VI МЕЖДУНАРОДНА НАУЧНА КОНФЕРЕНЦИЯ НА МЛАДИТЕ УЧЕНИ
11 – 13 юни, Пловдив, 2015г.

СЪЮЗ НА УЧЕНИТЕ
ПЛОВДИВ

Ambit-Tautomer Ranking System: Analysis and optimization

Part I. Speed optimization of the tautomer generation process and filtering methods

Vesselina Paskaleva[a]*, Nikolay Kochev,[a] Ognyan Pukalov[a], Atanas Terziyski[a], Nina Jeliazkova[b]

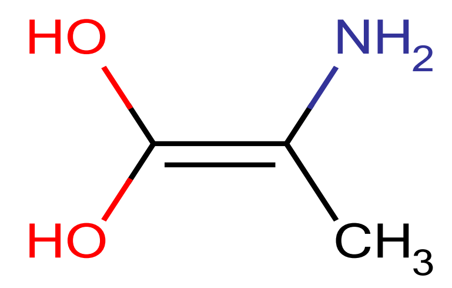
[a] University of Plovdiv, Department of Analytical Chemistry and Computer Chemistry
[b] Ideaconsult Ltd, 4 A. Kanchev str., Sofia 1000, Bulgaria

Ambit-Tautomer workflow and ranking system

Software characteristics

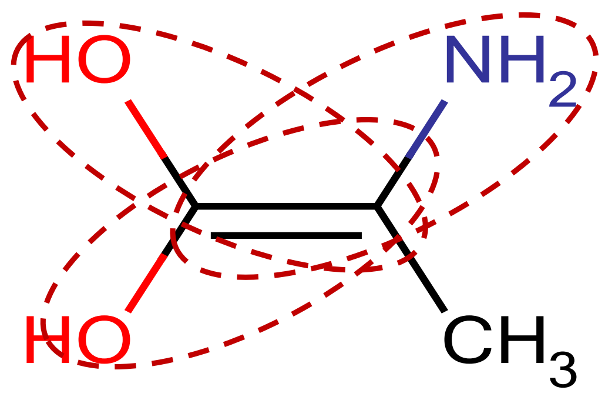
- CDK based structure representation
- Supports standard chemical formats: SMILES, InChI, MOL/SDF file, CML
- Exhaustive tautomer generation
- Customizable set of rules and post-generation filters
- Set of predefined rules
- Tautomer ranking based on simple empirical rules

Structure input
OC(O)=C(N)C



(CDK representation)

Substructure search



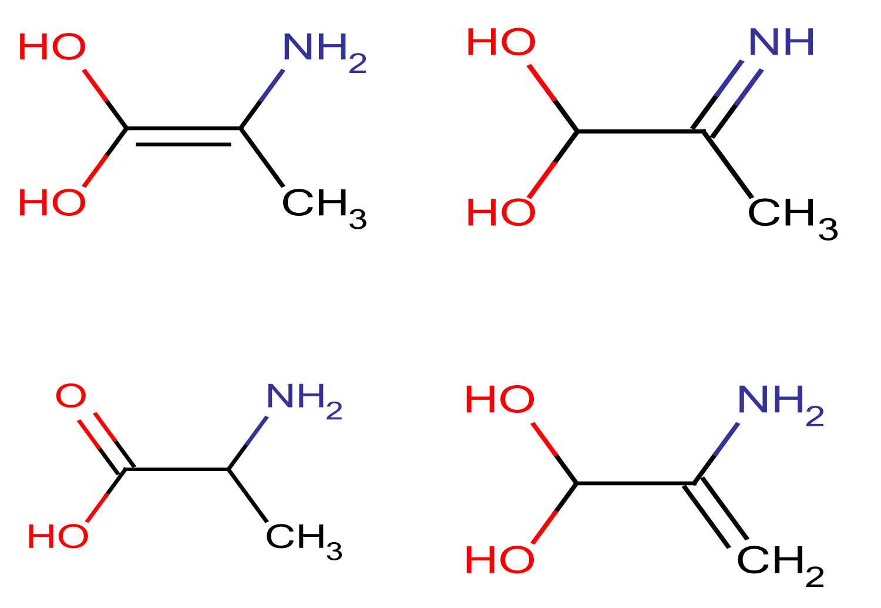
Initial rule list

Generation of all possible combinations of the rule states based on **Depth-first search** with refinement of the rule list at each step.

Post-generation filtering
duplicates, topological equivalency, allene atoms, incorrect structures…

Ranking

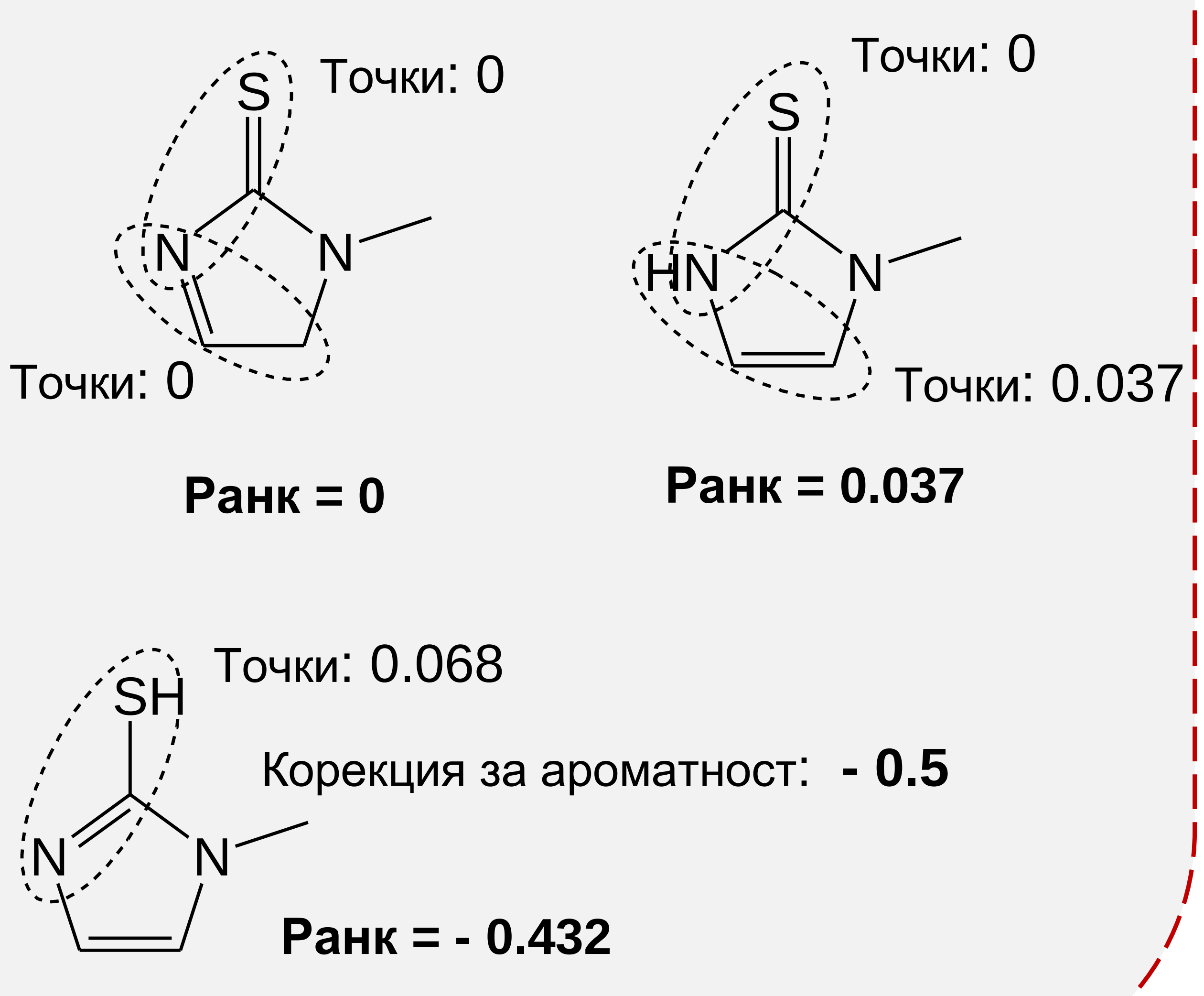
Result output



$$Rank = \sum_{i=1}^{Nrules} ES_i^{state(i)} + N_{aromAtoms} C_{arom}$$

Rule States	Energy scores (eV) ES_i^0 , ES_i^1
O=CC ↔ OC=C	0 ↔ 0.315
N=CC ↔ NC=C	0 ↔ 0.037
O=CN ↔ OC=N	0 ↔ 0.673
O=NC ↔ ON=C	0.025 ↔ 0
N=NC ↔ NN=C	0 ↔ 0.137
S=CC ↔ SC=C	0.246 ↔ 0
S=NC ↔ SN=C	0.983 ↔ 0
N=CN ↔ NC=N	0 ↔ 0.137
N=NN ↔ NN=N	0 ↔ 0.173
S=CN ↔ SC=N	0 ↔ 0.068
O=NN ↔ ON=N	0 ↔ 0.639

Example:



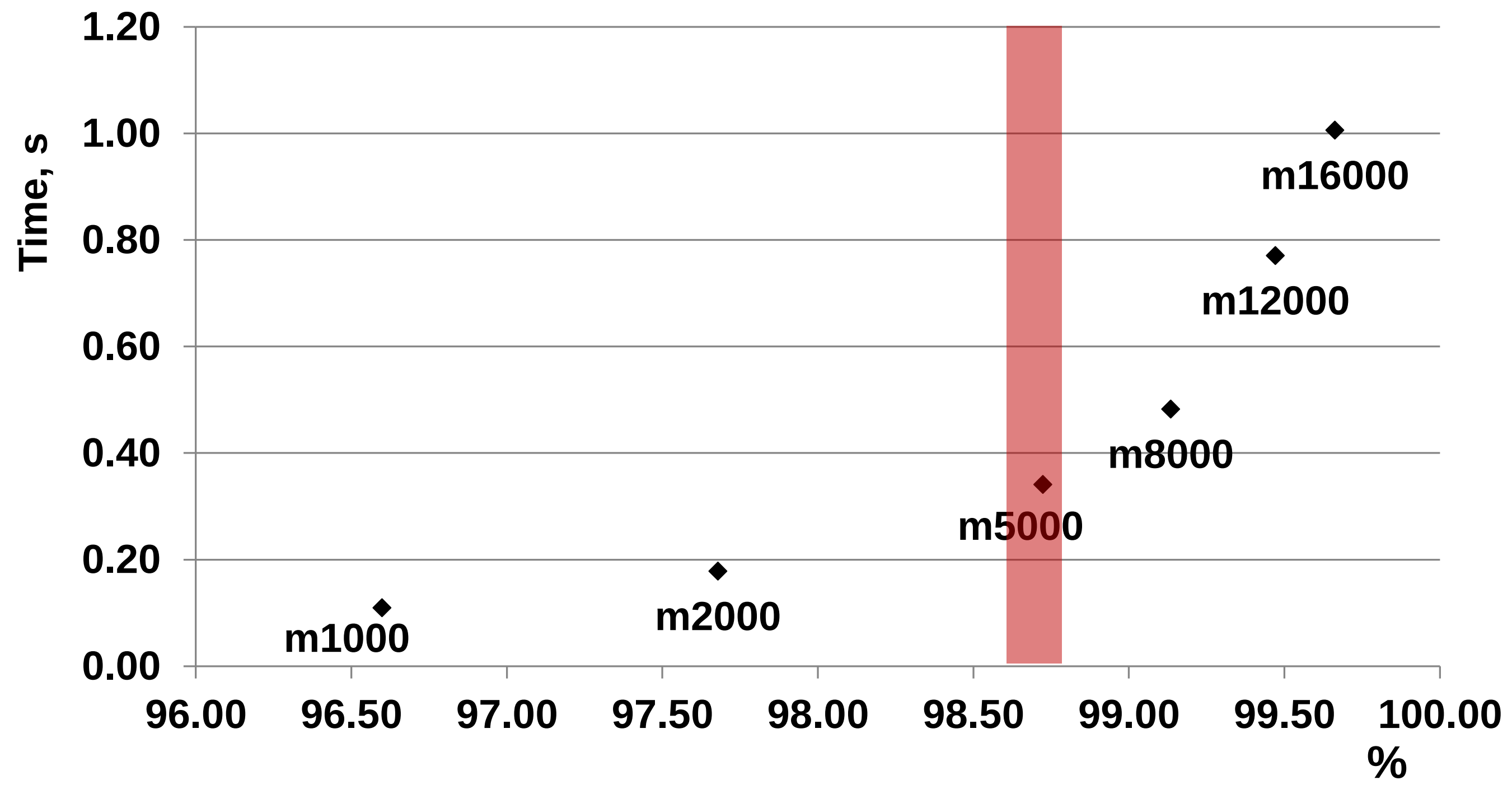
Ранк = 0

Ранк = 0.037

Ранк = - 0.432

Ambit-Tautomer [1] is a new open source software tool for automatic generation of all the tautomeric forms of a given organic compound. Tautomerization is important in a number of chemoinformatics routines such as structure representation, chemical database searching, molecular descriptor calculation, estimation of physicochemical properties, QSAR modelling, virtual screening and more. Ambit-Tautomer is part of the Ambit2 [2] built on top of the Chemistry Development Kit library [3]. Ambit-Tautomer utilizes a depth-first search algorithm, combined with a set of rules for tautomeric transformation. Each rule represents two possible states of the molecule part, which undergoes tautomerization.

Optimal number of backtracks for IA-DFS algorithm



Graph.1 Represents the results obtained for different used backtracks (m) as percentage according to generated time. The optimal number of backtracks is chosen to 5000 and is marked in red.

Time,s	1000	2000	5000	8000	2000	16000	20000
(1)	0.11	0.18	0.34	0.48	0.77	1.01	1.19
(2)	0.10	0.15	0.27	0.35	0.47	0.58	0.64
(3)	0.29	0.46	0.84	1.08	1.45	1.75	1.91
(4)	0.31	0.50	1.02	1.42	2.32	3.10	3.64
Diff	1000	2000	2000	8000	12000	16000	
(1)	1.08	1.02	0.86	0.71	0.44	0.21	
(2)	0.54	0.49	0.37	0.29	0.18	0.09	
(3)	1.62	1.46	1.10	0.84	0.50	0.24	
(4)	3.33	3.14	2.64	2.23	1.36	0.63	
%	w(1000)	w(2000)	w(5000)	w(8000)	w(12000)	w(16000)	
(1)	96.60	97.68	98.72	99.13	99.47	99.66	
(2)	97.04	98.05	98.95	99.32	99.57	99.72	
(3)	91.30	94.43	97.29	98.22	98.95	99.56	
(4)	90.30	93.64	96.83	97.87	98.75	99.48	

Legend:
(1) – all structures;
(2) – structures with tautomers less then 400;
(3) – (2) and 4-10 applied rules;
(4) – 4-10 applied rules, no limit of tautomer count

The implementation used for the current test can be found on:
https://github.com/ideaconsult/examples-ambit/tree/master/tautomers-example

Optimization of duplicates filter parameters

The default system parameter for removing duplicates uses isomorphism check (z) and filters the result isomorphic tautomeric forms. This was supposed to be the most slow part of the tautomer generation process. So we implied a new duplicate filter based on InChI. It finds and remove duplicates by comparing the InChI keys of the generated tautomeric forms. We suggested that the new filter will remove one of the kekule forms of the aromatic structures. For checking our suggestion we compiled a subset of publically available structures part of the ChemBL19.

	nOff, zOn, s [%]	nOn, zOff, s [%]
Total time	313	179
Generation time	261 [83]	124 [69]
IO/convert time	53 [17]	55 [31]

After removing the aromatic structures the final size of the subset was reduced to 2971 non-aromatic structures. This pre-filtering was necessary for providing the differences of applying the two post-filters for removing duplicates to be caused only from new generated aromatic tautomers. It is expected that for non-aromatic tautomers the two filters will give the same results. The speed was inspected and the results are given in the table in left.

References

[1] Kochev, N. T., Paskaleva, V. H. and Jeliazkova, N., Ambit-Tautomer: An Open Source Tool for Tautomer Generation. Mol. Inf., 32: 481–504, 2013

[2] AMBIT project, http://ambit.sourceforge.net

[3] Steinbeck C., Hoppe C., Kuhn S., Guha R., Willighagen E.L., “Recent Developments of the Chemistry Development Kit (CDK) – An Open-Source Java Library for Chemo- and Bioinformatics”. Curr. Pharm. Des. 2006; 12(17):2111-2120 (DOI: 10.2174/138161206777585274)

Acknowledgement to the project НИ15-ХФ-001 and ИТ15-ХФ-001 funded by the University of Plovdiv