Contribution to Structural Elucidation: Behaviours of Substructures Partially Defined from 2D NMR

EPOUHE, Celine[a] **FAN, Bo-Tao**[a] **(范波涛)** **YUAN, Shen-Gang**[*,b] **(袁身刚)** **PANAYE, A.**[a] **DOUCET, J. P.**[*,a]

[a] *ITODYS, Université Paris 7-Denis Diderot, CNRS UMR 7086, 1 Rue Guy de la Brosse, Paris 75005, France*
[b] *Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China*

Structural elucidation (automatic determination of the structure of a molecule from its spectra) is frequently hampered by combinatorial explosion when trying to assemble the identified substructures. We devised a new method which can avoid this pitfall by a systematic examination of allowed ^{13}C chemical shifts ranges for all substructures chemically possible and combined with a progressive pruning thanks to neighbouring relationships appearing from 2D NMR. This method is explained by a detailed example.

Keywords structural elucidation, ^{13}C NMR, 2D NMR, partially defined substructure

Introduction

Computer aided structural elucidation, *i.e.* automatic determination of the structural formula of a compound from its spectra is still an open challenge in spite of numerous attempts.[1,2] The first structure elucidation systems proposed were based on the mass spectroscopy[3] for the determination of structural fragments and the infrared spectroscopy[4,5] for the identification of present or absent features. The breakthroughs of nuclear magnetic resonance (and particularly ^{13}C NMR) made it, about thirty years ago, a tool of choice for the detection of the possible structural fragments, because of the great sensitivity of the resonant carbon to its environment.

The first generation structural elucidation systems[6-10] based on ^{13}C NMR used only 1D data (possibly in conjunction with other types of analytical information), but do not generally provide a general solution, due to the severe under determination of the problem when considering only chemical shift data. Second generation systems (either new or updated issues of existing systems)[11-13] introduced 2D NMR data (such as COSY *etc.*) to determine, from the correlation spots, links between neighbouring atoms, and various applications appeared.[14-20] A little apart from these systems, EPIOS[21,22] approach uses only chemical shifts. But, considering n-uplets of δ values (rather than isolated δ), it attempts to retrieve possible neighbouring relationships, like those detected in 2D spectra. However, these approaches still suffer some limitations or drawbacks:

(1) Structural elucidation systems working on ^{13}C NMR spectra first assign each resonance peak to structural fragments (substructures) corresponding to a central carbon (that the chemical shift of which is identified) and its environment. For practical reasons (the number of fragments to consider), in the substructures retained, the environment of the resonating carbon is somewhat limited (generally one or two bonds from the ^{13}C). This results in the fact that a given fragment is associated to a range of chemical shift values (depending on the nature of the remote environment not identified in the description of the substructure). As a consequence, in the elucidation process, each resonance peak can not be attributed to a unique substructure but to several (perhaps a lot of !) possible fragments. This leads to a combinatorial explosion when trying to assemble the recognised substructures to propose a candidate molecule, a problem already found in other spectroscopy domains.[23]

(2) The system, of course, can only retrieve the substructures stored in its database. If a substructure is not included in this database, it will never be assigned to a resonance peak and the relevant solution will be missed. In other words the efficiency of the system largely depends on the more or less comprehensive extent of the database.

It must be also kept in mind that an elucidation system can not rely only on 2D NMR, on account of the intrinsic limitations of the method. A correlation spot in a 2D spectrum indicates some kind of neighbouring, but the absence of a spot is not unequivocal. It may come from small coupling constants, equivalence from structural symmetry *etc.*, even if protons are neighbours. Interpretation may also be difficult when peaks overlap. To relieve these drawbacks, our objective is thus to propose a process which reduces the number of possible candidate substructures, and is not dependant of the completeness of a database at the same time.

* E-mail: yuansg@mail.edu.ac.cn; doucet@paris7.jussieu.fr
Received April 25, 2003; revised and accepted July 10, 2003.

Method

For developing the system, we therefore worked in two directions. Firstly, building up a knowledge base encompassing all the possible environments of a ^{13}C, and secondly exploiting as soon as possible all the information extracted from 2D NMR in the "assembling/pruning" process in the structure generation.

For building the knowledge base, we determined a list of all the possible substructures that have to be taken into account and evaluated the ranges of possible associated resonance peaks. This evaluation is based not only on the experimental chemical shifts that can be found in a database, but also on "predicted" values for substructures not included in the database, so as to cover all chemically possible environments. In this step we took advantage of our expertise in chemical shift/structure relationships, allowing us to infer chemical shifts ranges rather than deducing them from a database.

Furthermore, in the elucidation process, starting with a gross molecular formula, we will, obviously, systematically try to limit the number of substructure candidates to assembly. In the pruning step, a list of allowed or forbidden substructures is built, using both chemical shift ranges and 2D NMR information (in fact only the most easily available data,[24] from DEPT, $^1H^1H$ COSY, $^1H^{13}C$ COSY, HMBC *etc*.), and the connection table of the candidate molecule will be progressively updated .

In this paper, we focus our attention on the construction and updating of the connection table of the target molecule from substructure identification. Determination of allowed shift ranges associated to all possible substructures will be presented in detail elsewhere.[25] However some examples of these shift ranges, depending on the connectivity of the central ^{13}C, and of its immediate environment will be indicated in the tables of the following section (the resonating carbon is indicated in bold).

Suffices here to say that the retained substructures include only the resonating ^{13}C carbon (as focus) and its first neighbours (α rank atoms) or in few cases the β atoms (as in $H_2C=CH-N$). Bleaching the exact nature of environment atoms may be also introduced (creating equivalence classes). So A means any heavy atoms, Q any heteroatom, Ar any aromatic atom (Csp^2 or N *etc*.).

We will only consider, at that time, molecules containing C, H, O, N, S and the halogens F, Cl, Br, which correspond to the most widespread elements (phosphorous would be easily incorporated considering ^{31}P NMR). Charged structures are not treated. The total number of substructures to consider is indicated in Table 1 according to the connectivity of the ^{13}C.

Basically the process is as follows: starting from the gross formula, and the DEPT spectrum, one determines the number of carbon for each connectivity (CH_3, CH_2, …, C), the number of hetero-atoms and the number of hydrogens not borne by carbons. Additional information may be gained from Infrared spectroscopy, from usual correlation charts.[26]

Table 1 Number of possible structures

Resonating carbon	Number of substructures	
CH_3	6	
CH_2	26 20sp^3 6 sp^2	
CH	60	40 sp^3 10 sp^2 4 sp 6 aro[a]
C	140	49 sp^3 20 sp^2 22 sp 49 aro[a]

[a] aro: aromatic.

Then, for each carbon (depending on its connectivity) all the possible substructures are examined to determine which of them are allowed according to their chemical shift ranges. A second pruning step is carried out using the 2D information provided by HH COSY, and the combination of HH COSY and CH COSY, which gives, at least for hydrogenated carbons, information equivalent to that of an INADEQUATE spectrum.[18] Of course, if an INADEQUATE spectrum is available, it will be directly considered, but this information is generally not available due to sensitivity limitations. This allows for determining links between carbons and eliminating further substructures. And the process is repeated until no more changes occur. At last, the HMBC information is used to determine links in β positions, which is very useful for atoms on both sides by a quaternary carbon or a heteroatom. This process is now automated as to exploitation of COSY spectra and the extension to HMBC spectra is now underway.[27]

For the sake of clarity and conciseness, we chose to exemplify the process on a rather simple example where elucidation can be completely performed, giving a unique solution. In the majority of (more complex) cases, the same methodology will be applied but it will only lead to a (limited) set of extended substructures to assemble, avoiding, or at least significantly reducing combinatorial explosion, the most critical bottleneck in structure elucidation. A large number of other examples may be found in Ref. 27.

Elucidation of a structure from its ^{13}C spectra by using the method

The target molecule is indicated in Fig. 1. The input NMR data, besides ^{13}C chemical shifts, comprises the DEPT information indicating the connectivity of the carbons, the CH COSY spectrum, the combination of HH and HC COSY giving the links between hydrogenated carbons (as in a partial INADEQUATE spectrum) and the HMBC

spectrum indicating links between carbons either in α or β positions. Additional information is extracted from infrared spectroscopy. So, in this example, there is no allenic carbon, no nitro group, but at least one carbonyl group.

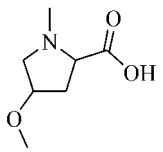
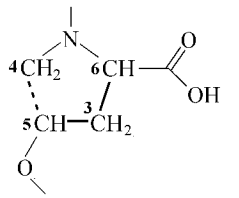


Fig. 1 Structure of target molecule.

From the gross formula and the DEPT spectrum, it appears that one hydrogen atom is borne by an heteroatom (here O or N). This leads to information **[i1]** to **[i7]** (see Table 2). It can also be inferred that **C**H groups **5** and **6** can not be aromatic. This will be indicated as conclusion **[c1]**.



(CH_2)**4** ···· (CH)**5** — (CH_2)**3** — (CH)**6**

C=O

Fig. 2 HH COSY and HC COSY indicate a string (CH)**6**—(CH_2)**3**—(CH)**5**—(CH_2)**4** with single bonds (in bold line) between groups **6**, **3** and **5**, but the nature of the bond between groups **5** and **4** remains undetermined (single or double, in dot line).

As a first step, HH COSY and HC COSY indicate a string (CH)**6**—(CH_2)**3**—(CH)**5**—(CH_2)**4** with single

Table 2 Initial data

Molecular formula: $C_7H_{13}O_3N_1$						
DEPT						
CH_3	**C**H_2	**C**H	**C**	O	N	H
2	2	2	1	3	1	1
CH COSY: correlated C and H atoms						
No.	δ	TYPE	δ_H			
1	49.1	CH_3	3.2			
2	55.0	CH_3	3.4			
3	38.2	CH_2	2.34, 2.66			
4	75.4	CH_2	3.45, 4.02			
5	67.7	CH	4.6			
6	77.8	CH	4.3			
7	170.6	C				
Combination of HH and HC COSY: correlated hydrogenated carbons						
3	CH_2	5	CH			
3	CH_2	6	CH			
4	CH_2	5	CH			
HMBC: a and b (long range) correlations						
1	CH_3	4	CH_2	b		
1	CH_3	6	CH	b		
2	CH_3	5	CH	b		
6	CH	7	C	a, b*		
Initial information						
From NMR				From IR		
[i1] Only one C				**[i5]** No = C =		
[i2] Three O				**[i6]** No NO2		
[i3] One N				**[i7]** At least one C = O		
[i4] Two CH_3						
Conclusion:						
[c1]: CH numbers 5 and 6 can not be aromatic, because of 2CH + 1C + 1N = 5 atoms, insufficient for 6 membered aromatic ring.						

* a, b indicate that the carbons are either in a or b positions by respect to each other.

Table 3 Examination of the shift ranges (δ) for the CH_3 information about neighbours is derived from HH COSY

Resonating carbon (in bold)	δ_{min}	δ_{max}	CH_3 No. **1**	**C**H_3 No. **2**
CH_3—C	−5	50	49.1	55.0
CH_3—O	35	75	0 Neighb.	0 Neighb.
CH_3—N	20	65		

Conclusion:
[c2] only one Me available for (H_3C)—C.

Table 4 Examination of the shift ranges (δ) for the resonating carbon of CH (a): information about neighbours from HH and HC COSY

CH No. **5**: 67.7 2 Neighb. CH_2 CH_2 3 4
CH No. **6**: 77.8 1 Neighb. CH_2 3

CH Resonating carbon (in bold)[a]	δ_{min}	δ_{max}				
CH < C(CH_3)(CH_3)	18	65	**[c2]**	/[b]	/	
CH < CC(CH_3)	3	70		//[c]	/	
CH < CCC	−15	95		Yes	Yes	
CH < CCO	30	115		Yes	Yes	
CH < CCN	10	110		Yes	Yes	
CH < COO	60	145		//	Yes	
CH < CNN	6	145	**[i3]**			
CH < OOO	85	130	**[i2 7]**			
CH < NNN	65	120	**[i3]**			
CH < CON	19	145		//	Yes	
CH < OON	75	129		/	//	
CH < ONN	68	125	**[i3]**			
CH_2 = **C**H—C	107	155		/	/	
CH_2 = **C**H—O	135	160		/	/	
CH_2 = **C**H—N	115	155		/	/	
A—**C**H = CH—A	54	180		//	//	
A—**C**H = C < AA	30	202		//	Yes	→*d*
A—**C**H = C = A	30	202	**[i5]**			
C—**C**H = O	160	225		/	/	→[c6]
O—**C**H = O	145	180		/	/	
N—**C**H = O	145	185		/	/	
C—**C**H = N	105	188		/	/	
Q—**C**H = N	**?**	**?**		//	//	
aro (CA/CA) > **C**H	64	180	**[c1]**	/	/	
aro (CH/CA) > **C**H	85	170	**[c1]**	/	/	
aro (CH/CH) > **C**H	105	160	**[c1]**	/	/	
aro (CA/N) > **C**H	110	190	**[c1]**	/	/	
aro (CH/N) > **C**H	135	180	**[c1]**	/	/	
aro (N/N) > **C**H	145	190	**[i3]**			
CH C—C	60	95		//	//	
CH C—O	20	45		/	/	
CH C—N	40	60		/	/	
32 − 6 = 26						

Conclusions:
[c4]: no **C**H for **C**H_2 = **C**H—A
[c5]: no **C**H for **C**H **C**—
[c6]: no **C**H = O, so a **C** = O

[a] The resonating ^{13}**C** is indicated in bold character. [b] / Forbidden from δ. [c] // Forbidden from the nature of the neighbours. [d] This substructure will be deleted in the next step, on examination of A—HC = **C** < AA, see text for pruning process.

bonds between groups **6**, **3** and **5**, but the nature of the bond between groups **5** and **4** remains undetermined (single or double).

Next step comprises to restrict the allowed substructures on account of possible shift ranges for the different ^{13}C connectivities. Only some steps carried out in the process will be developed here .

So starting from a $\mathbf{C}H_3$ resonating carbon (Table 3) three possibilities are examined ($H_3\mathbf{C}-C$; $H_3\mathbf{C}-N$, $H_3\mathbf{C}-O$). According to the shift observed, for $(H_3\mathbf{C})\mathbf{1}$ the three possibilities remain, whereas for $(H_3\mathbf{C})\mathbf{2}$, with a δ 55, only bonds to O or N may exist; leading to conclusion **[c2]**: only one methyl is available for a fragment $(H_3C)-C$, and excluding, for example, a gem di-methyl group $(H_3C)_2-C$.

The other resonating carbons (from $\mathbf{C}H_2$ to **C**) are then examined and the possible substructures interactively pruned, according to initial information **[i1]** to **[i7]**, previous conclusions (as **[c1]**, **[c2]** *etc*.), and allowed shift ranges. An example is given for CH groups **5** and **6** in Table 4.

For example, for a resonating carbon of **C**H there may exist 32 possible substructures, composed with carbon, nitrogen and oxygen. Some can be excluded from initial information, such as **C**H < NN (**[i3]**, only one nitrogen), others from previous conclusions: **[c1]**, ("no aro-

Table 5 Connectivity matrix from δ ranges, HH and HC COSY

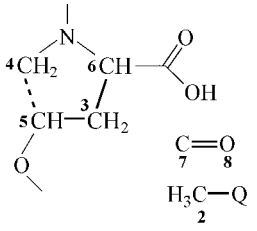
	1 CH_3	**2** CH_3	**3** CH_2	**4** CH_2	**5** CH	**6** CH	**7** C	**8** O	**9** O	**10** O	**11** N	**12** H
1 CH_3		0	0	0	0	0	? [1]	0	? [1]	? [1]	? [1]	0
2 CH_3			0	0	0	0	0	0	? [1]	? [1]	? [1]	0
3 CH_2				0	1	1	0	0	0	0	0	0
4 CH_2					1	0	0	0	? [1]	? [1]	? [1]	0
5 CH						??? [1]	? [1]	0	? [1]	? [1]	? [1]	0
6 CH							? [1]	0	? [1]	? [1]	? [1]	0
7 C								2	? [1]	? [1]	? [1]	0
8 O									0	0	0	? [1]
9 O										0	0	? [1]
10 O											0	? [1]
11 N												? [1]
12 H												

Notes: (1) The carbonyl oxygen is arbitrarily assigned number 8. (2) For each case, 0, 1, 2, …, ? in the upper line indicate surely no bond (0), one bond with its value (1: single; 2: double), or indeterminacy respectively. For the indeterminacy, the lower values between brackets indicate the authorised values. (3) The main diagonal are presented by filled cases in indicating that only the upper triangle matrix is used to represent the connectivity.

matic"), eliminates 6 fragments; **[c2]** eliminates a possible gem-di-methyl group. Others are excluded on account to the nature of their neighbours (such as **C**H < CC(CH_3), which can not match for **C5** with two CH_2 neighbours (// in the Table) or of their shift (/ in the Table) as **C6** (δ 77, superior to the upper border of 70).

From this table it is concluded that fragments CH_2 = **C**H—A are not allowed in this example (from the resonating carbon of **C**H), conclusion **[c4]**. The same is concluded for **C**H—C—C **[c5]** or **C**H = O **[c6]**. These conclusions will be now introduced in the treatment of the other resonating carbons. For instance, re-examining resonating carbon of **C**H_2 (carbons C**3** and C**4**) in a second phase, one can eliminate from **[c4]** fragment **C**H_2 = CH—A. Similarly, for the **C**H carbon number 6, one can use the conclusion previously drawn from the quaternary **C** fragments: no possibility for A—HC = **C** < AA leads to eliminating fragment A—**C**H = C < AA. And the process is repeated until no further updating is possible.

Finally, we obtain the recognized fragments shown in Fig. 3 and a connectivity matrix may be drawn (Table 5). For each case, in the upper line 0, 1, 2, … ?respectively indicate surely no bond (0), one bond with its value (1: single; 2: double), or indeterminacy. For the indeterminacy, the lower values between brackets indicates the authorised values.



Chain>(CH)**6**—(CH_2)**3**—(CH)**5**—(CH_2)**4**—Q
Fragment (CH_3)**2**—Q
Fragment C**7**=O**8** (arbitrarily the carbonyl oxygen is here noted O**8**)

Fig. 3 Fragments recognized after restricted the allowed substructures on account of possible shifts ranges (bonds in bold line are determined at this step).

Next step we use HMBC data in Table 2 to refine the structure. Group (**C**H_3)**1** is connected in β to (**C**H_2)**4** and (**C**H)**6**. A location in α is impossible, since it would have been detected in HH COSY. So, α atom can only be Nitrogen. Indeed, (divalent) O is impossible, as well as (divalent) C = O or OH which would lead to a fully connected structure. Systematic exploitation of HMBC spectra finally leads to the complete structural formula (Fig. 1).

Conclusion

At that time, our home written C++ software, SCOPES,[27] allows for an automatic treatment of shift ranges and 2D COSY data. The incorporation of HMBC data is currently underway.

As a conclusion, we would stress out that combining 2D information and allowed chemical shifts ranges (in a comprehensive exploration of all possible substructures) in iterative pruning steps, allows for limiting the number of putative substructures to combine, and so drastically accelerate the structural elucidation process.

It may be noticed that examination of chemical shift ranges may give useful information for quaternary carbons not involved in COSY spectra. Similarly, HMBC allows to locate long distance neighbourhood, and is particularly interesting when two hydrogenated carbons are located on both sides by a quaternary carbon or an heteroatom.

References

1 Williams, A. *Curr. Opin. Drug Discovery Dev.* **2000**, *3*, 298.
2 Munk, M. E. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 997.
3 Duffield, A. M.; Robertson, A. V.; Djerassi, C.; Buchanan, B. G.; Sutherland, G. L.; Feigenbaum, E. A.; Lederberg, J. *J. Am. Chem. Soc.* **1969**, *91*, 2977.
4 Woodruff, H. B.; Munk, M. E. *J. Org. Chem.* **1977**, *42*, 228
5 (a) Woodruff, H. B.; Smith, G. M. *Anal. Chem.* **1980**, *52*, 2321.
(b) Woodruff, H. B.; Smith, G. M. *Anal. Chim. Acta* **1981**, *133*, 545.
6 (a) Elyashberg, M. E.; Martirosian, E. R.; Karasev, Y. Z.; Thiele, H.; Somberg, H. *Anal. Chim. Acta* **1997**, *337*, 265.
(b) Elyashberg, M. E.; Martirosian, E. R.; Karasev, Y. Z.; Thiele, H.; Somberg, H. *Anal. Chim. Acta* **1997**, *348*, 443.
7 Miyabayashi, N.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18.
8 Munk, M. E.; Christie, B. D. *Anal. Chim. Acta* **1989**, *216*, 57.
9 Will, M.; Fachinger, W.; Richert, J. R. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 221.
10 Elyashberg, M. E.; Blinov, K. A.; Martirosian, E. R. *Lab. Autom. Inf. Manage.* **1999**, *34*, 15.
11 Bradley, D.; Christie, B. D.; Munk, M. E. *J. Am. Chem. Soc.* **1991**, *113*, 3750.
12 Funatsu, K.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 190.
13 Blinov, K. A.; Elyashberg, M. E.; Molodtsov, S. G.; Williams, A. J.; Martirosian, E. R. *Fresenius' J. Anal. Chem.* **2001**, *369*, 709.
14 Chen, P.; Yuan, S. G; Zheng, C. Z; Hui, Y. Z. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 805.
15 Chen, P.; Yuan, S. G; Zheng, C. Z; Hui, Y. Z.; Wu, H. M.; Ma, K. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 814.
16 Peng, C.; Yuan, S. G.; Zheng, C. Z.; Shi, Z.; Wu, H. M. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 539.

17 Peng, C.; Bodenhausen, G; Qiu, S. X.; Fong, H. S.; Norman, R.; Farnsworth, N. R.; Yuan, S. G.; Zheng, C. Z. *Magn. Reson. Chem.* **1998**, *36*, 267.
18 Nuzillard, J. M.; Massiot, G. *Anal. Chim. Acta* **1991**, *2427*, 37.
19 Steinbeck, C. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 1984.
20 Lindel, T.; Junker, J.; Koeck, M. *Eur. J. Org. Chem.* **1999**, *3*, 573.
21 Carabedian, M.; Dagane, I.; Dubois, J. E. *Anal. Chem.* **1988**, *60*, 2186.
22 Dubois, J. E.; Carabedian, M.; Ancian, B. *Comptes Rend. Acad. Sci.* (Paris), **1980**, *290*, 369 and 372.
23 Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Roberston, A. V.; Duffield, A.; Djerassi, C. *J. Am. Chem. Soc.* **1969**, *91*, 2973.
24 Jaspars, M. *Nat. Prod. Rep.* **1999**, *16*, 241.
25 Epouh, C.; Fan, B. T.; Yuan, S. G.; Panaye, A.; Doucet, J. P. to be published.
26 Lin-Vien, D.; Colthup, N. H.; Fateley, W. G.; Graselli, J. G. In *The Handbook of Infrared and Raman Characteristic Frequencies of Organic Molecules*, Acad. Press, **1991**.
27 Epouhe, C. *Ph. D. Thesis*, University Paris 7, Paris, **2002**.

(E0304251 ZHAO, X. J.)

{ } [c4]

{ } [c4]

{ } [c5]

{ } [c5]