Tautomerism and Expert Systems in Spectroscopy

Part I: The Problem, Ramifications for Spectral Databases, and Current Solutions

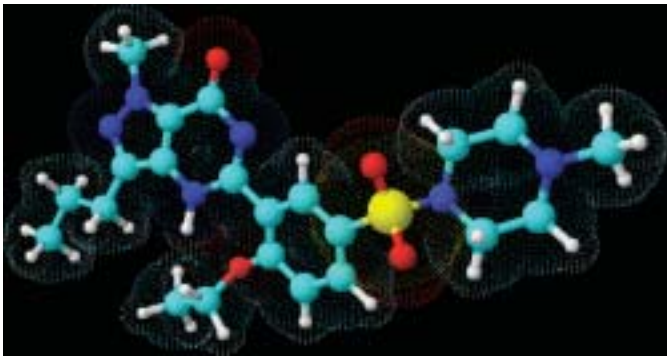
Because tautomers often cannot be isolated as single isomers, spectroscopists should always anticipate possible tautomeric forms that can complicate the interpretation of their spectra. This involves the use of spectral databases and tautomerically enabled search algorithms, which are reviewed here.

Michel R.J. Hachey

Tautomerism is an important area of chemical study that has long been a challenge computationally, experimentally, and intellectually. For example, in 1950, Watson and Crick were able to propose the structure of DNA because they had the insight to focus on the keto-form base pairs rather than the enol-form tautomers. Unfortunately, the complicating effects of tautomerism affect more than just the biochemical field; they also have a broad impact on analytical data interpretation and management for infrared (IR), Raman, UV/Vis, and nuclear magnetic resonance (NMR) spectroscopy, mass spectrometry (MS), and other analytical techniques.

One of the difficulties with tautomers is that under some conditions they cannot be isolated as single isomers, which forces spectroscopists to deal with a mixture of closely related analytes that have distinct properties and spectroscopic responses. When it comes to tautomers, a so-called pure sample can contain two, three, or more structural isomers that result from rapid interconversion through the migration of neutral groups — generally hydrogens. Consequently, the complete spectroscopic interpretation for tautomeric species cannot be rationalized fully unless each form present at equilibrium is accounted for. Another problem is finding which tautomers predominate. Spectroscopists should, therefore, always be ready to anticipate possible tautomeric forms that can impact their interpretation of the spectrum.

Failures to anticipate tautomerism frequently lead to spectral misassignments to the wrong isomeric forms. This can create both embarrassing and potentially costly situations. For example, a vendor of chemical libraries for high-throughput screening sold the same tautomer sample at different prices due to a duplicate in its library that had been assigned accidentally to different structural forms (1). Likewise, most spectral libraries, whether commercial or in-house, contain duplicates arising from alternative attribution to different tautomeric isomers. Assignments to different forms might or might not be justified — too often the latter is the case. As a result, simple structural and substructural searches of spectral libraries can fail to yield all chemically relevant hits if the search algorithm cannot check for possible tautomeric forms. Ultimately, trouble with the spectral interpretation of tautomers can extend beyond just the spectroscopic realm. For example, patent protection for a blockbuster drug can be put in jeopardy if the active tautomeric form is not directly included in the original patent (2).



ADVANCED CHEMISTRY DEVELOPMENT, INC.

Studying Tautomerism: Traditional vs. Software-Assisted Methods

Even vigilant scientists can make mistakes due to tautomerism, because tautomers do not always behave as expected and it is easy to overlook potentially relevant forms. These two difficulties can be illustrated by the series of pyridine-related tautomers shown in Figure 1.

Noticing all possible tautomers for a compound is not always trivial. For example, in contrast to pyridine, which can only occur in one form, uric acid has 24 theoretically possible tautomers. To invite speculation on the major

forms, uric acid is drawn purposely as a minor form in Figure 1. It is not always easy to determine, a priori, forms that might be present at equilibrium even in small molecules.

In the first tautomeric pair shown in Figure 1, there is a bias toward suggesting that pyridin-2-ol is the predominant form versus pyridin-2(1*H*)-one because of our familiarity with pyridine. This somewhat unusual predominance of the enol over the keto form is indeed observed, and is attributed to the stabilizing effect of the aromatic ring (3). However, let's now compare the next tautomeric pair, pyridin-4-ol and pyridin-4(1*H*)-one, which differ only by the location of the oxygen around the ring. Here the enol predominates in the gas phase, but the keto form predominates in solution (3), showing that general predictions are dangerous. Careful examination of the structural environment of the structural group and of the physicochemical environment is of paramount importance when contemplating what the major forms are.

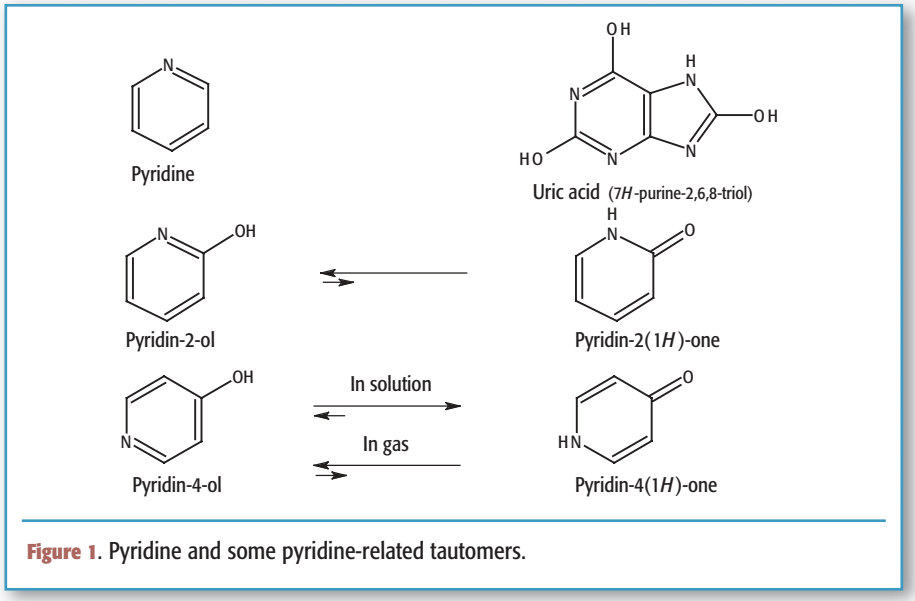


Figure 1. Pyridine and some pyridine-related tautomers.

In view of these difficulties, software that automatically generates different tautomeric forms and suggests relative stabilities should reduce the likelihood of making tautomeric mistakes. Because tautomerism is a nontrivial problem, software must be imbued with high degrees of chemical intelligence. This means that algorithms must not only validate proposed connectivity, valence and charges, they also must find ways of generating the different structural isomers that are needed through chemically valid transforma-

tions. Computer algorithms have been built that can generate possible tautomers automatically, such as CLIFF (4), or empirically suggest their likely major forms, such as ACD/Tautomers (5), so that mistakes are less likely to occur with these tools when assigning spectra or performing searches.

This article series provides a summary of the recent progress with a focus on practical applications to spectroscopy. This first part focuses on spectral databases and tautomerically enabled search algorithms.

Methods

Canonical and normalization-based computer algorithms for automatically searching tautomers that occur in databases of organic compounds were reviewed recently and need not be repeated here (1). However, this review is incomplete because it focuses only on structure registry systems, such as those provided by Chemical Abstracts Services (CAS) and Beilstein (see references in reference 1), and cheminformatic systems such as ChemoSoft (see references in reference 1). Some tautomer-enabled algorithms also have been applied by ACD/Labs to both cheminformatic and spectral data management systems using slightly different approaches that deserve mention here.

Unlike the canonical and normalization approaches (1), the algorithm used by ACD/Labs (5) generates the most likely tautomeric forms based upon empirical knowledge and identifies the presumed minor and major forms. The most reasonable and favorable forms are found through a set of hierarchical rules that describe functional groups in a generalized form with variable substitution. First, the generalized forms are sorted through a system of hierarchy and seniority ranking. Then, the relative dominance or equivalence of different forms is considered. The perturbations coming from the number of substituents and their nature (donor/acceptor, cyclic/acyclic, ring size) are factored into the prediction. Figure 2 illustrates a decision process for an acyclic substructure.

The empirical rules used above are calibrated to a physicochemical environment that best spans the normal range for the solution phase, with an emphasis on generating all tautomeric forms that reasonably can be expected to exist at equilibrium. Nonetheless, this algorithm still is somewhat useful in generating likely forms in the gas phase, although the proposed minor and major forms sometimes will be interchanged. Exotic theoretical forms are filtered out of the suggested tautomer list automatically unless they are used as the input structure. Because there is no direct account of specific condensed phase effects for different solvents, concentrations, and pH ranges, care should be exercised in critically evaluating the results.

Note that for gas-phase tautomerism, one also could use AM1 and other semi-empirical techniques — such as Agent 2.0 (6) — and more complex quantum mechanical software packages — such as Gaussian 03 (7) or Spartan '02 (8) — to determine the most stable forms, but these packages generally are complex and not directly integrated with spectral management applications and will not be discussed further.

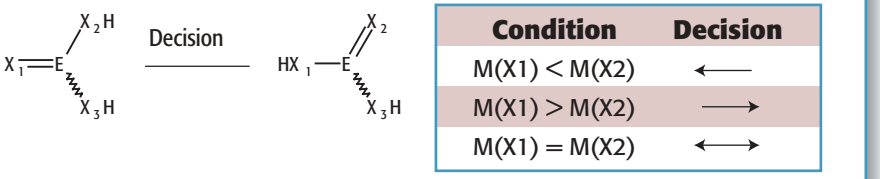


Figure 2. Decision process for an acyclic substructure. The wavy bond can have any order or can be absent.
General requirements:
- X_1, X_2, X_3 are from {X}, at least one must be non-carbon.
- E cannot be marked from aromatic rules.
- E is any nonmetal.

{X} is the description of allowed atoms (nature, charge, valence). M is a characteristic of the atom taking into account the nature of the atom and its substituents
Exceptions:
- Another rule applies if E is carbon and at least one of either X_1 or X_2 is carbon.
- Another rule applies if E is nitrogen and X_1 and X_2 are carbons.

Chemical Structures and Tautomeric Spectral Databases

When doing a search on a tautomeric molecule, some spectral management systems automatically prompt the user to confirm the tautomeric forms that should be used in the search. Some programs might ask you to narrow the search to only the given form, the most stable forms, all forms, or another select form. Figure 3 shows the dialog box that appears to help select alternative forms of quinoline-2,6-diol. Selecting all forms uses the tautomeric algorithm described in the "Methods" section to avoid missing relevant hits.

The ability to select all forms in a search is particularly useful when trying to eliminate duplicates. For example, the ACD/NIST MS Database (9) of mass spectra was searched for all tautomeric forms of anthralin using the 1,8-dihydroxyanthracen-9(10*H*)-one structure as input. As shown in Figure 4, two distinct hits were found with one entry assigned to a keto-diol and the other to a triol form, that is, anthracene-1,8,9-triol. Careful inspection shows the two spectra to be remarkably similar. Considering that electron ionization (EI) MS at nominal mass resolution cannot distinguish among diastereoisomers, regiomers, double-bond isomers, and tautomers, could one of the assignments be mistaken?

To answer this question, the IR spectra in the gas phase (10) and in Nujol (11, 12) were studied. Both IR spectra suggest the appropriate assignment is to the keto-diol form, which is confirmed by a very strong aromatic C=O stretch band at around 1610 cm^{-1}. As well, the carbon and hydrogen Fourier-trans-

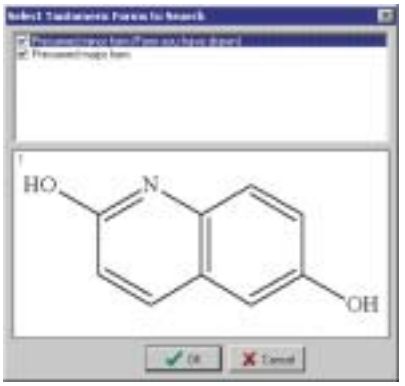


Figure 3. Dialog box defining the scope of the search with respect to tautomerism of quinoline-2,6-diol. This figure was taken from the ACD/SpecManager interface (5).

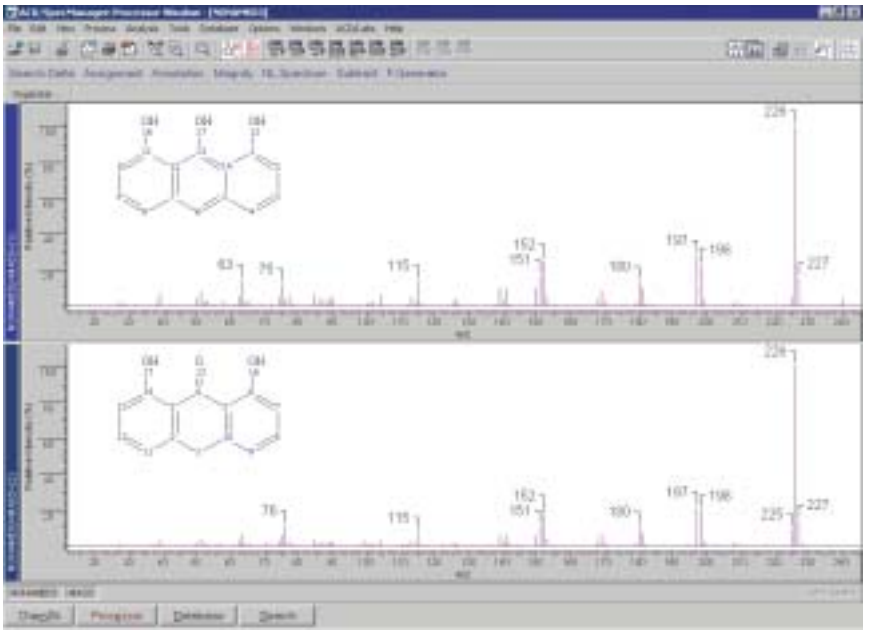


Figure 4. Possible duplicate spectra for anthralin in the NIST '98 database. Both spectra were extracted from the ACD/NIST MS database (9). The synchronized spectra of anthralin are displayed within ACD/MS Manager in Tile view with the triol form on top and keto-diol at the bottom.

form-NMR (FT-NMR) spectra (13) detect only the keto-diol form in chloroform-D. These are sufficient grounds to believe that anthralin exists predominantly in its keto-diol form and that the triol assignment in the mass spectral library was mistaken tautomerically. Therefore, it is likely that the two mass spectra are duplicates of the same tautomeric form.

Summary

Tautomerically intelligent algorithms are required for structure or substructure searches to find related tautomeric entries in a spectral library with structures. This is useful in avoiding missing relevant spectra and in finding duplicates. Because the spectrum–structure relationship is related intimately to the predominant tautomers present at equilibrium, the next installment of this series will discuss how expert systems help prevent misassignments.

Acknowledgement

The author wishes to thank Andrey Erin for his helpful comments while reviewing this article.

References:

1. S.V. Trepalin, A.V. Skorenko, K.V. Balakin, A.F. Nasonov, S.A. Lang, A.A. Ivashchenko, and N.P. Savchuk, J. *Chem. Inf. Comput. Sci.* **43,** 852–860 (2003).
2. C. Benson, *Monsanto vs. Merck,* www.legal500.com/devs/uk/ip/ukip_047.htm (accessed October 2003).
3. Michael B. Smith and Jerry March, *March's Advanced Organic Chemistry,* 5th ed. (John Wiley & Sons, Inc., New York, 2001), pp. 73–77.
4. CLIFF, *Generation of Tautomers* (Molecular Networks GmbH, www.molecular-networks.de/software/cliff/cliff_moreinfo6.html) (accessed October 2003).
5. Advanced Chemistry Development, Inc. (ACD/Labs), ver. 7.0, Toronto ON, Canada, www.acdlabs.com (accessed October 2003).
6. P. Pospisil, P. Ballmer, G. Folkers, and L. Scapozza, *Abstracts of Papers of the American Chemical Society* **224,** 211-COMP (2002).
7. *Gaussian '03 Online Manual* (Gaussian, Inc., Pittsburgh PA), www.gaussian.com (accessed October 2003).
8. Spartan '02, *Online Tutorial and Users Guide* (Wavefunction Inc., Irvine, CA), www.wavefun.com (accessed October 2003).
9. ACD/NIST MS Database: structure searchable form of the NIST '98 MS database using data licensed from the National Institute of Standard and Technology, Gaithersburg, MD (1998).
10. ACD/NIST IR Database: structure searchable form of the NIST/EPA Gas Phase Infrared Database using data licensed from the National Institute of Standard and Technology, Gaithersburg, MD (1998).
11. ACD/FDM FT-IR Spectra of Drugs/Canadian Forensic Spectra.
12. C.J. Pouchert, *The Aldrich Library of FT-IR Spectra,* 2nd ed., Vol. **2** (Sigma-Aldrich Co., 1997), p. 1900.
13. Aldrich/ACD Library of FT-NMR Spectra, www.acdlabs.com/products/spec_lab/exp_spectra/spec_libraries/aldrich.html (accessed October 2003). ■

Michel R.J. Hachey

is a technical marketing specialist and product manager for Advanced Chemistry Development, Inc. (Toronto, Ontario, Canada). E-mail: michel@acdlabs.com.