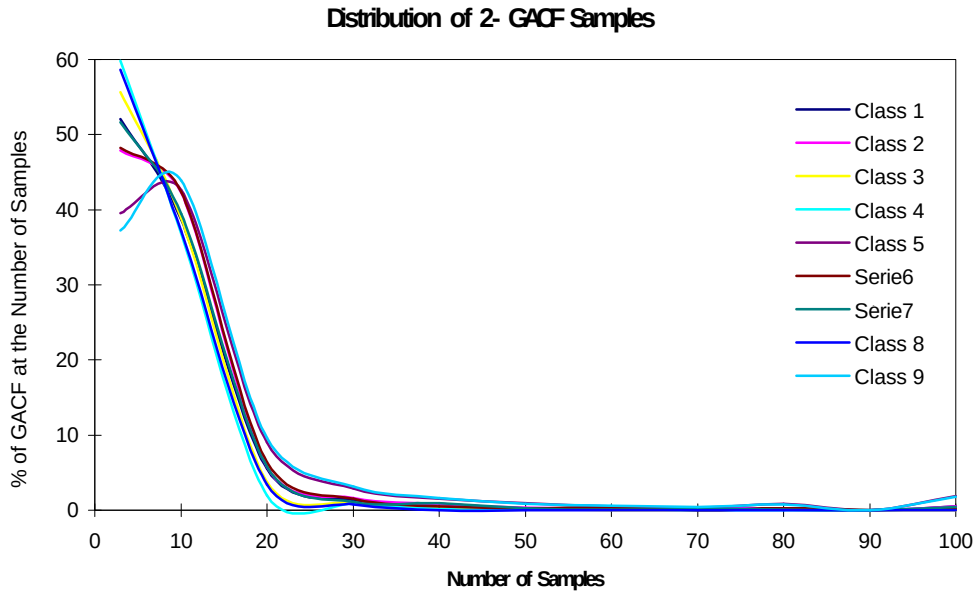


Figure 12. Distribution of 2-GACF *fn*.

Combinatorial: Structural Diversity Space

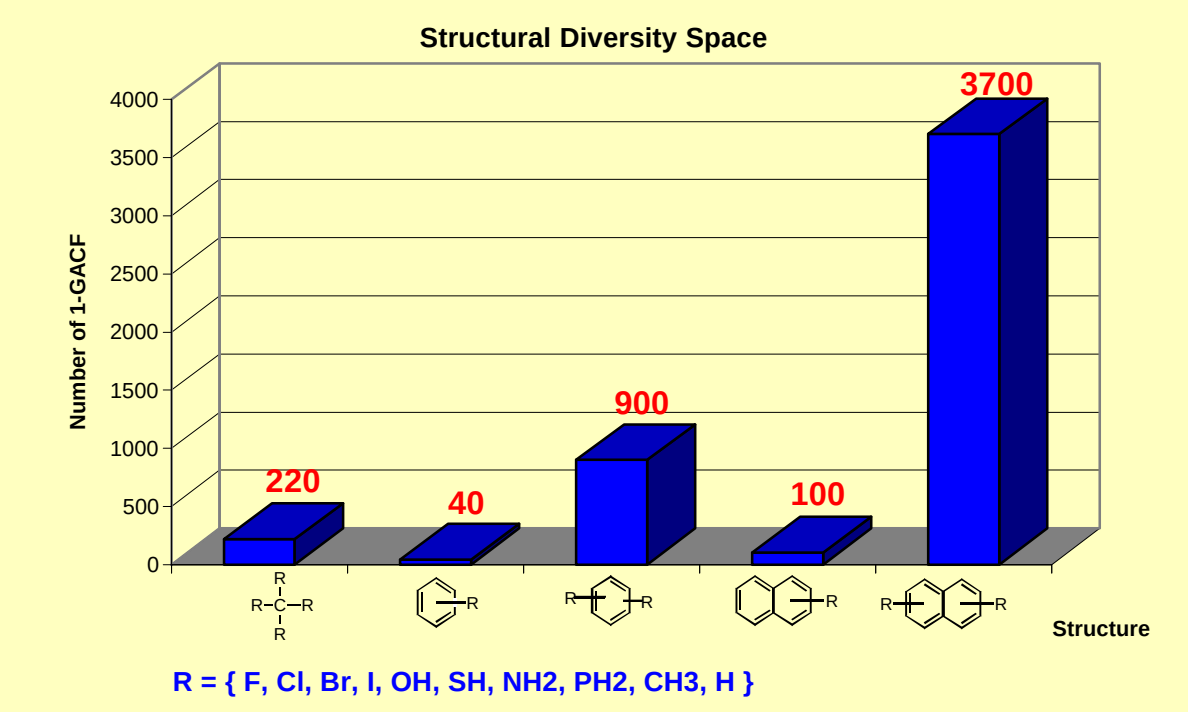


Figure 13. If R is defined as a 10-substitute group, the number of different types of carbon atoms is 220 (1-GACF substructures), for mono-substituted benzene, it is 40 due to symmetry, 900 for disubstituted benzene, etc. This Figure shows that the structural diversity space is a "combinatorial explosion" even if the scaffolds are simple and small.

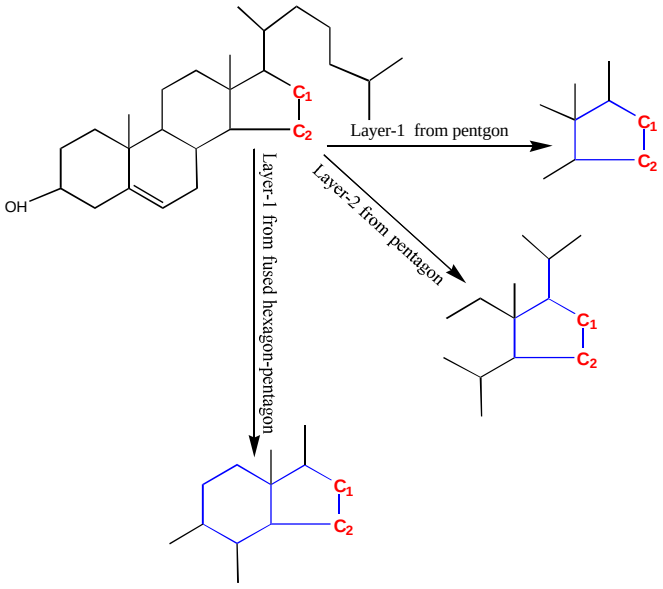


Figure 14. C_1 and C_2 have very different chemical shifts (C_1: 40 ppm, but C_2: 24 ppm). In order to predict correctly, the substructure measurement should distinguish them. Topologically, if the pentagon is considered as core layer, C_1 and C_2 cannot be distinguished up to the second layer. But, if the fused hexagon-pentagon is measured as the core layer, C_1 and C_2 can be distinguished. The hexagon is encoded as remote fused ring in GACF substructure measurement scheme.

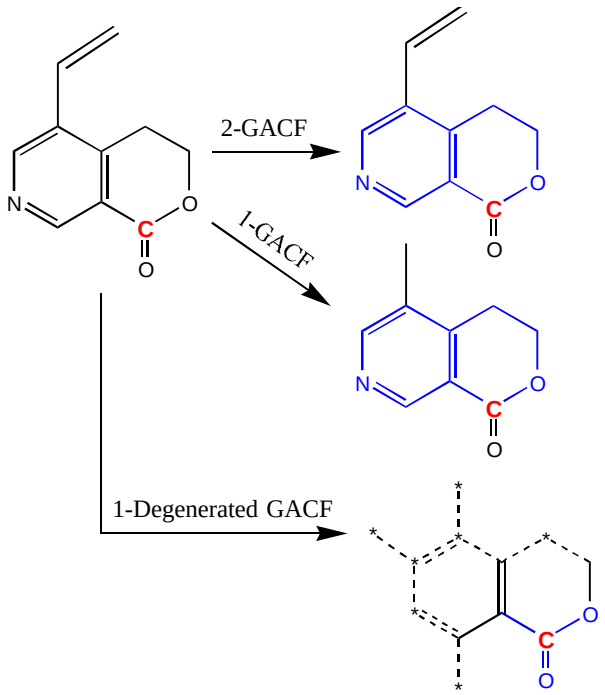


Figure 15. If 1-GACF and 2-GACF are not found in a knowledge base, the 1-Degenerated GACF may have more chance to match with a GACF in the knowledge base to get the closest estimation. "*" represents any atom; dashed bonds represent the bonds not in the Degenerated GACF.

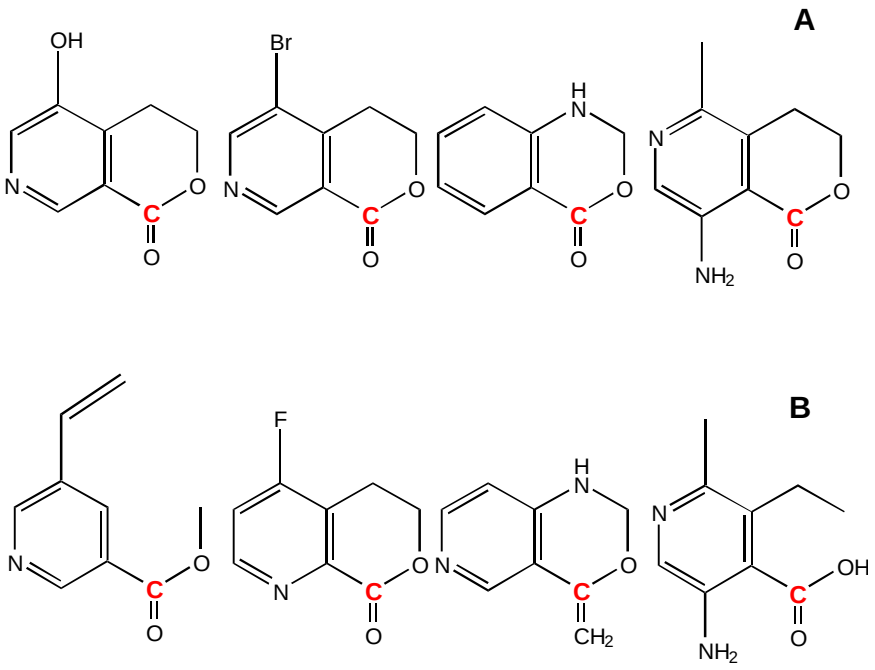


Figure 16. A: structures used to estimate the carbon atom (colored in red) chemical shift of 1-Degenerated GACF in Figure 15. B: structures which cannot be used for this estimation.

shown in Table 2 with an example. The "shift" is the chemical shift average value, "median" is the middle value in this range, "maximum" and "minimum" define the variable range (band). "σ" is the standard deviation, an "*fn*" is the number of sample chemical shifts which are used to produce the chemical shift average value, also called "frequency number".

The accuracy and generality of the ^{13}C NMR spectral prediction can be analyzed by studying the distributions of frequency numbers (*fn*). As shown in Table 2, *fn* should be larger than 3 to make statistical sense; however, if *fn* is too large, such as >1000, the accuracy will decrease. On the other hand, a ^{13}C NMR KB rule with larger *fn* may mean that it covers larger structural diversity, and therefore, has better generality. The accuracy and generality are conflicting, and have to be balanced. Figures 11 and 12 show the relationship of the *fn* distribution, accuracy and generality, where "the Number of Samples" is *fn*.

In Figure 11, about 47% 1-GACF rules have ~10 sample chemical shifts (*fn*). However, as shown in Figure 12, about 60% 2-GACF rules have 3~5 sample chemical shifts. Hence, 2-GACF rules will give more accurate predictions, but cover less structural diversity. In fact, the structural diversity space is extremely huge. It is almost not possible to build a knowledge base to cover the whole structural diversity with reasonable predicting accuracy. Figure 13 shows a way to estimate structural diversity space.

The ^{13}C NMR spectral prediction for a given structure is described in the following steps:

1. input a structure (draw a structure through a graphic user interface)
2. for each carbon atom, extract a 2-GACF (2 level GACF) substructure from the structure
3. search the 2-GACF against 2-GACF ^{13}C NMR KB
4. if this 2-GACF is found from the 2-GACF KB, the carbon's chemical shift is predicted
5. if not:
6. extract a 1-GACF (1 level GACF) substructure from the structure
7. search the 1-GACF against 1-GACF ^{13}C NMR KB
8. if this 1-GACF is found from the 1-GACF KB, the carbon's chemical shift is predicted with less accuracy
9. if not report: "cannot predict chemical shift for this type of carbon atom"
10. go to 2

Searching a GACF from 64,307 GACF substructures by means of atom-by-atom search will be time-consuming. Hence, all GACF substructures have been converted to hash codes and sorted. Therefore, the GACF structure search is very fast.

The accuracy and generality are conflicting. Discriminating similar substructures can enable more predicting accuracy, but less substructures can match the GACF in the knowledge base. In Figure 14, C_1 and C_2 (colored in red) should have very different chemical shifts. Taking pentagon atoms (colored in blue) as the core layer (super-atom center), either 1-bond-away or 2-bonds-away from the pentagon ring, C_1 and C_2 cannot be distinguished. In order to distinguish C_1 and C_2, the fused hexagon and pentagon atoms are taken as the super-atom center. With this GACF measure, C_1 and C_2 are distinguished in 1-GACF.

More discrimination, however, reduces generality. No matter how large the knowledge body is, many substructures are still not included. The mission of a prediction algorithm is to output the best estimation based upon existing knowledge collection.

In order to compromise the accuracy and generality, degeneracy technique (way to loosen structural pattern match restriction) is introduced. If a GACF substructure is not found in a higher level GACF knowledge base, this substructure can be degenerated to become a simplified GACF, and enable more chance to match. Figure 15 shows a way to degenerate a GACF substructure.

The degeneracy technique gains more generality for a knowledge base. Figure 16 A lists a set of structures used to estimated the carbon atom (colored in red) chemical shift of the 1-Degenerated GACF in Figure 15. Figure 16 B shows the structures not used for this estimation.

Results and conclusions

GACF-based ^{13}C NMR spectral prediction program has been implemented in C and Visual C++ in both UNIX (SUN or SGI) and NT/Windows platforms. The Knowledge Base occupies ~4 MB space. Average standard deviation of the prediction is 4.56 ppm. The average prediction time for a small structure (<255 carbon atoms) is less than a second. The program has been tested on a number of structures selected from other data resource by a third part chemist. Some of the testing results are listed in Table 3 for comparison.

General speaking, KB-based NMR spectral prediction is influenced by the following factors:

1. algorithms to correctly classify the center atoms and their chemical environment
2. the quality of the structures-spectra database

The atomic chemical environment classification includes: (1) aromatic and non-aromatic, (2) cyclic and acyclic, (3) hetero-ring and homo-ring, (4) ring size, (5) ring types, such as single, fused, bridged, spiro,

Table 3. Comparisons of observed chemical shift values, ACD/CNMR predictions and GACF predictions.

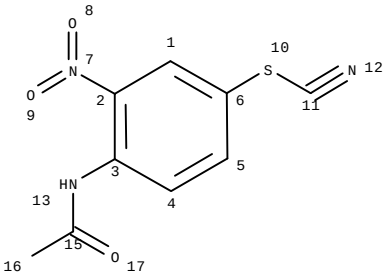
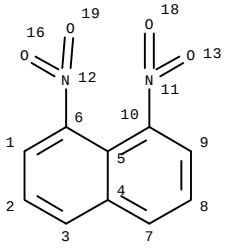
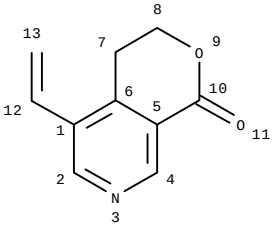
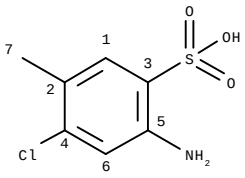
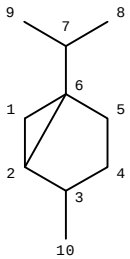
Structure	Label	Observed	ACD [a, b]	GACF [a]
	1	127.24	**120.12**	126.32
	2	139.43	136.95	139.12
	3	142.26	**134.19**	143.34
	4	126.46	**120.10**	121.05
	5	135.82	136.11	**130.06**
	6	120.68	123.68	123.73
	11	111.25	111.20	111.47
	15	168.79	**161.24**	168.68
	16	23.41	24.24	23.26
	1	127.11	129.03	124.61
	2	127.43	125.21	128.51
	3	134.97	**158.47**	**126.83**
	4	134.71	**126.58**	132.20
	5	115.21	**122.91**	**128.85**
	6	144.73	149.68	145.91
	1	130.80	131.89	128.44
	2	150.90	155.29	151.13
	4	150.20	148.88	147.80
	5	144.70	145.22	144.70
	6	120.90	119.74	116.95
	7	24.20	23.45	**33.90**
	8	66.10	**54.15**	66.41
	10	163.40	161.10	165.10
	12	129.50	130.10	**135.04**
	13	120.30	119.20	116.00
	1	130.36	129.45	128.55
	2	134.27	129.19	135.48
	3	129.36	**134.10**	129.82
	4	134.62	134.34	131.18
	5	138.37	**152.63**	143.34
	6	123.39	**111.78**	120.12
	7	19.36	20.24	19.08
	1	13.10	12.04	12.70
	2	30.10	28.83	27.63
	3	34.60	**40.63**	**39.95**
	4	28.80	32.49	30.11
	5	24.80	**30.82**	25.28
	6	33.90	35.51	32.51
	7	32.90	33.56	32.78
	8	20.40	20.22	19.35
	10	19.80	18.54	15.85

Table 3. (continued)

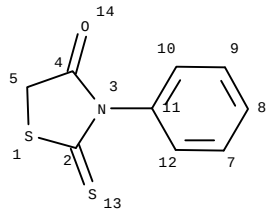
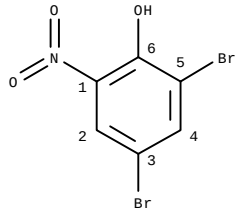
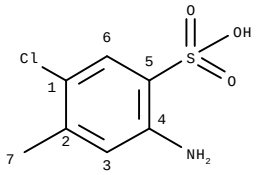
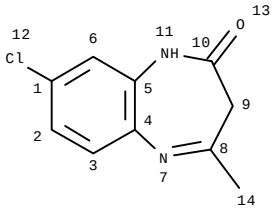
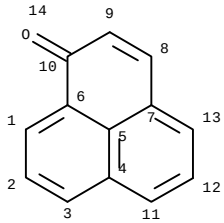
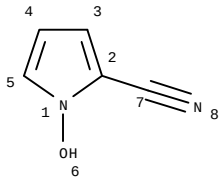
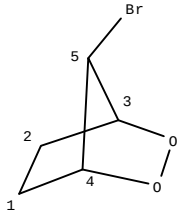
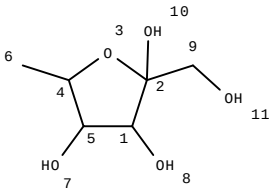
Structure	Label	Observed	ACD [a, b]	GACF [a]
	2	203.48	200.93	202.47
	4	173.90	173.73	172.40
	5	36.92	35.89	35.47
	8	129.11	126.45	128.40
	9	129.11	128.03	128.40
	10	128.57	127.49	126.82
	11	135.47	133.01	136.23
	1	143.40	141.61	143.98
	2	126.60	130.88	125.90
	3	111.40	114.76	**127.94**
	4	142.80	142.56	135.22
	5	114.40	113.75	**117.42**
	6	151.30	148.02	148.96
	1	131.51	128.76	131.18
	2	138.66	**133.45**	138.06
	3	128.12	**115.72**	125.44
	4	139.00	**151.67**	144.40
	5	126.10	**135.76**	129.82
	6	129.02	**134.69**	130.31
	7	19.84	20.22	19.26
	1	131.33	126.43	128.32
	2	125.28	121.68	128.87
	3	128.89	**121.81**	125.44
	4	138.40	138.03	133.41
	5	130.40	130.88	**136.32**
	6	121.48	120.44	120.12
	8	163.17	160.98	162.65
	9	43.70	42.14	43.80
	10	166.90	163.51	167.00
	14	28.09	**19.04**	25.49
	1	129.80	**123.86**	127.52
	2	126.70	129.85	128.40
	3	134.5	**127.83**	**127.43**
	4	131.90	131.73	132.20
	5	127.30	**132.60**	126.33
	6	129.30	127.92	130.05
	7	127.50	128.10	**135.47**
	8	141.30	142.78	140.42
	9	128.90	126.40	124.28
	10	185.00	**178.13**	180.52
	11	131.60	128.38	127.58
	12	126.40	128.84	128.40
	13	131.60	131.10	**127.52**

Table 3. (continued)

Structure	Label	Observed	ACD [a, b]	GACF [a]
	2 3 4 5 7	100.50 123.60 116.50 105.60 112.90	**123.15** **116.50** **108.07** **125.14** 112.90	**107.14** 122.00 **110.23** 105.60 112.89
	1 3 5	27.85 82.36 55.12	27.25 82.40 55.34	31.03 80.36 55.95
	1 2 4 5 6 9	81.10 106.00 75.70 77.50 15.80 63.40	79.51 107.09 72.93 78.00 13.00 63.19	80.17 105.38 76.86 76.16 18.81 62.57

[a] Results come from the ^{13}C NMR prediction product of Advanced Chemical Development (ACD), Inc.
[b] The results having absolute deviation larger than 5 ppm from the observation are in bold type.

(6) saturated and unsaturated, (7) conjugated and non-conjugated, and (8) atom layers. These classifications require a set of structural perception algorithms, which have been solved in this paper.

The quality of the structures-spectra database consists of three aspects: (1) number of assignments for a atom center fragment; (2) structural diversity of the database; and (3) correctness of the database. In order to improve the prediction, we have developed tools to review the general atom center fragments, to analyze the atomic diversity of the structure database, and to correct wrong assignments. These will be discussed in a separate paper later.

References

1. Pretsch, E.; Simon, W.; and Seibl, J. *Tables of Spectral Data for Structure Determination of Organic Compounds* 2nd Ed., Springer Verlag, Berlin Heidelberg, **1989**.
2. Clerc, J. T.; Sommerauer, H. *Anal. Chim. Acta* **1977**, *95*, 33.
3. Small, G. W.; Jurs, P. C. *Analy. Chem*. **1984**, *56*, 1314.
4. Schweitzer, R. C.; Small, G. W. *J. Chem. Info. Comput. Sci*. **1996**, *36*, 310.
5. Mitchell, B. E.; Jurs, P. C. *J. Chem. Info. Comput. Sci*. **1996**, *36*, 58.
6. West, G. M. J. *J. Chem. Info. Comput. Sci*. **1993**, *33*, 577.
7. Kvasnicka, V.; Skelenak, S.; Pospichal, J. *J. Chem. Info. Comput. Sci*. **1992**, *32*, 742.
8. Anker, L. S.; Jurs, P. C. *Anal. Chem*. **1992**, *64*, 217.
9. Bremser, W. *Anal. Chim. Acta* **1978**, *103*, 355.
10. Robein, W. *Mikrochim. Acta* , **1986**, 2, 271.
11. Chen, L.; Robien, W. *Chemom. Intell. Lab. Syst*. **1993**, *19*, 217.
12. Dubois, J. E.; Carabedian, M.; Dagane, I. *Anal. Chim. Acta* **1984**, *158*, 217.
13. Panaye, A.; Doucet, J.-P.; Fan, B. T. *J. Chem. Info. Comput. Sci*. **1993**, *33*, 258.
14. Munk, M. E.; Lind, R. J.; Clay, M. E. *Anal. Chim. Acta* **1986**, *184*, 1.
15. Xu, J. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 25.

Sample Availability: not available.