Extending SMILES to Encode Reaction Mechanisms

• Miguel Velez

Faculty Advisors: • Dr. Peter Gittins, Department of Chemistry • Dr. Jason Sawin, Department of Computer Science

This research was made possible in part by the Grants and Research Office

Introduction

The Simplified Molecular Input Line Entry System (SMILES), is a line notation that represents molecular structures using alpha-numeric characters[1]. SMILES can also be used to represent chemical reactions[3], but they focus on the net rearrangement of atoms rather than the reaction mechanism. The reaction mechanism is valuable information in understanding how a reaction takes place. To address this limitation, we created the Simple Mechanism Of Reaction Encoding System (SMORES) to represent and understand the mechanisms of organic reactions.

SMORES

SMORES is an extension of the SMILES language. The grammar of SMORES language[5] is:

transform	:	**molecule(s) '>>' mechanistic step ;**
molecule(s)	:	**SMILES**
mechanistic steps	:	**mtype , mechanistic steps | null ;**
mtype	:	**+b{class, class}| -b{class,class}| =b{class,class};**
SMILES	:	**a valid SMILES specification that uses explicit class tags**

We identified two types of electron movements in molecular reactions:

1. **Heterokinetic in which a pair of electrons move together**
2. **Homokinetic in which two electrons move independently**

These yield four types of mechanism steps:

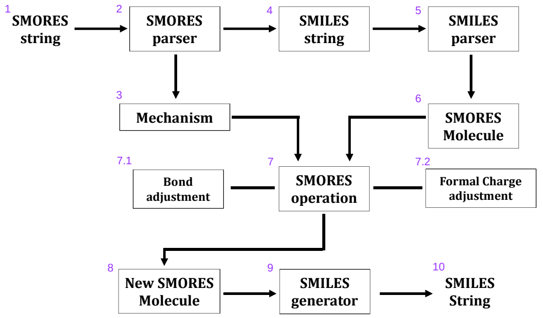
Heterokinetic making of bonds	:	***molecules>>+b{class,class}***
Heterokinetic breaking of bonds	:	***molecules>>-b{class,class}***
Homokinetic making of bonds	:	***molecules>>=+b{class,class}***
Homokinetic breaking of bonds	:	***molecules>>=-b{class,class}***

We extended the Chemistry Development Kit (CDK), which is an open source Java library specific to computer science that can manipulate SMILES strings. Then, we developed the SMORES parser that applies mechanisms to the specified molecule.

- The user enters a (1)SMORES String.
- The (2)SMORES parser processes and creates a (3)SMILES string and (4)mechanism.
- The (5)SMILES parser checks if the user entered a correct SMILES string and creates a (6)SMORES molecule.
- The (7)SMORES operation takes the SMORES molecule and mechanism and (7.1)adjusts bonds and (7.2)formal charges.
- A new (8)SMORES molecule is created.
- The (9)SMILES generator takes the new molecules and outputs to the user a (10)SMILES string with the new structure of the molecules.

Analysis

Logic structure of the program:



1 SMORES string

2 SMORES parser

3 Mechanism

4 SMILES string

5 SMILES parser

6 SMORES Molecule

7 SMORES operation

7.1 Bond adjustment

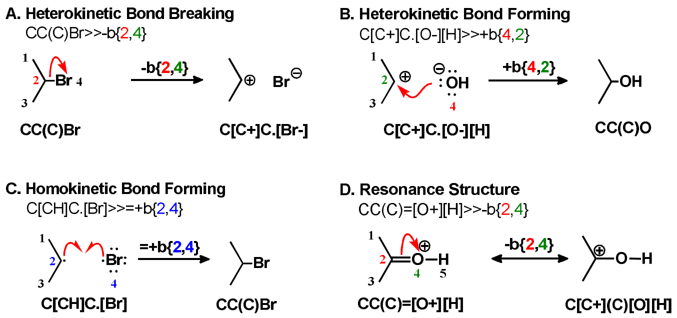
7.2 Formal Charge adjustment

8 New SMORES Molecule

9 SMILES generator

10 SMILES String

Results of our program:



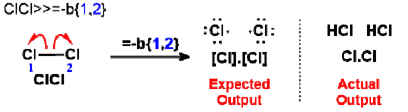
A. Heterokinetic Bond Breaking
CC(C)Br>>-b{2,4}

B. Heterokinetic Bond Forming
C[C+]C.[O-][H]>>+b{4,2}

C. Homokinetic Bond Forming
C[CH]C.[Br]>>=+b{2,4}

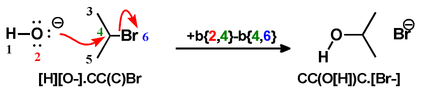
D. Resonance Structure
CC(C)=[O+][H]>>-b{2,4}

E. Homokinetic Bond Breaking: Homolytic Cleavage of Cl_2
ClCl>>=-b{1,2}



Due to the implicit interpretation of H atoms by the CDK, molecules to which homokinetic breaking of bonds are applied do not return the correct product.

F. Multi-substep Mechanism: An S_N2 Reaction
[H][O-].CC(C)Br>>+b{2,4}-b{4,6}



Due to the CDK renumbering of atoms within a molecule after a step in the mechanism, we cannot yet encode reactions with multiple steps or substeps.

Results and Conclusions

- We developed a robust software that processes SMORES strings, which allows users to understand reaction mechanisms.
- It can process heterokinetic making/breaking of bonds, homokinetic making of bonds, and single movement resonance structures.
- It can also process multiple step mechanisms discretely.
- The CDK implies H atoms in molecules, which became an issue when H atoms were directly involved in a reaction or during homokinetic breaking of bonds.
- The CDK also changes the numbering of atoms during chemical reactions which does not allow us to process multi-step mechanisms in a single run.

Future Work

- Automatically render the reactant and product SMILES strings to obtain a graphical representation of the molecules instead of using other software tools.
- Correctly keep track of the numbering assigned to atoms within a molecule before, during, and after reactions using atom mapping techniques.
- Modify the CDK to handle structures with atypical numbers of implied hydrogens and radicals more consistently.
- Extend the SMORES syntax to encode steps, sub-steps, and resonance stretches.

Acknowledgements

I would like express my deep appreciation to Dr. Peter Gittins for his continuous guidance in the progress of this research work. I would also like to thank Dr. Jason Sawin for his valuable advice in the development of this work.

References

1. Apodaca, Richard, Noel O'Boyle, Andrew Dalke, John van Drie, Peter Ertl, GeoffHutchison, Craig A. James, Greg Landrum, Chris Morley, Egon Willighagen, and Hans De Winter. "OpenSMILES Specification" Draft: 2007-11-13, http://www.opensmiles.org/spec/open-smiles.html (accessed October, 2013).
2. "Chemistry Development Kit" *Sourcefoi* 12/11/13, http://sourceforge.net/apps/mediawiki/cdk/index.php?title=Main_Page (accessed February, 2014).
3. Daylight Chemical Information Systems, Inc. "Daylight Theory Manual" 08/01/11, http://www.daylight.com/dayhtml/doc/theory/index.html (accessed February, 2014).
4. May, John. "New SMILES behavior – parsing (CDK 1.5.4)" http://efficientbits.blogspot.com/2013/12/new-smiles-behaviour-parsing-cdk-154.html (accessed May, 2014).
5. Pratt,Terrence, and Marvin Zelkowitz, *Programming Languages: Design and Implementation* (New Jersey: Prentice-Hall, 1995).