General Paper

A Tautomerization Software Based on Lewis Structures and Reaction Mechanisms

Ming Yu*

47 Hampstead Rise NW, Calgary, AB, Canada

E-mail: yuming086@gmail.com

(Received: December 3, 2021; Accepted for publication: April 13, 2022; Online publication: December 22, 2022)

A new method is proposed to translate molecule Lewis structures to machine-readable format data and a software based on Lewis structures has been developed to simulate the practice of chemists writing tautomerization mechanisms. For a tautomeric compound, from one Lewis formula entered by users, the software can generate its other equivalent tautomer(s), which share a common tautomerization intermediate stabilized by resonance. Not only atom balance but also electron balance are guaranteed in the tautomerism representation. The stereochemistry changes caused by tautomeric equilibrium can be tracked too. All the stereoisomers for the compounds containing tetrahedral carbon stereocenters and double bond stereocenters can be identified.

KEYWORDS: Tautomerization, Software, Stereochemistry, Lewis structure, Mechanism

1 INTRODUCTION

Tautomerism refers to a phenomenon in which a set of molecules can interconvert by movement of a hydrogen or group of atoms and/or molecular rearrangement. The presence of multiple tautomers is expected to increase the structural and chemical diversity. The determination of the precise structure of molecules which are potentially tautomeric is an important topic for many purposes [1–6] and the study of tautomerism will remain a challenging field for some time to come. The attempts to use the results of tautomerism reaction to assess the form of a particular compound are misguided [7]. The rational of some efforts [8] was the fact that even less abundant tautomers have been proved to play more important roles in reaction mechanisms. The review [9] covers specifically the instances where the minor tautomers have been shown to play critical roles in biochemical processes. In any case, the first step to address these tautomerism problems is to identify all chemically reasonable tautomers. Do not ignore a tautomer even if it is a minor one [10].

For a given initial structure, it is not difficult for experienced chemists to write its tautomer(s). However, for a tautomeric substance having many potential tautomers, the work to write manually all its tautomers may be a tedious labor work and it is likely that some tautomers could be missed unintentionally. According to the analysis [11] of 722500 organic compounds taken from a database, tautomerism based on hydrogen atom shifts between heteroatoms occurs in 11% of these structures. Although the number of different tautomers does not exceed 20 for most of the compounds, some substances have more than hundreds tautomers. To free chemists from the tedious tautomer generating work, a software tool that can simulate chemist’s tautomer generating practice and can generate all reasonable tautomers from a given initial structure may be desired.

There are many tautomer generation software tools available [1, 11–15] in chemoinformatics. Typically, they are encoded based on molecular line notations, such as InChI and reaction SMIRKS, and implemented based on a set of rules for the possible molecular transformations obtained by applying the common chemistry knowledge. Generally, they are relatively reliable [16]. These software tools can enumerate all possible tautomers, generate a canonical tautomorphic form of a compound, and recognize tautomerism (and handle it appropriately) for chemoinformatics purposes [11]. The major advantage of this approach is that it can calculate very fast everything that needs to be known about the tautomerism of an individual compound, and can be applied to very large numbers of compounds. Despite the flexibility and transparency of this approach, some new problems appear: many rules are companied with exceptions that in turn would require the development of additional rules. It was pointed out that [17] such rules are not usually based on, or verified by, specific experimental analyses. The outcome of such tautomer calculations is dependent on the exact rules used. From an organic chemists’ viewpoint [18], it is possible
that such rules for tautomer enumeration may be too aggressive and not realistic. It was suggested [14] that the tautomerism of the very same molecule can mean very different things to organic chemists and the developers of tautomer software.

Like other organic reaction software tools [18], another issue with these rule-based approaches is their challenge [16] in dealing with the stereochemistry change caused by tautomeric equilibrium, which is a significant and often overlooked effect of tautomerism. Their difficulty can be attributed to the molecular line notations on which the software are encoded. Although stereochemistry control is often the most challenging and interesting part of organic reactions, reaction stereochemistry is not included in the vast majority of the organic reaction simulation works, despite several decades of continuing efforts since the 1960s. This is not an accidental omission but maybe an inherent problem with data structures of computer molecular representation, which has been one of the long-standing problems in the field. It was suggested [19, 20] that this problem is related to the grammar of SMILE/SMARTS due to the lack of keeping track of the change in chirality.

The purpose of this work is to develop a software for tautomer generation with a focus on tracking stereochemistry changes in tautomerization. For a given initial structure, the software will simulate the practice of chemists writing tautomeric equilibrium, who can predict stereochemistry changes in tautomerizations. Unlike the other tautomer generation software tools that are encoded based on molecular line notations, the software is encoded based on molecular Lewis structures which include valence electrons explicitly so that a complete and step-by-step account of how a reaction of organic compounds takes place can be simulated, just like chemists writing tautomerization mechanisms on papers. Different from the other rule-based tautomer generators, which mostly consider only overall transformations and reflect only the net change of several successive mechanistic reactions and may obfuscate the underlying physical reality of chemical reactions, this work is based on mechanistic reactions, an approximation which considers reactions as discrete entities: concerted electron movements through a single transition state, e.g., Ingold’s mechanisms [21]. Mechanistic reactions can be drawn as “arrow-pushing” diagrams explicitly showing the concerted electron movements. Therefore, atom stereochemistry changes in each elementary step can be tracked. Although organic reaction mechanisms are not derived from physical first principles and may be far from physical reality, and they also have many exceptions, they in general are well rationalized from experimental observations and can summarize myriads of data in a simple, yet very effective way. In fact, they save chemists from having to fully memorize all reactions, and even help chemists work out reaction mechanisms that are unfamiliar. This article is organized as follows: a molecular representation method in computers based on Lewis structure is first introduced. Then several example of acid or base catalyzed tautomerization are presented. They were chosen to demonstrate how the stereochemistry changes caused by tautomerism equilibria can be tracked. Finally, an automation tool is implemented to facilitate tautomer enumeration and for users manipulating reaction regiochemistry.

## 2 MOLECULAR REPRESENTATION IN COMPUTER

The method used in this work is to mimic chemist’s view of molecules based on the covalent bond theory. A molecule is an electrically neutral group of two or more atoms, and an atom itself is made up of atomic cores and valence electrons, held together by covalent bonds. A covalent bond corresponds to a pair of valence electrons explicitly so that a complete and step-by-step account of how a reaction of organic compounds takes place can be simulated, just like chemists writing tautomerization mechanisms on papers. Different from the other rule-based tautomer generators, which mostly consider only overall transformations and reflect only the net change of several successive mechanistic reactions and may obfuscate the underlying physical reality of chemical reactions, this work is based on mechanistic reactions, an approximation which considers reactions as discrete entities: concerted electron movements through a single transition state, e.g., Ingold’s mechanisms [21]. Mechanistic reactions can be drawn as “arrow-pushing” diagrams explicitly showing the concerted electron movements. Therefore, atom stereochemistry changes in each elementary step can be tracked. Although organic reaction mechanisms are not derived from physical first principles and may be far from physical reality, and they also have many exceptions, they in general are well rationalized from experimental observations and can summarize myriads of data in a simple, yet very effective way. In fact, they save chemists from having to fully memorize all reactions, and even help chemists work out reaction mechanisms that are unfamiliar. This article is organized as follows: a molecular representation method in computers based on Lewis structure is first introduced. Then several example of acid or base catalyzed tautomerization are presented. They were chosen to demonstrate how the stereochemistry changes caused by tautomerism equilibria can be tracked. Finally, an automation tool is implemented to facilitate tautomer enumeration and for users manipulating reaction regiochemistry.

![Figure 1. Class design of molecules](image)
centers and the compounds containing carbon/carbon or carbon/nitrogen double bond stereocenters, all the stereoisomers can be presented together with their stereochemistry information (Cahn-Ingold-Prelog sequence rules). No single one is missed. In addition, using the sigma bond rotation functionality implemented in the software, users can also display the three-dimensional structures of different conformational isomers.

To enter a Lewis structure, users need to drag, by using the mouse, the elements from the Periodic Table menu to the left window, which is the user input window. The other windows, except the control panel, are the output windows, which present reaction mechanisms, the 3-D structures of a selected molecule, and stereochemistry information. To build a bond between two atoms, users need to right click mouse on an electron, and then click mouse on another electron of another atom, which is similar to the way humans writing Lewis structures on papers.

The software based on this design is atom-based, because computers do not need to know what a molecule is. It is the intention of this work that the approach should be applicable not only to the known molecules but also to the presently unknown molecules, as long as users can write reasonable Lewis structures. The software is designed to predict organic reaction products from the reactant Lewis structures. The major considerations of the design include:

1. The number of possible organic molecules is huge. It is very difficult if not impossible to include all the molecules in a software, not mention most of the possible molecules are unknown. But, they are all composed of the atoms of a small amount of elements. Therefore, atom-based approaches are more cost effective than molecular based methods in terms of coding effort.

2. Molecule Lewis structure can be coded and can be read by software application in this molecule design. The knowledge and intuition of chemists, if being Lewis structure based, such as the concept of resonance and electron push curved arrows, can be translated to machine readable format too. Computer can “look at” molecules in the same way chemists do.

3. Tautomerization is a chemical reaction. Atom balance and atom-mapping are the minimum critia for any reaction prediction software. Atom-based model and software, which include each and every atoms explicitly, may fit this purpose more naturally than molecular based, because all the atoms are accounted for in every step of reactions and no single one could be left behind.

4. Tautomerization may be considered as a stepwise redistribution of electrons between the atoms in molecules. Understanding the “location” and “movement” of electrons during reactions allows us to appreciate what controls reaction, how reaction proceed and highlight the similarities between seemingly unrelated reactions. The use of electron pushing arrows in chemistry is vital when depicting reactions. Including the valence electrons explicitly in the model provides a convenient way to track how electrons and their associated atoms redistribute as bonds are made and broken. Electron balance also helps to prevent the software from making meaningless predictions.

3 TAUROMERIZATION REPREIZATATION IN COMPUTERS

Tautomeromic compounds are dynamic and exist in multiple equivalent or interconvertible tautomers. From one Lewis formula of a tautomeromic compound, entered by the user, the software can present its tautomerization mechanism in a step-by-step manner with electron pushing and generate its other equivalent tautomers.

3.1 Example 1. Acid Catalyzed Sugar Tautomerization

In sugars, the linear and cyclic hemiacetal coexist. These tautomers share a common resonance stabilized intermediate. Figure 2 shows acid-catalyzed keto-enol tautomerism of an aldose containing five carbon atoms, the left part being simple linear tautomerism and right part being open chain and 5-membered ring tautomerism. The keto form is a straight-chain aldose, and the enol form can either be a straight-chain or a ring. The ring size might range from 3 to 6 for the aldose, with 5-membered ring being the dominant.

The initial aldose in this example contains 3 tetrahedral carbon stereocenters, C2, C3, and C4. Therefore, it can exist in 8 stereo-isomers. The isomer with the configuration of (S, S, S) shown in Figure 3 was chosen as the initial material in this example.

The linear enol tautomer, predicted by the software from the initial material, contains 2 tetrahedral carbon stereocenters (C3, C4) and one carbon–carbon double bond stereocenter. The stereochemistry of the tetrahedral centers (C3 and C4) in the products heritages their stereochemistry in the initial material, because the relative orientations of their substituents remain unchanged after the reaction. The double bond stereocenter between atoms C1 and C2 is newly formed due to the reaction. Because the sigma-bond between the two atoms in the carbocation intermediate is free for rotation, the formed alkene can be either a trans or a cis isomer. Therefore, the linear enol tautomer is a mixture of two isomers, as shown in Figure 4. It can be seen that the double bond has the trans configuration in the left and the cis configuration in the right windows.

The predicted cyclic hemiacetal tautomer contains 4 tetrahedral carbon stereocenters (C1, C2, C3, and C4), with the C1 being the newly formed stereocenter due to the tautomerization, as shown in Figure 5. The stereochemistry of C2, C3 and C4 in the products heritages their stereochemistry in the initial material, because the relative orientations of their substituents remain unchanged after the reaction. Because of the symmetric planner conformation of C1 in the carbocation intermediate, the oxygen of the hydroxyl group can attack it from either sides with equal probabilities. Therefore, the newly formed stereocenter center (C1) may take either the S configuration or the R configuration, and the cyclic hemiacetal tautomer is a racemic mixture. To guide the eye, the atom C1 is positioned at the center of the windows. It can be seen that atom C1 has the R configuration in the left and the S configuration in the right windows, according to the scheme of Cahn-Ingold-Prelog.

3.2 Example 2: Base Catalyzed 2-methyl Lamotrigine

It is suggested in ref [22] that there are eight possible tautomers and isomers of 2-methyl lamotrigine. Five of them can be found in the mechanism shown in Figure 6.
The initial material in the mechanism has a double bond stereocenter between atoms C9 and N4, it may take either cis or trans configurations. The cis configuration shown in Figure 7 was chosen in this example.

The second product in the mechanism has two double bond stereocenters. It is mixture of two isomers as shown in Figure 8. The double bond stereocenter between atoms C8 and N5 is the newly formed due to the reaction. Because the sigma bond between the two atoms are free for rotation in the initial material, the newly formed double bond can take either cis or trans configurations. Because the substituent relative orientations of atoms C9 and N4 do not change in the reaction, the stereochemistry of the double bond stereocenter between the two atoms in the product remains as the same as in the initial material. To guide the eye, the atom C8 is

**Figure 2.** Acid catalyzed monosaccharide tautomerization

**Figure 3.** One of the eight 3D configurations of the aldose

**Figure 4.** 3D configurations of the linear enol tautomers

**Figure 5.** Absolute Configuration

**Figure 6.** Absolute Configuration
positioned at the centers of the windows.

The third product of this mechanism has a double bond stereocenter between C8 and N5, that is newly formed because of the reaction. Again, the newly formed double bond can take either cis or the trans configurations. Therefore, it is a mixture of two isomers, as shown in Figure 9. And again, C8 is positioned at the centers of the windows.

4 FUNCTIONALITY ENHANCEMENTS

To facilitate tautomer enumeration, an automation tool was implemented. For a given initial structure entered by user, its tautomers based on the single step acid or base catalyzed interconversion are generated. The procedure is applied again to all newly generated structures with the same catalyst. This will be repeated, until no new structures appear. As this recursive is very exhaustive, a strategies to avoid redundancy was implemented by comparing the Lewis structures of newly generated tautomers and those of previously generated ones. For a tautomeric compound, it is the goal of the software to present all the tautomers on the one hand, and to avoid redundancy on the other hand.

Reaction regiochemistry is another interesting topic. If users disagree with the built-in model, they can specify their preferred reaction sites by right clicking mouse in the initial Lewis structure window. For example, for base catalyzed tautomerizations, users can specify which hydrogen atom will be removed by the catalyst base. For acid-catalyzed tautomerizations, users can specify which nucleophile atom will be protonated at. At the present version, the nucleophile atoms can be either sp2 carbon, or nitrogen, or oxygen atoms.

5 RESULTS AND DISCUSSION

A new design is proposed in order for computers to understand molecule Lewis structures. No attempt is made to convert Lewis structures to textual or other known machine-readable formats. Like the other molecular representation methods, such as line notations and connection tables, the molecular representation method used in the software is also based on covalent theory. The fundamental defect of covalent theory on electron localization can be corrected by using the concept of resonance. Because resonance structures are sets of Lewis structures that describe the delocalization of electrons, the software based on Lewis structures will be able to use the concept more effectively. Chemist’s knowledge on organic reaction mechanisms, if they are based on Lewis structures, can be understood by computers. This paper is on the tautomerization. For a user-entered Lewis structure, the software can present its acid or base catalyzed tautomerization mechanisms,
and generate its equivalent tautomers.

Beyond the other molecular representation methods, Lewis structure contains the lone pair electron information explicitly. Therefore, not only atom balance but also electron balance are guaranteed for the computer representation of tautomeric equilibria. In addition, atom hybridization and the local geometry in the environment of a particular atom, such as reactive sites and stereocenters, can be determined. For the purpose of tracking atom configuration changes in every elementary reaction step, our focus is on atom local geometry. The argument behind this is the fact the most chemical reactions are local events, and the stereochemistry around stereocenters depends only on their substituent relative orientations, which is local geometry too. Because the software is atom-based and all atoms are included in the model explicitly, all the reactive site and stereocenter atoms can be identified from the user entered Lewis formulas. The atom hybridization states, and their local geometries, as well as their stereochemistries can be tracked in a straightforward manner. In the present version, the stereochemistry tracking is limited to chain and simple ring molecules only, entangled and bridged rings are not included. Although the present implementation includes only the first two-row atoms, no fundamental difficulties are expected for the molecules containing atoms of other rows. However, extra coding is needed for their three-D structure estimation and presentation.

While mechanistic reaction representations, which are the basis of this work, are approximation quite far from the Schrodinger equation, it is expected that they are closer to the underlying reality than overall transformation approaches. One advantage of this approach is its transparency to users in terms of bond forming and bond breaking, as well as electron movement in every elementary reaction step.

6 SOFTWARE AVAILABILITY

The trial version of the software and sample cases as well as the software demos are available at https://www.rosoft.website.

7 SUPPLEMENTARY MATERIAL

More examples are provided as supplementary materials and available online.

8 APPENDIX

The data formats of molecules, atoms, and electrons stored in computers are briefly listed below:

```cpp
class CElectron: public CObject
{
    CElectron();
    ~CElectron();
    ................
};
class CAtom: public CObject
{
    CAtom();
    ~CAtom();
    CElectron* electron;
    ................
};
```

Figure 7. One of the two 3-D configurations of 2-methyl lamotrigine

Figure 8. 3-D configurations of the second product

Figure 9. 3-D configurations of the third product

Figure 10. 3-D configurations of the third product

Figure 11. 3-D configurations of the second product
class CMolecule: public CObject
{
    CMolecule();
    ~CMolecule();
    CAtom* atom;
    //......
}

REFERENCES

[1] M. Haranczyk, M. Gutowski, J. Chem. Inf. Model., 47, 686 (2007). DOI:10.1021/ci0602703 , PMID:17381178
[2] F. Oellien, J. Cramer, C. Beyer, W. D. Ihlenfeldt, P. M. Selzer, J. Chem. Inf. Model., 46, 2342 (2006). DOI:10.1021/ci060109b , PMID:17125178
[3] A. Ghosh, T. Wondimagegn, H. J. Nilsen, J. Phys. Chem. B, 102, 10459 (1998). DOI:10.1021/jp983010j
[4] Y. A. Zhang, V. Monga, C. Orvig, Y. A. Wang, J. Phys. Chem. A, 112, 3231 (2008). DOI:10.1021/jp0744026 , PMID:18341304
[5] T. Kalliokoski, H. S. Salo, M. Lahtela-Kakkonen, A. Poso, J. Chem. Inf. Model., 49, 2742 (2009). DOI:10.1021/ci900364w , PMID:19928753
[6] P. Pospisil, P. Ballmer, L. Scapozza, G. Folkers, Recept. Signal Transduct., 23, 361 (2003). DOI:10.1081/RRS-120026975
[7] J. A. Joule, K. Mills, Heterocyclic Chemistry, 5th Edition, WILEY, Singapore, 2010, pp. 12.
[8] D. D. Colasurdo, M. N. Pila, S. L. Laurella, P. E. Allegretti, D. L. Ruiz, Rapid Commun. Mass Spectrom., 35, 9000 (2021). DOI:10.1002/rcm.9000
[9] V. Vipender Singh, I. Bogdan, RNA, 21, 1 (2015). DOI:10.1261/rna.048371.114 , PMID:25516996
[10] J. Elguero, A. R. Katritzky, O. V. Denisko, Advances In Heterocyclic Chemistry; Katritzky, A.R. eds.; Academic: Calif., 2000; Vol. 76, pp. 1–64
[11] T. Thalheim, A. Vollmer, R. U. Ebert, R. Kühne, G. Schüürmann, J. Chem. Inf. Model., 50, 1223 (2010). DOI:10.1021/ci1001179 , PMID:20666407
[12] W. A. Warr, J. Comput. Aided Mol. Des., 24, 497 (2010). DOI:10.1007/s10822-010-9338-4 , PMID:20372974
[13] L. Guasch, M. Sitzmann, M. C. Nicklaus, J. Chem. Inf. Model., 54, 2423 (2014). DOI:10.1021/ci500363p , PMID:25158156
[14] D. K. Dhaked, W. D. Ihlenfeldt, H. Patel, V. Delannée, M. C. Nicklaus, J. Chem. Inf. Model., 60, 1253 (2020). DOI:10.1021/acs.jcim.9b01080 , PMID:32043883
[15] N. T. Kochev, V. H. Paskaleva, N. Jeliazkova, Mol. Inform., 32, 481 (2013). DOI:10.1002/minf.201200133 , PMID:27481667
[16] M. Haranczyk, M. Gutowski, J. Comput. Aided Mol. Des., 24, 627 (2010). DOI:10.1007/s10822-010-9344-6 , PMID:20361238
[17] L Guasch L., Yapamudiyansel, W., Peach, M. L., Kelley, J. A., Barchi, J. J. Jr. and Nicklaus, M. C., J. Chem. Inf. Model., 56, 2149 (2016).
[18] M. Sitzmann, W. D. Ihlenfeldt, M. C. Nicklaus, J. Comput. Aided Mol. Des., 24, 521 (2010). DOI:10.1007/s10822-010-9346-4 , PMID:20512400
[19] S. Szymkuć, E. P. Gajewska, T. Kluczniak, K. Molga, P. Dittwald, M. Startek, et al., Angew. Chem. Int. Ed., 55, 5904 (2016). DOI:10.1002/anie.201506101
[20] M. H. S. Segler, M. Preuss, M. P. Waller, Nature, 555, 604 (2018). DOI:10.1038/nature25978 , PMID:29595767
[21] C. Ingold, Structure and Mechanism in Organic Chemistry; Cornell University Press: Ithaca, NY, 1969.
[22] I. Alkorta, J. Elguero, A. Font, J. Galceran, I. Mata, E. Molins, et al., Tetrahedron, 70, 2784 (2014). DOI:10.1016/j.tet.2014.02.075
[23] K. Lin, Y. Xu, J. Pei, L. Lai, Chem. Sci. (Camb.), 11, 3355 (2020). DOI:10.1039/C9SC03666K