2002 年第 60 卷
第 7 期, 1291～1297

化 学 学 报
ACTA CHIMICA SINICA

Vol. 60, 2002
No. 7, 1291～1297

化学结构中互变键、交替键和芳香键的自动识别

姚建华　李　丰　罗时玮　袁身刚*　陈海峰　李　强　郑崇直*

(中国科学院上海有机化学研究所　中国科学院计算机化学开放实验室　上海 200032)

摘要　化学结构计算机处理中最常遇到的困难是互变现象、交替键和芳香键的处理. 尽管解决这些问题的方法早有报道, 但它们都有仅考虑计算机处理的方便, 而很少注意其化学应用的不足. 本工作在环系识别算法的基础上, 设计了新的识别算法, 使得识别的整体性能更好, 形成了拥有自主知识产权的软件. 简要介绍了这些算法, 并通过例子说明了它们的可能应用.

关键词　互变现象, 交替键, 芳香键

Automatic Identification of Tautomeric, Alternating and Aromatic Bonds in Chemical Structures

YAO, Jian-Hua　LI, Feng　LUO, Shi-Wei　YUAN, Shen-Gang*
CHEN, Hai-Feng　LI, Qiang　ZHENG, Chong-Zhi

(Key Laboratory of Computer Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032)

Abstract　The most usual problems encountered in the structure handling are tautomerism, alternating and aromatic systems. Though a number of methods have been reported in the literature, a common disadvantage is that only the convenience for processing were taken into account, but their chemical sense was rarely considered. Thus their application is limited. Based on the identification of ring system, novel algorithms have been devised in order to enhance the overall performance and a software with copyright has been developed. This paper introduces these algorithms and their potential applications by examples.

Keywords　tautomerism, alternating system, aromatic system

化学结构作为化学家使用最普遍的语言, 在化学知识的记录、传播和交流中的核心作用已为大家所认识. 随着计算机的普及, 广大化学工作者都希望能得到计算机的帮助, 实现这一愿望的首要困难就是化学结构的计算机处理. 这些困难主要有两类: 非确定 (non-specific) 结构和歧义化学结构的处理问题. 本文将就歧义化学结构中重要的交替键、芳香键和互变键的计算机处理问题介绍我们的最新研究成果. 尽管美国化学文摘社 (CAS)[1]、DRAC 系统[2]和 CAMEO 系统[3]等曾报道过类似工作, 但是它们或多或少都有某些不足. 例如, Shelley[4]就曾指出"无机、金属有机和互变异构的化学结构至今没有有化学意义的结构表达方式". 对于 CAS 只考虑计算机处理的方便而不顾化学应用的实际需要, 更是提出了批评:"像 CAS 目前采用的将互变异构体'规范化'的方法, 并不代表真实情况, 不是一个令人满意的解决办法"[4]. 但是, Shelley 的这个意见似乎并未得到重视, 因为八十年代末期以来几乎不再有关于歧义化学结构计算机处理方法的报道. 我们在建立新一代

* E-mail: yuansg@pub.sioc.ac.cn

Received November 23, 2001; revised January 30, 2001; accepted March 14, 2002.
973 计划 (No. G1998051115) 和国家自然科学基金 (Nos. 29832050, 29872048, 20073058) 资助项目.

分子信息系统[5]的过程中发现有必要吸取这些意见,紧扣化学应用的实际需要发展新算法.同时我们还注意充分利用以前自主开发的环系识别算法的结果,在它的基础上进行识别,使得系统整体性能更好,形成了拥有自主知识产权的软件[6].

1 化学结构中的互变异构、芳香键和交替键

1.1 互变异构(Tautomerism)

互变异构通常是指两种或两种以上化学结构间的一种平衡,这些化学结构基本相同,仅差在一个活动基团(通常是氢原子)的位置不同(图 1).

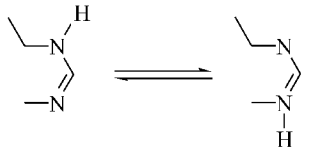

图 1 快速平衡中的两种不同结构构成了互变异构

Figure 1 Two rapidly equilibrating forms constitute the tautomerism

用图论的术语来表达,凡在化学结构中具有图 2 所示子图(结构片段)的即为互变异构,所包含的键称为互变键.

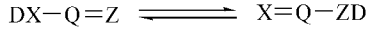

图 2 构成互变异构的通式

Figure 2 General schema for a tautomerism

图 2 中 Q 是互变中心,可以是除 H 以外的任何非金属原子;X 和 Z 是互变系统端点,可以是C(Ⅳ),N(Ⅲ),O(Ⅱ),S(Ⅱ),Se(Ⅱ)和 Te(Ⅱ);但三者不能同时为 C. D 一般为氢原子,但也可以是一个电荷(正负皆可,但其绝对值只能是 1).

1.2 芳香键(Aromatic Bond)

具有离域共轭π键的环系统,由于电子离域使得单双键平均化,而形成了一种既不同于单键也不同于双键的新键型,称作芳香键.尽管到现在化学界对于芳香性的定义仍有一定的争议,不过一般认为Hückel 的 4n + 2 规则最为简洁实用.根据 Hückel 规则和判别芳香性的一般原则,我们定义了芳香键的判别依据:

(1) 具有离域共轭π键的环状体系;

(2) 整个环系都能处于一个平面或至少非常接近于一个平面;

(3) π键中的π电子数目是 4n + 2 个(其中,n = 1, 2, 3, 4, 5……).

我们将具有这些特征的环系视为芳香环,对应的,该环系上的键就称为芳香键(图 3).

图 3 芳香系统和芳香键

Figure 3 Aromatic system and aromatic bonds

另外,一些杂环,由于杂原子提供的孤对电子也能参与体系共轭,或者带有可离域电荷,使得整个系统的π电子数满足 4n + 2 规则,也可以构成芳香系统和芳香键(图 4).

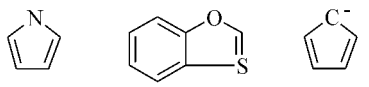

图 4 杂环和带电荷的芳香系统和芳香键

Figure 4 Aromatic system and aromatic bonds in heterocycles and charged ring

1.3 交替键(Alternating Bond)

当环系统中σ键和π键交替出现时我们称之为交替键(图 5).但是当整个系统满足我们定义的芳香键判据时,就定义为芳香键而不是交替键.因此,在我们的系统中芳香键可看作交替键中需要特别处理的特例,因为它们有许多独特的化学性质.

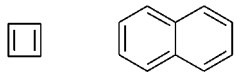

图 5 交替系统和交替键

Figure 5 Alternating system and alternating bonds

2 互变键、芳香键和交替键识别的复杂性

2.1 互变键的传递

在根据定义对互变键进行了一次判别后,如果找到互变键,说明一个原来标记为单键的键有可能转变为双键,原来标记为双键的键有可能转变为单键.这样,一个原来不在互变系统中的原子有可能成为互变系统中的原子.例如,图 6 中最左边结构中右边的 N 原子.通过互变传递,它和 C 原子与 C* 原子组成了新的互变系统,它与 C* 原子间的单键可转变

为双键,使得它与 C*原子间原来的键不能看作单键而应该看作互变键. 互变系统的这一特性使处理算法较为复杂.

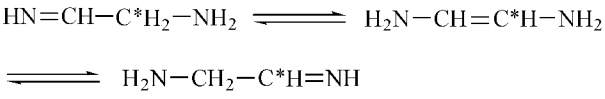

图 6　互变键的传递

Figure 6　Transfer of tautomeric bonds

假设用户输入的是图 6 中最左边的结构,在第一轮识别时,C*只带有单键,因此不会被识别为互变中心,但是由于与它相邻的 C 可以作为互变中心,所以 C—C*可以互变为 C═C*,此时 C*就可以作为互变中心了. 同理,由于互变也可能使得双键变成单键,从而造成互变键的传递. 例如,当用户输入的是图 6 中下边的结构,在第一轮识别时,C 只带有单键,因此不会被识别为互变中心,但是由于与它相邻的 C*可以作为互变中心,所以 C—C*可以互变为 C═C*,结果 C 也可以作为互变中心了.

因此,在互变键的识别过程中,只进行一次识别是不够的,必须依据前一次的识别结果,再进行下一轮的识别,不断重复,直到找不到新的互变键为止. 例如,图 7 所示的结构如只作一次识别不可能得到完全准确的识别,须三次后才能得到准确识别.

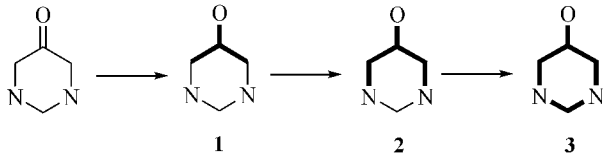

图 7　一个互变结构的识别过程(粗线表示标志为互变键)

Figure 7　A complete identification for tautomerism (bold lines designate the tautomeric bonds)

2.2　互变键对识别交替键和芳香键的影响

由于化学结构中往往会同时存在互变键和交替键(或芳香键),互变异构可能使原来的单双键间发生互相转变,使得它们的识别变得十分复杂. 例如图 8 所示三嗪三吩中由于互变异构现象的存在,其中左右两端的结构具有与苯环相类似的结构,应该具有芳香性,而中间的却变成了非芳香结构. 这就迫使我们必须考虑互变键和交替键(芳香键)之间的互相影响. 由于芳香性是比较重要的化学性质,它的存在使得整个系统更稳定,某些双键反应不能发生,而只能发生某些独特的反应,所以在我们的系统中给予芳香键以较高的优先级,凡通过互变异构变化可导致结构中出现芳香性的话就确定为芳香键. 在识别过程中,我们首先找出互变异构部分,并把互变键看作同时具有单双键两重性,使得在紧接的识别交替键和芳香键时不会漏识别. 因此,对图 8 所示的三嗪三吩,即使用户输入的是图中左边第二个结构,我们仍能把整个杂环的芳香性识别出来.

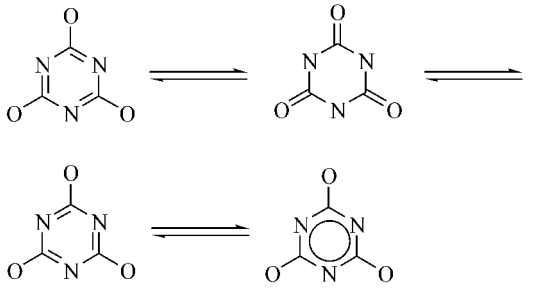

图 8　三嗪三吩的互变异构与芳香键

Figure 8　Tautomorism and aromaticity in [1,3,5]triazine-2,4,6-triol

3　互变键、芳香键和交替键的识别算法

3.1　互变键识别算法

互变键识别算法(参见图 2)步骤如下:

(1) 找一个可能的互变中心. 同时连有单双键的非金属元素(除氢)皆可能是互变中心,令为 Q. 若没有,则结束算法.

(2) 检查与 Q 相邻的原子,找出 X 和 Z 原子. 与 Q 用单键相连,且连有可移动基团(H 或一价电荷)的 C(Ⅳ),N(Ⅲ),O(Ⅱ),S(Ⅱ),Se(Ⅱ)和 Te(Ⅱ)原子,可作为 X 原子;与 Q 用双键相连的上述原子,可视作 Z 原子. 若 Q 周围不同时有 X 和 Z 原子,则 Q 不能作为互变中心;若 Q,X 和 Z 原子都是 C 原子,则 Q 不能作为互变中心;否则,Q 原子是互变中心(Q),将其与 X 和 Z 原子相连的键标记为互变键.

(3) 回到步骤 1.

本算法通过按照互变键的定义不断查找分子结构中的互变子结构,直到找不到为止. 一旦找到后,即将相应的键设置为互变键属性.

图 9 以实例简要说明了互变键识别算法的执行过程.

3.2　芳香键和交替键识别算法

芳香键和交替键的识别过程比较复杂. 在我们

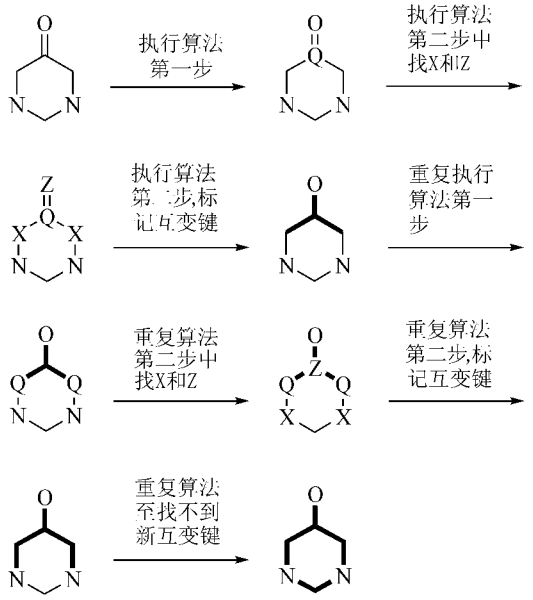

图 9 说明互变键识别算法执行过程的一个例子(粗线表示标志为互变键)

Figure 9 An example illustrating the algorithm for identifying the tautomeric bonds (bold lines designate the tautomeric bonds)

的系统中,所有结构都已做过环块识别的预处理[7]. 所谓环块即对应于美国化学文摘社(CAS)的环系统概念,更确切的定义可在参考文献[7]中找到. 考虑到互变键可能对识别结果的影响,所以需要准备一张互变块表,可以通过原子在原分子中的序号找到它所在的互变块. 同时考虑到芳香键和交替键必须在环上,所以只需对环块上的键进行识别. 为了进一步缩小识别范围,在芳香键和交替键识别前,首先启动候选环块识别算法剔除不可能是芳香或交替环系的环块. 经此算法后保留下来的称候选环块. 具体算法如下:

候选环块识别算法

(1) (在先进行互变键识别后)剔除所有链(非环键)和链(非环)原子.

(2) 剔除所有不可能是交替键和芳香键上的原子. 我们定义除同时连有环双键和环单键的 C 原子、连有环上互变键的 C 原子、三价的 N 族原子和二价的 O 族原子外,其余原子都不可能是交替键和芳香键上的原子,应予剔除.

(3) 由于执行步骤 2,可能使一些原来的环键成为非环键. 因此须回到步骤 1,直至在步骤 2 找不到交替键和芳香键上的原子(原子和键数都不再减少)时,结束算法.

算法结束后,剩余子结构中各个独立的连通块就是候选环块. 这样减小了问题的规模,只需逐个对这些候选环块进行芳香键和交替键的识别即可.

候选环块识别算法的执行过程可用图 10 简要说明.

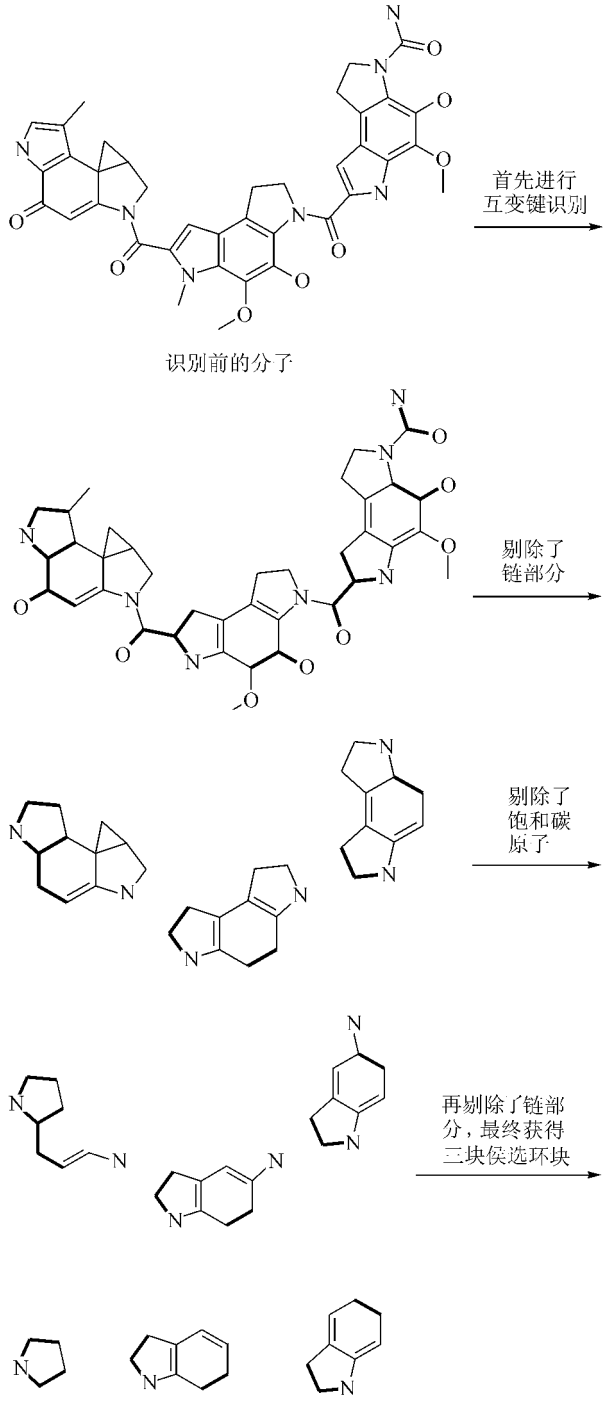

图 10 候选环块识别算法的一个实例(粗线表示标志为互变键)

Figure 10 An example illustrating the algorithm for identifying the candidate rings (bold lines designate the tautomeric bonds)

芳香键和交替键识别算法

(1) 检查环块中是否包含互变键,如果没有,则该环块上的必然都是交替键,直接进入步骤 4;否则,进入下一步.

(2) 对该环块中的所有互变键块进行如下计算:

互变键块中的 C 数目记做 N_C,总原子数记做 N_A,互变键块中最少的双键数目 m 和最多的双键数目 M,直接与该块相邻的双键个数 N_B. 如果所有互变键块的$[N_C, N_A]$数域和$[2m + N_B, 2M + N_B]$数域有交集,则环块上的键都是交替键,此时

$$令\begin{cases} i = \sum_z \min([N_C, N_A] \cap [2m + N_B, 2M + N_B])_z \\ j = \sum_z \max([N_C, N_A] \cap [2m + N_B, 2M + N_B])_z \end{cases},$$

其中,z 是互变键块的数目. 转步骤 5. 否则,转步骤 3.

(3) 如果环块中的键不都是交替键,必然有些不符合 2 中条件的互变块. 去掉这些互变块,将剩下部分的非环部分去掉,如果还有环块,则将这些环块加入到候选环块中. 退出本算法.

(4) 如果环块中没有互变键,则计算环块中的π电子数 N_π,计算公式如下:环块中 C 原子数目记作 N_C,总原子数记作 N_A,双键个数 N_B;则 $N_\pi = 2(N_A - N_B)$,如果 N_π 符合 $4n + 2$,则是芳香键.

(5) 如果环块中有互变键,则对非互变键部分的 N_π 进行计算,计算方法同步骤 4. 若数域$[N_\pi + i, N_\pi + j]$中,存在符合 $4n + 2$ 的自然数,则该环块是芳香键.

(6) 若环块是交替键而不是芳香键,如果该环块不是单环,则将组成该环块的单环加入到候选环块中.

(7) 结束.

图 11 分别列举了芳香环和非芳香环的两个识别实例.

4　结果 ——部分互变键、芳香键和交替键的识别实例

4.1 卟吩

卟吩是叶绿素、血红素中的核心部分. 化学实验表明,它的四个五元环都是芳香性的吡咯环. 如果应用一般最小环识别方法进行判断,只有右下角的一个五元环被判定为芳香环的. 应用本工作的方法,将首先识别出环块然后进行判断,则整个环系统(包括

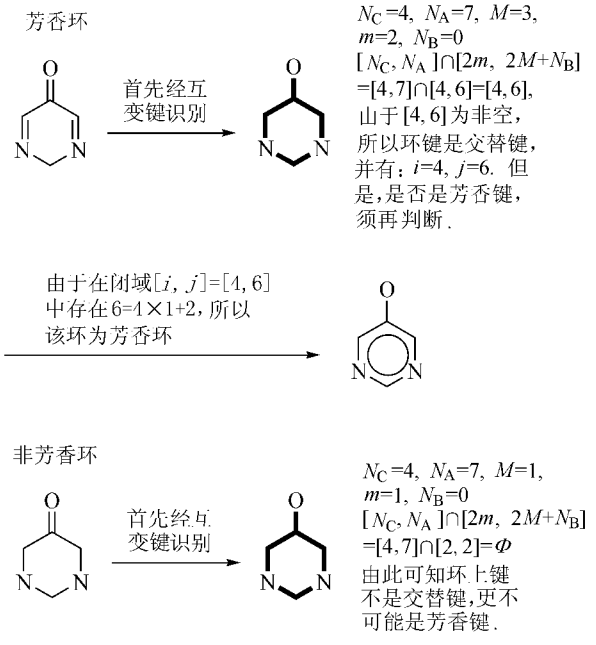

图 11　芳香环(上)和非芳香环(下)的识别实例(粗线表示标志为互变键,单键加圆表示为芳香键)

Figure 11　Examples of identification of aromatic and non-aromatic rings (bold lines and the single bonds plus a circle designate the tautomeric bonds and the aromatic bonds respectively)

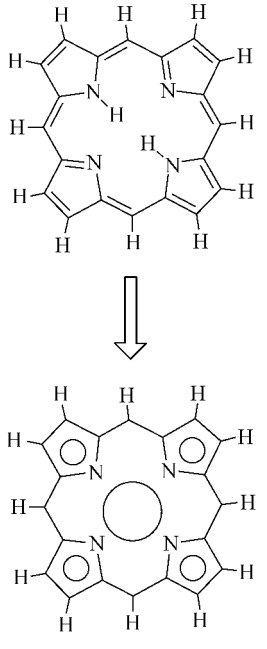

图 12　卟吩在识别前后的键的属性(上边结构是识别前的键属性,识别后用单键加圆表示整个环系统都是芳香键)

Figure 12　Bond nature for porphine before and after the identification (the upper structure designates the bond nature before the identification, in the lower one the single bonds plus a circle mean these bonds identified as aromatic bonds)

四个亚甲基)都是芳香性的(图 12). 这更符合实验的结果.

4.2 环辛四烯

这是一个反芳香性的分子,本工作中识别为交替键(图 13).

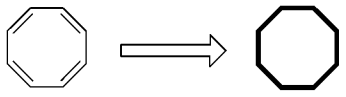

图 13 环辛四烯在识别前后的键的属性(左边结构是识别前的键属性,粗线表示交替键)

Figure 13 Bond nature for cyclooctatetraene before and after the identification (the bold lines in the right structure designate the alternating bonds)

4.3 三嗪三吩

这是互变键和芳香键混杂的例子(图 14)

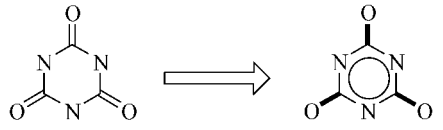

图 14 三嗪三吩的互变异构键与芳香键(左边结构是识别前的键属性,右边为识别后的键属性,粗线表示互变键,整个环系为芳香键)

Figure 14 Bond nature for [1,3,5]triazine-2,4,6-triol before and after the identification (the left and right structures designate the bonds before and after identification respectively, the bold lines in the right structure designate tautomeric bonds and the whole ring system is an aromatic system)

4.4 CC-1065

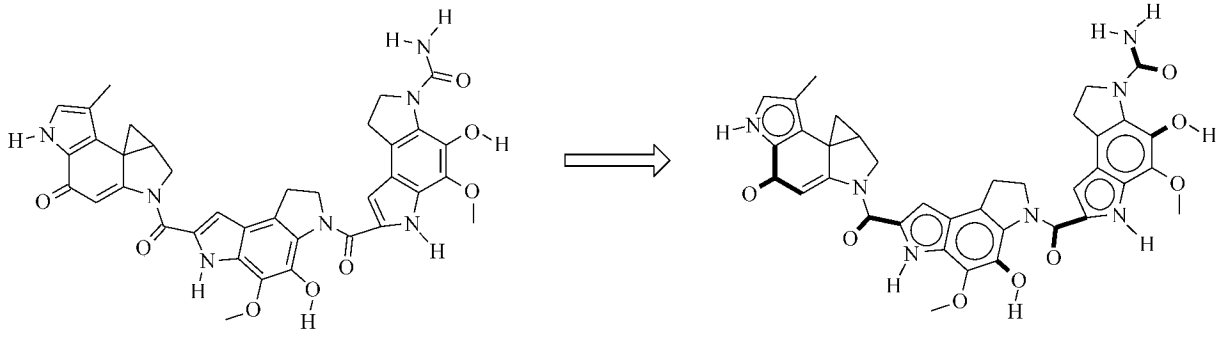

图 15 一种从链霉菌中提取的化合物(左边结构是识别前的键属性,识别后用单键加圆表示整个环系统都是芳香键,互变键用粗线表示)

Figure 15 A compound separated from *Streptomyces* (the left structure designates the bond nature before the identification, in the right one the single bonds plus a circle mean these bonds identified as aromatic bonds and the tautomeric bonds are designated by bold lines).

5 结论

化学结构的计算机处理问题,经过几十年来众多计算机化学研究者们的努力,取得了极大的成功,在化学信息的计算机处理方面已经给化学工作者带来了许多实惠. 各种形式的文献检索,特别是化学反应数据库和谱图数据库检索等,已成为日常必备的工具,在结构解析、构效关系研究的某些方面,也已为化学研究提供了极大的便利. 在化学研究的更深层次上,像分子设计、反应策划、组合化学等涉及复杂化学现象的结构处理方面,还不尽如人意. 本工作针对化学中最普遍存在的互变键、交替键和芳香键等复杂问题提出了自己的解决方案,并取得了成功. 但是,芳香性问题至今仍颇具争议,使得目前还不可能设计出一种完美算法,多多少少仍带有少许近似色彩. 通过一段时间的实际使用,基本上能满足结构数据库对化学反应结构规范化的要求,当目标化合物中有芳香系统时在合成路线设计中也能准确地找到应该采用的反应(谋略键). 可以预见,随着相关研究的不断深入,这些算法还将进一步得到完善和提高.

References

1 Mockus, J.; Stobaugh, R. E. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 18.

2 Roos-Kozel, B. L.; Jorgensen, W. L. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 101.

3 Panaye, A. *Dissertation of State Doctorate of Physical Sciences*, The Seventh University of Paris, Paris, **1976**.

4 Shelley, C. A. In *Computer-Supported Spectroscopic Database*,

Ed.: Zupon, J., Eillis Horwood Ltd., Chichester, **1986**, p. 19.

5 Yuan, S.-G.; Zhang, W.-Q.; Zheng, C.-Z. In *Proceedings of Scientific Data Banks and Information Technology*, Vol. 3, Chinese Science and Technology Press, Beijing, **1996**, p. 82. (**袁身刚, 张伟琪, 郑崇直, 科学数据库与信息技术论文集, 第三集, 中国科学院科学数据库中心编, 中国科学技术出版社, 北京, 1996**, p. 82)

6 *Chinese Software Copyright No. 2001SR2422*, accepted in **2001**.

7 Yao, J.-H.; Yuan, S.-G.; Zheng, C.-Z. In *Proceedings of Scientific Data Banks and Information Technology*, Vol. 3, Chinese Science and Technology Press, Beijing, **1996**, p. 201. (**姚建华, 袁身刚, 郑崇直, 科学数据库与信息技术论文集, 第三集, 中国科学院科学数据库中心编, 中国科学技术出版社, 北京, 1996**, p. 201.)

(A0111231 PAN, B. F.; DONG, L. J.)