Open Source Chemical Structure Generator

Julio E. Peironcely[1,2,3], M. Rojas-Chertó[2,3], Davide Fichera[4], Leon Coulier[1,3], Theo Reijmers[2,3], Jean-Loup Faulon[4], Thomas Hankemeier[2,3]

[1] TNO, Utrechtseweg 48, Zeist, The Netherlands
[2] Analytical Biosciences, Leiden University, Einsteinweg 55, Leiden, The Netherlands
[3] Netherlands Metabolomics Centre, Einsteinweg 55, Leiden, The Netherlands
[4] iSSB, Institute of Systems and Synthetic Biology, University of Evry, 5 rue Henri Desbrueres, 91030 EVRY Cedex, France

julio.peironcelymiguel@tno.nl, peironcely@chem.leidenuniv.nl

Netherlands Metabolomics Centre

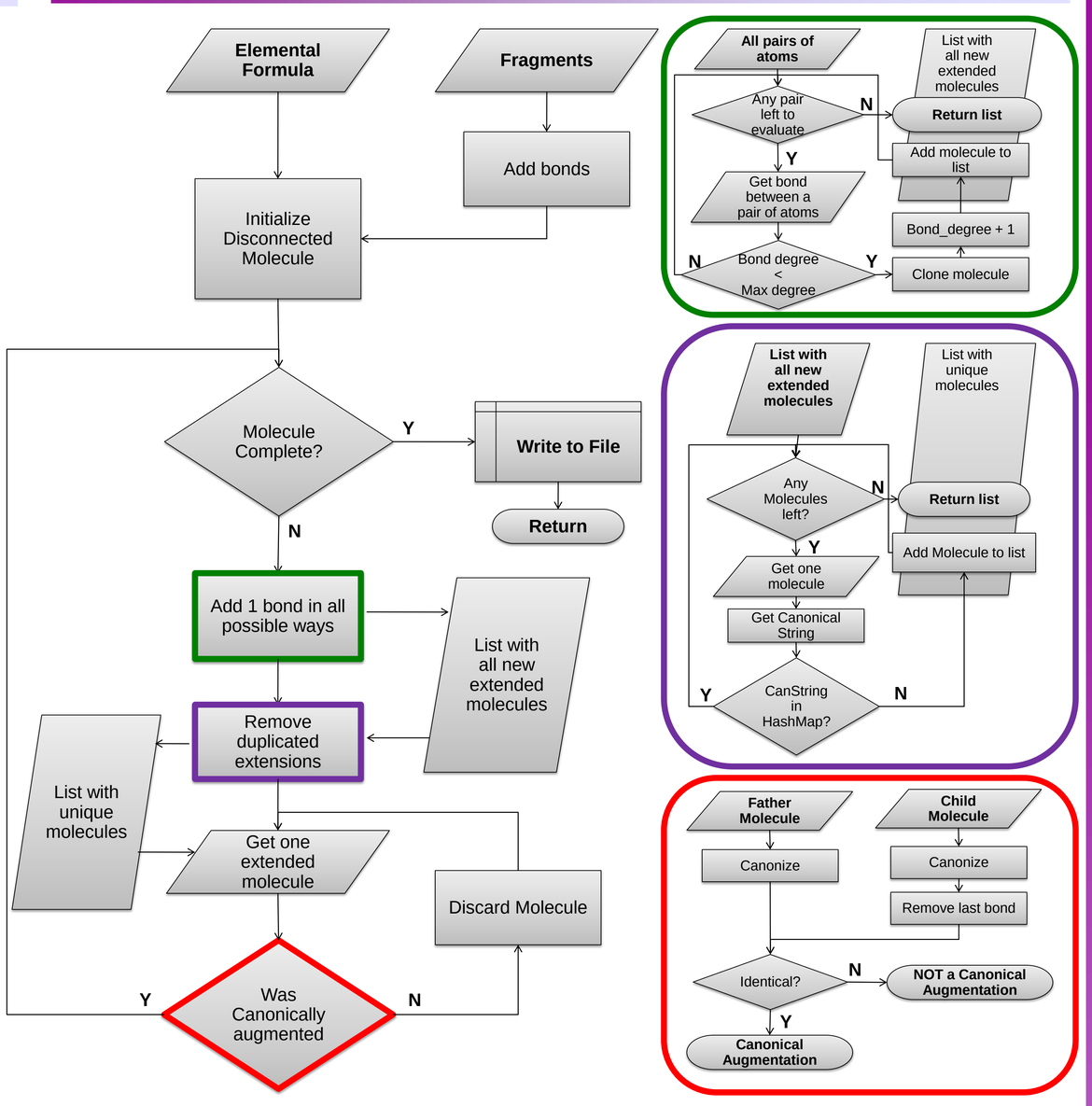
Overview

Computer Assisted Structure Elucidation has been used for decades to discover the chemical structure of unknown compounds. In this work we introduce the first open source structure generator, which for a given elemental formula produces all non-isomorphic chemical structures that match the formula. Furthermore, this generator can accept one or multiple non-overlapping prescribed substructures.

Methods

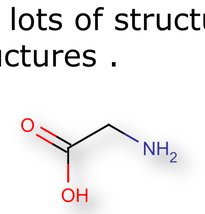
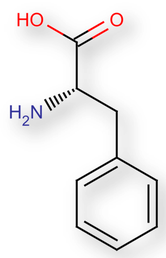
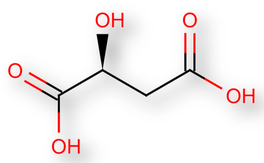
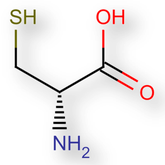
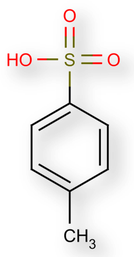
- We have used the approach of "Canonical Path Augmentation"[1](CPA) introduced by McKay to ensure that we exhaustively generate non-isomorphic chemical structures for a given elemental formula.
- CPA is a depth-first backtracking algorithm.
- The generation can be seen as a tree of intermediate chemical structures.
- A bond is added in all possible conformations for a given intermediate molecule.
- Each extended molecule is checked so that the extension is performed in a canonical way.
- A molecule is canonically augmented if the last bond of its canonical representative was the bond augmented.
- Canonical representatives are obtained using an adaptation of the Nauty canonizer.[2]
- This program has been implemented using the Chemistry Development Kit (CDK).[3]
- The program can receive as an input one or multiple non-overlapping fragments(not possible with MOLGEN).

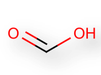
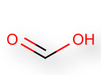
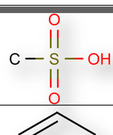
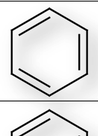
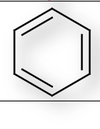
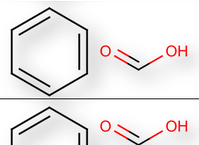
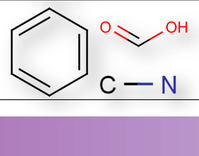
Algorithm



Results

- Our results are comparable to those of MOLGEN, a commercial structure generator.
- Only elemental formula as input = lots of structures.
- Multiple fragments = very few structures .

	Glycine	Phenylalanine	Malic acid	D-Cysteine	p-Cresol sulfate
Elemental Formula	C2H5NO2	C9H11NO2	C4H6O5	C3H7NO2S	C7H8O3S
# Output Molecules	84	277,810,163	8,070	3,838	10,203,389
1 Fragment as Input	 6	 4,037,499	 1,601	 100	 19,940
	 93,137				 948
2 Fragments	 584				
3 Fragments	 278				

MOLGEN Generates same # of molecules

Conclusions

- We implemented the first open source exhaustive structure generator by combining graph theory and an open source library for the development of chemoinformatics software.
- Our results show that the implementation of our algorithm is as complete as the best commercially available generator.
- We can use multiple non-overlapping substructures as constrains to reduce the number of output molecules, not possible in other tools.

Future plans:

- This algorithm is suitable for parallel and high-performance computing, and the inclusion of "while-generating" constraints.

1. McKay, B. 1998. "Isomorph-Free Exhaustive Generation." *Journal of Algorithms* 26:306-324.
2. McKay, B. 1981. "Practical graph isomorphism." *Congressus Numerantium* 30:45-87.
3. Steinbeck, C. 2003. "The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics." *Journal of Chemical Information and Computer Sciences* 43:493-500.
4. Kerber, A. et al. 1998. "MOLGEN 4.0." *Match Communications In Mathematical And In Computer Chemistry* 37:205 – 208.