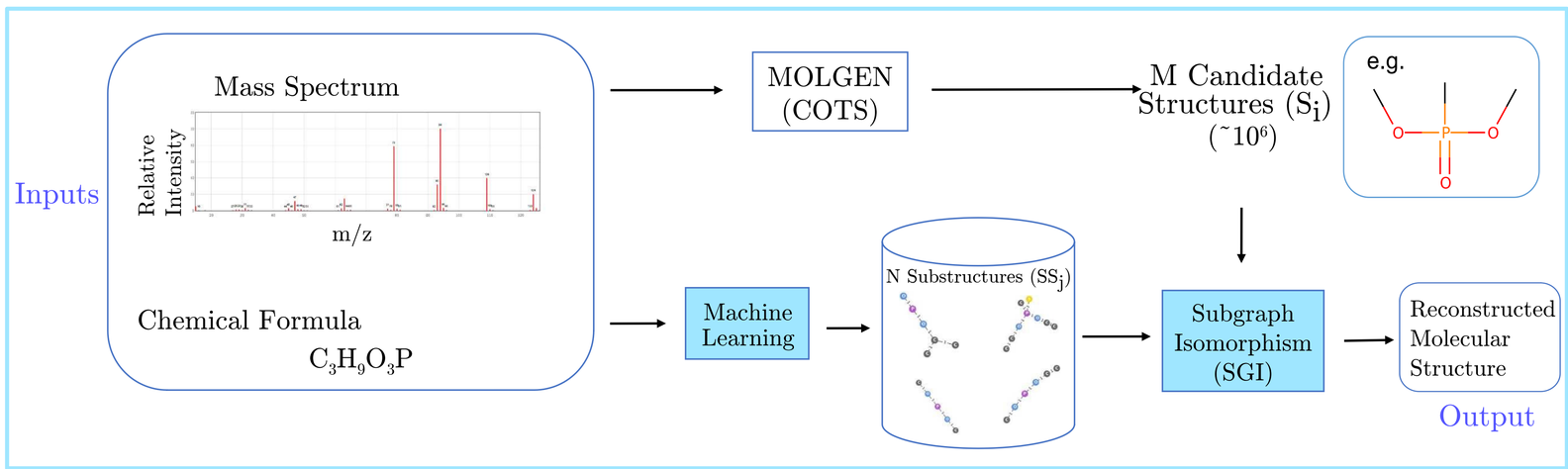
Scaling Up Subgraph Isomorphism

Cheong Sik Feng[1], Kong Xin Yang[1], Chieu Hai Leong[2], Wong Jialiang Joshua[2]
[1]NUS High School of Mathematics and Science, 20 Clementi Avenue 1 Singapore 129957
[2]DSO National Laboratories, 12 Science Park Drive Singapore 118225

Introduction

Chemical Structure Elucidation (CSE) is the process of determining the molecular structure of an unknown molecule.



Problem: SGI is NP-complete (fastest algorithm takes O(V!·V) time [1]) and we need to perform M*N SGI operations. Very computationally expensive!

Proposed Solution: By computing signatures for each S_i and SS_j in O(M+N) time, we can eliminate potential (S_i, SS_j) pairs with SGI if the graph signatures are not compatible.

Figure 1: CSE Process [2]

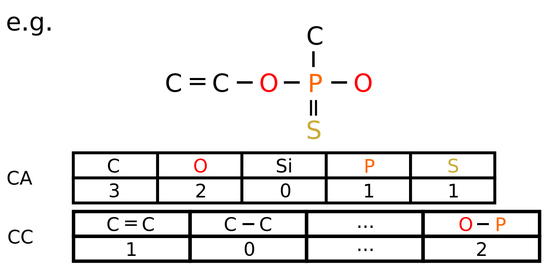
Methods

Graph Signatures we tested:

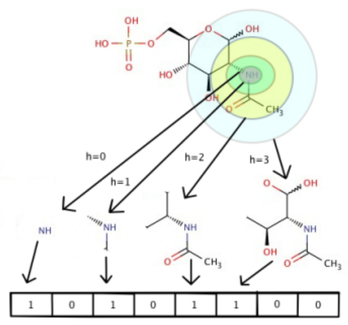
- Implemented by us:
  CA - counts the number of atoms of each element
  CC - counting of number of certain groups
  CF - generates a CountFingerprint
- Already implemented in CDK and RDKit
  BF - generates a BitFingerprint

Additionally, we also tested the performance of the following libraries:

- RDKit (C++ and Python)
- CDK (Java)
- igraph (Python)



(a) CA and CC



(b) BF and CF[3]

Fingerprint

Algorithm 1 Fingerprinter algorithm for BF and CF

Input: Graph G
Output: a BitArray or IntArray

```
for all vertices V in G do TRAVERSE(V)  ▷ Do Depth-First Search (DFS)
end for

function TRAVERSE(V_1)
    Add V_1 to visited
    k ← hash of visited
    kth bit set to 1 or value Incremented by 1
    if len(visited) < 7 then                ▷ Max path length of 6 atoms
        for all vertices V_2 adjacent to V_1 do
            if V_2 not in visited then TRAVERSE(V_2)   ▷ Continue DFS
            end if
        end for
        Unvisit V_1
    end if
end function
```

Results

	Phosphonothionate (PPTN)		Phosphonate (PPN)	
	Time (S)	# SGI Operations	Time (S)	# SGI Operations
none	38.90	10,248,847	1474.87	297,349,392
CA	32.23	8,702,199	1031.62	208,905,713
CC	36.96	8,354,016	1671.23	296,942,352
BF	9.11	252,469	186.61	7,895,048
CF	6.34	192,383	103.71	4,681,512
$CA+CC$	29.18	7,125,558	1218.80	208,905,713
$CA+BF$	8.48	245,219	165.95	7,845,089
$CA+CF$	**6.64**	**192,383**	**93.82**	**4,681,512**
$CC+BF$	9.77	252,469	184.47	7,895,044
$CC+CF$	6.92	192,383	111.20	4,681,512
$CA+CC+BF$	9.21	245,219	173.90	7,845,089
$CA+CC+CF$	7.10	192,383	98.89	4,681,512

Table 1: Comparison of run time and number of SGI operations with different graph signatures. All run time is reported using CDK.

PPTN test case contains 293 graphs and 34,979 subgraphs. PPN test case contains 1,713 graphs and 173,584 subgraphs.

Analysis

Our results show that a combination of CA and CF is the fastest method.

- CF and BF are able to greatly reduce the time taken for SGI, with CF being more effective
- Using CA with BF or CF is able to to reduce the time taken and the number of SGI operations slightly further
- CC is effective when given a smaller number of graphs and subgraphs.

Conclusions

- Graph signatures do indeed reduce run time of SGI
- The fastest method would be to use a combination of counting the atoms of each element (CA) and our CountFingerprint (CF)

Ranking of libraries from fastest to slowest for SGI is:

- CDK (Java)
- RDKit (C++)
- igraph (Python) and RDKit (Python)

References

[1] L. P. Cordella et al. "A (sub)graph isomorphism algorithm for matching large graphs". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.10 (Oct. 2004), pp. 1367–1372. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2004.75. URL: https://doi.org/10.1109/TPAMI.2004.75.

[2] Jing Lim et al. *Chemical Structure Elucidation from Mass Spectrometry by Matching Substructures*. 2018. arXiv: 1811.07886.

[3] SA Rahman. *Revisiting Molecular Hashed Fingerprints*. Oct. 30, 2011. URL: https://chembioinfo.wordpress.com/2011/10/30/revisiting-molecular-hashed-fingerprints/.