getting from the conventional name to the Fischer projection and is, therefore, a most valuable and much needed teaching aid for the student and time saver for the chemist and biochemist.

Finally, this system allows for an easy and direct method of representing sugar structure in a digital computer, suggesting possibilities for more sophisticated and elegant analyses of carbohydrates than have been done to date.

And so, since sugar (in one form or another) is by far one of the most abundant compounds in nature, it is one of nature's biggest blessings to mankind. Ergo, we may now literally "count our blessings" and draw the one we need, rather than drawing them all each time.

ACKNOWLEDGMENT

I wish to express my gratitude to Professor Henry Yuska, Chairman of the Department of Chemistry at Brooklyn College, Brooklyn, New York, for his help, understanding, guidance, and encouragement; and to Professor Henry G. Mautner, Chairman of the Departments of Biochemistry and Pharmacology at Tufts University School of Medicine, Boston, Massachusetts, for his valuable advice on the preparation of this paper.

REFERENCES AND NOTES

(1) S. Neelakantan, *Curr. Sci.*, **38**, 353 (1969).
(2) L. F. Fieser and M. Fieser, "Organic Chemistry", 3rd ed, Reinhold, New York, 1959, p 359.
(3) M. L. James, G. M. Smith, and J. C. Wolford, "Applied Numerical Methods for Digital Computation with FORTRAN", International Textbook Co., Scranton, Pa., Oct 1968, p 4.
(4) "McGraw-Hill: Encyclopedia of Science and Technology", Vol. 9, McGraw-Hill, New York, 1971, p 258.
(5) A. L. Lehninger, "Biochemistry", 2nd ed, Worth Publishers, New York, 1975, p 266.

ADDITIONAL REFERENCES

Feldman, A., *J. Org. Chem.*, **24**, 1556 (1959).
Morrison, R. T., and Boyd, R. N., "Organic Chemistry", 3rd ed, Allyn and Bacon, Boston, 1973.
Neelakantan, S., *Curr. Sci.*, **39**, 85 (1970).
Neelakantan, S., *Indian J. Chem. Educ.*, **2**, 15 (1971).
Ore, O., "Number Theory and Its History", McGraw-Hill, New York, 1948.
Pigman, W., Horton, D., and Herp, A., "The Carbohydrates", Vol. II-B, 2nd ed, Academic Press, New York and London, 1970.
Rosenblatt, D. H., *J. Chem. Educ.*, **42**, 271 (1965).

The Chemical Abstracts Service Chemical Registry System. VII. Tautomerism and Alternating Bonds

J. MOCKUS* and R. E. STOBAUGH

Chemical Abstracts Service, Columbus, Ohio 43210

Received December 3, 1979

The Chemical Abstracts Service (CAS) Chemical Registry System is a computer-based information system that uniquely identifies chemical substances on the basis of their molecular structure. Substances that have several possible chemically equivalent representations are difficult to portray precisely by a single structure diagram or connection table. Among the major causes of this problem are aromatic rings, whose alternating single and double bonds can be represented in more than one way, and tautomerism, an equilibrium involving single/double bond shifts coupled with hydrogen migration. The CAS Chemical Registry System handles the problem by algorithmically recognizing tautomeric and alternating bond structures, replacing the explicit single and double bonds with special normalized bonds, and associating the migrating tautomeric hydrogen with groups of atoms rather than just single atoms. This article describes the normalization techniques used in handling alternating bonds and tautomeric bonds, as well as substructure search aspects involving these bond types, and denormalization procedures required for algorithmic structure display and name generation.

INTRODUCTION

The Chemical Abstracts Service (CAS) Chemical Registry System is a computer-based system for the unique identification of chemical substances on the basis of structure.[1] The initial, experimental system, Registry I, began operation in 1964 and established the viability and validity of the registration concept for fully defined organic substances. In 1968, the scope of the system was increased as it began to handle additional classes of substances. The system, now known as Registry II, began to be integrated into the CAS indexing operation. In 1974, the most recent version, Registry III, made major adjustments in the Registry structure records to provide increased support to the process of generating names for the *Chemical Abstracts* (CA) Chemical Substance Index, and also to computer-based structure output operations through explicit identification of the ring systems present in a substance. As its use has expanded, the CAS Chemical Registry System has proven to be reliable and consistent as a structure identification method and has become an essential CAS production tool supporting CA index input and compilation. It has also found widespread interest and support in the scientific and technical community.

The foundation of the CAS Chemical Registry System is an algorithm that generates a unique and unambiguous machine-readable description of the molecular structure of a substance. The principal component of the machine record is a connection table, a detailed description of the atoms and bonds that comprise the basic structure of the substance. Other components describe stereochemical characteristics, isotopic labeling, and derivatives (salts, hydrates, etc.).

The representation of a chemical substance by a unique structure diagram or connection table poses problems to both chemists and chemical information systems when the substance has several possible representations, chemically equivalent but structurally distinct. Resonant or aromatic bonds which have characteristics of both single and double bonds are one major

0095-2338/80/1620-0018$01.00/0

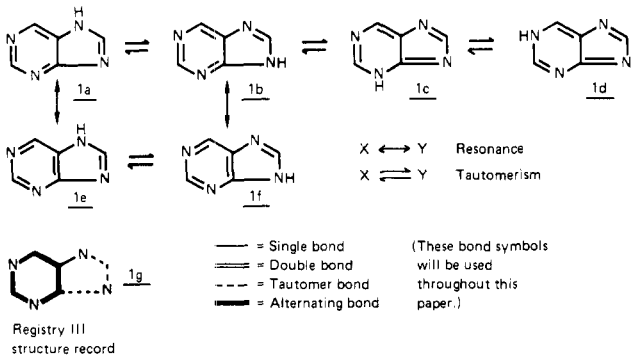

Figure 1. Purine representations.

cause of problems, and tautomerism, an equilibrium involving single/double bond shifts coupled with hydrogen migration, is the other (see Figure 1). These phenomena are quite common; about 70% of the structures in the CAS Registry Master File are aromatic, possessing rings containing alternating single and double bonds, and about 25% exhibit tautomerism.

To a chemist, the multiple representations of a substance resulting from tautomerism and alternating or aromatic bonds usually pose only minor problems. Their equivalence is recognized with little effort, as the result of chemical training and experience.

To a chemical information system based on structural diagrams (or their machine equivalents, connection tables), tautomerism and alternating bonds hinder the representation of a single chemical substance by a single diagram. The CAS Chemical Registry System handles the problem by normalizing (i.e., recognizing the equivalence of) tautomeric and alternating bond structures, replacing the explicit single and double bonds with special tautomer and alternating bonds, and associating the migrating hydrogen in a tautomer with a group of atoms rather than just a single atom. Since single/double bond patterns and specific migrating group locations have been replaced by normalized data, all forms of the tautomeric structure lead to the same Registry III structure record. Thus, the six possible representations of purine shown in Figure 1, each containing six single and four double bonds in differing arrangements, all lead to the same Registry III connection table containing one single bond, three tautomer bonds, and six alternating bonds (see structure **1g**).

OVERVIEW OF ALTERNATING BONDS AND TAUTOMERISM

Alternating bonds are a compromise approach to the chemist's concept of aromatic bonds, bonds which have characteristics of both single and double bonds. The chemist represents such bonds with circles or dotted lines, or with alternating single and double bonds. In the latter case, it is implicitly understood that the actual arrangement of single and double bonds is not critical as long as they alternate. In the purine example (see Figure 1), structures **1a** and **1e** would be considered simply as different representations of an alternating bond situation, as would **1b** and **1f**.

Tautomerism is a state of equilibrium of two or more molecular structures that differ in the location of a mobile group, usually a hydrogen atom. Bonding changes occur at the same time as migration of the mobile group. In the purine example, the **1a–1b–1c–1d** and **1e–1f** equilibria are due to tautomerism. Tautomer and alternating bond situations may overlap, as shown by the purine example. The implications of this overlap will be discussed later.

The CAS Registry III System definition of tautomerism also allows migrating positive and negative charges, even though these would be considered more properly cases of resonance

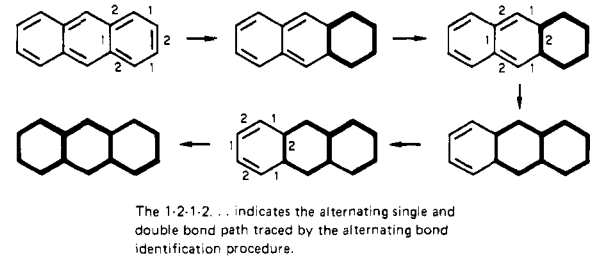

Figure 2. Identification of alternating bonds.

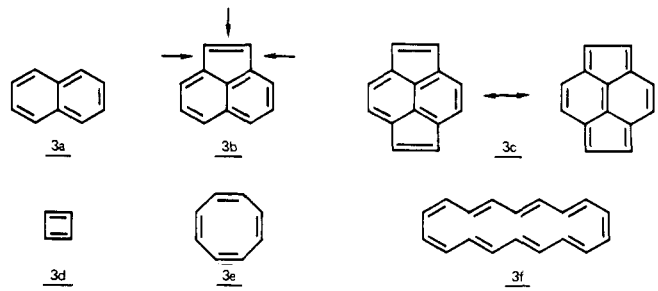

Figure 3. Alternating bond examples.

or charge delocalization. Only simple cases, such as the carboxylate or phosphate anions, are handled this way. More complex cases, such as the cycloheptatrienyl (tropylium) cation or the cyclopentadienyl anion, are handled with special delocalized charge procedures.

Alternating Bonds. The CAS Registry III System uses a path-tracing procedure to identify alternating bonds. It searches through a structure, backtracking when necessary, to find cyclic paths in which the bonds are alternately single and double, marking the bonds as alternating in such paths as they are found (see Figure 2). The search is exhaustive, tracing all possible paths.

The alternating bond procedure will accept a normalized bond in place of an explicit single or double bond on the assumption that the normalized bond could assume the required value. This approach is necessary for proper treatment of overlapping tautomers and alternating bonds, as will be discussed later. In addition, it speeds up the identification process by reducing the number and size of the paths that are traced.

At this point, it would be appropriate to consider the relationship between the chemist's aromatic bonds and the CAS Chemical Registry System's alternating bonds. Aromaticity is still an unsettled topic; for example, chemists would probably agree that the bonds in **3a–3c** (see Figure 3) were aromatic and those in **3d** and **3e** were not, and would argue about **3f**. The CAS Chemical Registry System procedures would find all bonds to be alternating except those emphasized in **3b**, the sole criterion being the alternating single/double cyclic path.

These structures illustrate two key points. First, the substructure searcher must be concerned with the CAS Chemical Registry System's alternating bond concept and must look at a chemist's aromatic substructure from this viewpoint. Secondly, the searcher must always consider the environment of the substructure. For example, the three emphasized bonds in **3b** are fixed single and double when the system is isolated, but are alternating bonds when the substructure is embedded in a larger system (**3c**).

Tautomerism. Tautomerism is a state of equilibrium in which a mobile group, typically a hydrogen, migrates between atoms with concurrent changes in bonding.[2] The basic generic tautomeric structure is

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

where M and Z are endpoints and Q is a centerpoint. In the

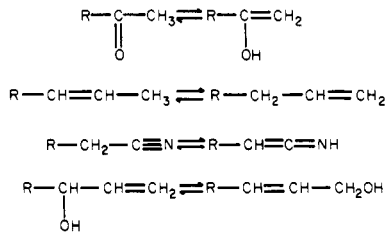

Figure 4. Tautomers not recognized by the CAS Registry III System.

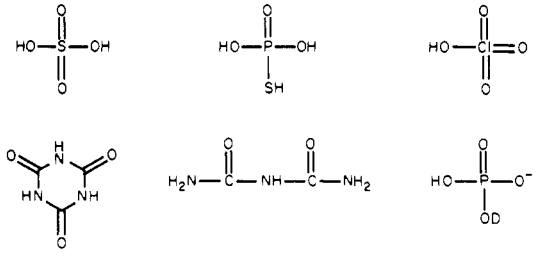

Figure 5. Tautomer examples.

CAS Registry III System, Q can be carbon or most any nonmetallic element, but M and Z are limited to nitrogen and chalcogen atoms. Olefinic and keto–enol tautomers are not recognized as such and are registered (and named by CAS) as distinct structures, as are other variations (see Figure 4).

This definition follows general chemical practice. The normalized tautomers are (1) substances whose names usually describe the structure as a whole, such as the trivial name Barbituric acid (cross-referred to the CA preferred name 2,4,6(1*H*,3*H*,5*H*)-Pyrimidinetrione); (2) structures where tautomerism affects only minor details of the name, such as substituent locants; or (3) functional groups where one form is invariably selected over the alternative, such as an amide over an imidic acid. Tautomers that are not normalized, such as keto–enol tautomers (the most common tautomeric system), are those whose alternative forms usually receive distinct names. Acetone, for example, is the trivial name of 2-Propanone but not its tautomer 1-Propen-2-ol, CH_2=C-(OH)CH_3.

Using the generic HM—Q=Z tautomeric structure, the CAS Registry III System requires that

- the centerpoint Q may be C, N, P, As, Sb, S, Se, Te, Cl, Br, or I, with any acceptable valence;
- the endpoints M and Z may be trivalent N or bivalent chalcogen (O, S, Se, or Te) in any combination;
- the centerpoint–endpoint tautomer bonds may be either cyclic or acyclic, or both types in combination;
- the mobile group H may be hydrogen (H or its isotopes D or T) or a −1 charge.

Tautomers handled by the CAS Registry III System are not limited to the basic three-atom HM—Q=Z substructure. Larger tautomers may be linear, as in HM—$(Q{=}N)_n$—Q=Z, where "N" is trivalent nitrogen, or cyclic, or branched. A centerpoint may have more than two attached endpoints, as in HM–$Q(=Z)_n$ or Z=$Q(-MH)_n$. Finally, the mobile groups in a tautomer may be all alike or may be different types in combination. (Examples are shown in Figure 5.) The description of a tautomer group in a Registry III structure record cites all of the endpoints in the group and the number of each type of mobile group that is associated with the endpoints. Centerpoints are not explicitly identified, but can be found readily by a check of the structure record connection table.

Tautomers are identified by a procedure that searches for potential endpoints, i.e., nitrogen or chalcogen atoms, that are doubly bonded to an atom acceptable as a centerpoint. When such a two-atom set is found, 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 (see Figure 6). As with alternating bond identification, a previously normalized bond may be used wherever a single or double bond is required. In addition, previously identified nitrogen endpoints may be reused to expand a tautomer group by the addition of endpoints attached to a new centerpoint (see Figure 7).

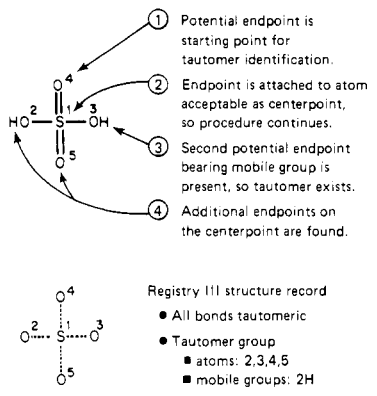

Figure 6. Tautomer identification.

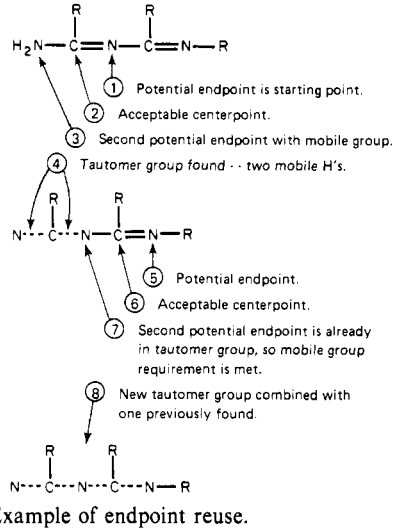

Figure 7. Example of endpoint reuse.

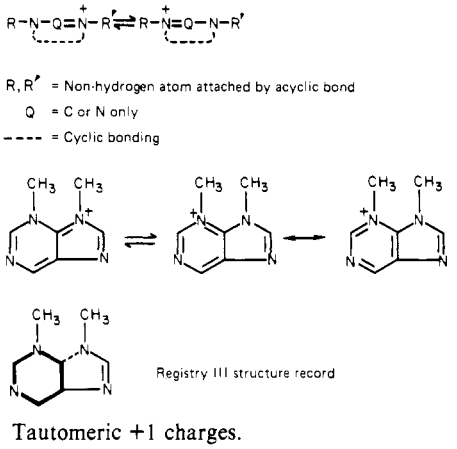

Figure 8. Tautomeric +1 charges.

The CAS Registry III System also recognizes a variety of tautomers in which the mobile group is a +1 charge (see Figure 8). This addition to the general tautomer definition was provided so that an "onium" substructure common in dyes could be normalized.

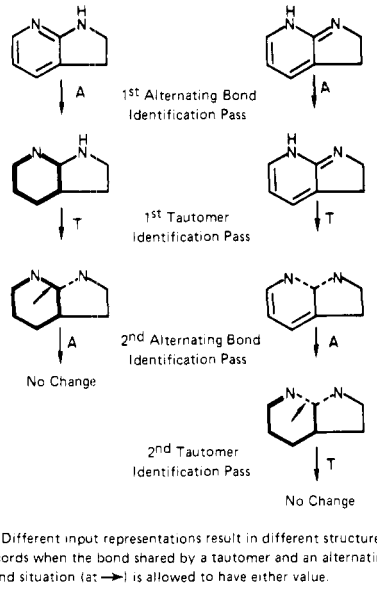

Figure 9. Multiple bond representations problem.

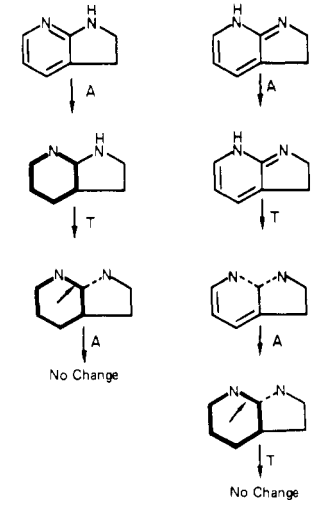

Figure 10. Alternating and tautomer bond identification.

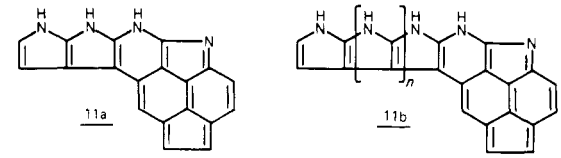

Figure 11. Structures requiring several normalization passes.

OVERLAP OF TAUTOMERS AND ALTERNATING BONDS

Tautomers and alternating bonds may overlap, as shown by the purine example (Figure 1). This possibility required careful consideration during the design of the CAS Registry III System so that multiple representations of a single structure would not be created by the normalization procedures.

Two distinct bond values for cyclic tautomer and alternating bonds are used in the connection table of the CAS Chemical Registry System structure record (in addition to the usual single, double, and triple bond values). Bonds that could be regarded as either tautomer or alternating bonds because of overlap (i.e., could be assigned either bond value) are arbitrarily classified as alternating bonds. If this were not done, the assigned bond values would depend on the bonding in the structure as input to the CAS Chemical Registry System (which could affect whether the bond was first recognized as tautomer or alternating), and this would lead to multiple representations of a single substance (see Figure 9).

Since tautomer and alternating bond situations may overlap, the two normalization procedures must be able to reuse previously normalized bonds of either type. The tautomer procedure, for example, must be able to use either tautomer or alternating bonds during its search for tautomers, not just tautomer bonds alone. If this were not done, the multiple representation problem would be quite severe.

Since tautomer bonds may be used in the identification of alternating bonds, and vice versa, the normalization procedures may need to be applied more than once to a given structure. After the identification of tautomers, for example, a new alternating bond path involving one of the just-normalized tautomer bonds might exist. If the normalization procedures were applied only once, some bonds might not get normalized, and multiple representations would result. In the CAS Registry III System, the alternating bond and tautomer procedures are applied alternately, in that order, until both procedures have been applied at least once and the last-used procedure has not found anything new to normalize (see Figure 10). (The alternating bond procedure is applied first simply because it is more likely to find bonds to be normalized.) Some structures require more than one or two passes. Structure **11a** (in Figure 11) requires four passes before all the bonds are normalized, the process ending after the fourth tautomer pass. Larger systems, shown generically by **11b**, would need still more.

TAUTOMERS, ALTERNATING BONDS, AND SUBSTRUCTURE SEARCH

Tautomers and alternating bonds have always presented problems for substructure search systems, and this will most likely always be the case. This is true whether the system requires detailed encoding by the searcher or whether the searcher can input a query via a structure diagram. (In the latter case, the system's query input conventions must be designed carefully so that valid retrievals are not lost due to an incorrect treatment of tautomerism and alternating bonds thus causing an incorrect interpretation of a query. The key point, as has been mentioned earlier, is that the searcher must consider potentially tautomeric or alternating substructures with respect to their possible surroundings in full structures. This usually means searching for both fixed-bond and normalized-bond variations of such substructures.

The searcher must keep in mind the CAS Registry III System definition of tautomerism while framing search questions. Thus, most bonds to nitrogen or chalcogen atoms must be regarded as potentially tautomeric. Only when there is clearly no possibility of a tautomer, as in an ether linkage (R–O–R′), should the searcher look for fixed-bond substructures alone.

Similarly, many cyclic bonds which by themselves are not alternating bonds become alternating bonds when the substructure is embedded in a larger system. Fixed-bond substructures alone should be sought only when alternating-bond substructures are clearly impossible.

DENORMALIZATION OF TAUTOMERS AND ALTERNATING BONDS

Denormalization is an algorithmic procedure which regenerates single and double bonds from normalized tautomer and alternating bonds and fixes the positions of mobile groups (see Figure 12). Bond and mobile group placement follows input structuring conventions and nomenclature rules. Denormalization algorithms have been developed at CAS for use in

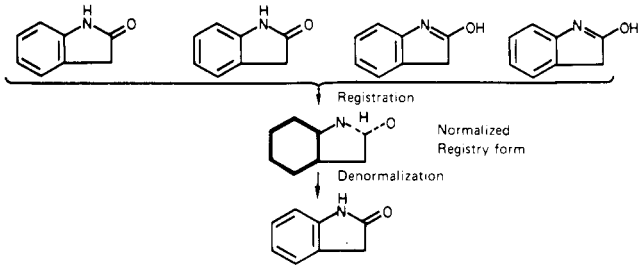

Figure 12. Example of denormalization.

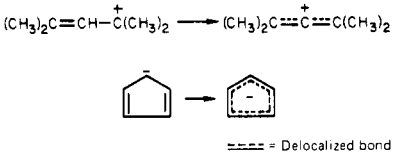

Figure 13. Examples of delocalized bonds.

algorithmic structure display and name generation.[3,4]

The denormalization algorithm used in name generation, which is the more recent and more accurate of the two, is incorporated in an algorithm which generates the systematic names of organic compounds for CA indexes from Registry III connection tables. It denormalizes the acyclic portions of tautomers prior to analysis for naming, since bond placement here primarily depends on structural considerations. Alternating bonds are also denormalized at this point, following input graphic standards for double bond placement. Cyclic tautomers and overlapping tautomer and alternating bond situations are denormalized during analysis for naming, when the preferred CA Index Name for the structure is being selected. Nomenclature rules such as "lowest locants for indicated hydrogen", "lowest locants for substituent prefixes", etc., determine double bond placement. The denormalization algorithm used in structure display operates similarly, although it must use a "best guess" approach to handle cyclic tautomers and overlapping tautomer-alternating situations since it does not have nomenclature rules to guide it.

OTHER ASPECTS

Delocalized bonds and charges are used to represent such species as allyl cations or cyclopentadienyl anions (see Figure 13). These bonds are identified by a chemist before the structure is input to the CAS Chemical Registry System, rather than by a machine procedure during registration.

Registry III includes a tautomer override feature that can be used to keep potential tautomers from being normalized. It is used only in those rare cases (only a few hundred to date) when a specific tautomer which is not the CA preferred form for naming is emphasized (see Figure 14).

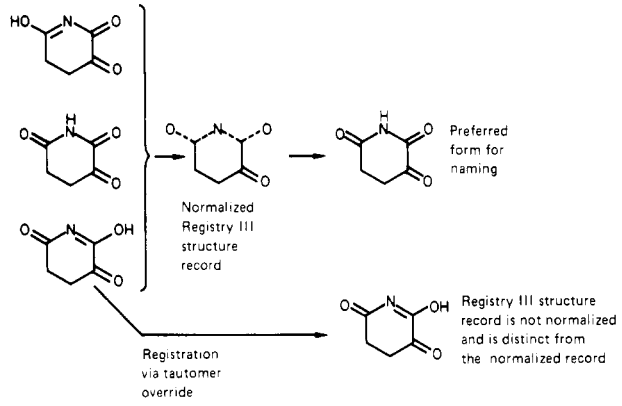

Figure 14. Application of the tautomer override feature.

SUMMARY

Tautomerism and alternating (aromatic) bonds hinder the representation of a chemical substance by a single structural diagram or its machine equivalent, a connection table. The CAS Chemical Registry System handles the problem by normalizing such structures: replacing the explicit single and double bonds with special normalized bonds, so that all input representations lead to the same unique Registry III structure record. Denormalization procedures to regenerate the single and double bonds in accordance with input structuring conventions and nomenclature rules have been developed for use in the generation of structure diagrams and CA Index Names from CAS Registry III structure records.

ACKNOWLEDGMENT

The development of the CAS Chemical Registry System was substantially supported by the National Science Foundation. Chemical Abstracts Service, a division of the American Chemical Society, gratefully acknowledges this support.

REFERENCES AND NOTES

(1) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Chemical Registry System. I. General Design", *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111-121.
(2) "*Chemical Abstracts Index Guide*", 1977, 100I-103I.
(3) Dittmar, P. G.; Mockus, J.; Couvreur, K. M. "An Algorithmic Computer Graphics Program for Generating Chemical Structure Diagrams", *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 186-192.
(4) Mockus, J.; Isenberg, A. C.; Vander Stouw, G. G. "Algorithmic Generation of *Chemical Abstracts* Index Names. I. General Design", in preparation.