the WINNOWER

CHEMISTRY

The Functional SMILES Perspective

MATTHEW S. MACLENNAN [1]

1. University of British Columbia

ABSTRACT

Simplified Molecular-Input Line-Entry System or SMILES is a notation scheme for representing chemical structures in a single line of text, encoding atom connectivity and stereochemistry, as well as charge and ring structures. There are a large number of possible SMILES notations for any one chemical structure, which has led to the development of the canonical SMILES notation. In contrast, I describe here a SMILES approach or “perspective” which encodes functional groups into valid SMILES strings. It is shown that this functional SMILES perspective further simplifies the human interpretation of SMILES strings, can be easily formed from reading IUPAC nomenclature, and has the ability to encode limited chemical reaction histories.

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:
msmaclennan@gmail.com

DATE RECEIVED:
June 24, 2015

DOI:
10.15200/winn.143518.87488

ARCHIVED:
June 24, 2015

KEYWORDS:
R, informatics, smiles

CITATION:
Matthew S. MacLennan, The Functional SMILES Perspective, *The Winnower* 2:e143518.87488 , 2015 , DOI: 10.15200/winn.143518.87488

© MacLennan This article is distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and redistribution in any medium, provided that the original author and source are credited.

FUNCTIONAL SMILES PERSPECTIVE

SMILES notation is fun to play with—another reason why SMILES is an appropriate acronym. Because SMILES is a graph/connectivity language in string format, there are various ways to enumerate bond paths and subgraphs in molecules. SMILES generally finds the longest chain of atoms in a molecule and proceeds to connect the loose ends to form rings. Yet, shorter paths can be found and bonds can be connected in a great many ways while still maintaining valid SMILES notation. Therefore, there are many "perspectives" one can take for generating valid SMILES strings.

For instance, the molecule N,N-diethylethylenediamine can be easily represented by the following SMILES (beginning at the primary amine N):

NCCN(CC)CC

There are, however, many other valid SMILES strings to represent this molecule:
CCN(CC)CCN
C(N)CN(CC)CC
N1.C12.C23.N345.C56.C6.C47.C7

and the list goes on. I call each of these valid SMILES strings "SMILES perspectives".

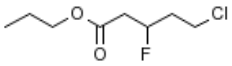
Molecules can be effectively represented with valid SMILES strings which are disconnected and reconnected versions of the functional groups in a molecule and this "SMILES perspective" encodes different (and possibly more) information than general SMILES. I call this the "functional SMILES perspective". The functional SMILES perspective can mirror IUPAC nomenclature but can also mirror the functional group perspectives of the individual chemist.

FUNCTIONAL SMILES PERSPECTIVE FOR REPRESENTING MOLECULES

Let's look at the molecule propyl 5-chloro-3-fluoropentanoate (whatever that is...). The molecule is likely represented with general SMILES as:

CCCOC(=O)CC(F)CCCl

which looks like this:[1,2]

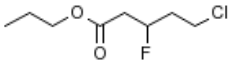


However, you can also represent the structure in a more verbose manner encoding each functional group from the name.

Propyl CCC
Pentanoate CCCCC(=O)O
Fluoro F
Chloro Cl

List all these separated by a period (order does not matter): CCC.CCCCC(=O)O.F.Cl
Finally, connect the fragments appropriately using numbers: CCC1.C3CC2CC(=O)O1.F2.Cl3

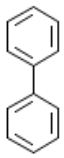
This produces a SMILES string whose molecule looks identical to the first structure:



Let’s emphasize the point with biphenyl.

Biphenyl general SMILES

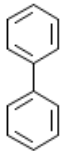
c1ccc(cc1)c2ccccc2



The perspective of this general SMILES string is the general perspective: find the longest continuous chain and link it to form the two rings at the appropriate locations.

Biphenyl Functional SMILES (example)

c1c3cccc1.c2c3cccc2



The perspective of this functional SMILES string is two phenyl groups c1ccccc1 and c2ccccc2 separated by a "." and connected at one carbon (denoted by the number 3). The two benzene or phenyl rings are encoded in the functional SMILES perspective and this strongly reflects the name "biphenyl", *whereas the general SMILES string does not clearly reflect the presence of two phenyls for the human interpreter.*

USING THE FUNCTIONAL SMILES PERSPECTIVE WITH CHEMICAL REACTIONS

The example of chemical reactions is also important. SMILES has a notation for chemical reactions which utilizes the ">" symbol in its notation. One can also encode reaction information using the "." symbol and numbers in the functional SMILES perspective. The following example of esterification reaction between alcohol and carboxylic acid (mediated by ethanol and HCl) is taken from the Daylight webpage.[3,4]

CC(=O)O.OCC>[H+].[Cl-].OCC>CC(=O)OCC

A functional SMILES perspective might want to write the reaction product not as CC(=O)OCC but as CC1(=O).O.O1CC, inserting a "." between the carboxylic C and carboxylic O to denote the breakage of that bond, as well as inserting a "1" after the ethanol O and the carboxylic C to denote the formation of a new bond between those two atoms. The resulting reaction SMILES is:

CC(=O)O.OCC>[H+].[Cl-].OCC>**CC1(=O).O.O1CC**

the product of which additionally includes the H2O by-product (".O."). It is important to note that using the functional SMILES perspective to denote the product allows for easy comparison of bond-breaking and bond-making in the course of the reaction, thus encoding the specific chemical reactions which have taken place.

CONCLUSION

The benefits of the functional SMILES perspective are manifold: SMILES strings become more easily readable; functional groups can be captured, as well as bond-breaking/bond-making reaction histories; the insertion of "." symbol and numbers is easily programmable for generating high-volume data. Putting SMILES strings this way may place a heavy burden on substructure searching, but, if desired, one could use OpenBabel to canonicalize these SMILES strings (i.e. make the SMILES perspective uniform).

References

NCI/CADD Group. (2015). "GIF/PNG-Creator for 2D Plots of Chemical Structures." http://cactus.nci.nih.gov/gifcreator/.

Craig A. Shelley.(1983). "Heuristic approach for displaying chemical structures." *Journal of Chemical Information and Computer Sciences,* 23 (2), 61-65, http://dx.doi.org/10.1021/ci00038a002.

Daylight Chemical Information Systems, Inc. (2015). “Reaction SMILES and SMIRKS.” http://www.daylight.com/meetings/summerschool01/course/basics/smirks.html.

Andrew R. Leach, John Bradshaw, Darren V. S. Green, Michael M. Hann, and John J. Delany III. (1999). “Implementation of a System for Reagent Selection and Library Enumeration, Profiling, and Design.” *Journal of Chemical Information and Computer Sciences,39* (6), 1161-1172, http://dx.doi.org/10.1021/ci9904259.