AUTOMATIC GENERATION OF TAUTOMERS

N. Kochev[a1], V. Paskaleva[a], N. Jeliazkova[b]*

a) University of Plovdiv, Department of Analytical Chemistry and Computer Chemistry,

b) Ideaconsult Ltd, 4 A. Kanchev str., Sofia 1000, Bulgaria,

ABSTRACT

A software module for automatic generation of tautomers is introduced. The software is implemented on top of Chemistry Development Kit (CDK) Library. The program enumerates all possible tautomeric forms of a given molecule. All places for double bond/hydrogen atom shift are identified and combined via exhaustive combinatorial algorithm. Different tautomerism cases are described as rules represented in the form: H-X-Y=Z ↔ X=Y-Z-H, where the states are coded as SMILES strings and the H atom positions are denoted as well. For a particular target molecule, each rule is applied by means of exhaustive substructure searching of the rule fragments against given target structure. As a result all possible locations for a shift are recognized. The tautomeric forms in this sense could be described as binary numbers where each digit represents the „shift“ state of each recognized location.

Key words: *tautomer, automatic generation, SMILES, substructure, isomorphism*

[1*] *correspondence author: nick@uni-plovdiv.bg*

INTRODUCTION

Tautomerism is a process of dynamic isomerization of one tautomer converting itself into another. The IUPAC definition [1] generalizes the Tautomerism in the form:

H-X-Y=Z ↔ X=Y-Z-H
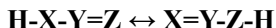
where X, Y and Z are typically C, O, N or S atoms and H is the moving group. When the moving group is H^+ the tautomerism is called „prototropic". Computer-based applications often handle tautomers as different structures because of their different topological representation (i.e. double bond positions are changed). Tautomerism can influence the calculations of pKa, LogP and solubility and also can change the stereochemistry of a compound. Different tautomers have different ligand-receptor interactions. Also structure similarity searching is affected because of the differentiations of tautomeric fingerprints [2]. Software for automatic generation of tautomers is a valuable tool for scientists working in the areas of structure-based drug discovery, design and optimization, structure elucidation and spectra prediction. The latter is the basic motivation for developing yet another software tool for automatic tautomer generation.

SOFTWARE IMPLEMENTATION

The program was implemented using object oriented language Java on top of the open source library CDK [3]. Our software uses the CDK model for structure representation. Figure 1 shows the flow chart of tautomer generation algorithm.

First the molecular structure is inputted into the system as a SMILES string. All possible fragment states for each rule are searched against the target structure thus all rule positions are identified using the AMBIT isomorphism algorithm [4].

Each rule has typically two states coded as 0 and 1. The found locations match one of the states (0 or 1) for each rule. The other state is generated by the software. Then all possible combinations of fragments' states are iterated to generate all possible tautomers (see figure 1).

Currently the software uses eleven rules described by means of SMILES line notation. They cover the range of 1,3 shifting of the moving group. For example the first rule describes the most common type of Tautomerism, keto-enol one, and has two possible states of the molecule. All other rules are listed in table 1.
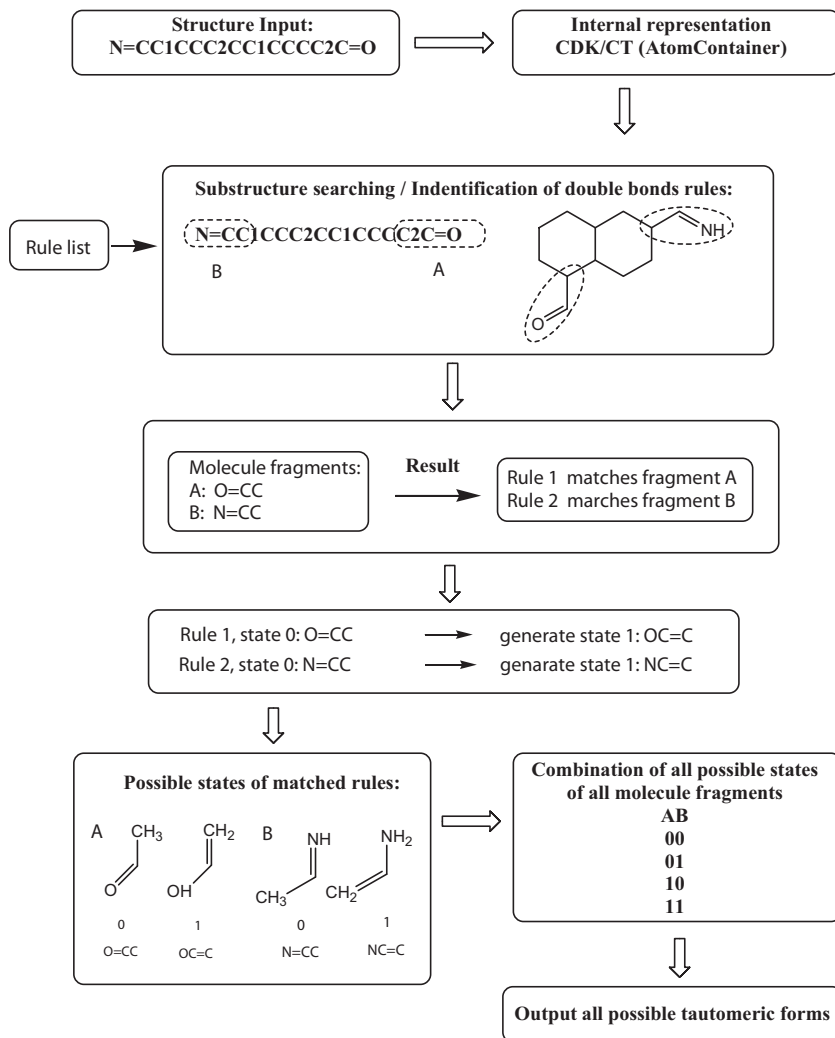
Figure 1. *Flow chart of the tautomer generation algorithm.*

Table 1. *Tautomer rule list*

|	NAME	STATES		H-position
1	keto/enol	O=CC	OC=C	3,1
2	amin/imin	N=CC	NC=C	3,1
3	amide/imid	O=CN	OC=N	3,1
4	nitroso/oxime	O=NC	ON=C	3,1
5	azo/hydrazone	N=NC	NN=C	3,1
6	thioketo/thioenol	S=CC	SC=C	3,1
7	thionitroso/ thiooxime	S=NC	SN=C	3,1
8	amidine/imidine	N=CN	NC=N	3,1
9	diazoamino/ diazoamino	N=NN	NN=N	3,1
10	thoamide/ aminothiol	S=CN	SC=N	3,1
11	nitrosamine/ diazohydroxide	O=NN	ON=N	3,1

EXAMPLE OF SOFTWARE APPLICATION

For illustration purposes we use the molecule of 6-(iminomethyl) decahydronaphthalene-1-carbaldehyde shown in Figure 2. It contains two fragments A and B which correspond accordingly to the rules: 1 – keto/enol rule and 2 – amino/imino rule.
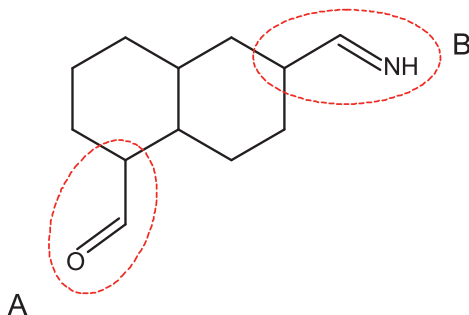

Figure 2. *Rule application for the molecule of 6-(iminomethyl) decahydronaphthalene-1-carbaldehyde. Fragments A and B are identified.*

Fragment A (keto/enol rule) is found in state 0 – „keto". Fargment B (amino/imino rule) is found in state 1 – „imino".

A 0 1
O=CC OC=C

B 0 1
N=CC NC=C
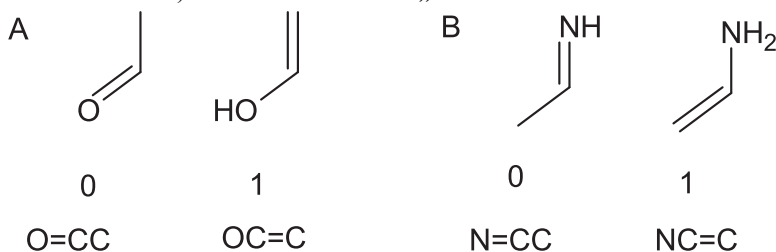
Figure 3. *Rule states of fragments A and B and their codes.*

To get all possible tautomeric forms for the molecule all combinations of states of fragments A and B are generated: *00 – keto, imino; 01 – keto, amino; 10 – enol, imino; 11 – enol, amino.*

00   10

01   11
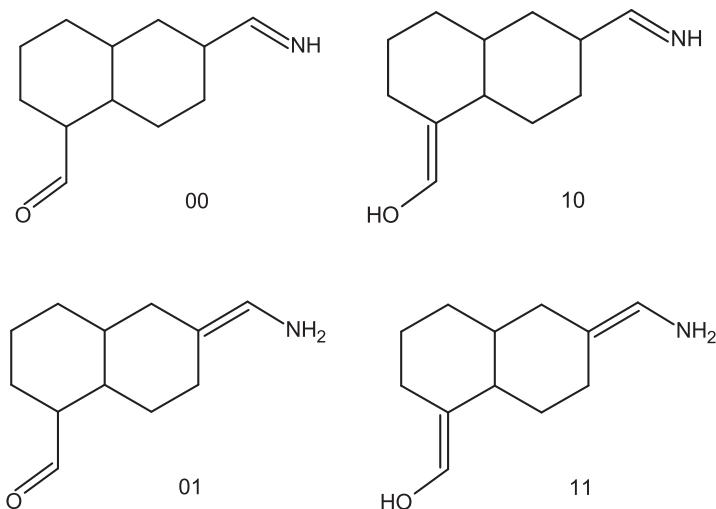
Figure 4. *Result (generated) tautomeric forms.*

CONCLUSIONS

Described software module handles well the basic forms of tautomerism and can be used for handling small organic structures. This is the first version

of a bigger project for automatic generation of tautomers. The software module is part of AMBIT software platform for Chemoinformatics. Current version of the software handles well molecules for which the described rules from table 1 are not overlapped. Also 1,5 shifts or higher distances shifts are not regarded.

REFERENCES

1. IUPAC Compendium of chemical terminology (electronic version). http://goldbook.iupac.org./T06252.html, *accessed* 18 Nov, **2011**.
2. Milleti F., Storchi L., Sforna G., Cruciani G.; Tautomer enumeration and stability prediction for virtual screening on large chemical databases. J. Chem. Inf Model, **2009**, 49(1), 68–75
3. Chemistry Development Kit (CDK); http://almost.cubic.uni-koeln.de/cdk/cdk_top
4. Nina Jeliazkova, Nikolay Kochev, AMBIT-SMARTS: Efficient Searching of Chemical Structures and Fragments; Molecular Informatics, 30 (8), **2011**, 707–720. http://ambit.sourceforge.net, *accessed* 18 Nov, **2011**.