中国科学
1992 年 3 月　　A 辑　　第 3 期

C-13 NMR 拓扑模型在有机化合物结构解析中的应用

袁身刚　彭琛　郑崇直
(中国科学院计算机化学开放实验室，上海有机化学研究所，上海 200032)

摘　要

本文介绍了一种利用计算机从 C-13 NMR 谱数据自动推断简单化合物结构的方法. 该方法基于子结构/部分光谱拓扑相关模型从未知物分子式及 C-13 NMR 数据推导出候选子结构以及对拼装这些子结构的约束条件，然后在这些约束条件之下利用部分重叠的方法将这些子结构拼装成候选结构. 解析过程中每一步推导到的子结构或结构都经过评估，以便及时删除与已知数据相矛盾的(子)结构. 评估的主要方法是模拟候选(子)结构的(部分) C-13 NMR 谱并与已知数据进行比较. 解析过程是自动进行的，同时使用者可以根据自己的经验和直觉对推理过程加以干预以提高解析效率. 对文献上报道的 247 个单官能团化合物作为未知物进行解析，成功率为 97.2%.

关键词：计算机辅助结构解析，谱图模拟，C-13 NMR

由有机化合物的性质，主要是它的光谱性质，推断其化学结构是有机结构分析中最为复杂的工作，至今主要还是运用化学家的经验和直觉的启发式方法. 以多达 224 个氨基酸残基组成的天花粉蛋白为例，其不对称中心成百上千，结构的确定花费了化学家几年的时间[1]. 随着计算机技术的飞速发展，人们自然而然地希望能借助计算机的庞大存贮容量和快速逻辑运算能力来帮助完成复杂化合物的结构解析. 早在 60 年代初，Stanford 大学就开始发展 DENDRAL[2] 系统，开创了后来人们称之为专家系统的人工智能中最活跃的一个领域.

目前最常用的计算机结构解析方法可分为三大类：查索型、模式识别和子结构/部分谱方法. 这些方法虽各有优缺点，但比较起来用得最为成功的还是基于子结构/部分谱关系的方法. 最近，我们在 C-13 化学位移的拓扑模型的基础上发展了一个由 C-13 NMR 数据推断化合物结构的计算机系统，我们的目标是对于较简单的未知物系统能迅速推导出唯一的正确结果；而对复杂的未知结构能推导出所有的与已知的化学和光谱数据相符的候选结构，为化学家进一步探索和推导未知物结构提供辅助线索.

本文 1990 年 11 月 22 日收到，1991 年 6 月 3 日收到修改稿.

一、拓 扑 模 型[3]

在这种模型中 C-13 共振峰的化学位移被看成是共振核和与它相距三根键以内的原子相互作用的结果。该碳核称为“中心碳”，而对它有作用的其它原子组成“环境”。环境又根据中心碳原子的连接度而分成若干子环境。每个子环境又看成由碳骨架和官能团所组成。官能团被看成一个特殊的原子，称为超原子。如果子环境中没有官能团，它称为烷基子环境，即把官能团看成甲基的子环境。在这个模型中不同的拓扑子环境对中心碳原子化学位移的影响是用它对应的拓扑参数 λ_R 来定量描述的，而且这些拓扑参数具有加和性[4,5]；而中心碳对环境的影响的敏感度 ω 只决定于它的连接度。这些特点使得整个模型十分简单，中心碳的化学位移 δ 可用有如下形式的公式来计算：

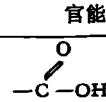
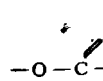
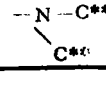
$$\delta = \omega \Sigma \lambda_R + \eta,$$

表 1　C-13 子结构拓扑模型的 32 种骨架*

编号	骨架	编号	骨架	编号	骨架	编号	骨架
1		9		17		25	
2		10		18		26	
3		11		19		27	
4		12		20		28	
5		13		21		29	
6		14		22		30	
7		15		23		31	
8		16		24		32	

*● 表示骨架碳原子，X 表示表 2 中的 15 种官能团中的任意一种.

表 2　C-13 NMR 子结构拓扑模型的 15 种官能团

编号	官能团*	编号	官能团	编号	官能团
1	—F				
2	—Cl	9		12	
3	—Br				
4	—I				
5	—OH	10		13	
6	$—NH_2$				
7	—NH—C**	11		14	=C**
8	$—CH_3$			15	—C≡C**

* 烷烃的“官能团”看成是$—CH_3$.

** 表示该原子的连接度未确定.

其中 η 是常数。目前已定义了由 32 种骨架(表 1)和 15 种官能团(表 2)所组合的 400 多种子环境的拓扑参数,只要确定中心碳周围三根键内的环境的组成,就能方便地运用这一模型计算它的 C-13 NMR 化学位移。

二、解析方法

本方法从未知物的元素组成和 C-13 NMR 谱数据(包括峰的化学位移和多重性)开始推导未知物的可能结构,推导过程主要包括图 1 所示的四个部分,分别讨论如下。

1. 子结构推导

我们的方法是枚举这 15 种官能团和 32 种骨架分别在以伯、仲、叔或季碳为中心碳时组合成的所有可能的子结构,同时用模型方法计算子结构的中心碳原子的化学位移和多重性,并将它们与未知物的已知谱图的所有共振峰进行比较。将在一允许的偏差之内(如多重性相同,$|\Delta\delta| \leqslant 2.0$ppm)的子结构作为共振峰的候选子结构。如果与所有的共振峰的偏差都大于允许偏差,则抛弃之。

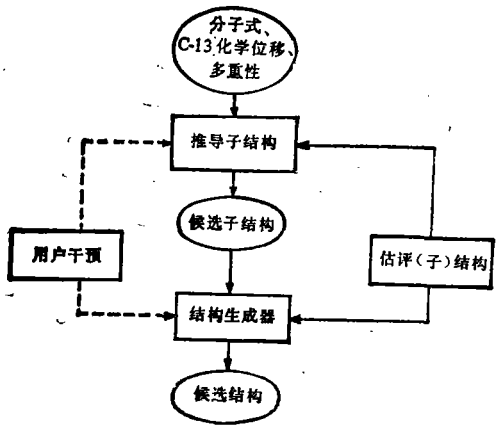


图 1 结构解析方法示意图

由于共振峰和候选子结构之间是"一对多"关系,在而后的子结构拼装过程中会导致"组合爆炸"。因此,如何提高"好子结构率[1)]",尽早地将子结构限制在一个合理数目之内,使后续的拼装能顺利进行成为本方法的关键问题之一。为此,我们在约束条件下推导子结构,即参与组合的每个骨架和官能团以及可能的子结构都要经过与这些约束条件的一致性检验,从而提高好子结构率。这些约束条件为:

筛选骨架和官能团时的约束条件:

(1) 骨架或官能团的各类原子数目≤未知物的各类原子数目;

(2) 官能团的不饱和度≤未知物的不饱和度;

(3) 骨架或官能团要求的 C-13 NMR 共振峰重数与已知谱图相符;

(4) 官能团的活泼氢(即不与碳原子相连的氢原子)数目≤从谱图推得未知物的活泼氢数目;

(5) 官能团中碳原子的化学位移范围与已知谱图相符合;

筛选候选子结构的约束条件:

(6) 子结构的各类原子数目≤未知物的各类原子数目;

(7) 从子结构模拟的部分 C-13 NMR 谱与已知谱图相符合,

其中(6)和(7)实际上是由后面所述的评估程序完成的。

表 3 通过一个例子说明了上述约束条件对候选子结构数目的影响。

---

1) 好子结构率定义为在子结构拼装过程中参与拼装的子结构的数目与所有的候选子结构数目之比。

表 3　不同约束条件下一个结构的各个峰的候选子结构的数目*

共振峰 \ 约束条件	0	1-1	1-2	1-3	1-4	1-5	1-6	1-7
13.9(q)	70	32	29	9	7	4	3	3
60.0(t)	21	6	6	5	3	2	1	1
173.5(s)	0	0	0	0	0	0	0	0
34.4(t)	272	100	94	28	14	10	6	3
24.8(t)	160	70	62	17	7	5	3	1
31.5(t)	252	104	95	43	27	18	9	6
22.4(t)	103	47	41	15	8	5	4	3
14.3(q)	72	34	32	9	7	4	3	3

* 未知物分子式为 $C_8H_{16}O_2$，结构为己酸乙酯．第二列列出了当仅用两个最基本约束条件时各个共振峰对应的候选子结构数目，这两个约束条件是：(1)推导出来的子结构中的碳原子数不应大于未知物中碳原子数目；(2)子结构的中心碳的模拟化学位移与实验值之间的差别不大于 3.0ppm．第三列至第九列分别对应于将正文中的第(1)至第(7)个约束条件加上去时推导得的子结构数目．

2. 结构生成

在给已知谱图中每个峰推导得一组可能的子结构之后，采用什么样的方法将这些子结构拼装成完整结构将是决定解析效率的关键．我们采用了子结构部分重叠的方法引导子结构的拼装，因为每个候选子结构代表了中心碳原子的 α 至 γ 层取代环境，作为目标结构中相邻两个碳原子的正确子结构的必要条件是它们能部分重叠．

首先从候选子结构集中选出一个子结构作为种子子结构（记为 SS0），SS0 用来作为结构组装的初始母体，与其它候选子结构进行部分重叠扩张．为了保证结构生成器的穷尽性，我们采用了类似于 Munk[6] 等的搜索树来表达拼装过程．如图 2 所示，从根结点 SS0 出发按深度优先周游搜索树：对某一层的一个子结构 SS1（如果是根结点，则为 SS0），搜索所有候选子结构 SS2 并与之根据具体情况做骨架或官能团重叠，如果重叠成功，将 SS1 的结构按 SS2 的结构复制，得到一个扩张了的子结构 SS3．SS3 必须经过后面所述的评估程序进行鉴定，SS3 经鉴定后可能有三种结果：

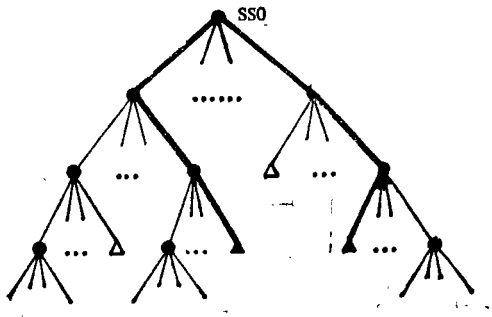


图 2　结构拼装过程的搜索树表示
（●代表 SS0 或者 SS1，△表示与已知数据矛盾的 SS3，▲表示正确的结构，图中的粗线表示成功的拼装路径）

(1) 与已知数据不相矛盾的子结构；

(2) 与已知数据不相矛盾的完整结构；

(3) 与已知数据相矛盾的子结构或完整结构．

对于(1)，将 SS3 进行规范化处理（处理方法见参考文献[7]），并与已记录的所有 SS3 进行比较．如果未曾记录过，则在搜索树中向深度走一步，即将 SS3 作为新的 SS1，搜索 SS2 与之进一步重叠扩张，并记录 SS3 的规范化连接表以便将来比较；如果 SS3 是一个重复的子结构，那么对它不再处理，因为相同的结点在搜索树中将导致完全相同的子树．对情况(2)，将该结构与已记录的候选结构进行比较，记录不重复的结构；然后与(3)一样，SS1 不变，在同一层

中继续向右搜索 SS2 进行重叠扩张.

在搜索树的某一层,当 SS1 的所有可能的重叠中心与所有候选子结构 SS2 都重叠完毕后,在搜索树中回溯到上面一层,即恢复前一个母体子结构 SS1,并在该层中搜索与它未曾重叠过的 SS2 进行重叠扩张. 当试图从搜索树的根上回退时,则对 SS0 的重叠扩张过程结束.

结构生成器将一个峰的所有子结构作为 SS0 进行上述拼装后,如果仍不能得到满意的候选结构,就挑选另一个峰的子结构集作为 SS0 集,重复上述过程,直到得到满意的候选结构或所有的子结构集都已试验过为止. 显然,SS0 集的选择直接影响到结构生成的效率,因此,我们确定了选择 SS0 集的优先规则:

(1) 数目最小的候选子结构集优先;

(2) 中心碳原子的连接度高的子结构集优先.

子结构拼装过程中的重叠方式分成骨架重叠和官能团重叠两种,所谓骨架重叠就是以 SS1 的中心碳原子 FO1 的 α 或 β 骨架碳原子为重叠中心 OP,与 SS2 的中心碳原子 FO2 进行重叠(见图 3(a)). 如果 OP 和 FO2 的连接度相同,且它们周围的对应的原子和键的性质都不相矛盾,则认为重叠成功(注意官能团被看成一个超原子). 这一判断过程实际上是用分别对 OP 和 FO2 的 $\alpha-\gamma$ 层环境进行规范化编码实现的[7]. 骨架重叠根据重叠中心是处于 FO 的 α 还是 β 位分为 α 骨架重叠和 β 骨架重叠. β 重叠的设置是因为对某些峰的推理可能是"哑"的,即未能找到准确的候选子结构,在这种情况下,β 重叠有可能跳过"哑"峰而推断出最终的结构. 是否进行 β 骨架重叠可由用户设定.

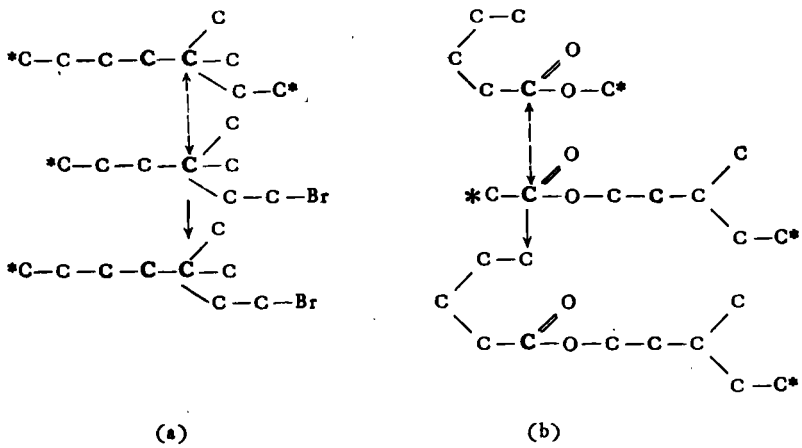


图 3 (a) α 骨架重叠和(b)官能团重叠示意图

(粗写的碳原子是子结构的中心碳原子,带 * 号的原子的连接度未确定)

当 SS1 含有可扩张的官能团时,必须进行官能团重叠,因为在骨架重叠时,官能团是被看成"超原子"而不加改变的. 官能团重叠时的条件是官能团相匹配且为官能团的连接度所允许,如官能团—CO—O—C* 可与两个骨架相连,可以与它重叠的官能团是—O—CO—C*;而—Cl 的连接度只能为 1,不能进行官能团重叠. 当官能团重叠成功时,将 SS2 的部分结构与 SS1 的官能团的"自由价键"相连,得到一个扩张了的(子)结构(图 3(b)).

3. (子)结构的评估

在子结构推导和结构生成过程中,每个子结构,都必须经过评估,以确定它是否与已知数

据相矛盾、是否是完整结构．首先将它的元素组成与分子式的相比较．假如某种原子的数目超过分子式的，它将被抛弃；假如其组成已经与分子式的相同，就作为完整结构（连接度未确定的原子补氢）进行评估；否则作为子结构进行评估．对通过元素组成比较的（子）结构的进一步评估主要是模拟（部分）C-13 NMR 谱图，与原始谱进行比较．

模拟 C-13 NMR 谱时，首先识别并比较组成每个碳原子的环境的骨架和官能团[7]，估计其 C-13 峰数，并利用模型方法计算其化学位移，根据连接度确定峰的多重性，将预测的谱图与实际谱图进行比较．在设定的阈值（可以由用户修改）内通过的（子）结构被认为是与已知数据相一致的（子）结构，否则是相矛盾的（子）结构．用更严格的阈值还可以判断完整结构是否"令人满意"．

4. 用户干预

用户往往对未知物的背景知识或部分结构有一定的了解，利用这些已知知识通常会大大提高解析效率，为此我们采用人机交互的解析方法，以便使用者根据自己对未知物的了解和直觉，对解析过程加以控制，以提高解析效率．用户的干预包括如下两方面的内容：

（1）允许用户用输入字符编码的方法输入已知的子结构，包括用户认为是可能存在、肯定存在或者不可能存在的三种类型（编码方法见文献[7]）．对这些用户说明的子结构，系统都将检查是否与已知数据矛盾，然后根据它们对候选子结构集进行修改．

（2）允许用户修改约束条件，包括推导子结构时允许的化学位移偏差，评估（子）结构时判断模拟谱和实验谱是否一致的阈值（如模拟 C-13 NMR 谱与实验谱图之间的化学位移、共振峰数和峰多重性的最大允许偏差等等），以及其他一些可选条件，如在结构生成时是否进行 β 骨架重叠等．用户可以通过改变这些条件以控制解析过程的深度和速度之间的平衡，直到满意为止．

三、结果与讨论

所述方法已用 FORTRAN 77 编程，并在 VAX11/780 计算机上调试通过．我们以在文献上找到的 247 个含 3—16 个碳的单官能团开链化合物的 C-13 NMR 数据作为测试实例，结果见表 4，总的成功次数为 240，成功率为 97.2%．对绝大多数实例在 CPU 时间 2min 以内能得到唯一的正确结果．失败的主要原因是当化合物结构过份拥挤时，模拟的化学位移和实验结果偏差增大．其他的测试结果表明当多个官能团相距较远（大于 6 条键时），这一程序也是适用的．

我们的方法在解释 C-13 NMR 数据时利用了已建立好的 C-13 NMR 拓扑模型，它与简单地利用化学位移/子结构对应表相比，浓缩了需要处理的信息，因此速度更快，贮存空间也更省．采用的约束条件在大部分情况下都能成功地将子结构数目限制在合理的范围之内，从而避免了结构拼装时可能出现的"组合爆炸"．在结构生成时所采用的新颖的 α，β 骨架重叠和官能团重叠相结合的重叠拼装方法较好地解决了由候选子结构集生成所有可能结构的问题．在（子）结构的评估和筛选中，同时采用了比较化学元素组成和比较模拟谱的方法，避免了需要装有大型子结构/化学位移的数据库，使得整个系统有可能在计算能力和贮存能力都比较差的微机上实现．

尽管本方法基于的子结构模型适用多官能团及某些环状化合物[3]，但作为一个初始的版

表 4 对一些单官能团开链化合物的测试结果

类型	未知物总数	碳数范围	成功次数	CPU 时间 (min)	成功率(%)	文献
羧酸	42	3—16	41	0.02—94.2	98	[8,9]
醇	15	5—12	15	0.02—1.5	100	[10]
伯胺	27	2—10	27	0.02—4.3	100	[11]
仲胺	16	4—9	16	0.02—1.5	100	[11]
叔胺	8	4—12	6	0.02—0.4	75	[11]
酯	31	3—15	27	0.02—4.1	87	[12]
烯烃	70	5—12	70	0.02—2.9	100	[13]
炔烃	6	5—7	6	0.02—0.2	100	[14]
酮	7	5—11	7	0.02—2.0	100	[15]
卤代烷	17	3—10	17	0.02—0.03	100	[16]
烷烃	8	6—13	8	0.02—0.06	100	[16]

本，我们的系统在推导候选子结构时只考虑了含单官能团的子结构，在结构生成器中，只考虑了非环化合物的生成，而且由于该模型目前只定义了有限的官能团和骨架所组成的取代环境的拓扑参数，因此限制了方法的适用范围。除了这些方面仍需改进之外，增加同其他(如红外、质谱等）谱图的联合解析以及处理立体化学的能力也是我们正在努力改进的一些重要方面。

徐越华给本工作提供了结构显示模块，陈维明对结构的规范化处理方法提供了不少有益的帮助，张伟琪完成了全部的测试工作，在此一并表示衷心的感谢。

参 考 文 献

[1] 顾梓伟等，化学学报，42(1984)，943.
[2] Duffield, A. M. et al., *J. Amer. Chem. Soc.*, 91(1969), 2977.
[3] Yuan, S. G., *These de Doctorat D'Etat es Sciences Physiques*, Universite de Paris VII, France, 1984.
[4] Dubois, J. E. & Doucet, J. P., *J. Chem. Res.*, (M) (1980), 1101, 1130.
[5] Dubois, J. E. & Carabedian, M., *Org. Magn. Reson.*, 14(1980), 264.
[6] Lipus, A. H. & Munk, M. E., *J. Chem. Inf. Comput. Sci.*, 25(1985), 23.
[7] 彭 琛、袁身刚、郑崇直，科学通报，35(1990)，1241.
[8] Marr, D. H., *Org. Magn. Reson.*, 13(1980), 29.
[9] Couperus, P. A. et al., *Org. Magn. Reson.*, 11(1978), 590.
[10] Ejchart, A., *ibid.*, 9(1977), 352.
[11] Eggert, H., Djerassi, C., *J. Amer. Chem. Soc.*, 95(1973), 3710.
[12] Ferent'ev, A. B. et al., *Org. Magn. Reson.*, 9(1977), 301.
[13] Couperus, P. A., et al., *ibid.*, 8(1976), 426.
[14] Breitmaier, E. & Voelter, W., *Carbon-13 NMR Spectroscopy*, 3rd ed., VCH Verlagsgesellschaft mbH, D-6940 Weinheim, 1987.
[15] Johnson, L. & Jankowski, W. C., *Carbon-13 NMR Spectra, A Collection of Assigned, Coded, and Indexed Spectra*, John Wiley & Sons, Inc., New York, 1972.
[16] Breitmaier, E. et al., *Atlas of Carbon-13 NMR Data*, Heyden & Son Ltd., London, 1975, 1976.