

Abstract

The chemical diversity in the world is surprisingly extensive and only expands over time, creating the problem of unambiguous and meaningful compound identification. Without unique names various issues occur such as:

- ▶ Definition of unambiguous chemicals in the literature;
- ▶ Searching compounds in databases;
- ▶ Spoken communication about chemicals.

This problem is solved by nomenclatures such as IUPAC. IUPAC describes the rules by which molecules are unambiguously named and molecular structure of a named molecule can be reconstructed.

Due to the complexity of this ruleset, creating a proper chemical name by hand is challenging. Free/libre open source software (F/LOSS) tools to automate this task are scarce [1]. Here we present *ChemOnomatopist* [2], a F/LOSS package to name chemical compounds according to the IUPAC nomenclature.

ChemOnomatopist

- ▶ *ChemOnomatopist* is F/LOSS tool written in Perl under BSD-3-Clause license.
- ▶ *ChemOnomatopist* method uses molecular graphs, graph theory based algorithms and data structures to implement IUPAC nomenclature rules, as opposed to database or machine-learned model-based tools like *STOUT* [3].
- ▶ *ChemOnomatopist* is able to read SMILES linear notation and output IUPAC names without supervision, opening a possibility to process large volumes of chemical data.

Algorithm

1. Read input in SMILES format;
2. Use *Chemistry::OpenSMILES* [4] module to convert SMILES to a molecular graph;
3. Mark functional groups in the graph;
4. Use graph algorithms to apply IUPAC naming rules;
5. Output chemical compound name based on IUPAC rules.

Example:

CC(C)C(CCC)CCC → **4-(1-methylethyl)heptane**

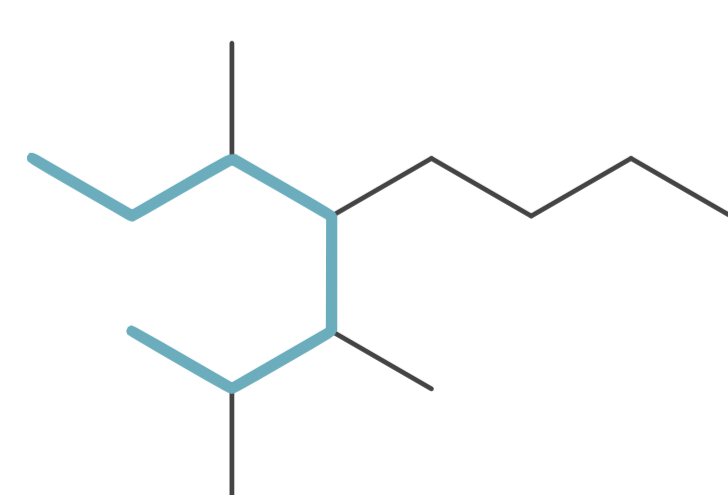
Current state

Currently *ChemOnomatopist* has saturated branched acyclic hydrocarbons as the main focus. Presently new functionalities and rules are being implemented to select the main chain from equal length chains of saturated branched acyclic hydrocarbons.

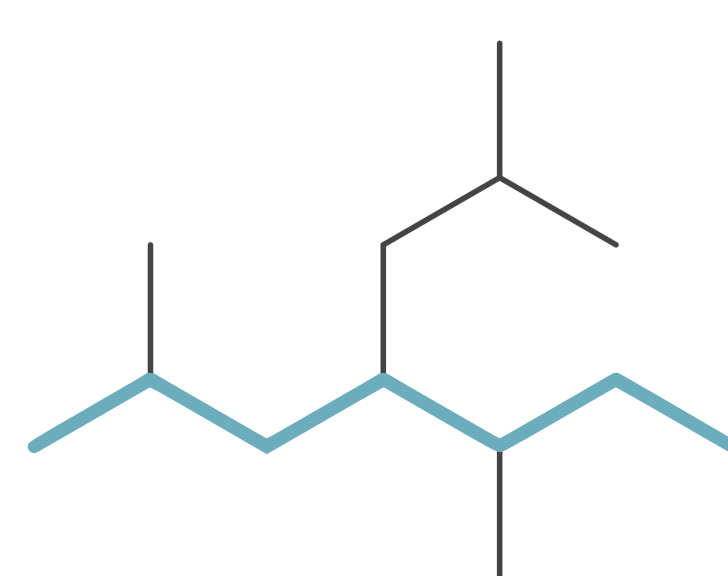
Main chain selection

To select correct main chain, when equal length chains are competing, IUPAC Blue Book [5] (nomenclature ruleset) suggests the following series of rules. In the examples, blue color marks the main chain according to a rule.

- ▶ The chain which has the greatest number of side chains;

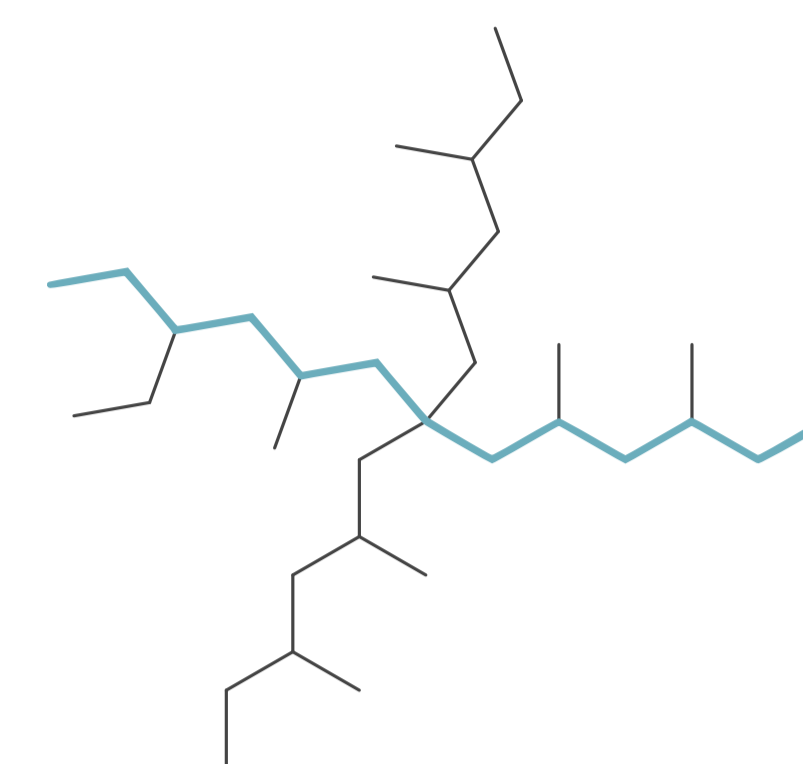


- ▶ The chain whose side chains have the lowest-numbered locants;

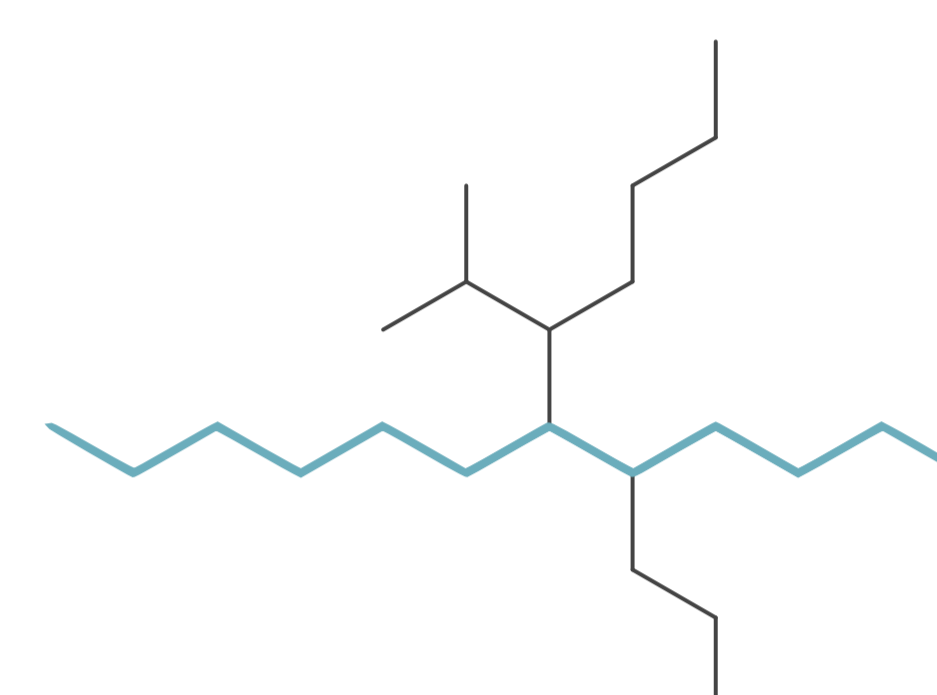


Main chain selection (continued)

- ▶ The chain having the greatest number of carbon atoms in the smaller side chains;

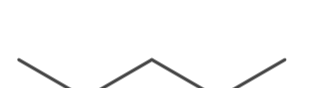
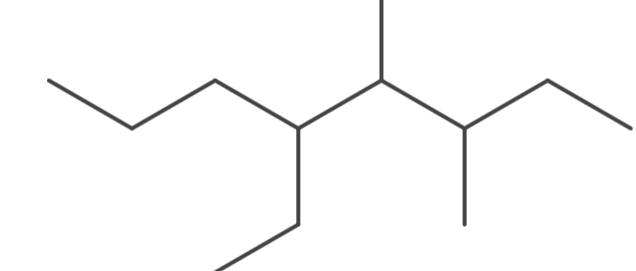
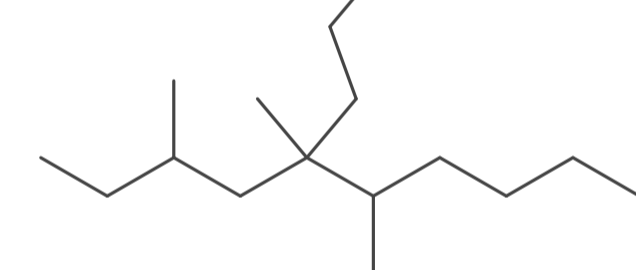
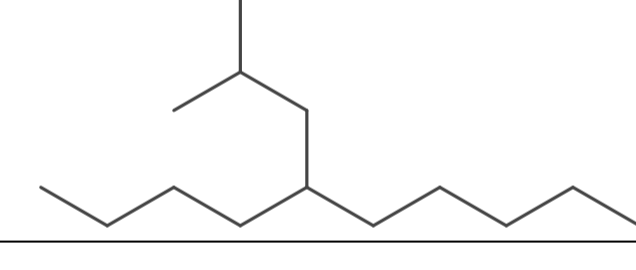
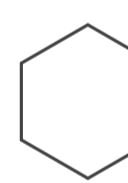
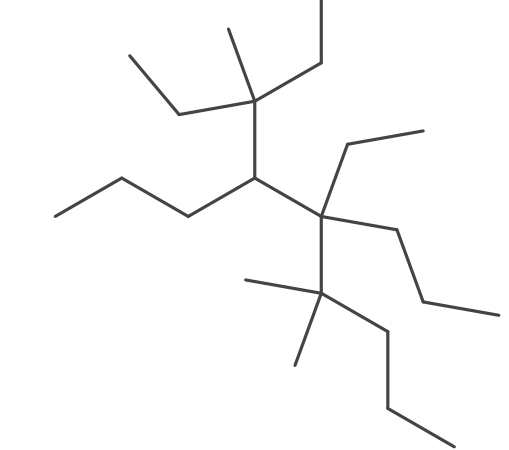


- ▶ The chain having the least branched side chains.



Results

Examples of *ChemOnomatopist* inputs (SMILES notation) and outputs (IUPAC name):

SMILES notation	Chemical graph	IUPAC name
CCCCC		pentane
C(C)C(C(C(C)C)C)CC		5-ethyl-3,4-dimethyloctane
C(CC)C(CC(C)C)C(C(C)C)C		3,5,6-trimethyl-5-propyldecane
CC(CC(C)C)CCCCC		4-(2-methylpropyl)decane
C1CCCCC1		cyclohexane
C(C)C(C(C)C)C(C)C(C(C)C)C(C)C(C)C		5,7-diethyl-4,4,7-trimethyl-5,6-dipropyldecane

Conclusions

Currently *ChemOnomatopist* provides correct names for the most of saturated branched acyclic hydrocarbons and some of simple cycloalkanes. However, the work is not yet finished. With additional time and effort, more and more molecule types will be faultlessly recognised and named with the intention of using this tool in places where commercial software is not an option, such as open science in databases like Crystallography Open Database [6].

Bibliography

- [1] Williams, Antony John, and Andrey Yerin. Automated systematic nomenclature generation for organic compounds. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3:150–160, Sep 2012.
- [2] Merkys et al. *ChemOnomatopist*. <https://github.com/merkys/ChemOnomatopist>.
- [3] Rajan et al. *STOUT*: SMILES to IUPAC names using neural machine translation. *Journal of Cheminformatics*, 13(1), Apr 2021.
- [4] Merkys. *Chemistry::OpenSMILES*. <https://metacpan.org/pod/Chemistry::OpenSMILES>.
- [5] Favre et al. Nomenclature of organic chemistry: IUPAC recommendations and preferred names. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2014.
- [6] Gražulis et al. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research*, 40, 2014.