化学结构二维子结构检索的开发

刘 冰， 周家驹

(中国科学院过程工程研究所, 北京 100080)

摘 要：将图形通用匹配 VF 算法应用到化学结构子结构检索中，用面向对象的方法，实现了化学结构数据库中二维子结构检索功能，程序各模块之间相互独立、可移植性强、健壮性好. 通过与3DFS 比较，结果正确，适于处理大型化学结构数据库，现已应用于中药数据库系统药物先导化合物的发现.

关键词：二维子结构检索；子结构匹配；VF 算法；面向对象
中图分类号：O6-39 文献标识码：A 文章编号：1009–606X(2003)04–0376–05

1 前 言

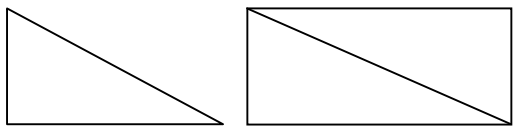
二维子结构检索对于创新药物先导化合物的发现，化学数据库的数据挖掘(Data mining)和知识发现(Knowledge discovering in database)工作都有着重要的意义. 它在化学结构数据库的应用中扮演着重要角色.

目前，已经有多种算法用于解决二维子结构检索问题. 例如 1957 年提出的 Ray and Kirsch 算法，1965 年提出的 Sussenguth 算法，1972 年提出的 Figueras 算法，1976 年提出的 Ullmann 算法[1]，1984 年提出的 von Scholley 算法，以及 1989 年提出的 GMA 算法[2,3](只针对化学结构)，1992 年提出的 Brown 算法[4,5]和 1996 年提出的 VF 算法[6–8](通用算法)等.

1995 年之前，Ullmann 算法被公认为是执行效率最高的子结构匹配算法，而后来的 VF 通用算法则实现了比 Ullmann 算法更高的执行效率和较低的复杂度[9]. 本文用 C++编程开发了通用 VF 算法对化学子结构的检索.

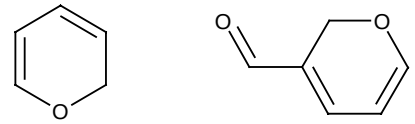
2 VF 算法

化学子结构匹配不同于普通的图形匹配方法，它不仅存在结点属性的区别(如 C, N, O, S 原子等)，边的属性亦有差异(如单键、双键、三键等). 在图 1 和图 2 中，(a)分别是(b)的子图, 很明显，图 2 较图 1 复杂度高许多.



(a) (b)

图 1 普通无向图
Fig.1 The general undirected graph



(a) (b)

图 2 化学结构示意图
Fig.2 The graph of chemical structure

收稿日期：2003–01–20, 修回日期：2003–03–10
作者简介：刘冰(1976–), 女, 山东省潍坊市人, 博士研究生, 计算机化学专业.

化学子结构匹配只能是提问药效团分子和数据库中目标分子的原子–原子、键–键的逐一比较，而且这种方法已被证明是一种 NP(Nondeterministic Polynomial)完全问题[8]，即是一种非常耗时的工作，并且匹配时间随着输入的提问药效团原子数目的增加而迅速增加. 因此，必须寻找一种高效可行的搜索算法来解决这个矛盾.

VF 算法是用于图的同构及子图匹配问题的回溯算法，它是一种不针对特定对象的通配图算法. 本文把这种思想应用于化学领域分子结构的二维子结构检索. VF 算法对图形采用了深度优先遍历

```
PROCEDURE Match(s)
    INPUT: an intermediate state s; the initial state s0 has M(s0) = Φ
    OUTPUT: the mappings between the two graphs
    IF  M(s) covers all the nodes THEN
        OUTPUT M(s)
      ELSE
        Compute the set P(s) of the pairs candidate for inclusion in M(s)
        FOREACH p ∈ P(s)
          IF  the feasibility rules succeed for the inclusion of pin M(s) THEN
            Compute the state s' obtained by adding p to M(s)
            CALL Match(s')
          END IF
        END FOREACH
    END IF
END PROCEDURE
```

的思想，其匹配过程如下[8]：

3　子结构匹配的实现

子结构检索的实现分为 4 个模块：结构读取模块、信息存储模块、匹配模块和结果输出模块. 由于使用了面向对象的设计思想，4 个模块之间相互独立，所以代码可重复利用率高，可以方便地嵌入到其它系统中.

3.1 结构读取模块

二维化学结构数据库的描述主要有连接表和线性编码两种方法. 前者不仅可以描述原子及其键的关系，还可以表达原子的空间坐标，常用的文件格式有 MDL 信息系统公司(MDL Information System)的 MOL[10]文件格式、TRIPOS 公司的 MOL2[11]文件格式、以及 CML(Chemical Markup Language，化学标记语言)[11−12]文件格式. 后者主要用字符串描述化学结构信息，它的存储空间远远小于连接表，但不能描述原子的空间坐标，也不能保证分子描述的唯一性，比较重要的有

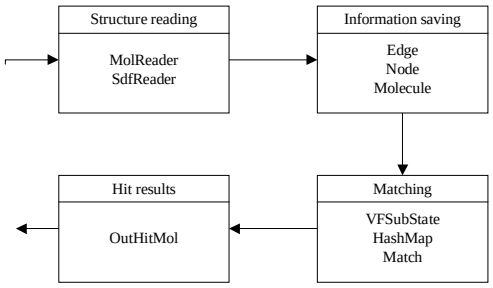


图 3 模块示意图
Fig.3 Module sketch map

Wisweser line notation(WLN)和 Simplified Molecular Input LinE System(SMILES).

鉴于连接表可以保证分子结构描述的唯一性，以及方便地显示分子各结点的坐标，提问分子的结构描述采用 MDL 公司的 MOL 文件格式(MolReader 类)，目标数据库也使用该公司的 SDF 数据库格式(SdfReader 类). 表 1 是图 2 中提问图 2(a)的 MOL 文件.

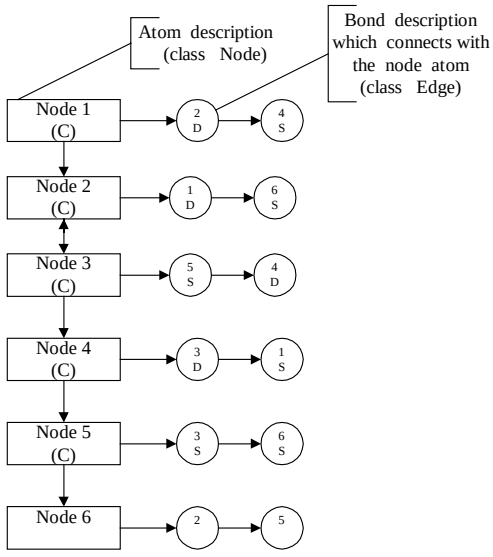
表 1 提问图 2(a)的 MOL 文件
Table 1　The MOL file of 2 (a) in Fig.2

```
–ISIS–   03040322022D

  6  6  0  0  0  0  0  0  0  0999 V2000
    4.9611   –5.8291    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.9599   –6.6565    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    6.3883   –5.8255    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.6729   –5.4164    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    6.3912   –6.6560    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.6739   –7.0672    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  5  3  1  0  0  0  0
  1  2  2  0  0  0  0
  3  4  2  0  0  0  0
  4  1  1  0  0  0  0
  2  6  1  0  0  0  0
  5  6  1  0  0  0  0
M  END
```

3.2 信息存储模块

信息存储模块主要分键性质存储描述、原子结点存储描述、分子存储描述 3 部分，(1) 用 Edge 类表达键，主要属性是相邻原子序号及键的性质[Edge(int nToNodeId, int nAttr)]；(2) 用 Node 类表达结点原子，主要属性是原子序号和原子符号[Node(int nNodeId, CString sAttr)]以及 Edge 类链表；(3) 用 Molecule 类描述整个分子结构，主要属性是一个 Node 类链表. 图 4 为图 2(a)分子信息存储图.



S: There is a single bond between two atoms
D: There is a double bond between two atoms

图 4 图 2(a)分子信息存储图
Fig.4 The molecular information storage graph of Fig.2(a)

3.3 匹配模块

以提问图为指导，依次提取数据库中化合物进行匹配，主要由 VFSubState 类、HashMap 类、Match 类组成. 其中在 Match 类中使用了回溯算法.

3.4 查询结果输出模块

成功匹配的分子以其 MOL 文件格式输出，由 OutHitMOL 类来完成.

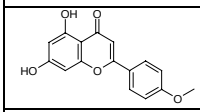
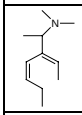
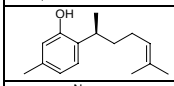
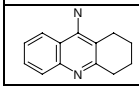
4 程序优化

由于子结构匹配中的原子–原子比较是一种非常耗时的过程，对数据库的分子进行初筛可以大大缩短这一过程. 本文中采取的措施：(1) 原子个数初筛. 数据库中分子的原子个数小于提问分子

原子个数的将直接被淘汰，而不进行匹配运算. (2) 杂原子起步. 如果分子中含有 N, P, S, O, Cl 等常用杂原子，则以其中一个作为匹配的起始原子，这样可以大大降低回溯的次数，从而优化程序执行效率.

5　效率及正确度检验

将本文的可执行程序与王亭[14]的 3DFS(三维药效团柔性搜索)程序在 NCI3D 的 126,705 化合物结构数据库进行搜索比较，结果均一致，特列举 4 例：

Structure	Name	Pharmacological activities	3DFS hit results	VF hit results
	Acacetin	Anti-inflammatory, reduces intestinal vascular permeability and brittleness, antispasmodic	12	12
	–	–	32	32
	(+)-Curcuphenol	Antimicrobial	1	1
	–	–	11	11

效率方面，总耗时随药效团分子的原子数目及杂原子数目的变化而不同. 在赛扬 900 MHz CPU，128 M 内存主机条件下，我们的 VF 算法搜索程序从 NCI3D 数据库中搜索一个例程大约耗时为 1~6 min. 在提问药效团原子数目较少时，3DFS 速度略快；而在提问药效团原子数目较多时，VF 算法占优势.

6　应用举例

图 5 是我们中药化学数据库中一种名为芍药醇的化合物，它可以从白牡丹、红花以及桑叶中提取，具有麻醉、抗菌、抗惊厥、抗高血压以及镇静催眠等多种药理活性.

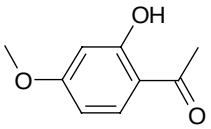


图 5 芍药醇
Fig.5 Paeonol

以此种芍药醇为提问图，在总数目为 9127 种化合物的中药数据库中搜索，共得到 436 个命中化合物，图 6 列出了其中的 10 个.

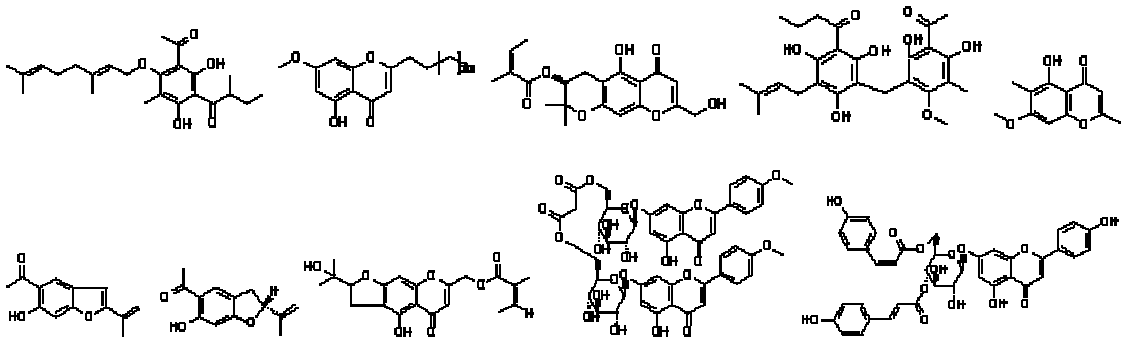


图 6 10 个命中化合物

Fig.6 Ten compounds of hit

图中化合物可作为治疗麻醉、镇静催眠等病症药物的先导化合物进行深入研究. 我们还对抗细胞毒素、抗真菌、抗癌菌素、抗肿瘤等药效团进行检索，并与 3DFS 比较，结果均一致.

参考文献：

[1] Ullmann J R. An Algorithm for Subgraph Isomorphism [J]. J. Ass. Comput. Mach., 1976, 23: 31–42.

[2] Xu Jun. GMA: A Generic Match Algorithm for Structural Homomorphism,Isomorphism, and Maximal Common Substructure Match and Its Application [J]. J. Chem. Inf. Comput. Sci., 1996, 36: 25–34.

[3] Wang T, Zhou J J. EMCSS: A New Method for Maximal Common Substructure Search [J]. J. Chem. Inf. Comput. Sci., 1997, 37: 828–834.

[4] Brown Robert D, Downs Geoffrey M, Willett Peter. A Hyperstructure Model for Chemical Structure Handling: Generation and Atom-by-atom Searching of Hyperstructures [J]. J. Chem. Inf. Comput. Sci., 1992, 32: 522–531.

[5] Brown Robert D, Gareth Willett Peter. Matching Two-dimensional Chemical Graphs Using Genetic Algorithms [J]. J. Chem. Inf. Comput. Sci., 1994, 34: 63–70.

[6] Cordella L P, Foggia P, Sansone C, et al. An Efficient Algorithm for the Inexact Matching of ARG Graphs Using a Contextual Transformational Model [A]. Proc. of the 13th International Conference on Pattern Recognition, Vol. III [C]. Wien Austria: IEEE Computer Society Press, 1996. 180–184.

[7] Cordella L P, Foggia P, Sansone C, et al. Subgraph Transformations for the Inexact Matching of Attributed Relational Graphs [J]. Computing, 1998, 12: 43–52.

[8] Foggia P, Sansone C, Vento M. An Improved Algorithm for Matching Large Graphs [A]. Proc. the 3rd IAPR-TC15 Workshop on Graph based Representations [C]. Ischia: Jean-Michel Jolion, 2001. 72–80.

[9] 李琰, 周家驹. VF 算法在化学结构检索中的应用 [J]. 计算机与应用化学, 2002, 19: 575–580.

[10] Arthur Dalby, James G N, Hounshell W D, et al. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited [J]. J. Chem. Inf. Comput. Sci., 1992, 32: 244–255.

[11] Murray-Rust P, Rzepa H S. Chemical Markup, XML and the Worldwide Web: 1. Basic Principles [J]. J. Chem. Inf. Comput. Sci., 1999, 39(6): 928–942.

[12] Murray-Rust P, Rzepa H S. Chemical Markup, XML and the Worldwide Web: 2. Information Objects and the CMLDOM [J]. J. Chem. Inf. Comput. Sci., 2000, 41(5): 618–634.

[13] Murray-Rust P, Rzepa H S. Chemical Markup, XML and the Worldwide Web: 3. Toward a Signed Semantic Chemical Web of Trust [J]. J. Chem. Inf. Comput. Sci., 2001, 41(5): 1124–1130.

[14] Wang T, Zhou J. 3DFS: A New 3D Flexible Searching System for Use in Drug Design [J]. J. Chem. Inf. Comput. Sci., 1998, 38: 71–77.

Research of 2D Substructure Search Program for Chemical Structure Database

LIU Bing, ZHOU Jia-ju

(*Institute of Process Engineering, Chinese Academy of Sciences, Beijing 100080, China*)

Abstract: This paper accomplished the substructure search program using object oriented method and VF algorithm. The program has good performance and transplantable character. The results are correct compared with 3DFS. It is suitable to deal with large chemical structure database. And it has been applied to data mining and finding leading drugs from our Traditional Chinese Medical Database.

Key words: 2D substructure search; substructure match; VF algorithm; object oriented