J Comput Aided Mol Des (2010) 24:497–520
DOI 10.1007/s10822-010-9338-4

# Tautomerism in chemical information management systems

**Wendy A. Warr**

Received: 8 March 2010 / Accepted: 16 March 2010 / Published online: 6 April 2010

**Abstract** Tautomerism has an impact on many of the processes in chemical information management systems including novelty checking during registration into chemical structure databases; storage of structures; exact and substructure searching in chemical structure databases; and depiction of structures retrieved by a search. The approaches taken by 27 different software vendors and database producers are compared. It is hoped that this comparison will act as a discussion document that could ultimately improve databases and software for researchers in the future.

**Keywords** Tautomer · Tautomerism · Chemical structure representation · Chemical information management · Registration · Substructure search · Software vendors · Database producers

## Introduction

Tautomerism has implications for many of the computational procedures used in drug discovery. The dividing lines between chemical information, cheminformatics and computational chemistry are by no means clear [1, 2] but this article aims to discuss only some "chemical information" aspects of tautomerism, namely novelty checking during registration into chemical structure databases; storage of structures; exact and substructure searching in chemical structure databases; and depiction of structures retrieved by a search. Implications of tautomerism in computational chemistry (for example in ligand preparation and property prediction) are deliberately not addressed.

The systems and databases that have been studied are listed in Table 1. The list is chosen to include most of the well known software companies; it is by no means comprehensive where databases are concerned. It is also true that some software organizations are missing: a few failed to reply to repeated requests for information, or had inadequate Web sites, and one or two may possibly have been overlooked. There are inter-relationships between the organizations and databases in Table 1; some of the software vendors sell databases built by other organizations and database vendors have selected their preferred chemical information management packages from the software vendors.

The aim of this article is not to provide a blow by blow account of every tautomer feature offered by every vendor, nor to make unfair comparisons among companies that serve very different markets (e.g., molecular modelers, chemical catalog companies and patent searchers). It should also be noted that vendors of "out-of-the box software" are in a rather different position from vendors of software toolkits. The latter may have the advantage of being able to offer multiple options from which customers can pick and choose, although those customers will have to expend some resource in building their own systems. Vendors of "out-of-the box" software may have to make decisions on behalf of a majority of customers, although in these days of open systems, it may be possible to plug in optional components.

Most of the facts in this article were collected by questioning software vendors and database producers individually. Some vendors were quicker than others in supplying detailed lists of chemical structures, but it should

W. A. Warr (✉)
Wendy Warr & Associates, Cheshire, England, UK
e-mail: wendy@warr.com

**Table 1** Software and database vendors

| | Web site | Comment |
|---|---|---|
| Accelrys | http://www.accelrys.com | Pipeline Pilot software |
| ACD/Labs | http://www.acdlabs.com | Software solutions that integrate chemical structures with analytical chemistry information |
| Beilstein/Reaxys | http://info.reaxys.com | Workflow tool integrating structure and reaction search with synthesis planning |
| CambridgeSoft | http://www.cambridgesoft.com | Enterprise solutions, desktop software, scientific databases, and professional services |
| Chemical Abstracts Service (CAS) | http://www.cas.org | Authoritative source of substance information (e.g., REGISTRY); software to access the CAS databases |
| Cambridge Crystallographic Data Centre (CCDC) Cambridge Structural Database (CSD) | http://www.ccdc.cam.ac.uk/ | CSD, the world repository of small-molecule crystal structures, and related software |
| Chemical Computing Group (CCG) | http://www.chemcomp.com | Applications for computational chemists, medicinal chemists and biologists |
| ChemAxon | http://www.chemaxon.com | Chemical software development platforms and desktop applications for the biotechnology and pharmaceutical industries |
| ChemoSoft | http://www.chemosoft.com/ | Cheminformatics solutions for drug design, and combinatorial and classical chemistry. |
| ChemSpider | http://www.chemspider.com | A free structure centric community for chemists |
| CWM Global Search | http://www.akosgmbh.eu/globalsearch/index.htm | Structure/name search of 35 databases on the Internet |
| Daylight Chemical Information Systems | http://www.daylight.com | Chemical information processing software |
| Dialog | http://www.dialog.com | DialogLink 5 interface to online information services |
| IDBS | http://www.idbs.com | Research data management and analytics solutions |
| InfoChem | http://www.infochem.de/ | Cheminformatics. Storage and handling of chemical structure and reaction information |
| InhibOx | http://www.inhibox.com | Computational methods for drug discovery. Virtual screening |
| John Wiley & Sons | http://www3.interscience.wiley.com/cgi-bin/mrwhome/104554785/HOME?CRETRY=1&SRETRY=0 | e-EROS Encyclopedia of Reagents for Organic Synthesis |
| Molecular Networks | http://www.molecular-networks.com/ | MN. TAUTOMER enumerates tautomeric states of a chemical compound |
| National Cancer Institute Chemical Structure Lookup Service (CSLS) | http://cactus.nci.nih.gov/lookup | Internet database and software |
| Open Eye Scientific Software | http://www.eyesopen.com/ | Software for molecular modeling and cheminformatics |
| Thieme | http://www.thieme-chemistry.com/en/home.html | Science of Synthesis and Pharmaceutical Substances |
| PubChem | http://pubchem.ncbi.nlm.nih.gov/ | Free database of chemical structures and biological activities |
| Questel | http://www.questel.com | Merged Markush Service (MMS) |
| Schrödinger | http://www.schrodinger.com | Molecular modeling and drug design software |
| SciTouch | http://opensource.scitouch.net/indigo/bingo | Open source cheminformatics toolkit |
| Symyx | http://www.symyx.com | Scientific information management (formerly MDL) |
| Thomson Reuters | http://www.thomsonreuters.com | Thomson Pharma, Web of Science, Thomson Innovation. Derwent World Patents Index, Derwent Chemistry Resource (DCR), Prous Integrity, etc. |
| Xemistry | www.xemistry.com | Chemical Algorithms, Construction, Threading and Verification System (CACTVS) |

not be assumed that other vendors have made a more cursory assessment of the subject, nor should it be assumed that anyone's solution is the one and only correct answer. This article should be viewed as a discussion document that could ultimately improve databases and software for researchers in the future.

## Chemical structure representation

Although there are very many databases and organizations, the number of "standards" for chemical structure representation is much smaller. Common ones are the Chemical Abstracts Service (CAS) connection table used in the REGISTRY system [3–15]; molfile, SDfile and other file standards developed by MDL, now Symyx [16, 17]; the International Union of Pure and Applied Chemistry (IUPAC) International Chemical Identifier, InChI [18]; and SMILES, developed by Daylight Chemical Information Systems [19, 20].

Through the use of a strict valence model, SMILES can represent molecular graphs, including tautomeric structures, with suppressed hydrogen structures, yielding very compact representations. These are suitable for database indexing and many related computational dictionary functions. Isomeric SMILES, which covers stereochemistry and isotopes, has further increased the utility of canonical SMILES. Note, however, that OpenEye canonical SMILES, Daylight canonical SMILES, SciTouch canonical SMILES and ChemAxon canonical SMILES are all independent unique descriptors. None of them can be used as interchangeable indices in cheminformatics.

The Morgan algorithm [3] underpins many of the systems in use today, and is the basis of the CAS REGISTRY database. It identifies atoms based on an extended connectivity value. The atom with the highest value becomes the first atom in the name, and its neighbors are then listed in descending order. Ties are resolved based on additional parameters, for example bond order, and atomic number. The original Morgan algorithm did not handle stereochemistry; SEMA (stereochemically unique naming algorithm) was developed to handle stereoisomers [21]. SEMA was adopted by MDL Information Systems (now Symyx Software). Symyx's NEMA (newly enhanced Morgan algorithm) produces a unique name and key for a wider range of structures than SEMA [22]. The work of Wipke et al. [23] identified the value of a constitutional key and a stereo key. This approach has been incorporated into NEMA.

InChI is an openly available, electronic format for exchanging chemical structure information over the Internet: a unique, linear identifier or "digital signature" [18, 24]. The InChI algorithm converts a chemical structure (in the form of its connection table) into a unique, alphanumeric string of characters. The program can also convert an InChI label back into a molecular structure. Two requirements must be fulfilled in doing this: different compounds must have different identifiers, with all the information needed to distinguish the structures; and any one compound must have only one identifier, including only the necessary information to identify that compound.

Since a given compound may be represented at different levels of detail, in order to create a robust expression of chemical identity the InChI team decided to create a hierarchical "layered" form of the Identifier, where each layer holds a distinct and separable class of structural information, with the layers ordered to provide successive structural refinement. In addition to basic connectivity and overall charge, the principal varieties of layers are mobile/fixed H-atoms (expresses tautomerism), isotopic composition and stereochemistry. The layered structure of the InChI allows future refinements with little or no change to the layers [25].

An InChI Key, a condensed digital representation of the identifier, can be generated based on a truncated SHA-256 hash [26] of the corresponding InChI layers. An InChIKey has two parts. The first block of 14 letters encodes the molecular skeleton (connectivity); the first eight letters of the second block encode stereochemistry and isotopes. Use of InChIKey allows searches based solely on atom connectivity (the first 14 characters). Tautomers have different structures, and different systematic names; those in Fig. 1 have identical InChIKeys but different NEMA keys. Mesomers do not exist separately and would ideally have the *same* identifier. Figure 2 is an example of mesomers with the same InChIKey but *different* NEMA keys.

The National Cancer Institute Computer Aided Drug Design (NCI/CADD) identifiers are calculated for the Chemical Structure Lookup Service (CSLS) [27]. They are based on hashcodes calculated by the cheminformatics toolkit CACTVS. CACTVS hashcodes represent a chemical structure uniquely as a 16-digit hexadecimal number (64-bit unsigned), have a high sensitivity to structural features of a compound, and change if the connectivity changes. Structure normalization is performed for any incoming structure set to be registered, or searched by, in CSLS. Each parent structure is then subjected to a hashcode calculation to generate the NCI/CADD identifier [28].

The normalization has adjustable levels of sensitivity. The Fragment Isotope Charge Tautomer Stereo (FICTS) identifier is a representation of the exact structure drawing, sensitive to all the five features. The FICuS identifier is not sensitive to tautomers ("u" stands for "unsensitive"), and comes close to how chemists perceive a chemical. The uuuuu identifier links closely related forms. Currently there are eight identifier variants defined for a structure: FICTS,

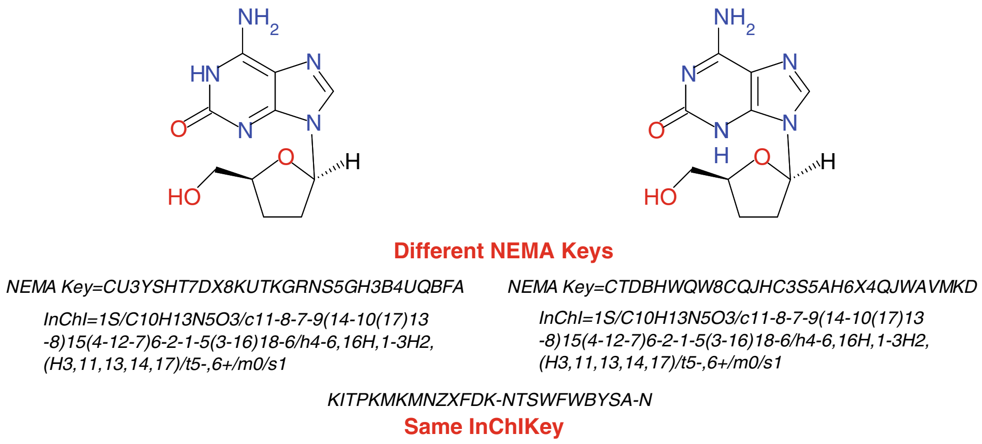

NH2
HN
N
N
O
N
O
H
HO
NH2
N
N
O
N
H
N
O
H
HO
Different NEMA Keys
NEMA Key=CU3YSHT7DX8KUTKGRNS5GH3B4UQBFA
NEMA Key=CTDBHWQW8CQJHC3S5AH6X4QJWAVMKD
InChI=1S/C10H13N5O3/c11-8-7-9(14-10(17)13-8)15(4-12-7)6-2-1-5(3-16)18-6/h4-6,16H,1-3H2,(H3,11,13,14,17)/t5-,6+/m0/s1
InChI=1S/C10H13N5O3/c11-8-7-9(14-10(17)13-8)15(4-12-7)6-2-1-5(3-16)18-6/h4-6,16H,1-3H2,(H3,11,13,14,17)/t5-,6+/m0/s1
KITPKMKMNZXFDK-NTSWFWBYSA-N
Same InChIKey

**Fig. 1** Tautomers with different NEMA keys but the same InChIKeys

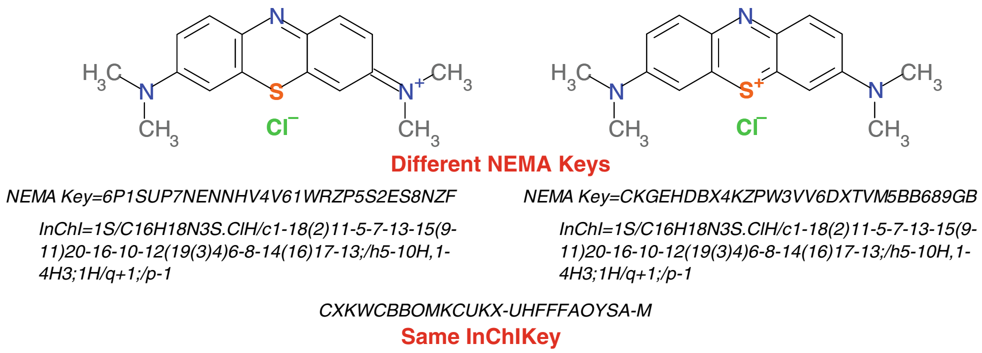

N
H3C
N
CH3
S
Cl−
N+
CH3
CH3
N
H3C
N
CH3
S+
Cl−
N
CH3
CH3
Different NEMA Keys
NEMA Key=6P1SUP7NENNHV4V61WRZP5S2ES8NZF
NEMA Key=CKGEHDBX4KZPW3VV6DXTVM5BB689GB
InChI=1S/C16H18N3S.ClH/c1-18(2)11-5-7-13-15(9-11)20-16-10-12(19(3)4)6-8-14(16)17-13;/h5-10H,1-4H3;1H/q+1;/p-1
InChI=1S/C16H18N3S.ClH/c1-18(2)11-5-7-13-15(9-11)20-16-10-12(19(3)4)6-8-14(16)17-13;/h5-10H,1-4H3;1H/q+1;/p-1
CXKWCBBOMKCUKX-UHFFFAOYSA-M
Same InChIKey

**Fig. 2** Mesomers with the same InChIKey but different NEMA Keys

FICTu, FICuS, FICuu, uuuTS, uuuTu, uuuuS, and uuuuu. Three of them, FICTS, FICuS and uuuuu are searchable for all the structure records in CSLS[28].

## Issues

When registering a compound into a chemical database system or registry, it is usual to check first whether its structure is novel. If the compound can exist in multiple forms, "novelty check" (or "duplicate search") must involve searching for all forms. This can be achieved in more than one way, for example, by storing all possible forms in the database (and probably indicating that they relate to just one compound), and doing an exact match search for the query molecule as drawn; or by storing just one form but ensuring that the existing and new structures are "normalized" in some way before comparison.

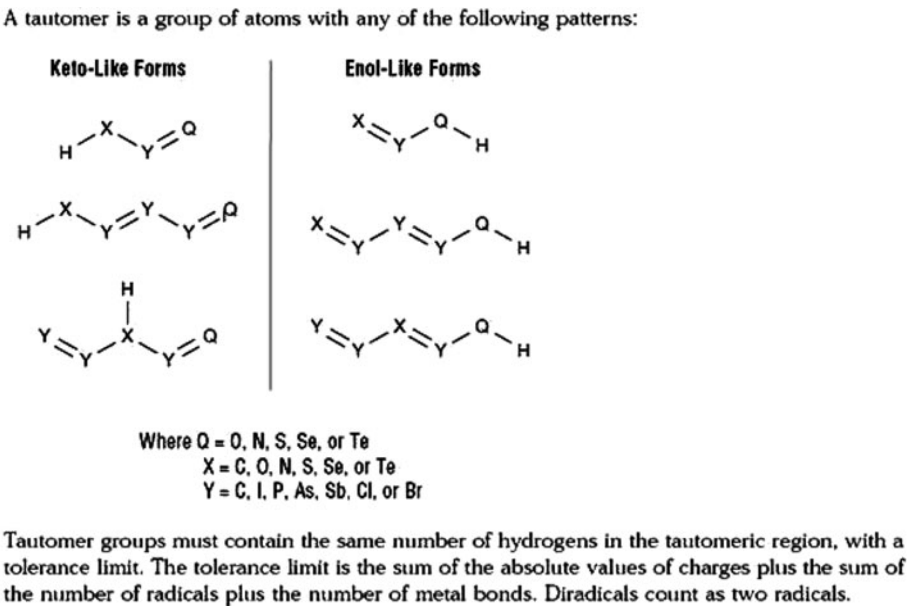
Whether or not all tautomers are stored, there may be reasons for selecting one of them as the preferred form. If one form only is required, how should it be chosen? Should it be the canonical tautomer described by some graph algorithm, or set of rules, or should the supposed major tautomer be stored? What algorithms and rules are currently in use?

If only one tautomeric form is stored, the query used in a substructure search might be modified to allow for the possibility of tautomerism. Alternatively the software developer may decide that it is up to the scientist doing the search to formulate queries that represent all the tautomeric forms being sought. Thus, substructure search is another challenge that can be addressed in more than one way. (Exact structure search is equivalent to the novelty checking procedure described above.)

Finally there is the decision of which tautomer or tautomers to display once the search is complete. Should this be the registered tautomer, the preferred tautomer (if any), the supposed major tautomer, the tautomer that best matches the query, or some combination of these options? Or should all possible tautomers be displayed, whether or not they are stored? The purpose of this article seeks to address all these issues, but before that can be done it is necessary to establish precisely what is meant by tautomerism.

## Definitions of tautomerism

The Symyx definition of tautomeric structures is given in Fig. 3. There can be multiple tautomeric groups in a single molecule. (Although Symyx software defaults to rules such as these, the rules are under user control and can be modified to suit local circumstances.) The Chemical Abstracts Service (CAS) definition [29] is similar (see Fig. 4) but has a broader range of elements. For tautomeric pyrazole

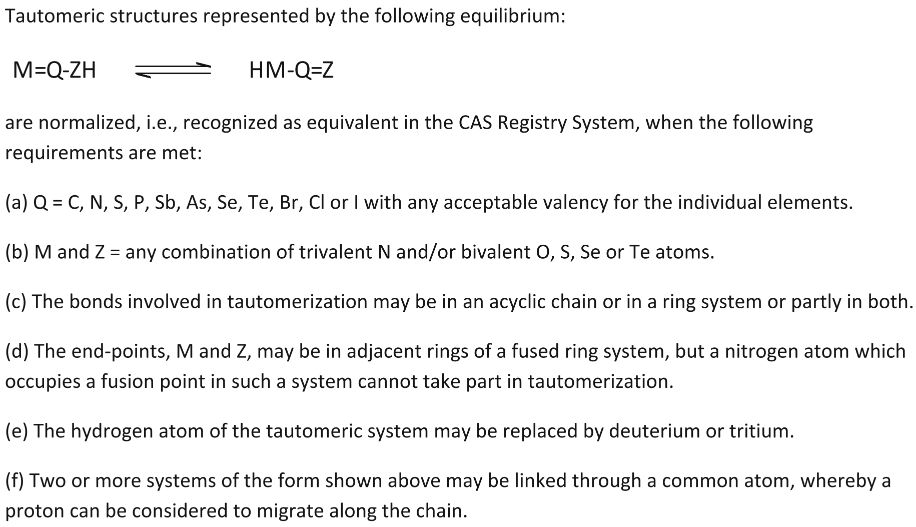


A tautomer is a group of atoms with any of the following patterns:

Where Q = O, N, S, Se, or Te
X = C, O, N, S, Se, or Te
Y = C, I, P, As, Sb, Cl, or Br

Tautomer groups must contain the same number of hydrogens in the tautomeric region, with a tolerance limit. The tolerance limit is the sum of the absolute values of charges plus the sum of the number of radicals plus the number of metal bonds. Diradicals count as two radicals.

**Fig. 3** Symyx definition of tautomerism



Tautomeric structures represented by the following equilibrium:

$$M{=}Q{-}ZH \rightleftharpoons HM{-}Q{=}Z$$

are normalized, i.e., recognized as equivalent in the CAS Registry System, when the following requirements are met:

(a) Q = C, N, S, P, Sb, As, Se, Te, Br, Cl or I with any acceptable valency for the individual elements.

(b) M and Z = any combination of trivalent N and/or bivalent O, S, Se or Te atoms.

(c) The bonds involved in tautomerization may be in an acyclic chain or in a ring system or partly in both.

(d) The end-points, M and Z, may be in adjacent rings of a fused ring system, but a nitrogen atom which occupies a fusion point in such a system cannot take part in tautomerization.

(e) The hydrogen atom of the tautomeric system may be replaced by deuterium or tritium.

(f) Two or more systems of the form shown above may be linked through a common atom, whereby a proton can be considered to migrate along the chain.

**Fig. 4** CAS requirements for the normalization of structures

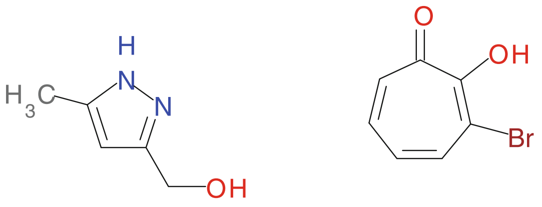
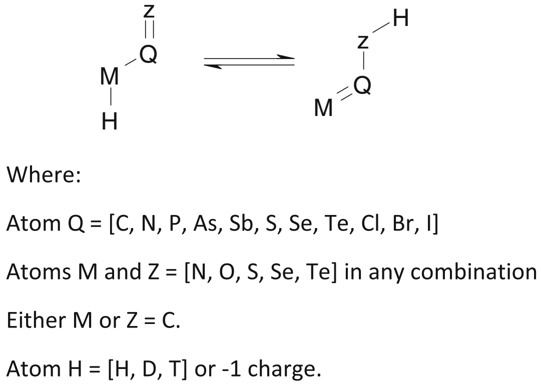


**Fig. 5** Preferred pyrazole and tropolone in CAS Registry System



Where:

Atom Q = [C, N, P, As, Sb, S, Se, Te, Cl, Br, I]

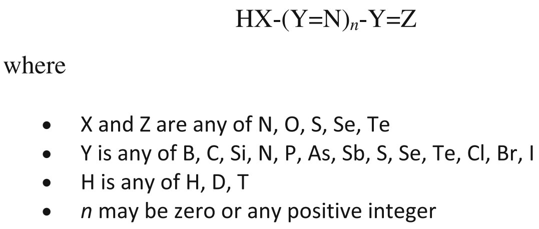
Atoms M and Z = [N, O, S, Se, Te] in any combination

Either M or Z = C.

Atom H = [H, D, T] or -1 charge.

**Fig. 6** IDBS definition of tautomerism

derivatives and for tropolones, Chemical Abstracts selects a single preferred structure and index name (using a lowest locant principle, see Fig. 5), and assigns a single CAS Registry Number, even though these systems do not conform to the general equilibrium in Fig. 4 and are not currently normalized by the CAS Registry System. The IDBS basic definition (Fig. 6) is very similar to CAS' except that M or Z can be carbon, and a negative charge can migrate instead of hydrogen (or a hydrogen isotope). CAS allows a positive charge to migrate.



$HX\text{-}(Y{=}N)_n\text{-}Y{=}Z$

where

- X and Z are any of N, O, S, Se, Te
- Y is any of B, C, Si, N, P, As, Sb, S, Se, Te, Cl, Br, I
- H is any of H, D, T
- *n* may be zero or any positive integer

**Fig. 7** Proton-shift tautomerism, without terminal carbons

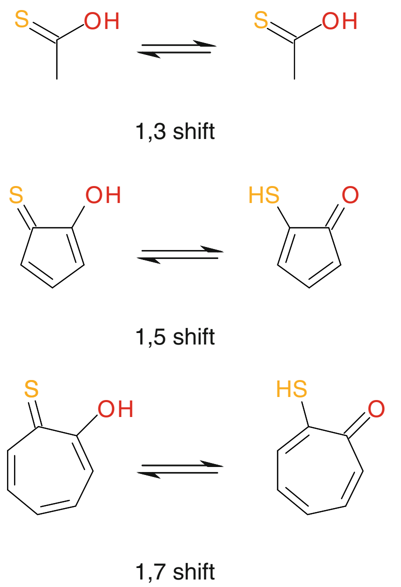


**Fig. 8** Size of recognized tautomeric systems

The CambridgeSoft representation of the simplest form of tautomerism, proton-shift tautomerism, without terminal carbons, is given in Fig. 7. The parameter *n* allows for 1,3-, 1,5-, 1,7-shifts (see Fig. 8) and beyond. (This is not to imply that other organizations do not recognize more distant shifts than 1,3.) CambridgeSoft can supply a very wide definition including proton-shift tautomerism without terminal carbons (Fig. 7), proton-shift tautomerism with one terminal carbon (Fig. 9), proton-shift tautomerism, with higher unsaturation (Fig. 10), ring-chain "tautomerism" (Fig. 11), valence tautomerism (Fig. 12), "charge tautomers" (Fig. 13), "unreasonable tautomers" (Fig. 14), and "hidden tautomers" (Fig. 15).

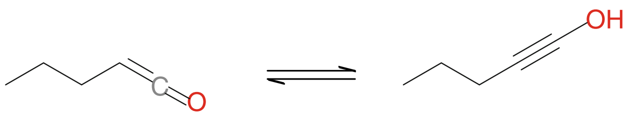


**Fig. 10** Proton-shift tautomerism, with higher unsaturation

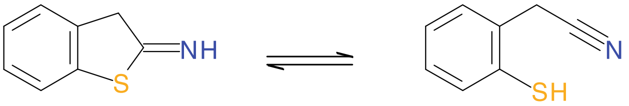


**Fig. 11** Ring-chain "tautomerism", involving ring opening

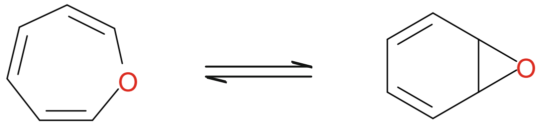


**Fig. 12** Valence tautomerism

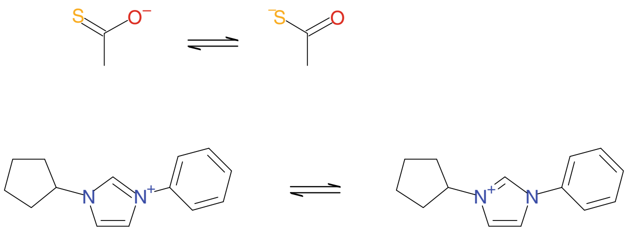


**Fig. 13** "Charge tautomers"

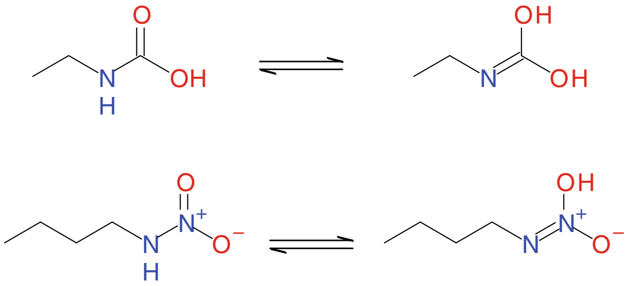
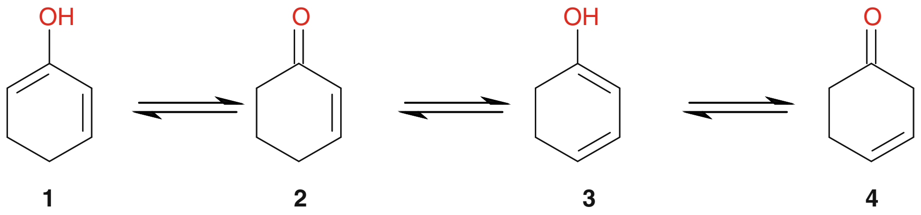


**Fig. 14** "Unreasonable tautomers"

IDBS' documentation supplies the specific examples in Fig. 16. In addition it describes overlapping systems (Fig. 17) and adjacent and non-adjacent forms (Fig. 18). Non-adjacent forms arise because of overlapping tautomeric systems. In the example in Fig. 18, the overlapping

**Fig. 9** Two examples of proton-shift tautomerism, with one terminal carbon



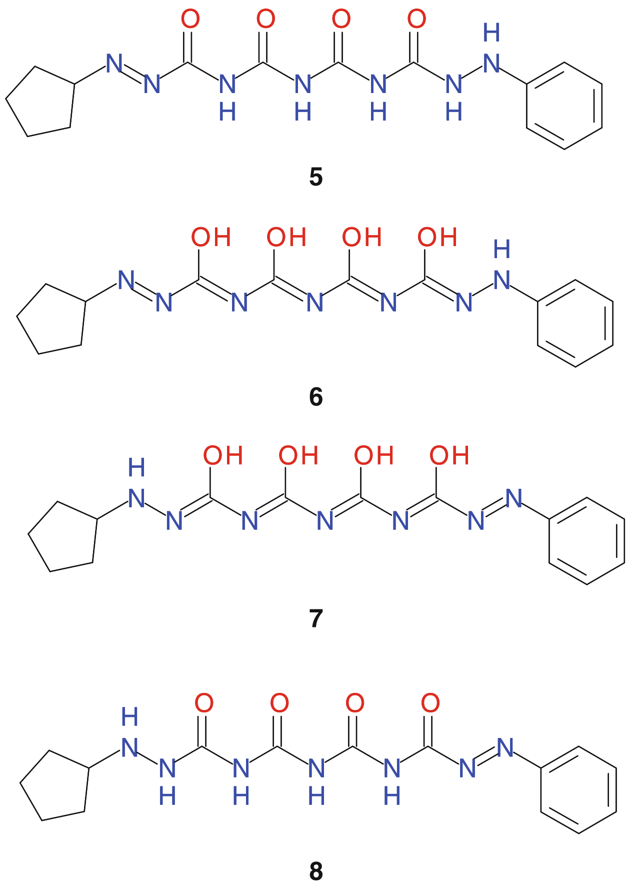


**Fig. 15** "Hidden tautomers"

systems belong to different tautomeric cases (imine-enamine and azo-hydrazone), but non-adjacency may also arise where the overlapping systems belong to the same tautomeric case (as in two overlapping keto-enol systems). Non-adjacent forms may only be interconverted via adjacent intermediate forms such as Structure **10** in Fig. 18. Unsaturated hydrazine **9** can only convert directly into hydrazone **10** (imine-enamine case). Azo **11** can only convert directly to hydrazone **10** (azo-hydrazone case). Thus **9** and **11** are non-adjacent forms that cannot convert directly into each other.

## Support for multiple types of tautomerism

We turn now to a discussion of which organizations recognize which sorts of tautomerism (assuming that they address the issue at all). Before elaborating on this, it is necessary to recognize that the approaches to tautomerism of experts in "informatics" (some might say "IT") are quite different from those of computational chemists. For example vendors of informatics software supply multiple chemical structure examples of the types of tautomerism they recognize; computational chemistry companies send lists of rules. Informatics experts think, at least partly, in graph theory terms, computational chemists talk of minimizing energies. This section thus tends to address informatics approaches. Organizations are discussed in a logical rather than alphabetic order in this section.

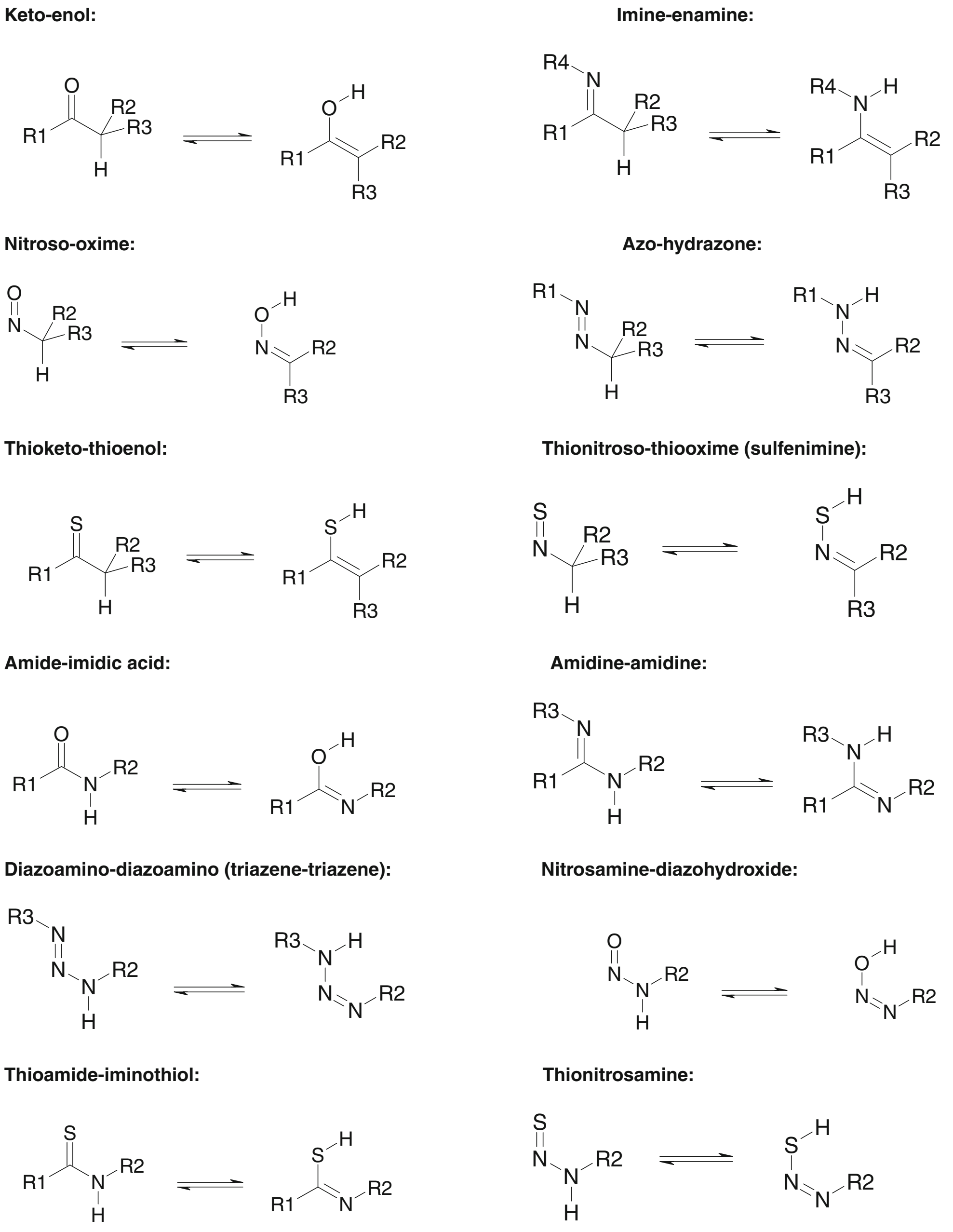
### Limited support

The Protein Data Bank (PDB) does not explicitly include tautomers in its databases; it only archives what is in the crystal structure complexes, so there is no IT structure for dealing with tautomers. The Beilstein database under the CrossFire system is being phased out. Beilstein is offered online on STN and it is included in Elsevier's new Reaxys system. In the past the database included a tautomer identifier (i.e., a compound was given a Beilstein Registry Number and individual tautomers were given further identifiers) but this feature has temporarily been disabled. Tautomer recognition is likely to be added to Reaxys before the end of 2010. Another system that still has limited tautomer support is InChI. InChI currently supports only 1,3-migration of hydrogen between heteroatoms; 1,5 migration and keto-enol tautomerism are currently being tested and may become available later as a special user-defined option. This has implications for Internet search engines which are dependent on InChI.

### CambridgeSoft

CambridgeSoft is at the other extreme, having studied a great many possibilities for tautomerism (Figs. 7, 8, 9, 10, 11, 12, 13, 14, and 15). CambridgeSoft implements the broadest form of the rule in Fig. 7. Tautomers involving 1,3-shifts are recognized by many software systems; CambridgeSoft products recognize tautomeric systems with no size constraints. There is nothing "magical" about 1,3-, 1,5-, or 1,7-shifts (see Fig. 8); from a chemical perspective, this sort of proton-shift tautomerism is characterized by a 1,*n* system of alternating single and double bonds, and sometimes it includes bonds in an aromatic system.

In the category "proton-shift tautomerism, with one terminal carbon", one of X or Z in Fig. 7 is a carbon atom. CambridgeSoft software recognizes keto-enol tautomers in this category but only for three-atom systems. They argue as follows. Consider the structures in Fig. 9. Structures **1** and **2** represent a keto-enol tautomeric pair. Structures **3** and **4** represent another keto-enol tautomeric pair. In contrast, structures **2** and **3** are related through a 1,5-shift where a hydrogen nominally shifts between the oxygen and the *para* carbon. If that sort of "extended keto-enol tautomerism" were to be allowed, then it would imply that



Keto-enol:

Imine-enamine:

Nitroso-oxime:

Azo-hydrazone:

Thioketo-thioenol:

Thionitroso-thiooxime (sulfenimine):

Amide-imidic acid:

Amidine-amidine:

Diazoamino-diazoamino (triazene-triazene):

Nitrosamine-diazohydroxide:

Thioamide-iminothiol:

Thionitrosamine:

**Fig. 16** Examples of tautomerism (Source: IDBS)

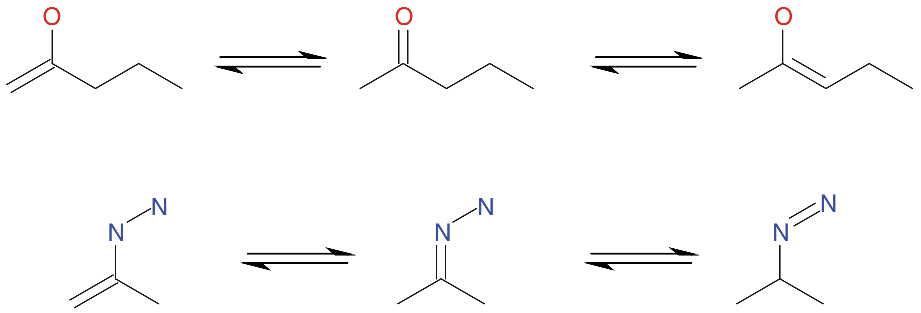


**Fig. 17** Overlapping tautomeric systems

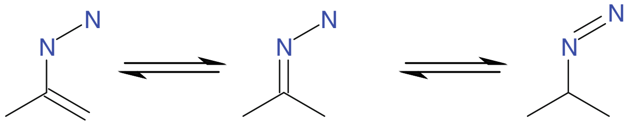


**Fig. 18** Adjacent and non-adjacent forms

structures **2** and **3** were also tautomers of each other, which is clearly unreasonable from a chemical perspective. Accordingly, tautomeric shifts involving a terminal carbon atom *must* be restricted to three-atom systems in CambridgeSoft logic.

CambridgeSoft software recognizes any amount of unsaturation across tautomeric systems of any size. Proton-shift tautomerism requires unsaturated bonds between the end points, but there are no restrictions from a chemical sense in the amount of unsaturation. From CambridgeSoft's point of view, a ketene-ynol tautomerism (Fig. 10) involves two units of unsaturation across a three-atom system.

Some forms of tautomerism involve ring opening (Fig. 11). Carbohydrates are a classic example, and also phenolphthalein, but the general category is much broader than that. CambridgeSoft products currently do *not* recognize ring-chain tautomerism. It is not clear from a chemical information management perspective whether chemists would want to see structures of this type recognized as tautomers, even though they do match CambridgeSoft's "dictionary definition".

Some structures are able to interconvert without any change in the hydrogen bonding pattern ("valence tautomerism", Fig. 12). CambridgeSoft claims that this is still a true tautomerism (rather than resonance) because the atoms of the structure do move relative to each other. Compounds of this type are said to have fluxional structures [30]. The prototypical example is bullvalene. CambridgeSoft products currently do *not* recognize valence tautomerism, even though it fits the dictionary definition.

Unlike tautomers, which are chemically distinct species that can be isolated under appropriate conditions (at least in theory), "charge tautomers" are simply representational variants of resonance forms (Fig. 13). They should be recognized as identical and not be treated as tautomers. CambridgeSoft software, however, does recognize these charge-shift resonance pairs, and as with tautomers there are no limitations to the size of the shift. In real compounds, and especially in dyes, the distances can be quite large.

There are certain types of tautomerism where the average chemist would agree that one compound in the "tautomer pair" could not possibly exist. Examples include "tautomers" that break a carboxylic acid or a nitro group (Fig. 14). While it is true that these tautomers are unreasonable, CambridgeSoft holds that this is not an issue that should be ignored: if a user did enter an "unreasonable" tautomer, the software should recognize that it is a tautomer of the more usual form. Some software is designed specifically to exclude recognition of this sort of tautomerism. CambridgeSoft software *does* recognize it, while admitting that it is indeed rarely encountered.

In the presence of multiple overlapping tautomeric centers the situation can get extremely complex. Consider the structures in Fig. 15. Few chemists would recognize that **5** and **8** are tautomers relative to each other. Structure **5** can interconvert with **6** through a series of keto-iminol tautomerisms. Structure **6** can interconvert with **7** through a 1,11-proton shift. Then **7** can interconvert to **8** through an inverse series of keto-iminol tautomerisms. So indeed **5** and **8** are tautomers, even though the net result is a shift of *two* hydrogen atoms from one end of the structure to the other. CambridgeSoft software recognizes this sort of tautomerism.

## ACD/Labs

ACD/Labs supports keto-enol tautomerism but forms that are very unlikely in practice are not proposed. Thus if any of the structures in Fig. 19a is drawn by a user, the other two are proposed as options; that in Fig. 19b is not proposed, but if a user draws this structure, the three in Fig. 19a will be proposed, and can be used for registration and search. For acetone the minor enol form is not proposed by ACD/Labs procedures as this form is really minor and hardly detectable. Tropolone tautomerism is handled because it also falls in the keto-enol category. The "length" of the tautomeric system is not a problem for ACD/Labs: all six tautomers in Fig. 20 are generated, two of them corresponding to N/NH tautomerism and the other four to N/CH tautomerism (imine-enamine), an analog of keto-enol tautomerism.

The double bond may be in an aromatic system. ACD/Labs' software recognizes the nitroso-oxime example in Fig. 21 but does not propose any tautomer for phenol since the keto form is hardly detectable. Both 2-hydoxypyridine and 2-pyridone are recognized and the pyridone is treated as the predominant form. Overlapping systems such as guanidine (Fig. 22) are handled.

## IDBS

In IDBS's software, tautomeric systems extending across a conjugated system of double bonds (along a chain or around a ring system) are detected, whether or not the conjugated system is aromatic. All forms in a tautomeric set, whether adjacent or non-adjacent (Fig. 18) will be

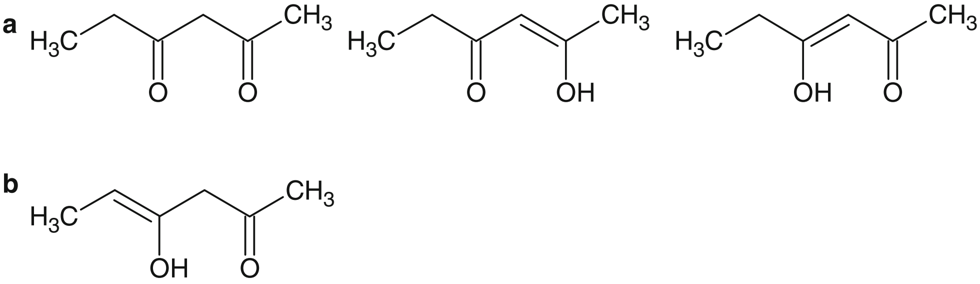


**Fig. 19** **a** Keto-enol tautomers generated by ACD/Labs software. **b.** Tautomer not proposed by ACD/Labs

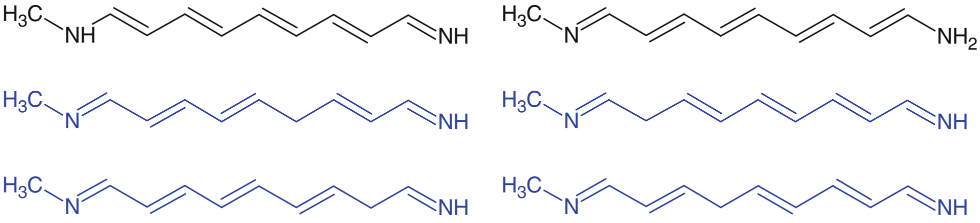


**Fig. 20** Tautomerism involving a 1,11 shift

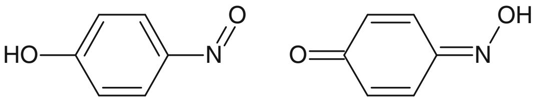


**Fig. 21** Nitroso oxime example

found in an “exact by tautomer” search, whichever form is input as the query. Many, if not most, of the examples suggested by CambridgeSoft are handled, but one difference is that IDBS *does* support 1,5-shifts of the keto-enol type, something that CambridgeSoft claims is inappropriate.

## Symyx

Symyx agrees with CambridgeSoft on proton-shift tautomerism, with or without terminal carbon, and on the unlimited size of a tautomeric system (Figs. 7, 8, and 9). The company agrees that proton-shift tautomerism with higher unsaturation is a valid example of tautomerism, but does not perceive it yet. Its customers have not complained, perhaps because energetics make the situation rare, but the company will consider implementing it in the future for completeness. Symyx considers Figs. 11, 12, 13, 14, and 15 to be stretching the definition of tautomerism and mostly degrading the usefulness of the term. Its software does, however, perceive “charge tautomers” as tautomers, even though the forms look more like mesomers, in line with CambridgeSoft’s treatment of these structures. The unreasonable tautomer where the carboxylic acid is “broken” (Fig. 14) is recognized by Symyx software but not the nitro group example. Symyx does not accept the “hidden tautomer” (Structures **5**, **8** of Fig. 15); it is seen as two products that might rearrange under the right conditions. The software does not recognize the two isomers as tautomers but it does recognize them individually as tautomers of the intermediate poly-enol in Fig. 23.

Symyx plans to widen the definition of a tautomeric region allowing the detection of larger collections of atoms with a mobile hydrogen. This will allow the software to detect tautomeric relationships such as “A is a tautomer of B and B is a tautomer of C, so A is also a tautomer of C” (see Fig. 18). Currently if the user standardizes on B as the reference structure all is well and the relationship is detected, but if A or C is selected as the standard format then only B is detected as a tautomer.

## OpenEye

OpenEye does not list specific forms of tautomerism because that is not the nature of its algorithm. It does not have a list of tautomeric forms recognized and successfully handled. Instead, it examines the atom types of the atoms in the molecule and their bonding patterns and then applies an algorithm designed to reproduce the chemistry of tautomerism. This algorithm attempts to be generally applicable and is not reliant on specific definitions of types of tautomers.

OpenEye’s tautomers program handles keto-enol tautomerism at the 1,3 level, and asked for the unique form,

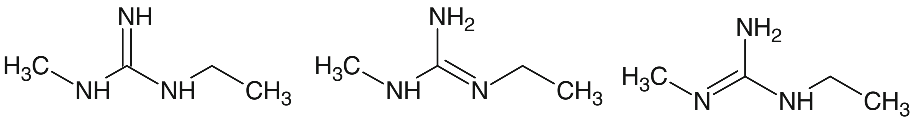


**Fig. 22** Guanidine with overlapping tautomeric systems

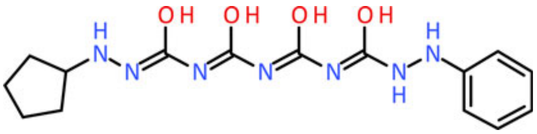


**Fig. 23** Poly-enol for comparison with Fig. 15

will give the same output tautomer for both input tautomers. The behavior for ketones is analogously successful. There is a flag to turn this feature on or off in the software as it can lead to a noticeable explosion of tautomers that is not always desirable. In line with CambridgeSoft, OpenEye does not cover extended keto-enol tautomerism (1,5 shifts and beyond) but it places no restrictions on the size of the tautomeric system in other cases. For the structures in Fig. 9, OpenEye products generate Structure **1** as the unique form from any of the four starting structures (as long as the flag indicating keto-enol tautomerism is set); if the flag is not set, none of the four forms is tautomerized.

OpenEye software perceives tautomer zones (atoms of which the electrons and protons transfer). Each zone has an independent state. When zones overlap, they are merged into a single zone and the algorithm is applied to the entire zone. Aromaticity (and its perception) is considered in the OpenEye algorithm, and a user can request the software to produce tautomers that retain maximum aromaticity.

OpenEye software both enumerates and canonicalizes the seven classes keto-enol, amide-imidic acid, amidine-amidine, diazoamino-diazoamino, nitrosamine-diazohydroxide, thioamide-iminothiol, and thionitrosamine examples in Fig. 16. Tautomeric systems extending across a conjugated system of double bonds (along a chain or around a ring system) are detected, whether or not the conjugated system is aromatic. Imine-enamine, nitroso-oxime, azo-hydrazone, thioketo-thioenol, thionitroso-thiooxime in Fig. 16 are not enumerated, and the non-adjacent tautomeric system of Fig. 18 is not recognized. Proton-shift tautomerism with higher unsaturation (Fig. 10), ring-opening (Fig. 11) and valence tautomerism (Fig. 12) are not addressed, but charge resonance forms are. Both "unreasonable" tautomers (Fig. 13) are recognized and the unique form is the one with the standard carboxylic acid or nitro group. CambridgeSoft is not alone in handling "hidden tautomers": OpenEye software generates a full set of 68 tautomers whichever one of the four structures in Fig. 15 is input, and it generates Structure **5** as the unique tautomer given any one of the four (or indeed 68) starting molecules.

## CACTVS

The CACTVS approach for generating tautomeric states of molecules is described in Ref. [31]. Molecular Networks' MN.TAUTOMER is a stand-alone encapsulated script of Xemistry's CACTVS, i.e., a rebranded CACTVS application. It is a rule-based enumerator of tautomeric forms of a chemical compound. The rules automatically detect substructures in a given molecule that can undergo a tautomeric transformation *via* SMIRKS transforms. In total, a set of 23 transformation rules is encoded in MN.TAUTOMER. Most of the important types of tautomerism are supported including:

- simple and long-range enol/thioenol exchange
- simple imine exchange
- nonsubstituted heteroaromatic exchange
- simple or long-range hetero atom hydrogen exchange
- keten/inol exchange
- nitro form/aci form of nitro compounds
- simple nitroso/oxime exchange or nitroso/oxime exchange of aromatic systems
- cyanuric acid, formamidinsulfonic acid, hydrogen cyanide/hydrogen isocyanide, phosphonic acid
- sulfonamides, sulfonylsulfane, sufonylphosphane.

In CACTVS, the tautomer rules are in principle configurable, and cover up to 1,11 H shifts, with long-range rules becoming more selective and less aggressive than short-range rules.

## Chemical Abstracts Service

CAS requirements for the normalization of structures [29] are shown in Fig. 4. All possible tautomers are identified by a computer procedure that searches for potential endpoints, for instance, nitrogen or chalcogen atoms, which are double bonded to an atom acceptable as a centerpoint. When such two-atom sets are identified, the remaining attachments of the centerpoint are checked. If a potential endpoint bearing a mobile group is found, all qualifying endpoints and their mobile groups are included in the tautomer group, and the centerpoint-endpoint bonds are marked as tautomer bonds. So that "onium" substructures common in dyes can be normalized, substances that have mobile groups with a plus (+) charge are also recognized as tautomers. Keto and enol forms of a substance are not normalized to a preferred form in the CAS REGISTRY. CAS scientists enter into the REGISTRY whichever form of the keto-enol tautomer is represented in the article being indexed.

## SciTouch

The Indigo organic chemistry tool kit from SciTouch [32] includes a data cartridge called Bingo. Bingo supports ring-chain tautomerism and user-defined constraints on tautomeric chains. The user can restrict the tautomer search by enabling conditions for boundary atoms in tautomeric chains. By default, there are three conditions:

- Each boundary atom in the tautomeric chain must be one of N, O, P, S, As, Se, Sb, Te
- Carbon not from an aromatic ring at one end of the tautomeric chain, and one of N, O, P, S at the other end
- Carbon from an aromatic ring at one end of the tautomeric chain and one of N, O at the other end

Users are allowed to use an arbitrary subset of these conditions for matching, and also add their own conditions, based, as the default three, on atom numbers and/or aromatic and aliphatic property.

### Schrödinger

In the 2010 release of Epik [33, 34], Schrödinger will cover 900 types of tautomerism with more than 4,000 tautomers listed (including many low population tautomers for identification purposes). While many types of non-aromatic tautomerism are covered, including various keto-enol tautomerisms and their nitrogen analogs, aromatic systems constitute more than 90% of the tautomerism covered, since these more commonly result in significantly populated tautomers in drug-like molecules. In addition, Epik's protonation state adjustments cover an even larger number of types of tautomers by removing a proton from one location on the molecule and subsequently adding a proton somewhere else.

### Others

ChemAxon tools allow for four different ways of handling tautomers; the types of tautomer recognized will depend on the method. In one method (customizing preferences in Standardizer) all transformations are defined by the user [35], an approach which allows for more esoteric transformations such as ring-chain tautomerism. A publication by Chemical Diversity Labs gives a detailed description of handling tautomerism in exact structure searching, in the ChemoSoft software environment [36]. Since then ChemoSoft has added "long distance" tautomer handling and other unrelated features. In Accelrys' Pipeline Pilot the default substructures considered to calculate a tautomer score include amide, enamine, keto, enol, diamide, thioamide, aromatic bonds, and exocyclic double bonds. Customization is possible. The procedure is discussed at the appropriate point later in this article.

## Registration

The recognition and treatment of tautomers is dependent on the specific task at hand. For example, the representation of a chemical substance in a corporate database might need to be independent of specific tautomeric form, whereas spectral properties often require the distinction between specific forms. Since tautomeric forms may have different signals in NMR spectra, ACD/Labs software generates all tautomers of an input structure and lets the user choose which tautomer(s) to use for the prediction of a spectrum.

Whether or not different tautomeric forms should be registered will depend on the use of the database and the business rules of the organization creating the system. Treating different tautomers as identical will always lose information. Whether that is appropriate depends on the given situation. A research laboratory specializing in ultra-low-temperature species almost certainly would want to keep tautomers distinct. On the other hand, a user keeping an inventory of a stockroom stored at room temperature would probably prefer to merge tautomers.

Reactions and mechanistic information are other cases. Chemists would be unlikely to use the unusual forms of carboxylic acid and nitro groups except in a reaction mechanism, and if such forms were used in a reaction, they probably should be registered unchanged. What is recorded in a reaction and what is put in a corporate registry are not necessarily the same. Symyx recommends registering the reaction as drawn, and also registering the structures after normalization, so that searching for what was done works, and queries that follow corporate business rules also work.

Some of the companies in this article provide software for systematic naming of compounds. Such software must be able to name any structure that is input; tautomerism is to some extent a non-issue. If the software can generate all tautomers it should be able to name all tautomers, should the user really want extra names in addition to the one for the input structure. If the software also has an algorithm for deriving a preferred tautomer, then it can name a preferred tautomer as well. The subject of chemical nomenclature is very complex and is outside the scope of this article. Registration approaches are discussed in this section in alphabetical order of vendor.

### Accelrys

In a registration system built using Accelrys's Pipeline Pilot, using the Pipeline Pilot Chemistry Cartridge to provide structure indexing and searching, there are two potential ways to recognize different tautomeric forms that may already be present in a database. In the first, the registration protocol takes the input molecule, enumerates all possible tautomeric forms of that molecule and stores all of these in an SDfile (along with the original input structure) as a single row in the database. In the second, the registration protocol calculates and stores the canonical tautomer for each molecule being registered into the database.

The componentized nature of Pipeline Pilot makes it possible for users to configure the system to support the tautomeric matching behavior that their business rules require. The Enumerate Tautomers component in Pipeline Pilot takes input molecules and can enumerate all possible tautomers or it can convert the molecule to a canonical tautomer. The tautomer algorithm that the component uses is based on that of Sayle and Delany [37].

### ACD/Labs

ACD/Labs software stores structures in the database as specific (fixed) tautomeric forms. The tautomer recognition tool generates possible tautomeric forms for a given structure according to a set of rules, estimates a preference for each specific form and ranks them as major, minor, or conditions dependent. For registration the software allows the user to choose the tautomeric form to store, but since the tautomer module is aimed at generation of preferred form(s), if a particular form is recognized as the major one, minor forms are not proposed for registration.

### CambridgeSoft

Because not everyone wants to treat different tautomers as identical, CambridgeSoft offers the choice of doing so or not. The software recognizes tautomeric regions during an "indexing" phase, and stores that perceived information in the database. That process is exactly analogous to the perception of other types of data, including aromaticity, stereochemistry, ring system membership, and so on. For determining an exact match, CambridgeSoft internally can consult a canonical representation of the chemical structure. By definition, the canonical representation will be the same no matter how a structure is drawn, and in the case of tautomers it will also be the same no matter which tautomer is drawn. The use of canonical codes allows very fast comparisons. If the registrar agrees that a new substance is the same as an existing substance, then only one structure diagram is maintained; multiple structure diagrams are not stored. If the registrar deems that a second diagram really is a different substance, it is entered as a new record with its own identifier and its own structure.

### Cambridge structural database

Entries stored in the Cambridge Structural Database (CSD) reflect the publication. The chemical structure described in the CSD, therefore, is a representation of the X-ray or neutron diffraction determination and, in the case of tautomers, it will correspond to the exact tautomer determined in the solid state. In cases where a different tautomer of the new entry has been determined in the past, and it is referred to in the publication, this information and the CSD identifier (refcode) of the previous tautomer may be indicated in the notes section of the entries. Tautomeric entries in the CSD will occur in different refcode families because they are different molecular species in the solid state.

The accuracy of location of hydrogen atoms can vary depending on the crystallographer, the quality of data and the temperature of the study. The positions of the hydrogen atoms are given following the crystallographer's indications. Most hydrogen positions are based on geometry. Programs like CCDC's Mercury [38] and Pluto [39] will recalculate and normalize hydrogen positions for the very reason that they can be poorly defined. Sometimes, disorder between the two tautomers might be shown and this will be indicated in the entry. In addition, any indications that the assignment of the tautomer may be incorrect will be noted.

### CAS

In CAS REGISTRY, all possible tautomers, normalized to the preferred tautomer, are included in a single substance record. If an author emphasizes and characterizes a specific tautomer other than the preferred one, it is included in REGISTRY as a separate record.

### ChemAxon

ChemAxon's tools allow for four methods of handling tautomers. Method 1 uses a "generic tautomer" or normalized structure for novelty checking. Normalization is akin to the methods described elsewhere as "flexmatch", in which tautomer regions between target and query are identified at search time, and compared specially during atom–atom mapping. In the tautomeric regions, the "sigma-skeleton" (common graph) is matched, but that is combined with considering other properties of the region (number of bonding electrons, movable hydrogens and hydrogen isotopes, etc.). Parts outside the tautomeric region are handled in the same way as non-tautomeric search. Here the ChemAxon method differs from some other approaches.

ChemAxon's second method is to allow for all tautomers at search time. Method 3 is a way of choosing a preferred tautomer, and method 4 is a way of allowing users to customize preferences in Standardizer. All four methods are suitable for registration, but the most efficient and robust method is the first approach, using generic tautomers. This method also handles tautomerism of molecules with hydrogen isotopes and the spontaneous racemization of stereo molecules as a result of tautomerism. Users can generate all tautomers and store them in the database if they wish, but this is not obligatory.

### Chemical computing group

The way MOE cheminformatics algorithms from Chemical Computing Group (CCG) handle tautomers is to store just the incoming structure, then enumerate when a comparison is required. The comparison method for determining whether two structures A and B are tautomers of each other is:

- If A and B are identical, accept
- If the total number of hydrogen atoms or charges are not identical, reject
- Examine the heavy-atom skeletons; reject if not identical
- Enumerate all tautomers for A; if any is the same as B, accept
- Enumerate all tautomers for B; if any is the same as A, accept
- Otherwise reject.

### ChemSpider

ChemSpider uses the InChI as the primary key in the database and as the basis of deduplication of data. As a result tautomers are collapsed according to the capabilities of InChI to recognize related tautomers. This is a historical choice from the beginning of the project and in the future the intention is to retain specific tautomers but use InChI as a way of interrelating specific tautomeric series. ChemSpider does not record the structure diagram of different tautomers. The tautomer stored is the first tautomer which was deposited for a given structure and it is depicted graphically as a thumbnail. Other tautomeric forms of that structure should be represented by the same InChI but this is of course limited by the ability of the InChI algorithm to recognize all tautomers. If the tautomers are identified then they collapse under the original thumbnail that was generated for the first tautomeric structure submitted to the database. It is possible to edit the displayed thumbnail to show an alternative or preferred tautomer as long as it has the same InChI.

### Daylight Chemical Information Systems

In Daylight's opinion, the most robust way to find tautomers in the database is to search based on common graph (connected atoms/bonds, ignoring bond order, aromaticity, and charge), total net charge, and total net hydrogen count. This will yield all the tautomeric isomers of a given molecule, although some of them will not be readily interconvertible and are not considered tautomeric (e.g., 1-butene and 2-butene). This will give false hits, but it is a very simple retrieval and is sure to produce all tautomers, plus some other structural isomers. After that, Daylight has algorithms which can identify mobile hydrogens and can interconvert tautomeric forms, which can be used to narrow down the search further for alternate tautomeric forms of a molecule in the database. Daylight does not impose restrictions on the handling of tautomers; users can choose to register tautomers individually, or to have all tautomers of a molecule be considered equivalent and only allow one preferred tautomer to be registered.

### IDBS

Using IDBS software, structures are stored in a single tautomeric form in the database which is the form entered by the user saving the structure. Before the structure is saved to the database a duplicate check is run that determines if the entered structure or a tautomeric form of it exists in the database. The user is presented with the list of tautomeric forms that exist in the database and is given the option whether to create a new registered structure, to cancel the save or to create an additional batch of the existing structure. A system wide configuration option allows the system to be implemented with or without tautomer detection in the duplicate structure check.

### InfoChem

In the routine use of InfoChem software, chemical structures are registered as they are drawn in the InfoChem database systems (the fast search engine IC*FSE* and the cartridge IC*CARTRIDGE*). Thus two different tautomers of one chemical compound are stored and registered as two entries in the database. If one of the tautomers already exists in the database and the second one is registered, no particular warning or constraint occurs. However a "tautomer search" option is also offered. This search takes only the "sigma skeleton" of a drawn query molecule into account and returns all tautomer structures. Users can choose if they want to perform a tautomer search before registering a structure.

The InfoChem search engine and cartridge also has an advanced mode (GENERIC), where generic structures can be stored and retrieved in the database. To store tautomeric structures in a database of type GENERIC the user may enter a tautomer independent structure, representing the possible tautomers, using special bond types (any bonds, single/double bonds, etc.), or may define specific rules for an automatic conversion of tautomeric structures into tautomer independent structures during the registration process, using the InfoChem structure standardization tool IC*CHECK*. A general rule set of common tautomeric definitions can be provided by InfoChem.

### OpenEye

For any molecule being registered in a database, a user of OpenEye software would carry out $pK_a$ normalization followed by generation of a unique tautomer structure. The canonical SMILES of this tautomer would be used as a database key. When a new molecule is assessed, an identical $pK_a$, resonance and tautomer normalization is carried out followed by creation of a canonical SMILES. Determination of whether the new molecule is unique in the database can now be carried out as a rapid comparison of compact strings. For a rapid database index, OpenEye recommends that the canonical tautomer is stored. If the customer prefers to use storage rather than computational expense, some or all of the enumerated tautomer forms can be associated with the index. On the other hand, if storage is at a premium, all the suitable tautomers can be generated from the single index tautomer.

### PubChem

Novelty checking in the PubChem database is carried out by taking the input structure and performing valence-bond canonicalization, i.e., structures are transformed into a standard tautomer on the connectivity level before they are registered as a unique *compound* with a compound identifier. Tautomer standardization is performed by OpenEye code. If the canonical valence bond form is in the database, the molecule is already in the database. The original input (PubChem Substance) record is kept but not shown by default: all original deposited structures are independently, and without transformation or standardization, stored in the *substance* database with simple ascending substance identifiers and can still be retrieved with their original diagram, by links between compound identifiers and substance identifiers. The original input is not searchable; only the canonical valence bond forms (PubChem Compound records) are searchable.

### Schrödinger

Schrödinger does not have a compound registration system as such, but it does have tools that can be used in what it describes as “the preparation process”. The company’s emphasis is on ligand preparation [34] not on basic informatics issues. One relevant tool, in Canvas, Schrödinger’s cheminformatics package, uses canonical (unique) SMILES to differentiate compounds. In Epik the empirical $pK_a$ prediction program [33, 40], Schrödinger strives consistently to convert the input structures from different sources, even high energy input tautomers, into a collection of low energy protonation states. Schrödinger software does a full enumeration of all energetically reasonable tautomers and ionization states using Epik and if any of the canonical SMILES patterns are already in the database then that compound can be skipped. Canvas is used for identifying duplicates, *via* canonical SMILES, after ligand preparation.

Schrödinger believes that that all energetically accessible states must be retained. Additionally, an accurate energetic penalty associated with each state should be retained in the database. All this would require some additional computational work upfront but should produce much better results in the end. The company even recommends vendors should show all possible tautomers in their compound catalogs. This is a different approach from what Schrödinger calls “the quick and dirty that people are used to”, but the company maintains that it is by no means impractical given the vast computer resources that are now available. It would only be computationally expensive to build the database, which is generally a one-time cost, and should not take more than a few days on a typical cluster. Once that is in place, additional tautomers and ionization states would need to be generated only for the query molecules, which would take only a few seconds at most.

### SciTouch

Indigo is an organic chemistry tool kit from SciTouch [32]. Within Indigo, Bingo is a data cartridge for Oracle databases: a storage and searching solution for chemical information. Bingo supports 2D and 3D exact and substructure searches, as well as similarity, tautomer, Markush, formula, molecular weight, and “flexmatch” searches. Different indexing options are available to optimize storage space requirements *versus* the speed of registration. For example, switching off the tautomer fingerprints will make the indexing faster. Indexes can be updated or re-built with no downtime for maintenance. The addition of new molecules/reactions does not require the rebuilding of the index. Bingo stores a single form of the molecule in the database; it does not record the structure diagrams of the different tautomers.

### Symyx

Symyx focuses on perceiving isomers, including tautomers, at search time to allow the greatest flexibility in database design. The software allows users to register any tautomer version they choose and it uses a search time approach to equate all tautomer forms *via* exact match (“flexmatch”). “Flexmatch” is similar to the techniques described as “sigma-skeleton” or “normalization” elsewhere in this article. Tautomer regions between target and query are identified at search time, and compared specially during atom–atom mapping. Depending on the state of various flexmatch flags, hydrogen count in the tautomer regions between query and target is required to match exactly, or within some specific tolerance defined by processing of the match switch settings. The default flexmatch option allows each tautomeric form to be registered separately; another option allows tautomeric forms to be equated, allowing only one (the first registered) form of a tautomer to be registered. In general Symyx does not impose preferred formats on users: it is up to users to determine what best fits their use cases. One form may work satisfactorily for structure databases but can cause annoyance in reaction databases when the reactive form is not the thermodynamically stable isomer (for example reactions that involve enols).

### Xemistry

The methodology in CACTVS is flexible and depends on the system [31]. The most common approach is to store the original form, but with an associated set of tautomer-unifying hashcodes (potentially with different parameters with respect to aggressiveness, handling of stereochemistry in tautomer paths, etc.).

## Canonical, dominant or preferred tautomers

Assuming that the structure of some preferred tautomeric form is to be stored in the database should the canonical tautomer be selected by some graph algorithm, without regard to energetics or what the chemist knows, or should it be the supposed major tautomer? Opinions vary, although some companies do not think that these are independent concepts. Consistent rules, consistently applied in many areas are essential for informatics and often the philosophy is taken, at OpenEye and at other companies, that as long as a consistent set of rules is needed, it is sensible to make those rules resemble as closely as possible what a chemist thinks is reality. Thus in choosing a unique tautomer from among a set, why not select a low, or the lowest energy form, and bias the consistent rules toward chemistry?

### OpenEye

OpenEye has a tautomer algorithm available in application form (QUACPAC) and also as a toolkit (QuacPac TK). It covers the primary capabilities necessary to address tautomers in cheminformatics, namely tautomer enumeration; representative, or unique, tautomer selection; and generation of a canonical representation of the unique tautomer.

Given any tautomer of a molecule, OpenEye can enumerate a small set of energetically accessible tautomers. This is a reentrant process: regardless of which tautomer is given as input to the algorithm, the same set of tautomers comes out. While it would be ideal to enumerate tautomers in energy sorted order, this task is difficult for the most expensive of computational methods and beyond what is reliably possible for cheminformatics methods. Instead OpenEye generates tautomers which are low in energy and include the lowest energy state, but do not have specific tautomer ratios assigned. OpenEye's tautomer algorithms also address enumeration and canonicalization of the valence states of resonance forms. Although these are not tautomers strictly speaking, this utility is useful for certain database applications.

The second task is to select a unique or canonical tautomer. Ideally, one would select the dominant tautomer as the unique tautomer, but since this is not strictly feasible in a timeframe suitable for cheminformatics, OpenEye applies rule-based heuristics to select the tautomer with the lowest estimated energy from among a large set of enumerated tautomers. This algorithm allows OpenEye tools to identify a unique tautomer consistently when given any one of the tautomers for a given charge state of a molecule. While the tautomer selected by this simple estimation will almost certainly not always be the lowest energy tautomer, for the purpose of cheminformatics, being absolutely consistent in the choice of tautomer, regardless of input tautomer, format or numerical instabilities is of primary importance. (As an aside, note that normalization of the formal charge of a molecule generally should be carried out before addressing tautomerism; OpenEye also has algorithms to address charge normalization.)

A final tool for tautomer handling is a means to represent the unique (or other) tautomer compactly and losslessly in a form suitable as a computational index. OpenEye uses a canonical SMILES string for this. OpenEye plans to support the InChI format in future versions of its software. OpenEye believes that the structure stored as a canonical database index in most cases should be the one that provides the most efficient implementation. It is ideal if the canonical tautomer is identical to the major tautomer, and OpenEye strives for this to be the case. Nevertheless, the major tautomer is dependent on experimental conditions, and predictions may change as algorithms improve, yet it is of high importance for canonical representations to remain as stable as possible to support back-compatibility.

### PubChem

Tautomer handling for PubChem is performed using the OpenEye QuacPac canonicalizer. The PubChem team has software built on top of this (based on OpenEye software) that scores each tautomer. The first tautomer with the highest score is kept. Due to combinatorial explosion, there are limits on how many tautomers are considered. This is a maximum of 250,000 tautomers for normal molecules, 2,500 for difficult molecules, and zero for problematic structures (those that consume massive amounts of CPU time per iteration or where the internal count of conjugated atom limits of QuacPac are reached). CACTVS tautomer hash codes are computed as a property.

### Accelrys

Accelrys' recommended solution requires all possible tautomers to be stored in the database. If the user wishes to store canonical tautomers, these can be generated using a Pipeline Pilot protocol. In Pipeline Pilot, canonical tautomers are calculated using a simple, customizable scoring function. If the users wished to generate the major tautomer, they could write a protocol or component that would calculate this based upon their own rule definitions.

To generate the canonical tautomer for an input molecule, Pipeline Pilot enumerates all of the tautomeric forms and calculates a score for each one. The canonical form is the one with the highest score. In case of ties, the top structures with the same score are sorted using their canonical SMILES to break the tie. The tautomer score is calculated by counting the number of times specific substructures are present in the molecule, multiplying by a weight factor for each substructure (positive for some, negative for others), and adding the products to obtain a total score. The default substructures considered to calculate the score include: amide, enamine, keto, enol, diamide, thioamide, aromatic bonds, and exocyclic double bonds. The canonical tautomer score calculation can be customized by the user by defining a set of substructures and their corresponding weights to use instead of the default set.

When enumerating tautomers to calculate the canonical tautomer, Pipeline Pilot provides the users with a number of options that can be used to control the enumeration process, including: "ConsiderCarbonAsDonor", "MakeAllSp2AtomsAcceptors", "Amides Tautomerization" and "Perceive Charge Tautomerization". The first of those options covers when to consider carbon atoms as donors in a tautomeric fragment. The carbon atom needs to be sp3 and have at least one hydrogen that is implicit or explicit. The second defines whether or not to consider sp2 hybridized atoms as acceptors when looking for tautomeric forms. The third is concerned with how to control tautomerism of amide groups. Here, the options include not tautomerizing any amide group, tautomerizing only amide groups in which the NH group is bonded to two or more sp2 atoms (as in R–C(=O)NC(=O)-R), and tautomerizing all amide groups. When "Perceive Charge Tautomerization" is set to *True*, the tautomer algorithm considers mobile charges in addition to mobile hydrogens. Groups such as R = (N +)(R)(R) are considered charge donors. Groups such as R-N(R)(R) are considered charge acceptors.

### ACD/Labs

Tautomer treatment in ACD/Labs tools includes recognition of different types of tautomerism; recognition that specific tautomers represent the same chemical substance; estimation of relative preference between major and minor tautomeric forms; the ability to predict properties for a specific tautomer, and for a tautomeric mixture; and the ability to choose a preferred tautomeric form according to the users' needs or criteria. The software stores structures in the database as specific (fixed) tautomeric forms. The tautomer recognition tool generates possible tautomeric forms for a given structure according to a set of rules, estimates the preference of each specific form and ranks them as major, minor, or conditions dependent.

### CACTVS

CACTVS enumerates all possible tautomers by applying encoded transformation rules to the input structure. Thereafter, an empirical scoring function is used in order to estimate roughly for each tautomer the probability of occurrence in solution relatively to the other tautomers. This empirical scoring function only takes into account structural features and properties of a tautomer. The standard version tries to balance thermochemical stability against the need to obliterate stereocenters and stereobonds in a hydrogen shift. The aspect of stereo preservations acts as a factor to boost scores of forms which do exhibit stereochemistry. This effectively results in an auto-adjusting score handling stereo compounds more carefully than basic chemicals.

Finally, all tautomers are ranked according to their scores. The highest ranked tautomer is called the "unique" tautomer, i.e., the form assumed to be one of the most prevalent in solution. By default, the unique tautomer is written to the output file. However, CACTVS can be forced to write out all possible tautomers that can be generated within the scope of the encoded rules, or a ranked subset. Furthermore, the total number of tautomers generated and the types of applied tautomeric transformations rules can be defined by the user.

Tautomer-unifying hashcodes (potentially with different parameters with respect to aggressiveness, handling of stereochemistry in tautomer paths, etc.) are computed from a standard form, which is selected to be close to the low energy form by scoring function. The standardized form can also be stored as a structure blob in a database, either parallel to the original form, or as replacement.

### ChemAxon

ChemAxon's methods (apart from the customization aspects) are based on the tautomerization plugin [41–43] which allows all tautomers, or dominant and canonical tautomers, to be generated. The dominant tautomer model includes an energy ($pK_a$) filter [43] that removes the transformations that are unlikely in solution. Tautomerism also depends on other environment factors such as phase (solid or solution), solvent, temperature, etc., but these are not considered in ChemAxon's methods. Tautomerism involves handling of dearomatization and stereochemistry, but this article will not go into such a depth of detail. The canonical tautomer, which is calculated by empirical rules, is not always the dominant tautomer, although ChemAxon tried to create the rules in a way that it tends to be. Options to customize tautomer generation include choosing dominant tautomers, with the option to set operating pH; setting the maximum distance (number of bonds) of a single proton migration; protecting structural features, such as aromaticity, charge, stereochemistry, and stable functional groups; and excluding unstable antiaromatic compounds. The filtering of stable functional groups, such as esters, to focus on practical tautomerism processes is an ongoing development at ChemAxon.

Thus far we have considered companies that have methods for scoring tautomeric forms and selecting a major form. Some companies, however, are not in favor of trying to establish a major tautomer, as the following discussion shows.

### CCG

MOE uses a tautomer enumeration process which is derived from the rules published by Selzer and co-workers [31] with some fine tuning and performance optimization. By default, the results are culled to exclude resonance-equivalent structures. Duplicates which are degenerate on the grounds of symmetry can be culled or retained; the latter is useful when applying to 3D structures such as bound ligands, which generally have lower symmetry than their molecular graph. MOE extends the tautomer enumeration to include *protomers*, by adding a set of rules for when tautomers can gain or lose protons. The addition and removal of protons is integrated into the enumeration process, so it can affect the tautomer rules which are applicable during intermediate steps [44].

Tautomer enumeration can also be useful for eliminating stereocenters. After an enumeration procedure, MOE checks for all explicit R/S or E/Z stereocenters which have been involved in a change of hybridization during the enumeration, and removes the stereochemical constraint, since it is deemed to be fluxional in solution.

CCG feels that the idea of a “canonical tautomer” has problems, with current implementations. If a registration system calculates a list of reasonable tautomers, then selects the tautomer by some arbitrary means, e.g. the canonical SMILES string with the lowest alphabetical sorting order, then the result would be a somewhat effective system for ensuring that duplicate tautomers always correspond to the same entry, but the canonical tautomer idea cannot currently be implemented properly, because methods for reliably determining which tautomer is most stable in aqueous solution have yet to be demonstrated.

### InhibOx

InhibOx believes that good practice is to store only what you know and to store it only once. An important principle is consistency. Identifying the major tautomer is not straightforward for many complex molecules, and may not even be well defined, as it depends on, for example, the solvent system in use. Therefore the best approach is probably to use a rule-based system that is unambiguous even if this means that in some cases the stored tautomeric form is not the major species in aqueous solution. However, this is achieved technically, the important point is how this impacts on searching; the ideal is not to miss hits because the user has drawn a particular tautomeric representation or to swamp the user with multiple different tautomeric representations of the same molecule.

### Daylight

Daylight recognizes that there are heuristics that can be used to identify (in most cases) the major tautomers in any given system. They are typically a combination of broad heuristics plus some specific rules to handle unusual cases. Daylight, however, usually relies on the ability to retrieve alternative tautomers efficiently rather than using a preferred tautomer for any molecule, although Daylight can do both. Some Daylight users include in their database design the canonical tautomeric form as a cross-reference field. This allows them to navigate to the additional tautomers of a molecule without having to search.

### InChI

Recognition of equivalence of specific tautomeric forms is supported by InChI procedures *via* generation and encoding of the canonical tautomer. At the same time InChI allows specific forms to be coded by inclusion of a special layer for fixed hydrogen positions. Standard InChI and the corresponding InChIKey assumed as identifiers of a substance do not include the fixed hydrogens layer and are independent of specific tautomeric form.

### Schrödinger

Schrödinger's overall approach is geared toward predicting protonation states rather than tautomers alone, since in its experience the protonation state of a ligand that binds to a protein is nearly always one of the low energy protonation states in bulk water. Schrödinger primarily uses LigPrep [45] and Epik for ligand preparation [33, 34, 40]. Canonicalization of tautomeric states has not been a primary interest: it is felt that the best type of canonicalization is to mimic as closely as possible, compound by compound, the ensemble of state(s) that the entity actually takes under physiological conditions. All of Schrödinger's canonical SMILES code is from Canvas. While the SMILES strings will be canonical within this software package, they will not be identical to what might be generated by other software packages, such as Daylight.

### SciTouch

According to the user manual [32] Bingo canonical SMILES is, according to Daylight and ChemAxon terminology, unique SMILES with isomeric information, or absolute SMILES. All significant molecular features, such as isotopes, charges, radicals, stereocenters, stereogroups, cis–trans bonds, and aromaticity, are encoded into SMILES in a canonical form. Bingo does not yet provide an option to canonicalize tautomer forms, but SciTouch has a separate product in Indigo, called Cano, which calculates canonical SMILES and (very recently) InChI codes, but without the mobile hydrogens layer.

### Others

Symyx has the ability to canonicalize structures to produce a preferred form. CambridgeSoft is not convinced that it is possible to define a "major tautomer", especially in the presence of multiple overlapping tautomeric regions. There are some cases with cyclic dependencies which could be very difficult to untangle. From CambridgeSoft's perspective, that has never been necessary to get the desired results, so it has never been pursued.

## Substructure search

Similarity search is outside the scope of this article. The approaches to substructure search are: (a) solve the problem at registration stage by storing all possible tautomers (b) apply a tautomer-aware search strategy, which may or may not be algorithmic (c) offer a combination of methods, and (d) generally ignore the problem. These approaches are dealt with in the order (a)–(d) in the text below. Method (b) is the most used.

### Enumerate all tautomers

Accelrys believes that the most appropriate way to identify molecules that could contain a sought substructure would be to store all possible tautomeric forms for each target molecule in the underlying database before performing a substructure search. This would ensure that all molecules that could tautomerize to give the sought substructure would be present in the target database before the search was initiated. CCG does substructure search on individual tautomers, so the tautomers have to be enumerated first. The company does not have a specific product which claims to do "tautomer substructure searching" (i.e., enumerating the tautomers during the actual substructure matching process) but there many ways in which CCG's tools could be composed in order to achieve this. The SMILES or 2D structure that is read into Schrödinger's general cheminformatics package Canvas is the query structure that will be used, so Schrödinger recommends expanding tautomer and ionization states in the database with Epik before using Canvas.

### Tautomer-aware search strategies

CambridgeSoft software allows for tautomer-aware substructure searching. The process must allow for partial overlap of tautomeric regions, including the proper matching of non-tautomeric regions in a query against a portion of a target structure that is tautomeric only because of other features that are not present in the query structure.

Tautomers and aromatic bonds are automatically accounted for by the SciFinder structure search algorithm in CAS Registry. Alternative tautomeric forms of the drawn structure, including the keto-enol forms, are automatically retrieved. SciFinder retrieves tautomers and aromatic systems on the basis of the single and double bonds drawn in the structure. By default in STN structure searching of REGISTRY, the structure drawing program will treat drawn tautomeric bonds as exact/normalized bonds. As a result, regardless of which form is drawn, all forms would be retrieved. However, STN also offers the user the power to override the default and set any bond(s) to have specific bond character that would allow for more precise retrieval. This should only be done with knowledge of the indexing policies of the target database(s). A unique case exists for the extremely common keto/enol tautomer situation. In this case, the structure drawing program in STN leaves the bond character specifically as drawn, by default, because usually only one form is desired.

In order to search for the possible tautomers of a given structure, CCDC's program ConQuest offers the possibility of introducing variable bonds and attachment points in the search so that tautomers for that given structure can be easily retrieved. An example of a search strategy which will successfully retrieve tautomeric structures is provided on the CCDC website [46]. The crystal structure will contain a specific tautomer and the safe procedure is to formulate the principal tautomers and search for them all. Searching the CSD for all possible tautomers is more complicated and requires advanced searching techniques and the support of additional software [47].

In CACTVS, bonds which are part of a tautomeric system are flagged (as tautomeric system identifiers, precomputed) in the database blobs. In tautomeric search modes, hydrogens which are themselves potentially mobile in a query structure are either marked or automatically perceived and can match at a different location in the database structure from the core fragment part, as long as it is still in the same tautomeric system. Bond orders in matching tautomeric fragments are ignored, but the sum of matched bond orders must agree with a (default) maximum deviation of 1 between fragment and database structure. For tautomeric substructure searches without explicit hydrogens, tautomeric bonds in the query must be explicitly marked.

ChemAxon's JChem software enumerates the query tautomers and searches each tautomer individually. During enumeration, explicit H atoms are migrated on the graph to make sure that the query feature is considered in the tautomers correctly. Tautomerism will only be recognized if the query contains the full tautomerizable region. The canonical tautomer generation algorithm requires a full molecule to properly consider energetics and the local structural environment of tautomerizable functional groups. It is therefore not ideal for substructure search.

Daylight typically uses more general search strategies to find tautomers rather than identifying specific tautomeric groups. Searching by common graphs plus additional charge/hydrogen information is a very efficient way to find potential tautomers. It is also possible to enumerate tautomeric queries and perform the searches, although this can get slow for complex systems. Substructure queries become difficult because a query may not contain sufficient information to indicate that the query can tautomerize, and if a system relies on heuristics for registration, there may be a mismatch between the preferred tautomer of the query and the preferred tautomer of full molecules stored in the database. Daylight suggests that this means that all systems need to be able to search multiple tautomers and not rely on storing preferred tautomers.

InfoChem provides a "Tautomer Exact Structure Search" (TXSS) option which finds all isomers, where hydrogens, radicals or metal atoms may be on variable positions, but where the (sigma)-skeleton and the total number of hydrogen atoms must be identical after removing all bonds to metal atoms, all charges and all radical electrons. Stereochemistry will be ignored during the search; isotopic labels will not be ignored. Tautomeric search can also be performed within a substructure search by using appropriate query features (MDL/Symyx query feature standards). More details of how to specify tautomer searches can be found in the InfoChem search tutorial [48].

For PubChem, identity and substructure searching performs valence bond canonicalization on the user query structure. Substructure searching can be performed with identification of tautomeric variations. PubChem Structure Search has a non-default sub- or super-structure option to "match tautomers". (The PubChem substructure and full-structure search system are based on CACTVS, so the explanations above about the CACTVS match system also apply to PubChem.)

SciTouch Indigo has an algorithm for tautomer substructure matching, and special fingerprints for screening large databases in tautomer substructure search. Unfortunately the algorithms are not yet documented, but the source code is available.

Symyx does not perceive tautomer regions for substructure queries. By definition, a substructure query is a substructure of a larger (unspecified) piece. Symyx does not perceive or alter a query to "broaden" it to allow hits on the tautomeric chemical space for that scaffold, but the software does provide ample query features that allow a user to be very specific, or very general in the type of hits returned. Symyx also provides tools (in its Cheshire software) that can be used to "tautomerize" a substructure query automatically, and it plans to expose more explicitly tautomer regions within a structure.

## Two approaches

ACD/Labs agrees that the storage of all possible tautomers for any entry, or the ability to store unspecified tautomeric forms, may be a more general solution for substructure searching. Alternatively if the ACD/Labs search program recognizes that a substructure may have several tautomeric forms, the user is allowed to choose to find either a specific tautomer, or all forms generated by the rule-based procedure. The constraint of this method is that if the substructure does not include the full tautomeric unit, all tautomers may not be found.

OpenEye addresses the problem by either enumeration of the queries or by enumerating the database structures, though the former has many subtle complexities. Enumeration of database molecules can be carried out beforehand or on-the-fly as part of the search. In either case, the variable and fixed regions of the molecule can be noted in order significantly to reduce redundant searching of the fixed portions of the molecules. These capabilities could be implemented with the QuacPac and OEChem toolkits.

### Other

At present, ChemSpider does not consider tautomeric variations when doing substructure searching. IDBS substructure searches also do not take into account tautomeric forms of a structure. CWM Global Search converts a query into three different InChI strings (because "standard InChI" differs from others) and into SMILES and molfiles. Two InChIKeys are also generated: the key with connectivity alone and the key allowing for stereochemistry etc. The query is then submitted to multiple databases on the Internet. Substructure search is possible (but not using InChI). The query should be formulated to allow for all possible tautomers because multiple tautomeric forms will be available in the various databases searched. CWM Global search does not currently generate different InChI keys for tautomers.

## Hit display

The most common approach is to display the originally input, registered structure. Other approaches are to show the standard form, but perhaps only as an option; or to display the matched tautomer, as an option; or to give an even greater degree of flexibility.

With OpenEye software, the structure stored as a canonical database index can be independent from the tautomer shown to users by OpenEye. Most OpenEye users prefer to see what they expect is the major tautomer. There is no right or wrong answer: it depends upon the subjective opinion of the user as well as the specific application. If the user wants to highlight the substructure match within the molecule being displayed, then the matched tautomer would be most appropriate. On the other hand, if the system is displaying experimental results that are associated with another specific tautomer, then perhaps the registered tautomer or both tautomers are appropriate for display. With OpenEye's tautomer and cheminformatics toolkits, users can implement a customized solution.

In a database built with Accelrys's software all possible tautomers can be stored alongside the original input structure. After a search the system can either return the original input structure (the registered tautomer) or the actual tautomeric match, depending on what the user desires.

In CACTVS the default display form is normally the one the query was performed on; in the case of parallel storage of deposited and standardized forms, this is often user-selectable. In display highlighting, the hydrogen actually matched on the database structure is indicated, i.e., the connectivity actually highlighted can be different from the query structure, and potentially even become disconnected when a mobile hydrogen matches at a part of the database structure tautomeric system which is not directly connected to the match region of the core query fragment.

ACD/Labs software displays the registered tautomer. CambridgeSoft software shows the registered structure in all cases. It also features hit highlighting to show how the query is mapped to the target. The display issue is not restricted to tautomers: if the user searches for acetic acid and finds sodium acetate, CambridgeSoft also shows the structure as registered, without suppressing the counterion from the display. In InfoChem's system, in a hit list received after a search, all tautomer structures appear as separate entries, provided they match the query (assuming duplicates have been registered). Bingo shows the registered tautomer, and the tautomer chains that transform it into a matching one are highlighted. Symyx always shows the explicit registered form. ChemAxon usually displays the originally inserted form, even if a canonical form was stored in the database, but it is also possible to show the standardized form.

The default structure search result display form in PubChem is the standardized form, with links to the original depositions.

## Selected commercial databases

Many database vendors are using ChemAxon software. Examples are SureChem (patents); Aureus Pharma; Thomson Reuters (ChemAxon is the sole cheminformatics platform for Prous Integrity); and Informa, Taylor and Francis and CRC Press (the online system and CD-ROM for the *Handbook of Chemistry and Physics*, and several other titles). WOMBAT and Eidogen-Sertanty's Kinase Knowledgebase use ChemAxon's Instant JChem for distribution. Many non-commercial databases such as Binding DB, Ligand Depot, and DrugBank also use ChemAxon software. InhibOx uses ChemAxon tools for registration.

Database vendors who are using either InfoChem's fast search engine or the chemistry cartridge IC*CARTRIDGE* include Thieme (in Science of Synthesis and Pharmaceutical Substances), Dialog (in DialogLink 5), and some companies with online catalogs. For electronic Encyclopedia of Organic Reagents, Wiley is still using InfoChem's older search engine (ICSE) which does not support tautomer searching.

In the process used by Thomson Reuters for its internal specific compounds database Derwent Chemistry Resource

(DCR) registration is not totally automatic. Analysts indexing a document will perform an initial search based on names and/or SMILES comparison. If the analysts then believe this to be a new structure, it is forwarded to a registration group which checks using Symyx ISIS structure searching. The registrar would be expected to check for alternative tautomers, where these may not be retrieved by an appropriate structure or similarity search.

DCR is used for several customer-facing databases. Where the customer-facing database is used in-house or online on STN, Thomson Reuters tends to use mainly standard conventions either enforced by the software or by the multi-file environment. There is also software that converts the structures into the "correct" tautomers for inclusion into other databases, such as Merged Markush Service (MMS) or for the generation of fragment codes in Derwent World Patents Index (DWPI). MMS includes both generic and specific chemical structures, including chemical structures from DCR. So, the same structure can look different in different file environments, e.g., in DCR on STN and in MMS on Questel. Markush structure searching such as Markush DARC [49] is beyond the scope of this article.

Reactions databases built by Thomson Reuters tend to use the conventions specified by Symyx, to facilitate multi-file searching. The company's drug databases usually stick to the form published by the innovator, which means that the registration team has to check carefully for tautomers, stereoisomers etc. and decide if these are "different" enough to create a separate entry.

For patents, internal software is used to convert a structure into the correct tautomer for the database. Derwent Markush TOPFRAG is a tool for searching the chemical structures and structure information found in Derwent's online databases. It automatically generates fragment search strategies for structure information in DWPI and for chemical structures in MMS. Markush TOPFRAG software contains tautomer recognition and conversion routines to help the searcher. Key elements are training and user aids to help users know what type of search (e.g., relaxed, similarity, substructure etc. in Symyx ISIS) to use or how best to draw a structure for the database being used (e.g., DCR or MMS).

Index Chemicus and Current Chemical Reactions both appear on Web of Science. For these databases, Thomson Reuters uses Oracle, Symyx Isentris and Symyx Draw. Before registering compounds the "Flexmatch = All" option is used to check for duplicates, including tautomers. If the molecule is already in the database, and a tautomer is later drawn, it will be flagged as a duplicate. When deciding which tautomer form to index, Thomson Reuters follows what is presented in the paper being indexed from the literature. When there is insufficient information, the registrars have editorial guidelines on what to draw, for example, aromatic takes preference over keto over enol. There are always exceptions. Well known, established structures are not changed to comply. When an author does indicate the prevalent form, then Thomson Reuters indexes it. An author may report that steric hindrance favors one form, so it is not changed. For the reaction database, Thomson Reuters follows the author's depiction, and highlights and maps the reaction as described by the author.

Web of Science uses Hampden Data Services' software [50] for structure searches; the "bond type = unspecified" option will retrieve tautomers. The search results will display exactly what was indexed originally using Symyx software. The production system uses Symyx software for all of the indexing, registration, processing and product creation. The data are processed each week to create a Hampden Data Services format database for Web of Science. For users who take the data as in-house databases, Thomson Reuters provides Symyx (REACCS, ISIS or Isentris) databases. Web of Science is mainly delivered over the Internet, but there are some intranet users.

## Conclusion

Throughout this article the differences between computational chemistry companies and informatics companies are very clear. As a generalization it might be said that the computational chemistry wants to study the shape and electrostatics of a molecule to see if it will fit in a receptor, whereas the informatician wants to check that it will *not* fit into a competitor's patent. The informatician is driven by financial decisions concerning novelty, unnecessary sample purchasing, and expensive repetition of synthetic work. The computational chemist also wants to make sure that the right compounds are made or purchased, but uses a more "rigorous" approach. Informaticians can supply pages of structure diagrams to prove precisely which tautomers their software can handle; computational chemists think in terms of rules and general algorithms, protomers, and ligand preparation.

The terminology involved is not so clearly divided between the two disciplines: a "canonical", "dominant", "preferred" or "unique" form means different things to different people. Some people maintain that if a "unique" tautomer is to be stored then it should be as near as possible to the real-life form; others claim that this is pointless since "reality" depends on many factors and the real form cannot be determined in reasonable CPU time. Informatics companies tend to be interested only in the pragmatic concept of "novelty" when registering a compound; they start by applying the concept of a "normalized" structure or basic skeleton. Some computational chemists, on the other hand, feel that novelty is best determined by generating and storing all possible tautomers.

In practice a registry system uses techniques such as hashcoding and indexing to make exact match searches faster. These technologies are beyond the scope of this article although a certain amount has been said about canonical SMILES representations (which are proprietary and not unique) and InChI which is open and unique (at least in its standard form). Systems that store all possible tautomers allow a chemist safely to input any desired query substructure; allowing for tautomers in a substructure query can be done algorithmically but is usually done, in the case of chemical registries, by allowing alternative multiple bond types, and training chemists how to design a query.

The most common method for the depiction of a hit from a substructure search is to display the structure that was originally registered. An alternative is to show the tautomer that matched the query. The latter is probably the more meaningful if the atoms and bonds in the query structure are highlighted in the hit structure. Displaying the "unique" tautomer is a less common choice. Some systems give the user options for which tautomer to display.

Above all, it is important to take account of the nature of the database and the process for which it is designed. A system that predicts NMR spectra, for example, cannot work on the basis of just one tautomer. A system that deals with chemical reactions may need to allow for unusual tautomers that are specific to the reaction mechanism. A system that is mainly concerned with ligand-receptor docking needs a highly refined method for preparing the ligand. Systems that are designed for public use over the Internet may have little control over what is registered into the database(s) they search. The variety of needs means that it is dangerous to lay down the law about what is right and wrong in the approaches to tautomers. This discussion document seeks only to outline all the options.

**Acknowledgments** I am indebted to the many people who supplied data for this article, and in particular I would like to thank Jonathan Brecher of CambridgeSoft and Geoff Skillman of OpenEye for helpful discussions, Russ Hillard and Keith Taylor of Symyx for drafting two figures, and Thomson Reuters for some careful proof reading.

## References

1. Warr WA (1999). Balancing the needs of the recruiters and the aims of the educators. Book of Abstracts, 218th ACS National Meeting, New Orleans, Aug. 22-26, 1999: COMP
2. Warr WA Extract from 218th ACS National Meeting and Exposition New Orleans, Louisiana, August 22-26, 1999. http://warr.com/warrzone2000.html. Accessed 28 Feb 2010
3. Morgan HL (1965) J Chem Doc 5(2):107
4. Dittmar PG, Stobaugh RE, Watson CE (1976) J Chem Inf Comput Sci 16(2):111
5. Freeland RG, Funk SA, O'Korn LJ, Wilson GA (1979) J Chem Inf Comput Sci 19(2):94
6. Blackwood JE, Elliott PM, Stobaugh RE, Watson CE (1977) J Chem Inf Comput Sci 17(1):3
7. Vander Stouw GG, Gustafson C, Rule JD, Watson CE (1976) J Chem Inf Comput Sci 16(4):213
8. Zamora A, Dayton DL (1976) J Chem Inf Comput Sci 16(4):219
9. Stobaugh RE (1980) J Chem Inf Comput Sci 20(2):76
10. Mockus J, Stobaugh RE (1980) J Chem Inf Comput Sci 20(1):18
11. Moosemiller JP, Ryan AW, Stobaugh RE (1980) J Chem Inf Comput Sci 20(2):83
12. Ryan AW, Stobaugh RE (1982) J Chem Inf Comput Sci 22(1):22
13. Hamill KA, Nelson RD, Vander Stouw GG, Stobaugh RE (1988) J Chem Inf Comput Sci 28(4):175
14. Stobaugh RE (1988) J Chem Inf Comput Sci 28(4):180
15. Blackwood JE, Blower PE Jr, Layten SW, Lillie DH, Lipkus AH, Peer JP, Qian C, Staggenborg LM, Watson CE (1991) J Chem Inf Comput Sci 31(2):204
16. CTfile formats. http://www.symyx.com/downloads/public/ctfile/ctfile.jsp. Accessed 28 Feb 2010
17. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J (1992) J Chem Inf Comput Sci 32(3):244
18. The IUPAC International Chemical Identifier (InChI). http://www.iupac.org/inchi/. Accessed 28 Feb 2010
19. Weininger D (1988) J Chem Inf Comput Sci 28(1):31
20. Weininger D, Weininger A, Weininger JL (1989) J Chem Inf Comput Sci 29(2):97
21. Wipke WT, Dyott TM (1974) J Am Chem Soc 96(15):4834
22. NEMA key based exact match searching. http://www.symyx.com/solutions/white_papers/nema_whitepaper.jsp. Accessed 28 Feb 2010
23. Wipke WT, Krishnan S, Ouchi GI (1978) J Chem Inf Comput Sci 18(1):32
24. Heller SR, McNaught AD (2009) Chem Int 31(1):7
25. InChI Technical Manual. http://www.iupac.org/inchi/download/index.html. Accessed 28 Feb 2010 (The file InChI_TechMan is included in stdinchi-1-doc.zip, inside the folder DOC-101)
26. Secure hash standard. http://csrc.nist.gov/publications/fips/fips180-2/fips180-2withchangenotice.pdf. Accessed 28 Feb 2010
27. The National Cancer Institute Computer-Aided Drug Design Group Chemical Structure Lookup Service (CSLS). http://cactus.nci.nih.gov/lookup. Accessed 28 Feb 2010
28. Sitzmann M, Ihlenfeldt W-D, Nicklaus MC (2010) J Comput-Aided Mol Des. doi:10.1007/s10822-010-9346-4
29. Anon (2007) Naming and indexing of chemical substances for Chemical Abstracts. Chemical Abstracts Service, Columbus
30. IUPAC Compendium of Chemical Terminology—the Gold Book. http://goldbook.iupac.org/index.html. Accessed 28 Feb 2010
31. Oellien F, Cramer J, Beyer C, Ihlenfeldt W-D, Selzer PM (2006) J Chem Inf Model 46(6):2342
32. Indigo, an organic chemistry tool kit from SciTouch. http://opensource.scitouch.net/indigo/bingo, http://opensource.scitouch.net/indigo/bingo/user_manual#tautomer_search, http://opensource.scitouch.net/downloads/bingo_user_manual.pdf. Accessed 28 Feb 2010
33. Epik. http://www.schrodinger.com/products/14/4/. Accessed 28 Feb 2010
34. Greenwood JR, Calkins D, Sullivan AP, Shelley JC (2010) J Comput-Aided Mol Des. doi:10.1007/s10822-010-9349-1
35. ChemAxon Standardizer. http://www.chemaxon.com/products/standardizer/. Accessed 28 Feb 2010
36. Trepalin SV, Skorenko AV, Balakin KV, Nasonov AF, Lang SA, Ivashchenko AA, Savchuk NP (2003) J Chem Inf Comput Sci 43(3):852

37. Sayle RA, Delany JJ (1999) Canonicalization and Enumeration of Tautomers. Paper presented at EuroMUG99, Cambridge, UK, 28–29 Oct 1999
38. Mercury. http://www.ccdc.cam.ac.uk/free_services/mercury/. Accessed 28 Feb 2010
39. Pluto. http://www.ccdc.cam.ac.uk/support/documentation/#rpluto. Accessed 28 Feb 2010
40. Shelley JC, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M (2007) J Comput-Aided Mol Des 21(12):681
41. ChemAxon tautomerization plugin. http://www.chemaxon.com/jchem/marvin/help/calculations/isomers.html#tautomer. Accessed 28 Feb 2010
42. ChemAxon tautomer duplicate filtering search option. http://www.chemaxon.com/jchem/doc/user/query_searchoptions.html#tautomer_duplicate. Accessed 28 Feb 2010
43. Tautomer generation. pKa based dominance conditions for generating dominant tautomers. http://www.chemaxon.com/conf/Tautomer_generation_A4.pdf. Accessed 28 Feb 2010
44. Clark AM, Labute P. SD File Processing with MOE Pipeline Tools. http://www.chemcomp.com/journal/sdtools.htm. Accessed 28 Feb 2010
45. LigPrep. http://www.schrodinger.com/products/14/10/. Accessed 28 Feb 2010
46. CCDC search example. http://www.ccdc.cam.ac.uk/products/csd_system/conquest/faqs/scientific_faq.php#tautomers. Accessed 28 Feb 2010
47. Cruz-Cabeza AJ, Schreyer A, Pitt WR (2010) J Comput-Aided Mol Des. doi:10.1007/s10822-010-9345-5
48. InfoChem search tutorial. http://www.infochem.de/content/downloads/icfsetutorial.pdf. Accessed 28 Feb 2010
49. Markush DARC. http://www.questel.com/customersupport/userdoc/docpdf/Thomson_Markush_Darc_User_Manual.pdf. Accessed 28 Feb 2010
50. Town WG (1989) In: Warr WA (ed) Chemical structure information systems. ACS symposium series, vol 400. American Chemical Society, Washington, p 68