Tautomer Enumeration and Stability Prediction for Virtual Screening on Large Chemical Databases

Francesca Milletti,[§] Loriano Storchi,[†] Gianluca Sforna,[†] Simon Cross,[†] and Gabriele Cruciani*,[§]

Laboratory for Chemometrics and Cheminformatics, Department of Chemistry, Università degli Studi di Perugia, via Elce di Sotto 10, 06123 Perugia, Italy, and Molecular Discovery Limited, 215 Marsh Road, Pinner, Middlesex, London HA5 5NE, United Kingdom

Received September 18, 2008

Tautomeric rearrangements affect the results of cheminformatics applications that depend on the knowledge of the 2D or 3D structure of a compound, such as tools for database searches, fingerprint generation, virtual screening, and physical-chemical properties prediction. In this paper we present TauThor, a tool to enumerate tautomers and predict tautomer stability in the aqueous medium. The enumeration is based on a recursive process that generates tautomers according to the general scheme HX-Y=Z ⇋ X=Y-ZH. The stability of a tautomer is calculated by using a library of 145 fragments associated with experimental tautomeric percentages in water and a pK_a based-method that utilizes pK_a values predicted by MoKa. Predicted tautomeric ratios based on pK_a calculations were benchmarked against literature data for a set of eleven compounds. The FDA approved drugs database, the NCI database and two vendor databases - Specs Screening Library and Asinex Gold Collection - were used to illustrate the impact of tautomerism on chemical libraries and to evaluate the relative occurrences of alternative tautomeric forms.

INTRODUCTION

According to IUPAC recommendations,[1] tautomerism can be defined by the general equilibrium shown in Figure 1, in which the two tautomers are readily interconvertible and X, Y, and Z are represented typically by any of C, N, O, or S, while H is transferred during the isomerization. Because the hydrogen atom is involved in the rearrangement, this type of tautomerism is referred to as prototropic tautomerism, which is the only type of tautomerism treated in this paper. In aromatic and other conjugated systems π-electron delocalization assists the displacement of H over larger distances, as exemplified by the 4-pyridone/4-hydroxypyridine tautomerism.

Numerous chemical groups may undergo prototropic tautomerism, and common examples include the keto−enol, imine-enamine, and lactim-lactam tautomerism. Different tautomeric forms of a molecule might differ in shape, conformation, functional groups, surface, and hydrogen-bonding pattern. For example, the enol group, a H-bond donor and acceptor, is a H-bond acceptor only in the correspondent keto form, and the NH group, a H-bond donor, becomes a H-bond acceptor when it tautomerizes to the imino form.

The prediction of physical-chemical properties of a compound can be inaccurate if the chemical structure used for the prediction does not correspond to the most stable tautomer in the medium of interest, as Figure 2 shows for ALogP[2] versus experimental logP[3] values. However, it is important to stress that this limitation of prediction tools can be overcome either by normalizing the input structure or by enriching the training data with multiple tautomeric forms, whenever this approach is feasible. For example, the pK_a prediction software MoKa[4] recognizes alternative tautomeric forms of the same compound, and it also identifies automatically unstable forms.

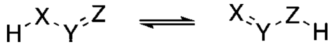


Figure 1. Tautomers must be readily interconvertible and X, Y, and Z are represented typically by any of C, N, O, or S.

Usually tautomers are registered as distinct structures in chemical databases and are associated with distinct CAS numbers. Canonicalization algorithms to retrieve different tautomeric forms of the same structure are useful to address tautomerism in database searches,[5] and while in large registry databases such as SciFinder tautomers can be found automatically, finding different forms of a tautomeric structure is still a problem in cheminformatics. The existence of distinct tautomeric forms is also an issue in the generation of molecular fingerprints and in similarity searches.

Tautomerism is important in ligand-protein interactions,[6] thus virtual screening methods for drug discovery require tools for tautomer enrichment, because the generation of multiple tautomeric forms lowers the risk of losing an important hit. However, tautomer enrichment has also the potential of increasing the number of false positives,[7] a problem that is difficult to address because of potential changes of the water tautomeric ratio upon interaction of the ligand with the protein.

The interconversion between two tautomeric forms is quantitatively described by the equilibrium constant K_T, which is defined as the concentration ratio of the two tautomers in equilibrium. Ab initio as well as theoretical calculations at semiempirical and density functional levels[8] are useful to determine with accuracy K_T, but their application

* Corresponding author phone: +390755855629; fax: +3907545646; e-mail: gabri@chemiome.chm.unipg.it.

[†] Molecular Discovery Limited.

[§] Università degli Studi di Perugia.

10.1021/ci800340j CCC: $40.75 © 2009 American Chemical Society
Published on Web 01/06/2009

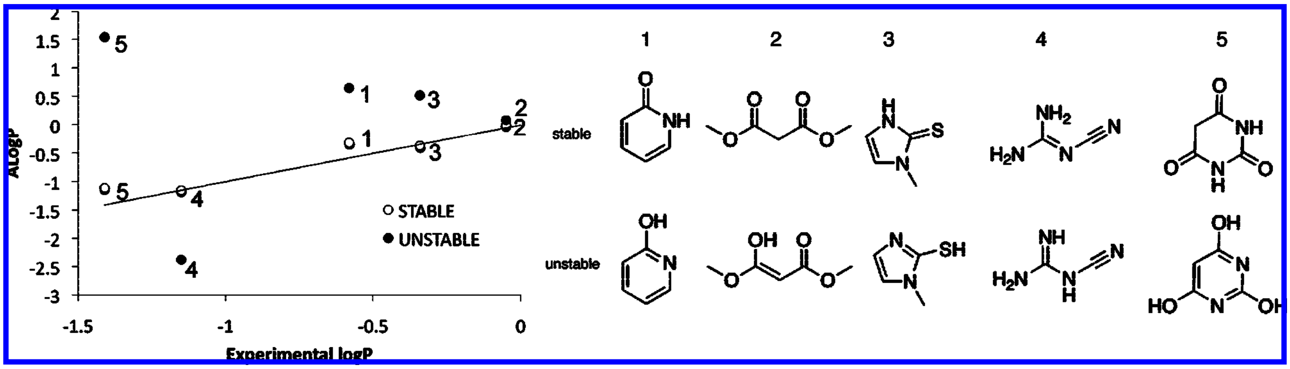


Figure 2. AlogP versus experimental LogP values. AlogP values of the stable form (in water) of the tautomers shown are predicted with higher accuracy than those of the corresponding unstable form.

over large databases is limited because such methods are time-consuming. Empirical methods can be used to obtain an estimate of K_T much faster, but at present only a few tools, such as Marvin,[9] predict tautomer stability, while most programs perform only tautomer enumeration (QuacPac,[10] ProtoPlex,[11] Pipeline Pilot,[12] TAUTOMER,[13] AGENT,[14] and LigPrep[15]). A common approach to enumerate tautomers is the use of libraries that define reasonable chemical rearrangements.[16] This approach is fast, but it has the potential of losing a tautomer when the rearrangement is not among those predefined.

The most straightforward experimental method to determine K_T is low temperature ^{1}H- NMR.[17] This technique requires that the proton transfer be slow enough to observe separate signals for both tautomers and that the equilibrium not be shifted too much toward one of the tautomers (at least 5% of the less stable tautomer). It is also possible to use ^{15}N-NMR and ^{13}C- NMR[18−20] spectroscopy by using chemical shifts of the NH tautomers and of the corresponding N-methyl derivatives.

Experimentally, another method to obtain K_T is the use of pK_a values of the fixed model compounds of a tautomeric pair.[21] In this method it is assumed that the methylation process does not change the relative basicities of individual tautomers, and K_T is calculated by using eq 1:

$$K_T = \frac{[T_2]}{[T_1]} = \frac{K_a^{T_2}}{K_a^{T_1}} \qquad (1)$$

It is important to stress that solvent effects are relevant in tautomer stability, since polarity differences among tautomers can induce significant changes in their relative energies in solution.[21] Therefore, the validity of experimental and computational results is limited to the medium considered.

Here, we describe (1) an algorithm for enumerating tautomers that is based on a recursive generation of molecular structures according to general scheme HX-Y=Z ⇌ X=Y-ZH and (2) a fast method to estimate tautomer stability by using a pK_a-based approach and a database of fragments associated with experimental tautomeric percentages in water. For a set of eleven compounds tautomeric percentages obtained by predicted pK_a values were benchmarked against literature data and yielded a good agreement. The algorithm here described was implemented in TauThor, a module of the software suite MoKa,[4] and was used to analyze the occurrences of tautomers in four chemical databases (683,862 compounds overall). Results highlighted that 29% of the structures are potentially tautomeric and that 7.8% of the compounds are represented in a form whose percentage in water is predicted to be below 95%. The contribution of alternative forms of thirteen tautomeric pairs was also analyzed to understand to what extent the major or minor tautomer of a molecule is used in chemical libraries. Both tautomer enumeration and stability prediction is fast, and the program can be used for virtual screening applications.

METHODS

Tautomer Enumeration. The algorithm implemented in TauThor generates tautomers by following the general approach summarized in Figure 3: (1) the parent tautomer (Figure 3a) undergoes any rearrangement that satisfies the scheme of Figure 1; (2) in the next iterative step (Figure 3b), the newly generated structures undergo other possible rearrangements; (3) the branches of the tree are pruned if a new structure is identical to one of the structures generated previously, a condition that is verified by matching the InChi[22] codes of each tautomer, because InChi codes allow the representation of a molecular structure by a unique string; and (4) the enumeration of tautomers is complete when all branches have been pruned.

Tautomer Filtering by Knowledge-Based Rules and by Aromaticity Rules. Tautomerism concerns both thermodynamics and kinetics; therefore, isomers must meet two conditions to be considered tautomers.[21] First, the energy difference between the two forms must not be too high (approximately ΔG < 28 kcal/mol), because in this event one form is virtually not present. Second, also the energy barrier to interconvert one form to the other must not be too high (approximately $\Delta G'$ < 25 kcal/mol), as this condition would render the reaction too slow to be considered a tautomeric rearrangement.

To identify energetically unfavorable displacements and avoid an exponential growth of tautomers, in the tautomer enumeration TauThor applies restrictions based on empirical knowledge and on aromaticity rules, as summarized in Figure 4. An example of empirical knowledge rule is presented in Figure 3, which shows that hydrogen atoms bonded to sp^2 carbon atoms do not undergo tautomerism because the corresponding rearrangement would generate a C=C=C within the six-membered ring. Additional rules have been implemented to prevent displacements such as those involving O=C-OH, O=S-OH, O=P-OH, and S=C-SH for symmetry reasons.

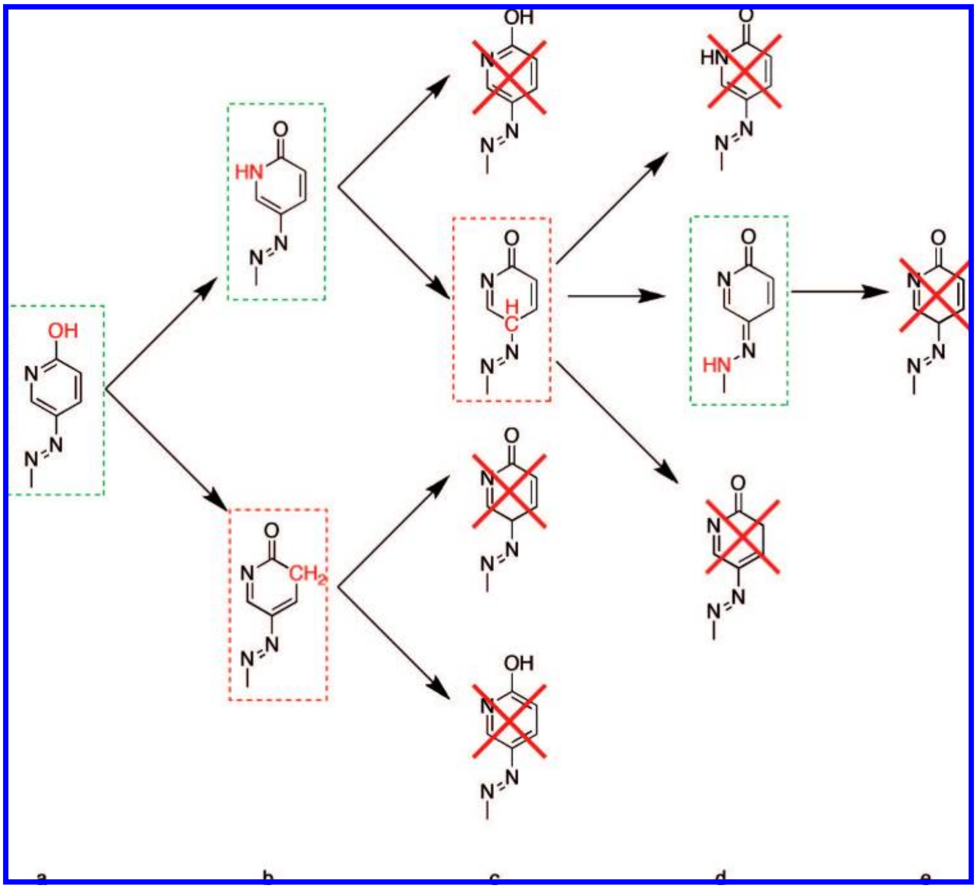


Figure 3. The tree-diagram illustrates the tautomer enumeration algorithm. New tautomers are generated according to the general scheme of Figure 1. A branch is pruned when a new structure is identical to a structure generated in one of the previous iterations. The algorithm ends when all branches have been pruned. Tautomers enclosed in red rectangles are excluded according to the aromaticity-based filter. Structures enclosed in green rectangles are the only tautomers enumerated.

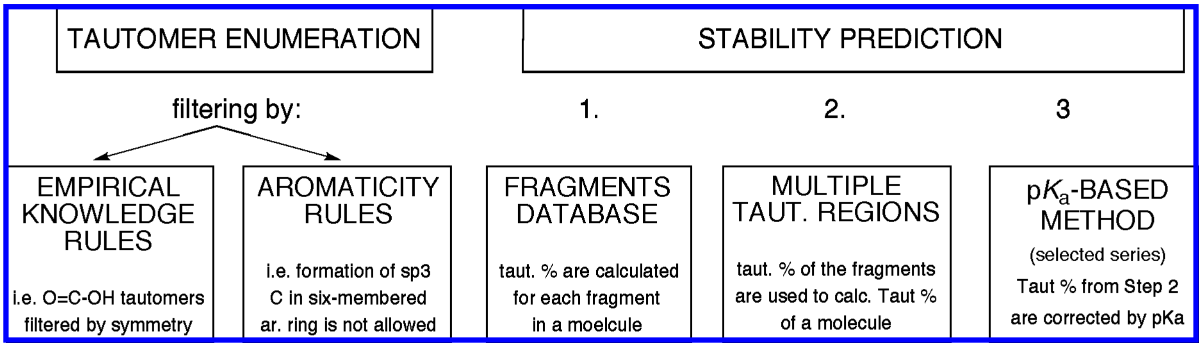

TAUTOMER ENUMERATION
STABILITY PREDICTION
filtering by:
1.
2.
3
EMPIRICAL KNOWLEDGE RULES
i.e. O=C-OH tautomers filtered by symmetry
AROMATICITY RULES
i.e. formation of sp3 C in six-membered ar. ring is not allowed
FRAGMENTS DATABASE
taut. % are calculated for each fragment in a moelcule
MULTIPLE TAUT. REGIONS
taut. % of the fragments are used to calc. Taut % of a molecule
pKa-BASED METHOD
(selected series)
Taut % from Step 2 are corrected by pKa

Figure 4. Flowchart for tautomer enumeration and stability prediction.

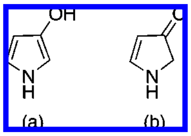


Figure 5. Although aromaticity is not preserved, compound (b) predominates in the aqueous medium as a result of higher dipole moment and extra-stability associated with the formation of the carbonyl bond.

Figure 3 presents also an example of aromaticity rule by showing that the introduction of an sp^3 carbon atom in the ring results in structures (enclosed in the red rectangle) that are filtered out because aromaticity is not preserved. As extensively explained in recent reviews,[23,24] the effect of π-electron delocalization is important to explain the preference of a tautomeric form to another when the displacement of a hydrogen atom generates nonaromatic tautomers. For example, phenol has no tendency to tautomerize to its corresponding keto form, because this would result in the loss of aromaticity of the ring; on the contrary, 2-hydroxyfuran tautomerizes to the keto form, which is more stable than the corresponding hydroxy form, because the energy destabilization due to the loss of aromaticity is much lower than in the case of phenol (aromaticity scale: benzene > furan), and such loss of resonance stability can be compensated by the formation of a strong carbonyl bond. Because there is experimental evidence that there are five-membered aromatic and hydroxyl-substituted rings,[25] such as the one shown in Figure 5, that are more stable in the keto and nonaromatic form, a rearrangement that introduces an sp^3 carbon atom in a five-membered aromatic ring results in a structure that is not subject to any filtering, as opposed to

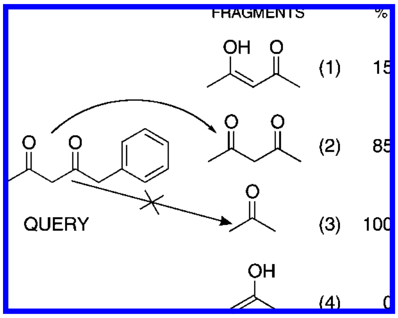


Figure 6. The query shown on the left is matched against a subset of the fragments in the TauThor library. Numbers in parentheses indicate the priority of the fragment, and the % represents the experimental percentage in water of a fragment. When the query matches the fragment that has priority 2, the matching atoms of the query are flagged so that they cannot be matched with fragments that have lower priority. The percentage of the query shown is 85%.

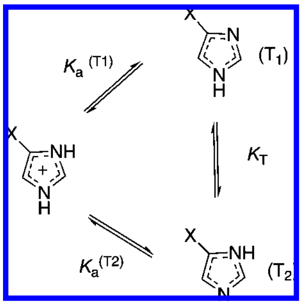


Figure 7. The relative percentage of the two tautomers T_1 and T_2 depends on their relative basicities. The less basic imidazole corresponds to the most stable tautomer, as illustrated in eqs 1 and 3.

rearrangements that introduce an sp^3 carbon atom in a six-membered aromatic ring.

Tautomer Stability. When the enumeration of tautomers is complete and forms identified as unstable have been filtered out according to rules based on the loss of π-electron delocalization, the stability of the remaining tautomers is predicted by using a library of 145 fragments associated with experimental tautomeric percentages in water. Each fragment is also associated with a priority score, as shown in Figure 6, which represents a subset of the internal library. The priority score is assigned to each fragment because a fragment of the library can be the substructure of other fragments. Fragments that have a higher priority score are retrieved first, and the atoms matched in the query tautomer are flagged so that they cannot be matched again against fragments that have lower priority.

Because a molecule can contain multiple tautomeric regions, the tautomeric percentages of each region are combined to give the tautomeric percentage of the whole molecule. Thus, the tautomeric percentage P of a tautomer is obtained by combining the tautomeric percentages p of the N regions i as shown in eq 2:

$$P = \prod_{i=1}^{N} p_i \tag{2}$$

Effect of pK_a on Tautomer Stability. The database of fragments cannot be used to account for the effect of any substituent on the K_T of a tautomeric pair because experimental K_T values are available only for a very small number of substituents. However, K_T depends on the acidity constant K_a of the individual tautomers as shown in eq 1. Therefore, it is possible to predict K_T from the pK_a of the proton acceptor sites of two tautomers, by using eq 3:

$$pK_T = pK_a^{T_2} - pK_a^{T_1} \tag{3}$$

The more basic site takes the proton, and the corresponding tautomeric form predominates in solution (Figure 7).

The pK_a prediction tool MoKa,[4] trained by over 25,000 pK_a values, is used to predict pK_a values to calculate K_T. A limitation of this approach is that predicted pK_a are obtained from statistical models, thus the applicability of such pK_a values to calculate K_T depends on whether the models used were trained to predict pK_a accurately for both tautomers. For example, this approach is not suitable to estimate the relative stability of the 2-hydroxypyridine/2-pyridone pair, because experimental pK_a values are available only for 2-pyridone derivatives, which, virtually, are the only form present in solution. This pK_a-based approach is used to calculate K_T of triazoles, pyrazoles, imidazoles, and pyrimidones. For tetrazoles, the predominance of the 1-H form in aqueous solution is influenced mainly by lone-pair repulsions;[21] therefore, an appropriate fragment is used in the internal library of TauThor to recognize the 1-H form as the most stable, and pK_a calculations are not used.

The tautomeric percentages for the pair of tautomers T_1 and T_2 are corrected according to the pK_a-based method described above by using eqs 4 and 5, where P is calculated from eq 2 and K_T from eq 3:

Table 1. Experimental and Calculated Values of Tautomeric Percentages in Water for Compounds in Figure 8 According to the Scheme $T_1 \rightleftharpoons T_2$

ID		-R$_1$	-R$_2$	R$_3$	pK_a (T_1)	pK_a (T_2)	exp (%T_2)	calc (%T_2)
1	pyrazole	-CH$_3$			2.85	2.92	53.9[21]	46.0
		-OEt	-CH$_3$		3.66	2.40	96.8[21]	94.8
		-Br	-CH$_3$		1.29	0.45	95.8−96.9[21]	87.4
2	imidazole	-CH$_3$			6.89	6.96	52.4[46]	46.0
		-Ph			6.04	5.65	90.1−97.4[46]	71.1
		-NO$_2$			0.69	−0.37	99.8[46]	92.0
		-Br			4.70	3.93	98.9[46]	85.5
		-Cl			4.26	4.00	98.9[46]	64.0
		-I			4.64	4.67	95.0[46]	48.1
		-NO$_2$		Br	−1.71	−2.77	99.5[46]	92.0
3	1,2,3-triazole				1.68	0.93	66.7[47]	84.9

$$\%T_1 = \frac{(P_{T_1} + P_{T_2})}{K_T + 1} \qquad (4)$$

$$\%T_2 = \frac{(P_{T_1} + P_{T_2}) \times K_T}{K_T + 1} \qquad (5)$$

Tautomeric percentages of species that not suitable for the pK_a-based method described here correspond to the percentages calculated from eq 2.

RESULTS AND DISCUSSION

Validation of K_T Calculations from Predicted pK_a Values. To validate the method presented here for estimating K_T from pK_a predictions, calculated tautomeric percentages in water were benchmarked against experimental data (Table 1) for a series of azoles (imidazole, pyrazole, and triazole) that differ in the position of one or more substituents in the ring (Figure 8). TauThor calculated the tautomeric percentage automatically using pK_a values predicted by MoKa to solve eqs 3 and 4. The results obtained demonstrate that the agreement between the experimental and predicted K_T value is reasonably good (RMSE = 20%) by considering the discrepancies between measurements for the same compound (i.e., 2-phenylimidazole).

Tautomer Occurrence in Chemical Databases. Tautomers generated by TauThor, along with predicted tautomeric stability, were used to analyze the FDA approved drugs[26] database, the NCI[27] database, and two vendor databases - Specs Screening Collection[28] and Asinex Gold Collection.[29] The speed of computation was on average 0.05 s/tautomer on a Core2 Q6600 Fedora 9 machine. Here, we define as "major tautomer" the tautomer whose percentage in water is the highest and as "minor tautomer" any of the tautomers whose percentage in water is not the highest. When the predicted tautomeric percentage is the same for all tautomers of a compound (such as for imidazole), the structure in the database is regarded as major tautomer.

Table 2 shows that 29.2% of the structures are potentially tautomeric; 25.0% of them are represented in a form that is also predicted as the major tautomer, while 4.2% are represented in a form predicted as one of the minor tautomers. Table 2 also shows that the compounds represented in a form whose percentage in water is predicted to be below 95% are 7.8% of the total: 4.2% are compounds represented in a minor form, and 3.6% are compounds represented in the major form, which is in equilibrium with other species.

Relative Occurrences for a Set of Tautomeric Pairs. A substructure search was run on each database by using internal software to retrieve examples of structures that are often represented in alternative ways. The occurrences of thirteen diverse tautomeric pairs relevant to medicinal chemistry are shown in Table 3, which also reports literature data (including K_T, when available) about the stability in water of each tautomer. TauThor includes all the queries and tautomeric stabilities of Table 3 in its fragment library and predicts the percentage of the molecules retrieved from the search accordingly. Overall, 35% of the molecules retrieved were represented in a minor form. Figure 9 summarizes the findings of the database search, which is discussed here in detail on a case-by-case basis.

Keto−Enol Tautomerism (ID 1−3). For simple ketones, the keto form predominates over the enol form.[30] The percentage of enol in 1,3-diketones is strongly influenced by the solvent, both in terms of polarity and hydrogen bond acceptor/donor properties.[31] While open chain, *cis*-enolizing 1,3-diketones exist in water predominantly as diketo tautomers, *trans*-enolizing 1,3-diketones (ID 1) show the opposite behavior in aqueous solution and exist predominantly as keto−enol tautomers.[32,33] However, our data show that 1,3-cyclohexandione is represented in the diketo form in 82% of the cases. 4-Hydroxyl-2*H*-pyran-2-one (ID 2) is the preferred form as a result of conjugation of the carbonyl with the oxygen atom in the ring,[34] and it is consistently represented in this form in the databases analyzed. 3-Hydroxypyrrole (ID 3) is more stable as a keto tautomer[35] as a result of its high dipole moment and little loss of resonance energy; however, it is represented in the hydroxy form in 18% of the cases. The query shown for tautomer 3 excludes from the search structures that have a -C=O as substituent in positions 2 or 4 because experimental data indicate that this group alters the tautomeric equilibrium by stabilizing the hydroxy form via intramolecular hydrogen bonding.[21]

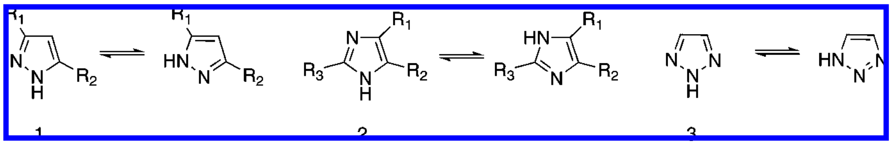
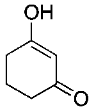
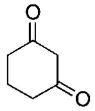
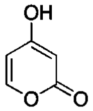
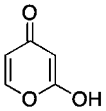
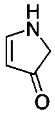
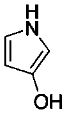
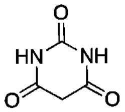
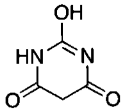
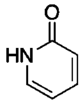
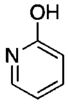
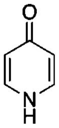
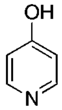
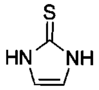
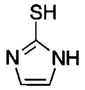
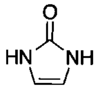
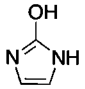
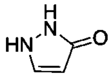
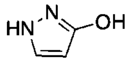
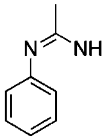
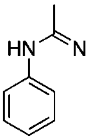
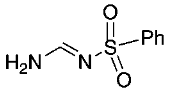
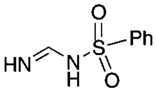
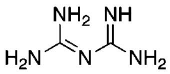
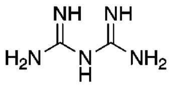
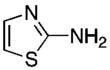
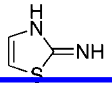


Figure 8. Tautomeric equilibria ($T_1 \rightleftharpoons T_2$) for compounds in Table 1.

Table 2. Statistics about Tautomerism in the Four Databases Analyzed

				Specs		Asinex		NCI		FDA		total	
				N.	%	N.	%	N.	%	N.	%	N.	%
total compds				198840		233554		250251		1217		683862	
tautomeric compds				56330	28.3	74876	32.1	67877	27.1	314	25.8	199397	**29.2**
	class	major tautomer	%										
	1a	*yes*	>95	45054	22.7	56968	24.4	44227	17.6	256	21.0	146505	21.4
	1b	*yes*	<95	6685	3.3	9511	4.1	8169	3.3	31	2.5	24396	3.6
tot 1				51739	**26.0**	66479	28.5	52396	**20.9**	287	**23.5**	170901	**25.0**
	2a	*no*	<5	3962	2.0	7183	3.1	14617	5.8	22	1.8	25784	3.8
	2b	*no*	>5	629	0.3	1214	0.5	864	0.4	5	0.4	2712	0.4
tot 2				4591	**2.3**	8397	3.6	15481	**6.2**	27	**2.2**	28496	**4.2**

Table 3. Occurrences of Tautomeric Pairs in the Three Databases Analyzed[a]

		Major Tautomer	N. of occurrences				Minor Tautomer	N. of occurrences			
ID	pK_T	Query A	Specs	Asinex	NCI	FDA	Query B	Specs	Asinex	NCI	FDA
1	NA[33]		1	0	5	0		3	5	19	0
2	NA[34]		89	31	251	4		1	0	2	0
3	0.89[35]		8	66	0	0		2	9	5	0
4	NA[36]		571	944	649	12		0	0	0	2
5	3.0[37]		382	1693	7	2		166	469	980	0
6	3.3[37]		57	64	497	0		95	208	425	0
7	NA[39]		7	5	228	0		59	112	0	0
8	NA[40]		43	251	133	1		5	6	0	0
9	NA[41]		11	3	57	2		12	30	5	0
10	2.4[42]		18	1	116	2		1	0	19	0
11[a]	7.0[43]		15	21	12	0		20	2	19	0
12	NA[44]		0	17	2	0		13	11	670	4
13	4.3[45]		132	359	16	1		2	1	229	0

[a] NA = not available. For tautomers 3 and 9 the query excludes from the search results structures that have a -C=O as a substituent in positions 2 or 4.

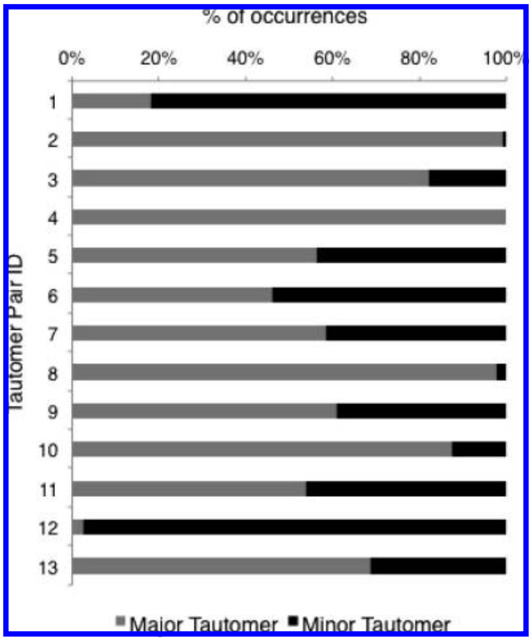


Figure 9. Percentage of occurrences of the major and minor tautomer (predicted) in the four databases analyzed for the thirteen tautomeric pairs of Table 3.

Lactam-Lactim Tautomerism in Aliphatic Compounds (ID 4). Barbituric acid derivatives (ID 4) predominate in the triketo form in water,[36] as a result of the stability of the carbonyl bond and its higher polarity compared to the hydroxy tautomers. Multiple alternative tautomers are possible, but in the databases analyzed the triketo form predominates compared to the alternative form in Table 3.

Lactam-Lactim Tautomerism in Six-Membered Aromatic Compounds (ID 5−6). Although phenol prefers the enol form because the proton transfer from the hydroxy group to the aromatic carbon atom destroys the aromaticity of the ring, heteroaromatic compounds such as 2- and 4-hydroxypyridines are more stable in the keto form,[21] which preserves the π-electron delocalization of the six-membered ring. 2- and 4-pyridone derivatives are represented in the less stable hydroxy form in 46% of the cases. However, it is noteworthy that very electron-withdrawing substituents may change the position of this tautomeric equilibrium; for example, in 6-chloro derivatives of 2-pyridones the equilibrium is shifted toward the hydroxy form (tautomeric ratio hydroxy/keto 61:10).[37,38]

Lactam-Lactim Tautomerism in Five-Membered Aromatic Compounds (ID 7−9). 2-Thioimidazole[39] (ID 7), 2-hydroxyimidazole[40] (ID 8), and 3-hydroxypyrazole[41] (ID 9) exist predominantly in the thione or keto form. 2-Thioimidazole is consistently represented in the thione form in the NCI database, while the mercapto form largely predominates in the Specs and Asinex databases (89% and 96% of the cases, respectively). 2-Hydroxyimidazole is consistently represented in the keto form (98% of the cases), while 3-hydroxypyrazole is represented alternatively in both forms (61% of the cases in the keto form). To avoid bias, the query corresponding to 3-hydroxypyrazole (ID 9) excludes from the search results structures that have a -C=O as substituent in positions 4 because experimental data indicate that this substituent shifts the tautomeric constant in favor of the hydroxy form.[21]

Amidine and Guanidine Tautomerism (ID 10−12). The tautomerism of amidines can be rationalized in terms of the basicity of the imino nitrogen atom. As illustrated in eqs 1 and 3, the less basic species corresponds to the most stable tautomer. As a result, phenyl- and arylsulfonyl-amidines (ID 10−11) exist predominantly in the tautomeric form that has the imino group closer to the electron withdrawing substituent (arylsulfonyl or phenyl).[42,43] Whereas phenylamidines are represented in the minor form only in 13% of the cases, arylsulfonylamidines are represented alternatively in both forms. Diguanides (ID 12) are more stable in the less basic, conjugated form;[44] however, they are represented in the nonconjugated form in 97% of the cases.

Amino-Imino Tautomerism (ID 13). In the amino-imino tautomerism of 2-aminothiazole, the amino form predominates in water.[45] While the Specs and Asinex databases consistently represent this compound in the amino form, in the NCI database 2-aminothiazole is represented in the imino form in 93% of the cases.

CONCLUSIONS

In this paper, we have presented a tool for enumerating tautomers and predicting tautomer stability in water. The algorithm used enumerates tautomers recursively in a tree-structured process in which each new branch contains tautomers generated from the previous step according to the general rearrangement HX-Y=Z $\leftrightharpoons$ X=Y-ZH. The distinctive feature of the program is the ability to generate tautomers regardless of predefined tautomeric rearrangement schemes, and this is advantageous because it minimizes the risk of missing tautomeric rearrangements not taken into account. Predictions of water tautomeric ratios are accomplished by using a 2-fold approach: the program first predicts the tautomeric abundance by using an internal library that includes literature data of tautomeric percentages in water, and then, for selected substructures, the tautomeric percentage thus obtained is corrected by using a pK_a-based method. To evaluate the impact of tautomerism on chemical databases, TauThor was used to enumerate tautomers and predict tautomer stability for 683,862 structures from four databases. Results indicate that 29% of the compounds are potentially tautomeric and that 7.8% are represented in a form whose water percentage is predicted to be below 95%. The frequency of occurrence of thirteen selected tautomeric pairs relevant to medicinal chemistry highlights that in the subset analyzed 35% of the structures are represented in the minor form.

Supporting Information Available: Tautomers and predicted percentages in water for compounds in the FDA approved drugs database. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) IUPAC Compendium of Chemical Terminology, Electronic version. http://goldbook.iupac.org/T06252.html (accessed Sept 10, 2008).

(2) VCCLAB, Virtual Computational Chemistry Laboratory. http://www.vcclab.org (accessed Sept 10, 2008).

(3) LOGKOW databank, Sangster Research Laboratories. http://logkow.cisti.nrc.ca (accessed Sept 10, 2008).

(4) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and Original pK a Prediction Method Using Grid Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *47* (6), 2172–2181.

(5) Trepalin, S. V.; Skorenko, A. V.; Balakin, K. V.; Nasonov, A. F.; Lang, S. A.; Ivashchenko, A. A.; Savchuk, N. P. Advanced Exact Structure Searching in Large Databases of Chemical Compounds. *J. Chem. Inf. Model.* **2003**, *43* (3), 852–860.

(6) Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. Tautomerism in Computer Aided Drug Design. *J. Recept. Signal Transduction* **2003**, *23*, 361–371.

(7) Brenk, R.; Irwin, J. J.; Shoichet, B. K. Here Be Dragons: Docking and Screening in an Uncharted Region of Chemical Space. *J. Biomol. Screening* **2005**, *10* (7), 667–674.

(8) Ibon Alkorta, J. E. Theoretical estimation of the annular tautomerism of indazoles. *J. Phys. Org. Chem.* **2005**, *18* (8), 719–724.

(9) Chemaxon. www.chemaxon.com (accessed Aug 25, 2008).

(10) OpenEye Scientific Software. www.eyesopen.com/products/applications/quacpac.html (accessed Aug 25, 2008).

(11) The New Tripos. www.tripos.com (accessed Aug 25, 2008).

(12) Accelrys Software Inc. www.accelrys.com (accessed Aug 25, 2008).

(13) Molecular Networks GmbH. www.molecular-networks.com (accessed Aug 25, 2008).

(14) Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. 15th European symposium on Structure Activity Relationships (QSAR) and molecular modelling, Istanbul, Turkey, Sept 5-10, 2004.

(15) Schrödinger. www.schrodinger.com (accessed Aug 25, 2008).

(16) Oellien, F.; Cramer, J.; Beyer, C.; Ihlenfeldt, W. D.; Selzer, P. M. The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46* (6), 2342–2354.

(17) Claramunt, R. M.; Lopez, C.; Santa Maria, M. D.; Sanz, D.; Elguero, J. The use of NMR spectroscopy to study tautomerism. *Prog. Nucl. Magn. Reson. Spectrosc.* **2006**, *49* (3−4), 169–206.

(18) Bojarska-Olejnik, E.; Stefaniak, L.; Witanowski, M.; Hamdi, B. T.; Webb, G. A. Applications of ^{15}N NMR to a study of tautomerism in some monocyclic azoles. *Magn. Reson. Chem.* **1985**, *23* (3), 166–169.

(19) Wofford, D. S.; Forkey, D. M.; Russell, J. G. Nitrogen-15 NMR spectroscopy: prototropic tautomerism of azoles. *J. Org. Chem.* **1982**, *47* (26), 5132–5137.

(20) Chenon, M. T.; Pugmire, R. J.; Grant, D. M.; Panzica, R. P.; Townsend, L. B. Carbon-13 magnetic resonance. XXVI. Quantitative determination of the tautomeric populations of certain purines. *J. Am. Chem. Soc.* **1975**, *97* (16), 4636–4642.

(21) Elguero, J.; Katritzky, A. R.; Denisko, O. The Prototropic Tautomerism of Heterocycles. In Advances in Heterocyclic Chemistry; Academic Press: New York, NY, 2000; Vol. 76, pp 1−64.

(22) International Union of Pure and Applied Chemistry. www.iupac.org/inchi/ (accessed Sept 18, 2008).

(23) Raczynska, E. D.; Kosinska, W.; Osmialowski, B.; Gawinecki, R. Tautomeric Equilibria in Relation to Pi-Electron Delocalization. *Chem. Rev.* **2005**, *105* (10), 3561–3612.

(24) Balaban, A. T.; Oniciu, D. C.; Katritzky, A. R. Aromaticity as a Cornerstone of Heterocyclic Chemistry. *Chem. Rev.* **2004**, *104* (5), 2777–2812.

(25) Parchment, O. G.; Green, D. V. S.; Taylor, P. J.; Hillier, I. H. The prediction of tautomer equilibria in hydrated 3-hydroxypyrazole: a challenge to theory. *J. Am. Chem. Soc.* **1993**, *115* (6), 2352–2356.

(26) ZINC - A free database for virtual screening. zinc.docking.org/catalog/fda.in (accessed July 14, 2008), database version March 2005.

(27) NCI database. cactus.nci.nih.gov/ncidb2/download.html (accessed Jul 14, 2008), SMILES string database version Aug 2000.

(28) Specs. www.specs.net (accessed Jul 14, 2008), database version Nov 2007.

(29) Asinex. www.asinex.com (accessed Jul 14, 2008), database version Nov 1, 2007.

(30) Keeffe, J. R.; Kresge, A. J.; Schepp, N. P. Keto-enol equilibrium constants of simple monofunctional aldehydes and ketones in aqueous solution. *J. Am. Chem. Soc.* **1990**, *112* (12), 4862–4868.

(31) Reichardt, C. Solvent Effects on the Position of Homogeneous Chemical Equilibria. In Solvents and Solvent Effects in Organic Chemistry, 3rd ed.; VCH: Weinheim, Germany. 2003; p 108.

(32) Luchmuller, C. H.; Maldacker, T.; Cefola, M. The bromometric determination of bulk enol content of β-diketones that undergo rapid tautomerization. *Anal. Chim. Acta* **1969**, *48*, 139–144.

(33) Schwarzenbach, G.; Felder, E. Bestimmung von Keto-Enol-Gleichgewichten in Wasser. *Helv. Chim. Acta* **1944**, *27*, 1044–1060.

(34) Stanovnik, B.; Tisler, M.; Katritzky, A. R.; Denisko, O. V. The Tautomerism of Heterocycles: Substituent Tautomerism of Six-Membered Ring Heterocycles. In *Advances in Heterocyclic Chemistry*; Katritzky, A. R., FRS, Ed.; Academic Press: New York, 2006; Vol. *91*, p 65.

(35) Capon, B.; Kwok, F. C. Tautomerism of the monohydroxy derivatives of five-membered oxygen, nitrogen and sulfur heterocycles. *J. Am. Chem. Soc.* **1989**, *111* (14), 5346–5356.

(36) Jeffrey, G. A.; Ghose, S.; Warwicker, J. O. The Crystal Structure of Barbituric Acid Dihydrate. *Acta Crystallogr.* **1960**, *14*, 881–887.

(37) Katritzky, A. R.; Pozharskii, A. In *Handbook of Heterocyclic Chemistry*; Pergamon: Amsterdam, The Netherlands, 2000; pp 48−49.

(38) Katritzky, A. R.; Rowe, J. D.; Roy, S. K. Potentially tautomeric pyridines. Part IX. The effect of chlorine substituents on pyridine-hydroxypyridine tautomerism. *J. Chem. Soc. B* **1967**, 758–761.

(39) Bojarska-Olejnik, E.; Stefaniak, L.; Witanowski, M.; Hamdi, B. T.; Webb, G. A. Applications of ^{15}N NMR to a study of tautomerism in some monocyclic azoles. *Magn. Reson. Chem.* **1985**, *23* (3), 166–169.

(40) Hofmann, K. The Oxo-and Hydroxyimidazoles and their Sulfur Analogues. In Chemistry of Heterocyclic Compounds; Interscience Publishers, Inc.: New York, NY, 2007; Vol. *6*, pp 55−110.

(41) Parchment, O. G.; Green, D. V. S.; Taylor, P. J.; Hillier, I. H. The prediction of tautomer equilibria in hydrated 3-hydroxypyrazole: a challenge to theory. *J. Am. Chem. Soc.* **1993**, *115* (6), 2352–2356.

(42) Cook, M. J.; Katritzky, A. R.; Nadji, S. Substituent effects in tautomerism. Part 11. para-Substitution in N-phenylamidines. *J. Chem. Soc., Perkin Trans. 2* **1976**, 211–214.

(43) Chua, S.; Cook, M. J.; Katritzky, A. R. Substituent effects in tautomerism. Part I. Acyl- and sulphonylamidines. *J. Chem. Soc., Perkin Trans. 2* **1974**, 546–552.

(44) Clement, B.; Girreser, U. Characterization of biguanides by ^{15}N-NMR spectroscopy. *Magn. Reson. Chem.* **1999**, *37* (9), 662–666.

(45) Forlani, L.; De Maria, P. Tautomerism of Aminothiazoles. pK_{BH+} Values of 2-Aminothiazoles and of Some Model Imines. *J. Chem. Soc., Perkin Trans. 2* **1982**, 535–537.

(46) Grimmett, M. R. Five-Membered Rings With Two Heteroatoms and Fused Carbocyclic Derivatives. In *Comprehensive Heterocyclic Chemistry II*; Shinkai, I. Ed.; Pergamon Press Inc.: Oxford, U.K., 1996; Vol. 3, p 77.

(47) Albert, A.; Taylor, P. J. The tautomerism of 1,2,3-triazole in aqueous solution. *J. Chem. Soc., Perkin Trans. 2* **1989**, 1903–1905.

CI800340J