第10卷 第1期
1984 年 1 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 10, No. 1
Jan., 1984

化合物子结构的标记法识别

徐 建 钱 诚 李介谷
（上海交通大学）（中国科学院上海有机化学所）（上海交通大学）

摘 要

本文运用标记法理论讨论和处理了计算机化学结构信息处理中的化合物子结构识别问题. 文中首先给出并定义了属性、互协调标记对、全协调标记集等一系列术语和概念；重新定义了标记法中原已有的约束关系集、相容性、协调性等概念；指出了化合物子结构匹配等价于全协调标记集的求取. 在此基础上引出了求取全协调标记集的有效算法. 最后，对随机采集的近两千例样品进行了试识别，并与 Ullmann[2] 的算法作了比较，结果是肯定的.

一、引 言

在计算机化学结构信息处理中，化合物子结构识别是非常基本又极其重要的手段之一. 例如当要求搞清楚某化学基团具有何种物化特性时，较为严格的方法是找出适当数量的带有此基团的化合物做相关分析. 由于化合物的种类极多(以数百万计，且还在不断增加)，要在适当的时间间隔内对如此庞大的化合物结构信息库做一次较全面的子结构判别检验，是颇为艰巨的. 化学界处理这类问题通常用对两化合物结构图逐点逐边进行精确比较，这种方法较为盲目. 在数学界，有人从图 (Graph) 的子同构出发给出了较为有效的算法，如 Cheng[1]，Ullmann[2] 等. 本文进一步运用结构模式识别的标记法解释和处理化合物子结构的识别问题.

二、建立化合物子结构识别的标记法模式

首先假定所讨论和处理的是已经有色图化了的化合物结构（图 1）. 有色图的定义为:

有色图 $G=((V, C), (E, \varepsilon))$.

其中: $V=\{v_1, v_2, \cdots, v_n\}$为 G 的节点集; $C=\{c_{v_1}, c_{v_2}, \cdots, c_{v_n}\}$ 为 V 中各元素的色性值集; $E=\{(i, j)|i, j\in V, i, j$ 之间有边相联$\}$ 为 G 的边集; $E=\{\varepsilon_{ij}|(i,j)\in E\}$ 为 E 中各元素的色性值集.

两有色图子同构的定义为: 设有两有色图 G_A, G_B. G_A 的节点集为 $V_A=\{1, 2, \cdots, n\}$; $\#V_A=n$. G_B 的节点集为 $V_B=\{\lambda_1, \lambda_2, \cdots, \lambda_m\}$; $\#V_B=m$. $\#V_A\leqslant \#V_B$($\#V_A$ 为 V_A 中元素的个数，称做 V_A 的基数，$\#V_B$ 与之类似). 若存在某一集合 $U_B=\{l_1, l_2, \cdots, l_n\}\subseteq V_B$, U_B 的元素与 V_A 的元素成一一对应关系 l: $i\longleftrightarrow l_i$, $i\in V_A$,

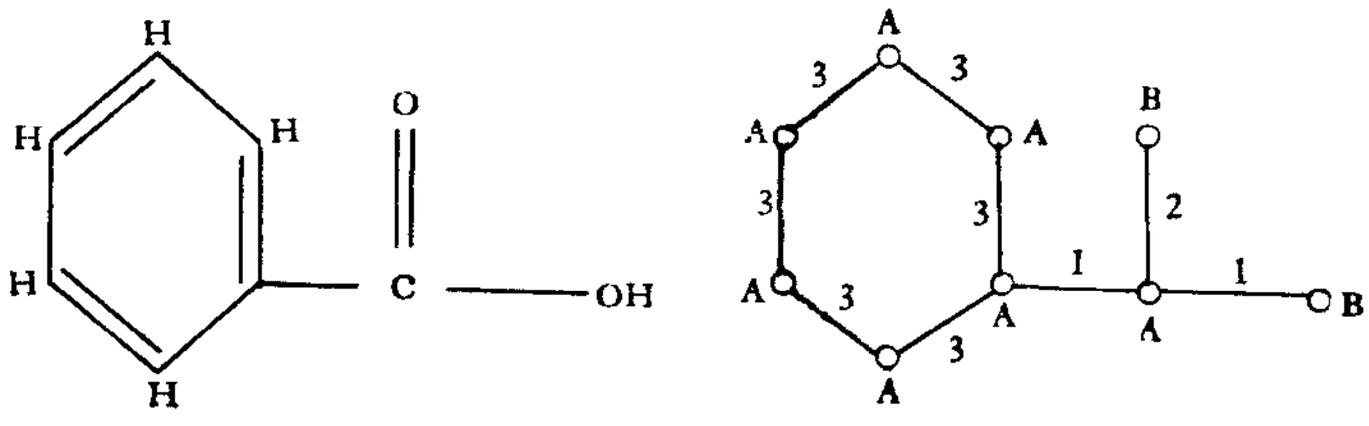

图　1

A 为碳原子，　*B* 为氧原子；　1 为单键，　2 为双键，　3 为芳香键.

$l_i \in U_B$,同时命题 $\forall i, j \in V_A, \exists l_i, l_i \in U_B.\ (i,j) \in E_A \Rightarrow (l_i, l_j) \in E_B \wedge (c_i = c_{l_i},\ c_j = c_{l_j}) \wedge (\varepsilon_{ij} = \varepsilon_{l_i l_j})$ 为真,则称 G_A 子同构于 G_B.

设有两个以有色图表述的化合物结构 G_A, G_B. 判断 G_A 是否为 G_B 的某一子结构，这可化作如下标记法问题来处理：令 G_A 的节点集 V_A 为目标集，G_B 的节点集 V_B 为标记集,要求判断能否替 V_A 中每一元素 i 找到一个无二义的标记 $l_i \in V_B$，使 G_A 按此种节点对应关系子同构于 G_B.

为解决这一问题,先给出一个六元类：$(V_A, V_B, C_A, C_B, T, \Pi)$. 其中 V_A, V_B 称之为目标集和标记集，分别由 G_A, G_B 的节点组成；C_A, C_B 为 V_A 和 V_B 的色性值集,分别称之为目标特性集及标记特性集；$\Pi = (\pi_{e_0}, \pi_{e_1}, \cdots, \pi_{e_\Omega})$ 为标记的约束关系集. 其中 $\pi_{e_0} = \{(\lambda, \lambda) | \lambda \in V_B\}$; $\pi_{e_k} = \{(\lambda, \mu) | \lambda, \mu \in V_B, \varepsilon_{\lambda\mu} = e_k\}, 1 \leqslant k \leqslant \Omega$. 其中 e_k 为标记之间约束形式的类别符.

为完备起见,定义同一标记之间也有某种约束形式,其类别符为 e_0. 因此，对于一个具有 Ω 种不同边色性的化合物结构 G_B 来说,相应地就有 $1 + \Omega$ 种约束形式. 显然，Π 中各元素互不相容,即

$$1 \leqslant k \neq k' \leqslant \Omega,\ 有\ \pi_{ek} \cap \pi_{ek'} = \phi.$$

目标的约束关系集 T 的定义与之类似,不再复述.

目标集及标记集中各元素都有一些与周围环境无关的性质,称为该元素的基本属性. 那些凝聚到被考察元素上而受周围环境制约的信息称之为该元素的附加属性.

当标记 $e \in V_B$ 的属性同目标 $i \in V_A$ 的属性之间满足某种关系时，称 e 为 i 的相容标记. 由 i 的相容标记所组成的集合 $L_i^s \subseteq V_B$ 称为目标 i 的相容标记集. 以 V_A 的所有元素的相容标记集为成员而组成的集合叫做 V_B 关于 V_A 的相容标记集. 记为

$$L^s = (L_1^s, L_2^s, \cdots, L_n^s);\ (1, 2, \cdots, n) = V_A.$$

定义 1.　$i, j \in V_A$, $l, k \in V_B$, 当不违背条件

$$(i, j) \in T_{\varepsilon_{ij}} \Rightarrow (l, k) \in \pi_{\varepsilon_{ij}},$$
$$(l, k) \in \pi_{e_0} \Rightarrow (i, j) \in T_{e_0},$$

则称标记对 (l, k) 为关于目标对 (i, i) 的互协调标记对.

定义 2.　设有 $U_A = \{i_1, \cdots, i_k, \cdots, i_H\} \subseteq V_A$, U_A 中任一元素 i_k 都有一非空的标记集 $L_{i_k} \subseteq V_B (k = 1, \cdots, H)$. 以这 H 个标记集组成 $U_B = (L_{i_1}, \cdots, L_{i_k}, \cdots, L_{i_H})$. 若 $\forall l_{i_k}, l_{ik'}, l_{ik} \in L_{ik}, l_{ik'} \in L_{ik'}, i_k, i_{k'} \in U_A$ 有 $(l_{ik}, l_{ik'})$ 关于 $(i_k, i_{k'})$ 成互协调标记对，

则称 U_B 为关于 U_A 的一个局部协调标记集. U_B 的任一标记集中的标记称为关于其相应目标的协调标记.

命题 1[1]. 若 U_B 是关于 $U_A \subseteq V_A$ 的局部协调标记集，则 U_B 与 U_A 之间的下述关系成立: $\forall L_{ik} \in U_B$ 有 $\# L_{ik} = 1$，且任取 $l_{ik} \in L_{ik}$, $l_{ik'} \in L_{ik'}$, i_k, $i_{k'} \in U_A$，有 $i_k \neq i_{k'} \Rightarrow l_{ik} \neq l_{ik'}$.

命题 2. 若 V_B 关于 U_A 的局部协调标记集 U_B 同时又是 U_A 的相容标记集，则由 U_A 及其目标特性集和约束关系集所构成的有色图子同构于由 U_B 及其标记特性集和约束关系集所构成的有色图.

若某局部协调标记集 U_B 的基数 $\# U_B = \# V_A$，则称 U_B 为 V_B 关于 V_A 的一个全协调标记集.

最后，给出协调性必要条件: 若 l 为某一局部协调标记集 U_B 中关于 $i \in V_A$ 的协调标记，则 l 必满足

$$(\{l\} \times L_j) \cap \Pi \varepsilon_{ij} \neq \phi, \ \forall j \in V_A, \ i, \ j \text{ 之间有约束关系.}$$

三、算 法

判断 V_B 关于 V_A 是否存在全协调标记集的全过程可以分成三个阶段:

第一阶段，建立 $V_A = \{1, 2, \cdots, n\}$ 的相容标记集 $\mathscr{L}^s = (L_1^s, \cdots, L_n^s)$. 这是一种直接算法，其效能取决于属性的选择. 这里选用了一个基本属性和四个附加属性.

设 $i \in V_A$, $l \in V_B$，当 i 及 l 的五个属性分别满足下述关系时，$l \in L_i^s$.

基本属性. 描述被考察点的点色性. 要求 $c_l = c_i$.

附加属性 1. 描述被考察点的连接度. 要求 $\mathrm{dgree}(l) \geqslant \mathrm{dgree}(i)$ ($\mathrm{dgree}(x)$ 为被考察点 x 的连接度，即与 x 邻接的边的条数).

附加属性 2. 描述与被考察点邻接的各点的点色性. 要求对 i 的所有 h 个不同的邻接点 $j_1, j_2, \cdots, j_h$, 能分别找到 h 个不同的 l 的邻接点 $k_1, \cdots, k_h$，且有 $c_{k_\mu} = c_{j_\mu}$, $\mu = 1, 2, \cdots, h$.

附加属性 3. 描述与被考察点相邻的各点的连接度. 要求对 i 的所有 h 个不同的邻接点 $j_1, j_2, \cdots, j_h$, 能分别找到 h 个不同的 l 的邻接点 $k_1, k_2, \cdots, k_h$，且有 $\mathrm{dgree}(k_\mu) \geqslant \mathrm{dgree}(j_\mu)$, $\mu = 1, 2, \cdots, h$.

附加属性 4. 描述被考察点的边色性. 要求 $\mathrm{Ne}_k(l) \geqslant \mathrm{Ne}_k(i)$, $k = 1, \cdots, \omega$ ($\mathrm{Ne}_k(x)$ 为与被考察点 x 邻接而色性值为 e_k 的边的条数).

在建立相容标记集的过程中以及在建立了相容标记集之后，一经发现相容标记集中某一成员为空集，则立即可判定 V_B 关于 V_A 不存在任何全协调标记集.

第二阶段为预检验，这是一种松弛算法. 图 2 为算法框图.

第三阶段是进行全协调性检验. 这是一种 OFS 算法，步骤如下: 1) 置层号 $k \leftarrow 0$ ("$\leftarrow$"为赋值号). 2) 从 $\mathscr{L}^k = (L_1^k, L_2^k, \cdots, L_n^k)$ 中找基数最小的成员 L_i^k，且要求

---

1) 命题 1，2 的证明见: 徐建，化合物子同构的标记法识别，1982.

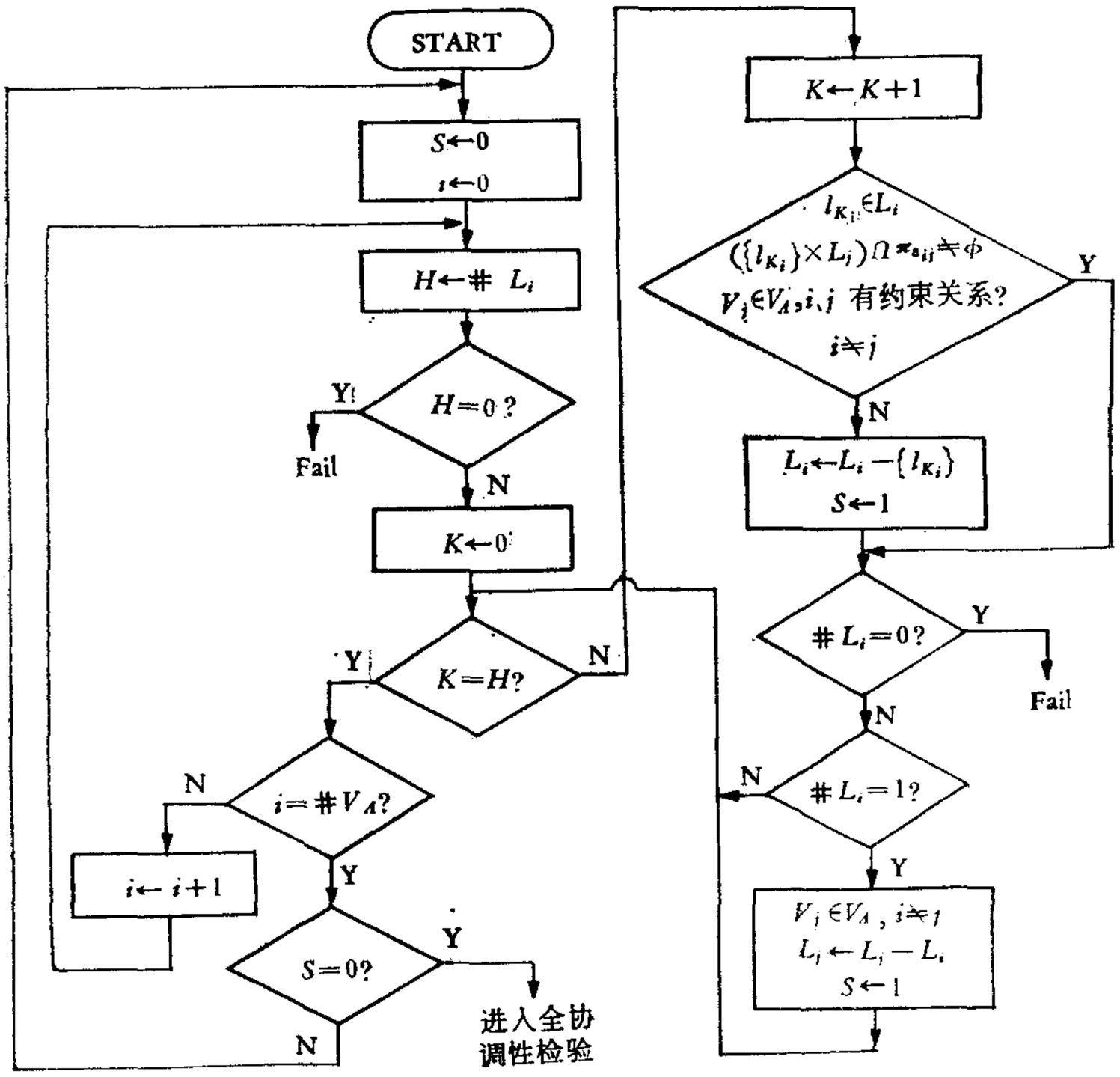

图　2

$i \in V_A$ 未被赋过唯一标记．3）考察 L_i^k 是否为空集，若 $L_i^k = \phi$，转步骤 9），否则从 L_i^k 中任取一标记 l_i，转步骤 4）．　4）进行集合的差分运算 $L_i^k \leftarrow L_i^k - \{l_i\}$．　5）对 l_i 做协调性必要条件检验，若 l_i 满足条件 $(\{l_i\} \times L_j^k) \cap \Pi\varepsilon_{ij} \neq \phi$，$\forall j \in V_A$，$i$，$j$ 之间有约束关系，$i \neq j$ 则转步骤 6），否则转步骤 3）．6）$k \leftarrow k+1$，考察 k 是否等于 $\# V_A$，若不等于 $\# V_A$ 则转步骤 7）；反之则转出口 Success，判定 G_A 子同构于 G_B．　7）对所有 $j \in V_A$ 做下述操作：$L_j^k \leftarrow L_j^{k-1} \cap \Lambda^{(\varepsilon_{ij})}{}_{l_i}$，$i$，$j$ 之间有约束关系，$i \neq j$；$L_j^k \leftarrow L_j^{k-1} - \{l_i\}$，$i$，$j$ 之间无约束关系；$L_j^k \leftarrow \{l_i\}$；$R(k) \leftarrow i$．式中：$\Lambda^{(\varepsilon_{ij})}{}_{l_i} = \{l \mid l \in V_B, (l, l_i) \in \pi_{\varepsilon_{ij}}\}$；$R(k) = i$ 意味着在第 k 层 i 被赋于唯一标记；“—”为集合的差分运算符．由此建成第 k 层的标记集 $\mathscr{L}^k = (L_1^k, \cdots, L_n^k)$．8）查询是否有与 $i \in V_A$ 有约束关系却又未被赋过唯一标记的目标 j，若有则 $i \leftarrow j$，转步骤 3），否则转步骤 2）．　9）判 k 是否为零，若不为零进步骤 10）；若是转出口 Fail，判定 G_A 不子同构于 G_B．　10）$k \leftarrow k+1$．考察 L_i^k 是否为空集，若是转步骤 9），否则 $i \leftarrow R(k)$，转步骤 3）．

用于化合物子结构识别时，此算法截废枝的能力要较 Ullmann 算法强得多，而内存的占用量却又较 cheng 的算法少得多．

下述三个结论是正确的[1]：

1）在全协调性检验过程中，所有已产生出来的唯一标记构成一个局部协调标记集．

1）其证明见徐建，“化合物子同构的标记法识别”文中对命题 3，4，5 的证明．

2) 若 V_B 关于 V_A 通过了相容性检验和全协调性检验，则有色图 G_A 必子同构于有色图 G_B.

3) 若 G_A 子同构于 G_B，则 V_B 关于 V_A 必然能通过相容性检验、预检验及全协调性检验. 结论 1)，2) 保证了算法不会发生择伪错误，结论 3)保证了算法不会出现弃真错误.

四、实 验 结 果

为了检验本算法的性能，从 Saldter 光谱集记载的近六万个化合物中随机选用了 97 个化合物作为被查询结构. 在这 97 个结构中，又抽取了 20 个子结构作为查询结构，这些子结构在化学上都具有一定特征. 根据这些结构及子结构，在 Burroughs 1955 型计算机上分别用本算法及 Ullmann 算法进行了 20 次查询，即 $20 \times 97 = 1940$ 次匹配检验. 查询结果是匹配(指具有子同构关系)的 39 次，不匹配(指不具有子同构关系)的 1901 次，与实际相符. 由于 Ullmann 算法仅处理无色图子同构问题，因此上述检验都是在图褪色之后做的. 两算法的 cpu 操作时间平均值按匹配及不匹配两种情况列于表 1 和表 2.

仅从这近两千例匹配检验已可看出此算法较 Ullmann 的算法为快，且随着查询结构及被查询结构非氢原子数的增长，Ullmann 算法的匹配时间均值的增长率远大于本算法，实验也证明了这一点. 还用 B1955 机的伪随机信号发生器产生一系列随机无色图，然后用本算法和 Ullmann 算法分别进行了试算，试算结果示于图 3. 试算中查询结构的节

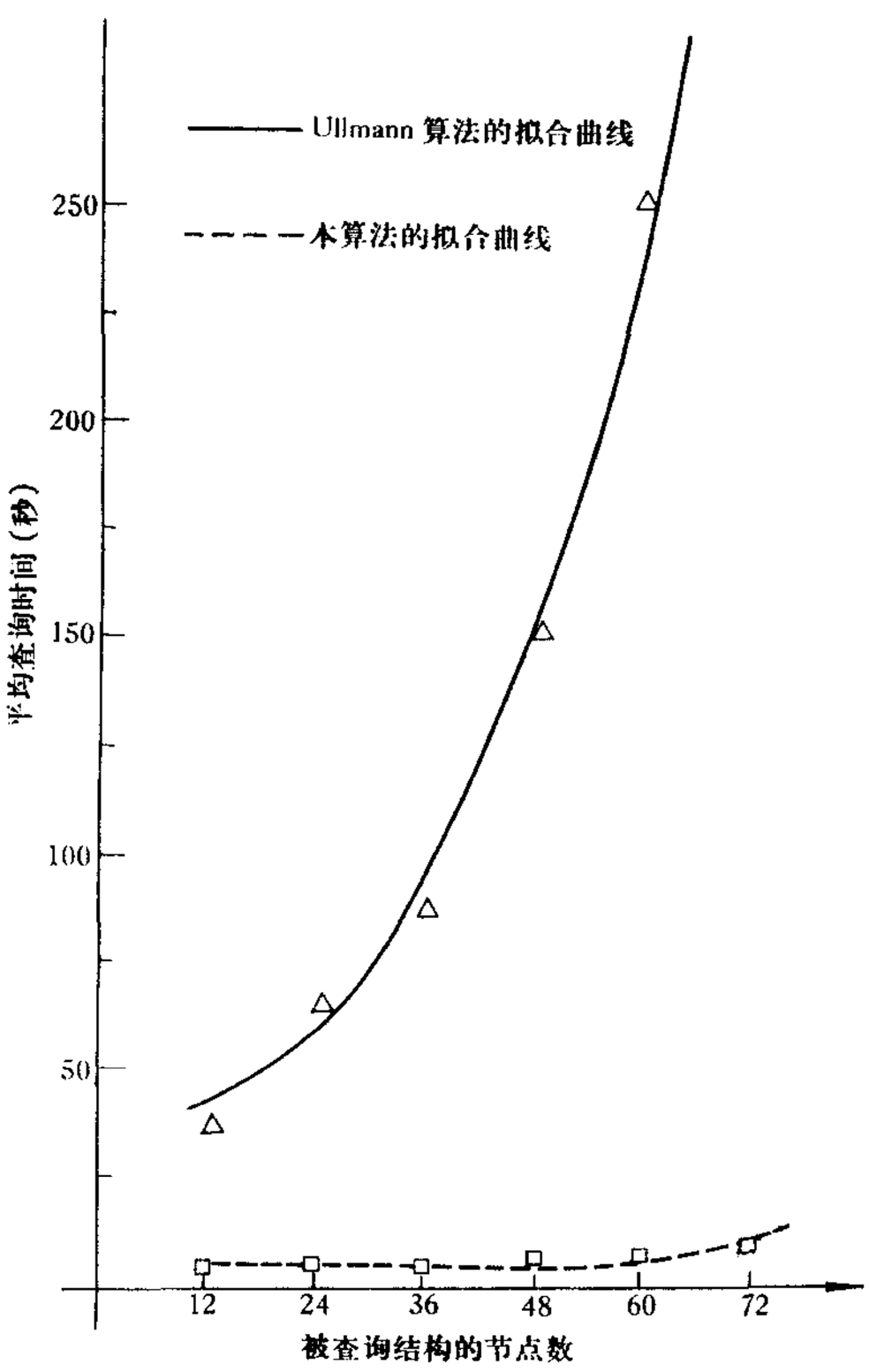

图 3

表 1　匹配的 39 例

	本算法 CPU 时间均值(秒)	Ullmann 算法 CPU 均值(秒)
辅助性操作	1.39	1.16
主要操作	1.40	8.96

表 2　不匹配的 1901 例

	本算法 CPU 时间均值(秒)	Ullmann 算法 CPU 均值(秒)
辅助性操作	0.99	0.90
主要操作	1.03	5.27

点数始终保持为 12，而被查询结构的节点数按 12，14，…，72 等六种情况依次递增．每一种情况分别进行廿次匹配检验，取其平均匹配时间．由图 3 可以看出两种算法在被查询结构节点数增大时，程序平均执行时间的变化情况．

参　考　文　献

[1] Cheng J. K. and Hung T. S., A Subgraph Isomorphism Algorithm Using Resolution, *Pattern Recognition*, **13**, (1981), pp. 371—379.

[2] Ullmann J. R., An Algorithm for subgraph Isomorphism, *J. ACM.*, **23**, (1976), pp. 31—42.

[3] Rosenfeld A., Hummel R. A. and Zucker S. W., Scene Labeling by Relaxation Operations, *IEEE Trans. SMC.* **6**, (1976), pp. 420—433.

THE SUBSTRUCTURE MATCHING OF CHEMICAL COMPOUNDS BY LABELING METHOD

Xu Jian
(*Shanghai Jiaotong University*)

Qian Chen
(*Shanghai Organic Chemistry Institute, Academia Sinica*)

Li Jiegu
(*Shanghai Jiaotong University*)

Abstract

In this paper, the problem of chemical substructure matching is handled within the frame-work of labeling method. First, some definitions such as attributes, mutually compatible labeling pairs and completely compatible labeling sets and so forth are given; some concepts such as constrain relation sets, consistency and compatibility, which have appeared in the early labeling theory, are redefined. And the equivalence between monomorphism and the achieving of completely compatible labeling sets is demonstrated. Then, an efficient algorithm to deal with this problem is provided. Finally, two thousand test matches, collected randomly are carried out to compare this algorithm with Ullmann's. The results obtained are satisfying.