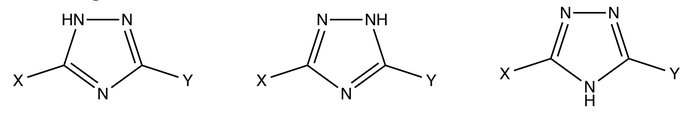
EVALUATING TAUTOMER STANDARDISATION RULES USING TAUTOBASE

Ingvar Lagerstedt and Roger A. Sayle
NextMove Software Ltd, Cambridge, UK.

Introduction

Standardising chemical structures may be done for several reasons and can encompass many steps, including but not limited to: functional group representation; salts, decompose or not, and how to represent bonding; resonance neutralisation; formal charge location molecular and elemental radicals; neutralisation for salts and to adjust for pK_a; and tautomer representation.
We are here looking at tautomer representation. Depending on the use case, different representations may be useful. For lookup purposes, aiming for high symmetry simplifies the problem, e.g., 4*H*-1,2,4-triazole.



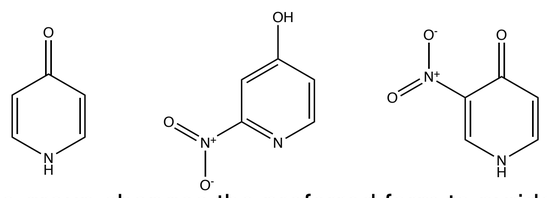
1*H* 2*H* 4*H*

If standardising for lookup, choosing 4*H* avoids making a decision between 1*H* or 2*H* when X and Y differ.

When looking at reactions, using the dominant tautomeric form in the reaction medium (neat, dissolved in chloroform, etc.) is preferable, and if interested in how the molecule interacts in the body, dissolved in water, or ideally in 0.9 % saline solution is useful. Another approach is to enumerate all possible tautomers, and see if any matches. Here we focus on standardising for molecules dissolved in water.

Substituent effects

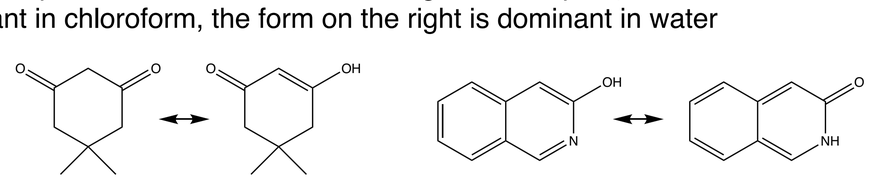
Substituents can significantly change the tautomer distribution, though to what degree varies. For 4-pyridone in water, 2-substitution has a much larger effect than 3-substitution



A 2-nitro group changes the preferred form to pyridol in water, but a 3-nitro group has little impact.

Solvent effects

Two examples where the solvent has a significant impact, in each the form on the left is dominant in chloroform, the form on the right is dominant in water



Some available tools

Tools for standardisation are available from various vendors and toolkits, and what form to adopt is to some degree a matter of style, e.g., charge separated or symmetric nitro groups. In some cases an underlying toolkit may influence the choice, e.g., RDKit's charge separated hypervalent halogens.

PubChem has a standardisation pipeline[3] based on OpenEye's toolkit[8]. Their public web server only allows user to upload one structure at a time.

Some pharma companies have internal pipelines, Eli Lilly and Co. made their tools publicly available, LillyMol[6]. It is quite flexible, we used the following command line:

```
preferred_smiles -g rvnv5 -g guan -g Rguan -g azid -g msdur -g
Rn+n- -g imidazole -g pyrazole -g triazole -g tetrazole -g
isoxazole -g aminothazole -g pirazolone -g arguan -g ltlt -g ltltr.
```

A team from IBM used their Transformer deep learning protocol to assign the major tautomer.

Data sources

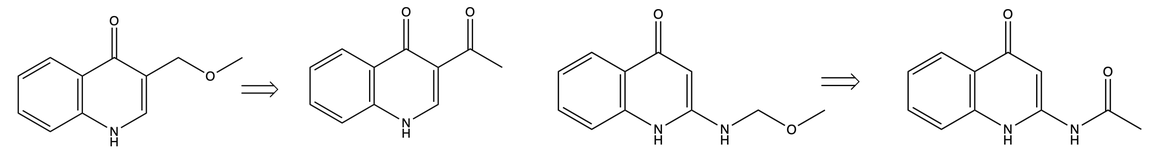
Finding enough reliable data on tautomers to create rules is not easy. We have used Tautobase[1], which is based on paper by Taylor and Kenny[2]. A useful older source is the book by Elguero et al[5].

Tautobase[1]

Tautobase contains a collection of 1680 tautomer pairs in a variety of solvents taken from Taylor and Kenny[2]. Of those 925 is in water and another 403 has unknown/undefined solvent/state. The assignments are done from a mixture of measurements, LFER calculations, or calculation of the relative energy of the relevant tautomers.

Always check your sources

We noticed that some of the Tautobase entries had functional groups that was somewhat unexpected, so decided to check the entries in the source. We have so far identified twelve cases where the entry in Tautobase do not match Taylor and Kenny.



Tautobase 598 Taylor & Kenny Tautobase 601 Taylor & Kenny

Tautobase 1167 Taylor & Kenny

Three Tautobase examples that differs from their source (Taylor and Kenny)

Methodology

We have so far added rules covering 270+ Tautobase pairs. The rules are written as SMARTS patterns used by the application rxnfix, which is part of HazELNut[9]. HazELNut can use our in-house toolkit as well as OpenEye's[8] and RDKit's[7] toolkits.

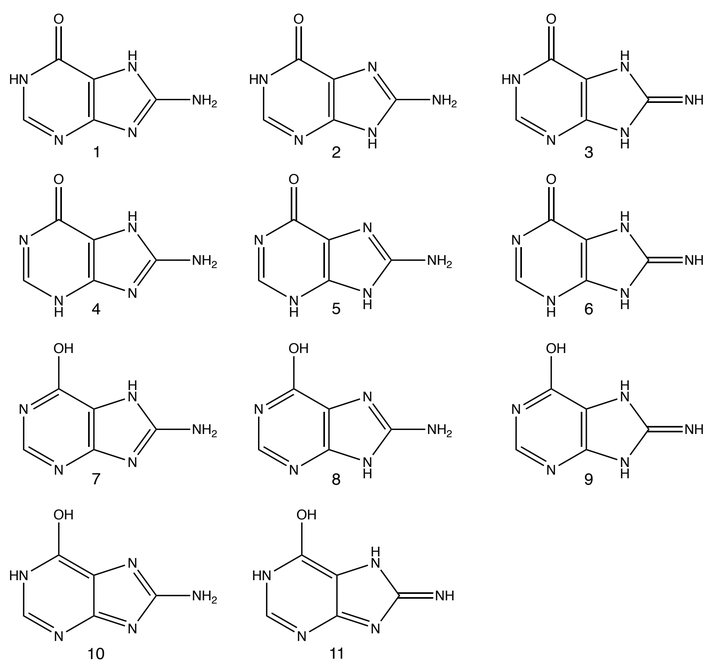
An example:

```
# 1,2,4-TRIAZIN-3-ONE: 2H preferred

[OD1h1+0:1][CD3h0+0:2]1=[ND2h0v3+0:3][ND2h0v3+0:4]=[CX3v4+0:5]
[CX3v4+0:6]=[ND2h0v3+0:7]1>>[Oh0:1]=[C:2]1-[Nh1:3]-[N:4]=[C:5]-
[C:6]=[N:7]-1
```

We found that our transforms work better if they are applied to localised structures. This makes it easy to control that the valence stays the same. There is a comment in the PubChem paper they had seen something similar.

Tautomers of 8-aminohypoxanthine



11 tautomer forms of 8-aminohypoxanthine.

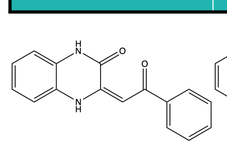
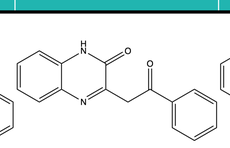
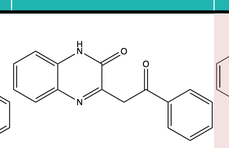
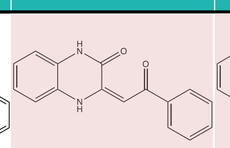
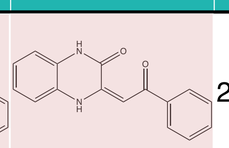
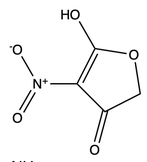
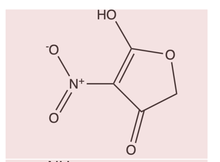
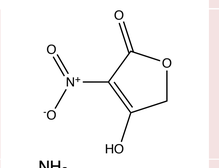
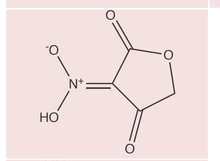
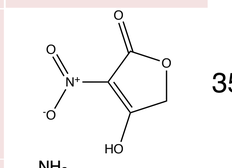
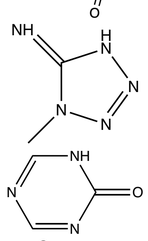
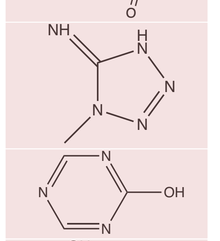
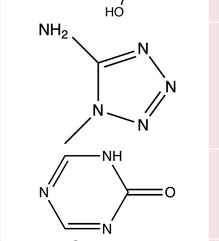
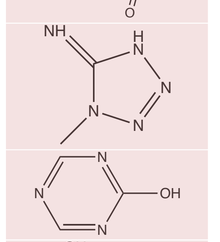
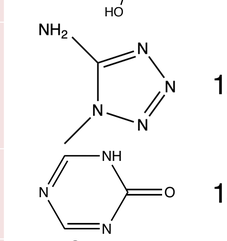
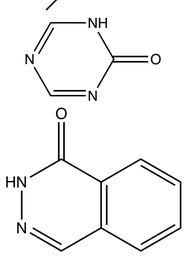
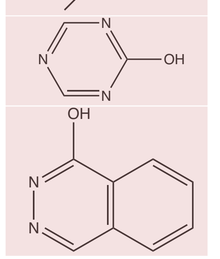
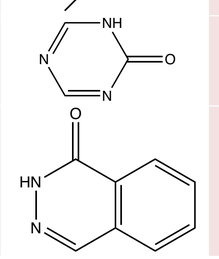
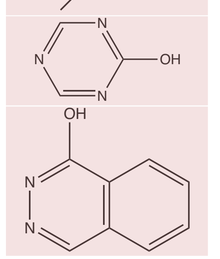
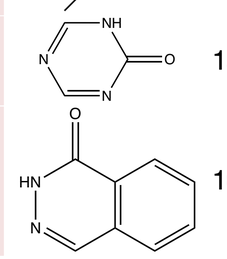
rxnfix and PubChem standardises all 11 to 1.

LillyMol standardises 1, 2, 4, 5, 7 and 8 into 1; 3, 6 and 9 to 3; and leaves 10 and 11 as is.

ChemDraw uses 7 for 8-aminohypoxanthine

Transformer deep learning[3]

The group from IBM used Tautobase as their data source making it interesting to compare results. Their ChemRxiv paper include five examples. Tautobase report results in three ways: either that one tautomer is dominant, a percentage for the first tautomer, or as a log K value. For the three cases that had log K values, the IBM group show the wrong tautomer as dominant. If Tautobase was arranged so that all entries on one side always indicated a tautomer with at least 50 %, then this type of error would be less likely.

Test case	IBM reported truth	Tautobase major tautomer	IBM prediction	rxnfix prediction	Tautobase rule
					226
					353, 354
					1242
					1503
					1005

Shaded entries either indicate wrongly extracted major tautomer, or wrongly predicted major tautomer. LillyMol handled the last two cases correctly, but left the top three as is. PubChem handled the last four cases correctly, the first test case was turned into a conjugated enol. The solvent in the second case is actually DMSO.

Bibliography

1. Oya Wahl and Thomas Sander, Tautobase: An Open Tautomer Database, Journal of Chemical Information and Modeling 2020 60 (3), 1085-1089, DOI: 10.1021/acs.jcim.0c00035, https://github.com/WahlOya/Tautobase
2. Peter J. Taylor† and Peter W. Kenny, The Prediction of Tautomer Preference in Aqueous Solution (Version 1.0), 2019, https://figshare.com/articles/preprint/The_Prediction_of_Tautomer_Preference_in_Aqueous_Solution_Version_1_0_/8966276
3. Hähnke, V.D., Kim, S. & Bolton, E.E. PubChem chemical structure standardization. J Cheminform 10, 36 (2018). https://doi.org/10.1186/s13321-018-0293-8, https://pubchem.ncbi.nlm.nih.gov/standardize/standardize.cgi
4. Cretu MT, Toniato A, Thakkar A, Debabeche A, Laino T, Vaucher AC. Standardizing chemical compounds with language models. ChemRxiv. Cambridge: Cambridge Open Engage; 2023; https://doi.org/10.26434/chemrxiv-2022-14ztf-v2
5. J Elguero, C Marzin, AR Katritzky, P Linda, The Tautomerism of Heterocycles, Academic Press, New York, 1976
6. LillyMol, https://github.com/EliLillyCo/LillyMol
7. RDKit, https://www.rdkit.org/
8. OpenEye toolkits, https://docs.eyesopen.com/toolkits/python/index.html
9. HazELNut, https://www.nextmovesoftware.com/hazelnut.html

NextMove Software Limited
Innovation Centre (Unit 320)
Cambridge Science Park
Milton Road, Cambridge
UK CB4 0WG
www.nextmovesoftware.com