分子结构上的偏序关系与结构感知算法*

张懋森　　徐　峻

（中国科学技术大学结构中心，合肥 230026）

关键词　偏序关系、结构感知、结构匹配、结构通式的识别

计算机辅助分子设计、合成设计是现代化学研究的最重要内容．要让计算机来辅助化学设计，就必须让它理解分子图形，包括图在计算机内部的表达（矩阵或关联表）及外部表达（二维图形、或线型编码）、认知特定的化学结构（如：有合成意义的子结构、环等）．

本文在建立分子结构的偏序关系基础上，试图突破这个限制计算机辅助化学设计以及结构数据库发展的一大障碍[1,2]．引入人工智能技术中的探试-回溯机制，破除 A-A 匹配算法中传统的穷举性，从而在 VAX11/785 机和 IBM PC 上实现了一组效率很高的感知分子结构的算法，并解决了 Markush 结构的表达与匹配问题，实现了合成设计系统的分子图形学系统、^{13}CNMR 波谱模拟系统及分子结构发生器．

一、偏序关系和结构同构与同态

结构的同构与同态又称 A-A 匹配（Atom-by-atom match），在图论中属 NP-完全性（NP-completeness）难题．例如，在图 1 中，要问查询图 QG 是否为超图 SG 的子图，就是要找从 QG 到 SG 上的映射，使得 QG 中在每一个节点都在 SG 中有对应的象．在最坏的情形下，设N是 QG 的节点数．M 是 SG 的节点数，有可能要检查 $N!/(M-N)!$ 个映射[2]，这就是"组合爆炸"的原因．



图 1

正如 Balaban 等人[3]所指出，造成这种情况的本质是 QG 在匹配到 SG 上时节点上的序的归一性已失去，亦即子图在 SG 中所处的化学环境发生变更．虽然 SG 与 QG 之间常常没有共同的拓朴线性序，但是却有共同的偏序关系（Partial ordering）存在，因为这是它们同构或同态的基础．行走序（Walking order）就是一种偏序．自结构图的任意一点出发，遍历图上所有的边，叫做对图的一次行走，行走时，按节点被访问的先后次序依次给各节点定

本文 1989 年 4 月 7 日收到．1990 年 8 月 17 日收到修改稿．

* 中国科学院结构分析开放研究实验室基金资助课题．

序，这个序就是行走序．如果 QG 与 SG 同构或同态，那末，它们必然有相同的偏序关系．这种偏序关系就建立起 QG 到 SG 上的映射关系．

同构与同态算法可分两步进行，第一步，在 QG 上获得一个反映 QG 的图论特征的偏序集；第二步，在超结构 SG 上进行探试回溯寻求在 SG 上是否也存在这种偏序关系．算法 HBA 实现这种探试-回溯而完成结构匹配过程，并根据需要给出同构或同态映射的象．

有了上述两个算法，可使计算机具有各种感知化学结构的能力．结构同构与同态的探试回溯算法的数学分析及运行效率已在文献[5,6]中作了详细讨论．

二、各种匹配算法

在化学结构数据库及专家系统的具体应用中，对匹配的要求是各种各样的．在数据库查询中，只要知道一组超结构中有哪些含有子结构 QG，而对匹配的节点间的具体映射并不关心，关键在于速度要快，这叫子结构搜索；在分子图形建造系统中，要找到一组最基本环的基本集合，即要找出多个环的映射，而映射表中的节点不同时出现在两个以上映射关系中，称非冗余性多重匹配．因算法 HBA 已提供了丰富的入口参数[13]，故可通过对入口参数的不同引用来满足各种匹配要求．

三、结构感知算法

结构感知问题内容十分丰富[7]，但主要以图的同态匹配为基础．本文讨论感知分子对称性、手征性、拓朴变换结果、结构生成、基本环集、分子模型建造、合成设计等方面的感知算法．

1. 感知分子的对应性　在合成设计中，要计算机认出分子是关于哪个化学键对称的，并且，将该键切断之后，分子将分为相等的两部分．

2. 感知不对称碳原子　可以用基于偏序关系的同构算法来找出复杂分子中的不对称碳原子．算法基本思想是，与不对称碳原子相联结的四个分支子结构互相不具有公共的偏序关系．因此算法 CHIRAL 这样来寻找不对称碳原子：分别扫描分子中所有的四度节点，取出与其相联结的四个分支子结构 A、B、C、D，用 HBA 检查，如果A和 B，A 和 C，A和 D，B 和 C，B 和 D，C 和D均不同构，那么，该节点就是不对称中心．

3. 感知基本环集合　在合成设计与分子三维模型的自动建造系统的研制工作中，需要计算机认识出分子中含有哪些环结构．定义一组基本环集合，包括基本的螺、桥环及单环，在分子图形学系统中，我们认为复杂的环都是由这些基本环组合生成的．这些基本环集均以邻接关系表及偏序关系的形式存贮在知识库之中．

4. 拓扑操作　化学反应的本质就是旧化学键之断去与新化学键之生成．在图论观点看来，就是执行拓扑操作，包括：切断边、修改节点，边的颜色、把其它子图接到指定的节点上．在合成设计软件系统中，计算机是通过修改邻接关系来完成拓扑操作的．执行拓扑操作之后，原来的邻接关系表就可能表达的不是一个连通图，而是一组连通图，计算机必须认出它表示了哪几个分子，并产生它们各自的邻接关系表，以作下步合成设计之用．算法 FIND-CONNECTED-BRANCH 寻找结构图的一个连通分支；算法 FIND-A-SET-OF-GRAPH：从非连通结构图中找出所有连通子结构图且建立其邻接表，它们对于计算机辅助合成设计研究是十分重要的．

5. 结构生成器　在计算机输助结构解析研究工作中，由波谱子图—结构子图相关表，可

以推断出未知物结构可能有的子结构碎片集，在这子图集中各集合元素之间互相拼接，生成一个结构集合，未知物的结构就在其中．这就是结构生成问题．结构生成问题可简述为：给定一子图集 S，通过 S 中各元素间的按拓朴及化学原则组合，产生一个图集，该图集中各元素均无可拼接的位置．为此，要把每一个子图上可拼接位置及色性表达出来．我们对拓朴邻接表予以扩充，产生了此类结构表达方法．

6. 结构通式（Generic structure）的表达与感知　在标准化学文献中，一般用 Markush 结构式来表达结构通式[8]，Markush 结构式是具有某种公共特性（包括公共的结构框架和取代基之间的约束关系）的一类结构图．Markush 结构的感知问题可以化为多次使用同构或同态匹配的算法．在合成设计中，让计算机选用的合成设计的规则都必须是通式，因此，实现 Markush 结构的匹配算法就至关重要．我们首先用形式语言嵌入结构的线型编码来表示 Markush 结构，在此基础上建立了匹配算法．

四、结果与讨论

本文所提出之算法均在 VAX-11/785 小型计算机上用 PASCAL 语言实现．这里列出若干运行结果来说明本算法的效率．

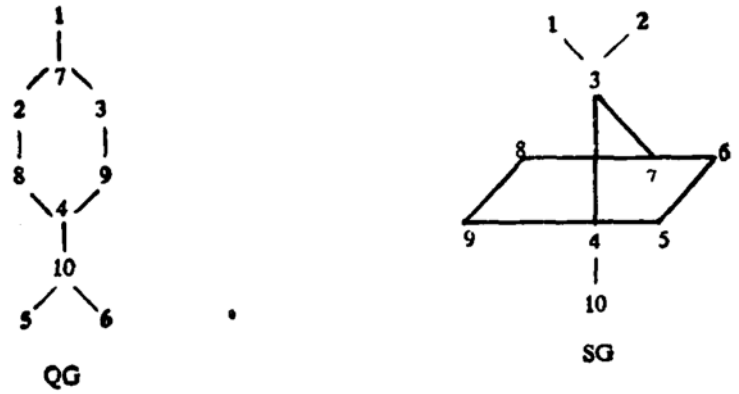
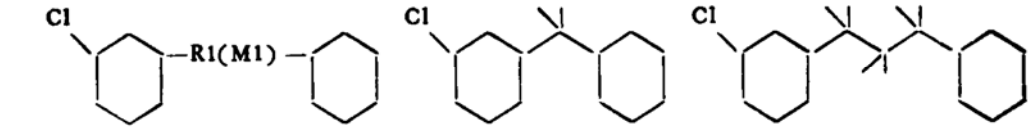


图 2

1. 冗余性多重匹配映射　形如图 2 的图匹配问题，Wipke 等人的并行算法在DEC-2060系统上求得 QG 在 SG 上的四重冗余性映射，用了 CPU 时间 14.45s[2]，最长的任务进程耗时 0.688s．本文在 DEC VAX 11/785 机上的结果为0.02s．所用的化学结构线型编码是作者设计的[10]．

2. 分子模型自动建造系统中的环感知　为建造血卟啉分子模型，必须让计算机"认出"分子中所有的基本环，然后运用图拼接技术及化学约束，计算出它的三维构象．作者主持实现的 USTC-分子图形学系统为感知该分子所有的基本环集用了 DEC VAX-11/785 CPU 时间为2.57s．其效率远远超过了 Wipke 教授等人实现的 AIME 系统[11]．

3. 通式匹配　如图 3 所示，要检查 QG1 和 QG2 是否属于 Markush 结构所表达的分子，正确的答案为：QG2 是，而 QG1 不是．Schalley 提出的松弛（Relaxation）算法分别花 0.66s 和 0.79s 的 CPU 时间，而且对 QG1 还错误地指出其匹配[12]．本文结果得出完全正确结果，分别用 CPU 时间为 0.01 和 0.02s．此处作者所用 CPU 与 Scholley 的相同．



M1 = [3] R1 = —C(C)(C)—　　QG1　　QG2

Markush 结构式　　查询结构

图 3

结构感知算法关键在于同构与同态匹配算法。由于本文的算法也要从一个节点出发作探试回溯行走，故易于在有并行处理能力的计算机上实现并行算法，这时，速度还会提高。

参 考 文 献

[1] Stobaugh, E. R., *J. Chem. Inf. Comput. Sci.*, 25(1985), 271—275.

[2] Wipke, W. T., Rogers, D., *J. Chem. Inf. Comput. Sci.*, 24(1984), 255—262.

[3] A. T. Balaban 著，金晓龙、陈志鹤译，图论在化学中的应用，科学出版社，1983，362—364.

[4] Xu Jun (徐峻), Zhang, M. S. (张懋森), VIII ICCCRE, CCS(1987), A-24.

[5] Xu, Jun (徐峻), Zhang, M. S. (张懋森), Xu, Y. J., Representation of Nondeterministic Graphic Knowledge & A Linear Notation for Graphs, Mini & Microcomputers and Their Applications. *The International Society for Mini and Microcomputers* (*ISMM*) (Ed. E. Luque), 261.

[6] 徐峻、张懋森，计算机辅助有机合成设计的原理，化学通报，1989，8: 58—64.

[7] Fujita, S., *J. Chem. Inf. Comput. Sci.*, 28(1988), 1—9.

[8] Lynch, M. F., Barnard, J. M., Welford, S. M., *ibid.* 25(1985), 264—270.

[9] 徐峻，中国科学技术大学博士学位论文，1988.

[10] Xu Jun (徐峻), Zhang, M. S., Xu, Y. J., *ISMM*, 35(1986).

[11] Wipke, W. T., Hahn, M. A., *Artificial Intelligence Applications in Chemistry*, ACS 306 (1986), 136—146.

[12] Scholley, A. V., *J. Chem. Inf. Comput. Sci.*, 24(1984), 235—241.

[13] Xu Jun, Zhang, M. S., *Tetrahedron Computer Methodology*, 2 (1989), 75—83.