THE STORAGE AND SUBSTRUCTURE SEARCH OF 2-DIMENSIONAL CHEMICAL STRUCTURE

Xin Li, Wan-Yi Sun, Xian-Feng He

Institute of Process Engineering
Chinese Academy of Sciences
Beijing 100080

Purpose

- CODATA Physics & Chemistry Data Center of China

- Scientific Data Center: www.nsdc.cn

- Large-scale chemical structure database

Outline

- Retrieval of chemical structure

1. Structure/substructure search (SSS)
2. Aromatic ring perception

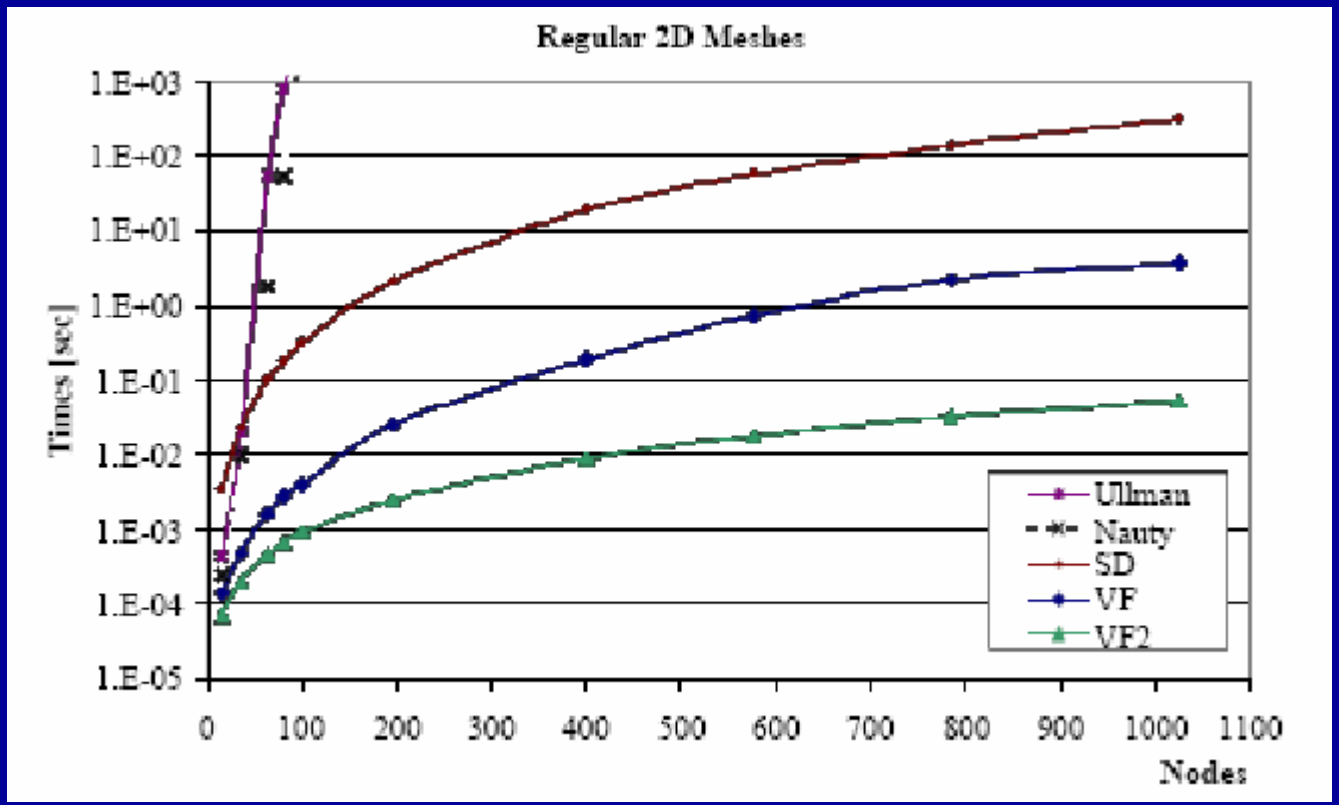
- Storage of chemical structure

3. Mol-file Compression
4. Uniquelization coding

Graph/subgraph matching Algorithms

- Ullman: J.D. Ullman, 1976
- SD: D.C. Schmidt, 1976
- Nauty: B.D. McKay, 1981
- VF: P. Foggia, 1999
- VF2: P. Foggia, 2001

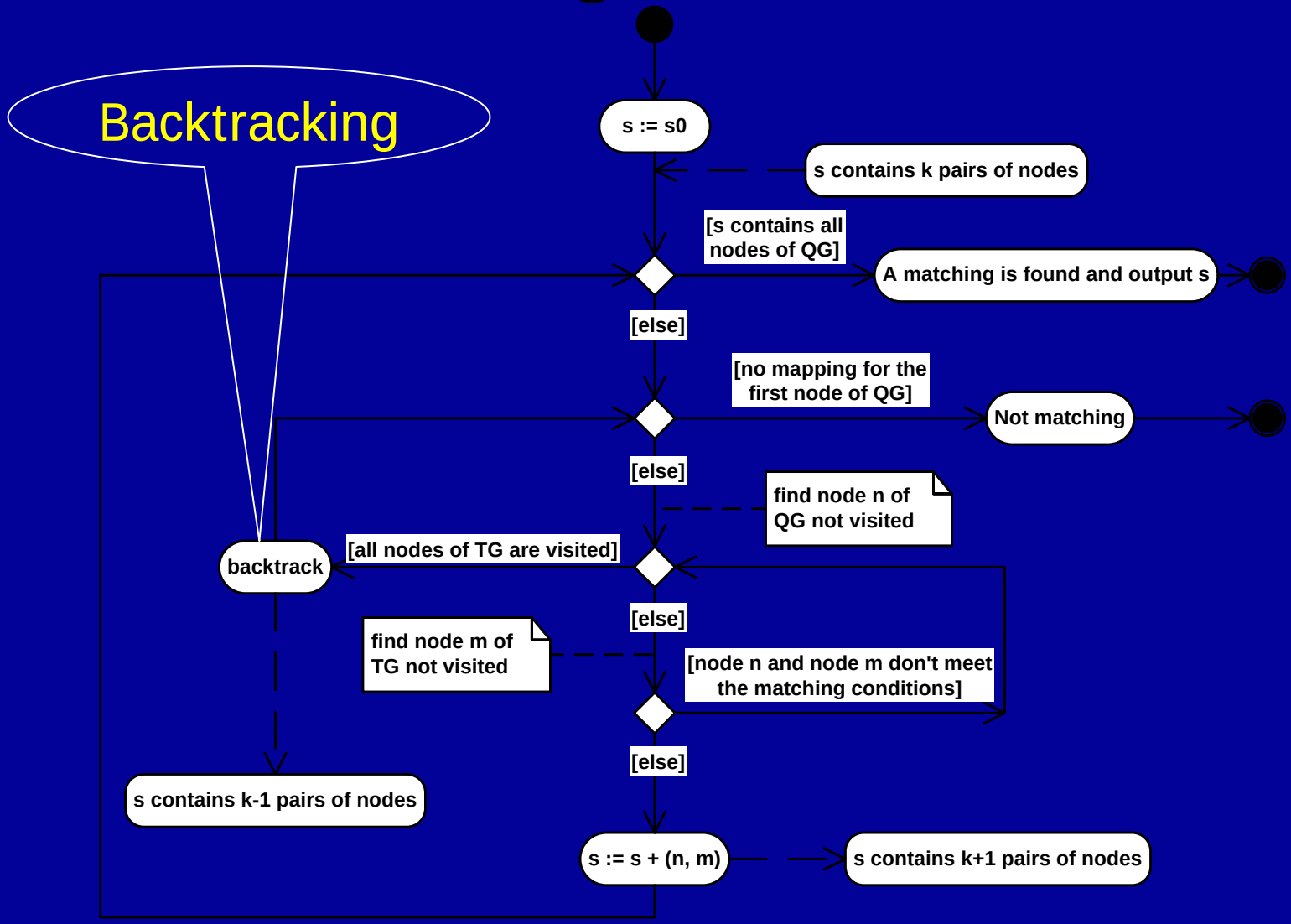
Graph Match Algorithm Comparison



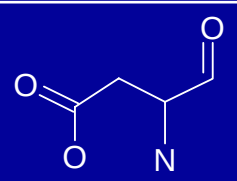
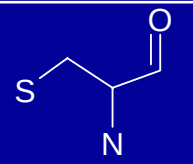
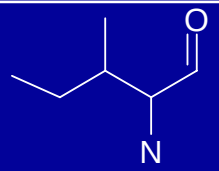
The performance of the five algorithms on *Regular 2D Meshes*

Comes from ”A Performance Comparison of Five Algorithms for Graph Isomorphism”, P. Figgia, 2001

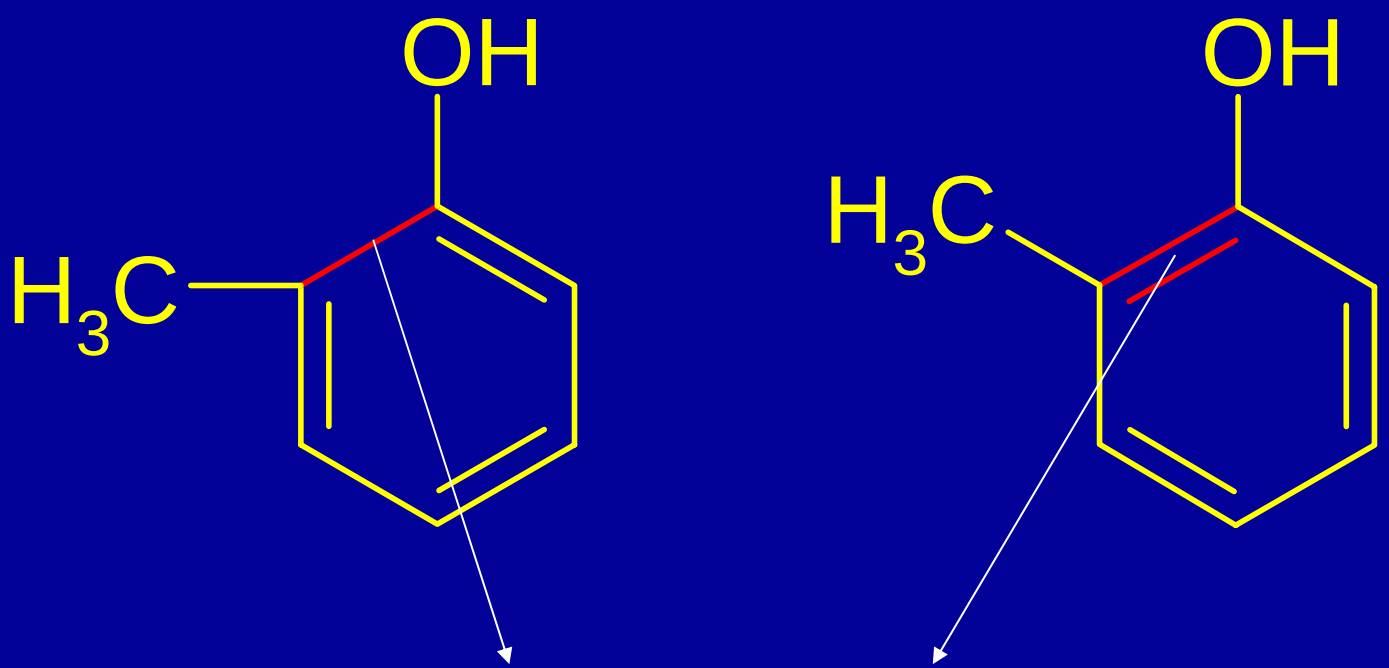
The VF2 Algorithm Flowchart



Comparison with ISIS/Base

Query Structure	Time consumed in VF2/ms	Matching Number in VF2	Matching Number in ISIS/Base
	3596	19	19
	3606	91	91
	3435	12	7

Aromatic ring problem in SSS



Single Bond? Double Bond?

Aromatic Bond

Aromatic Ring Perception

Huckel Rule:

$\sum(\pi \text{ electrons}) = (4n + 2)$

where n is any positive integer

Open Source Programs: Open Babel

Introduction

- Retrieval of chemical structure
  - Structure/substructure search (SSS)
  - Aromatic ring perception

- Storage of chemical structure
  - Mol-file Compression
  - Uniquelization coding

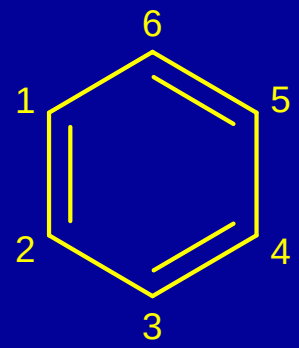
Chemical Structure Compression

Text Plain  Binary Code

Uniquelization Coding

Mol-file  Unique Code

Connection Table (CTab)




```
-ISIS-  04030610472D

  6  6  0  0  0  0  0  0  0  0999 V2000
    0.9652   -1.8166    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.9641   -2.6440    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.6789   -3.0569    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.3953   -2.6435    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.3925   -1.8130    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.6771   -1.4039    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  3  4  2  0  0  0  0
  4  5  1  0  0  0  0
  2  3  1  0  0  0  0
  5  6  2  0  0  0  0
  6  1  1  0  0  0  0
M  END
```

Counts Line

Atom Block

Bond Block

Properties Block

Mol-file Compression Result

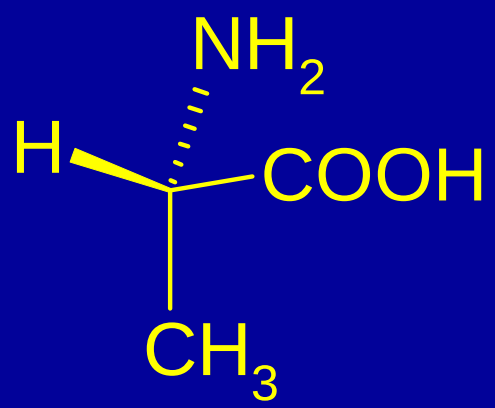
File Before Compression /MB	File After Compression /MB	Compression Ratio /%	Compression Time /sec
38	3.8	90	6
193	30.9	84	57
514	61.5	88	119

Atom Index Table

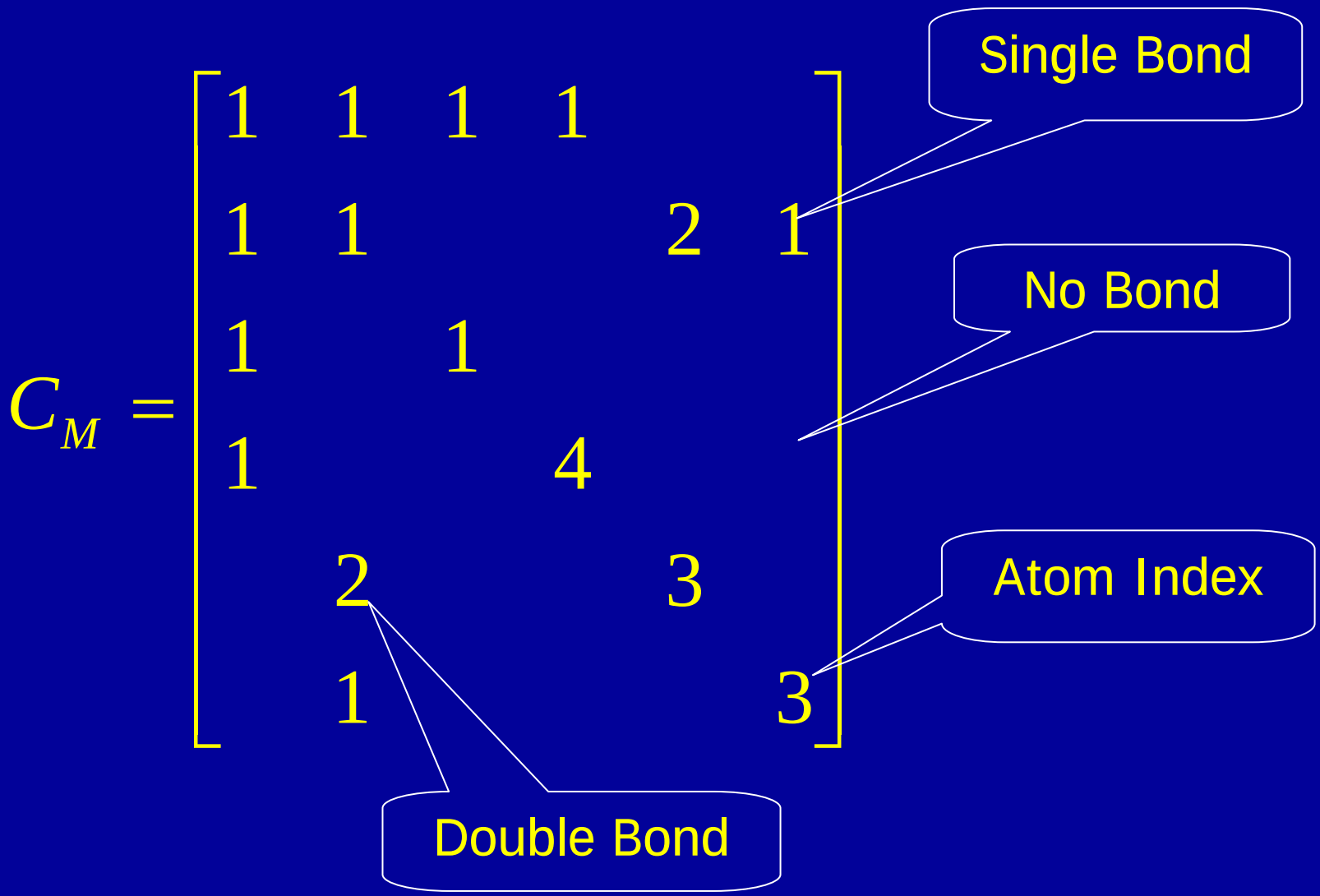
Atom	C	H	O	N	S	Cl	F	Br	…
Percent	**58.3**	**19.7**	**10.3**	**7.64**	**1.37**	**1.21**	**0.879**	**0.233**	**…**
Index	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**…**

L-Alanine

- 0.3458 -1.6208 0.0000 C 0 0 2 0 0 0
- 1.4375 -2.1542 0.0000 C 0 0 0 0 0 0
- 0.3458 -0.0625 0.0000 C 1 0 0 0 0 0
- -1.0917 -2.2458 0.0000 N 0 3 0 0 0 0
- 1.4375 -3.5292 0.0000 O 0 0 0 0 0 0
- 2.6250 -1.5292 0.0000 O 0 5 0 0 0 0
- 1 2 1 0 0 0 0
- 1 3 1 1 0 0 0
- 1 4 1 0 0 0 0
- 2 5 2 0 0 0 0
- 2 6 1 0 0 0 0



Connection Matrix



$$C_p = C_M^4$$

$$C_M^4 = \begin{bmatrix} \underline{54} & 49 & 19 & 109 & 52 & 26 \\ 49 & \underline{135} & 14 & 35 & 172 & 86 \\ 19 & 14 & \underline{10} & 30 & 12 & 6 \\ 109 & 35 & 30 & \underline{316} & 18 & 9 \\ 52 & 172 & 12 & 18 & \underline{241} & 80 \\ 26 & 86 & 6 & 9 & 80 & \underline{121} \end{bmatrix}$$

Uniquelization Coding Result

- The elements on the opposite corner line of Cp matrix are:

  316、241、135、121、54、10
- Square Variance is: 13126
- Unique code:

  316-241-135-121-13126

Conclusion

- Results
  - Programs written in Standard C++
  - Search chemical structure using VF2
  - Compress mol-file and extract unique code
- Problems
  - Efficiency
  - Large-scale Chemical Structures Database Test

Thanks
for attention