Automatic Assignment of Molecule Keywords†

Martin J. Schweiger

CHEMODATA Computer-Chemie GmbH, W-8038 Gröbenzell, Germany

Received June 12, 1992

For use in the ChemInform RX databases, molecule keywords are assigned automatically. These keywords are based on the thesaurus derived from the ORAC database system. The overall procedure is discussed. Details of the assignment algorithm are shown for one branch of the process.

INTRODUCTION

Since the beginning of 1991, *ChemInform*,[1] published by the FIZ-CHEMIE in Berlin, has been produced completely electronically. This allows the production of printed issues as well as the generation of an ADABAS database[2] using the same input. As ChemInform RX,[3] this database will be available as an in-house system to be used in a REACCS[4] and ORAC[5] environment.

While preparing the information about abstracts, molecules, and reactions for this database, an automatic addition of several kinds of data is performed. These data are, for example, keywords for reactions, molecules, reaction sites, and mappings for reactant atoms to product atoms.

The keywords assigned to molecules are derived from their structures; those for the reactions are derived from the reaction site information.

GENERAL NOTES ABOUT KEYWORDS

Before the discussion of the automatic assignment of the molecule keywords, some general notes about the keywords used for the ChemInform project will be reviewed.

Use of Keywords. The molecule keywords are a handy tool to characterize classes of compounds. They represent certain sections of a molecule (substructures). These sections are independent from other substructure elements that may occur in the same molecule.

The major use of the molecule keywords will be in the search of molecules and reactions thereof. While formulating a query based on molecule keywords, the user only needs to known the structural element that produces the keyword.

In contrast to a search based upon real substructures, there is no structure input necessary. This reduced and faster query input is especially profitable while searching for a certain molecule (solvent and reagent list) as part of a reaction.

The extraordinary advantages given by the use of keywords in searches were also described by Finch in context with the Chemical Reaction Documentation Service.[6] However, in that system, the keywords have been assigned manually during the input procedure.

Selection of Keywords. The keywords in our project had been selected for automatic assignment on the basis of the ORAC database thesauri for reactions, solvents, and reagents.[7]

The molecule keywords are assigned by the examination of the connection table (CT) of a molecule. Therefore, all keywords based on no-structure information, where no defined CT is available, are excluded from automatic assignment. Examples for these keys are "petroleum ether", "nujol" or "silicon oil".

† Dedicated to Prof. Wolfgang Beck on his 60th birthday.

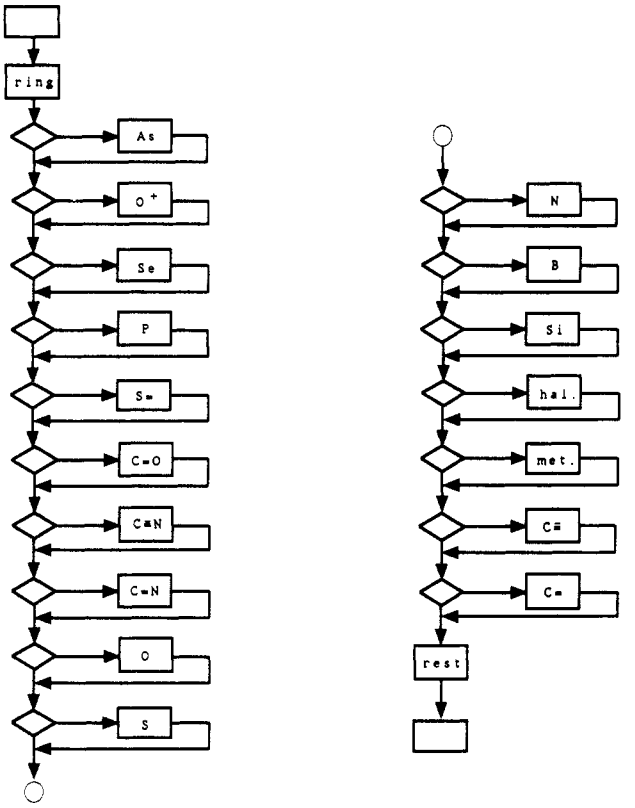

Figure 1. Hierarchy of root elements for keyword assignment.

Three categories of molecule keywords beside the list of reaction keywords are distinguished:

1. Molecule keywords describing fully defined compounds such as "water" or "THF".
2. Molecule keywords showing ring systems like "adamantane" or "benzene".
3. Molecule keywords derived from structural elements such as "amide" or "thio acid".

All keywords from category 3 correspond to the functional groups of molecules. The names of these functions have been standardized according to the IUPAC rules for organic nomenclature.[8]

If there were doubts about the structural elements of a less common keyword, they were derived from textbooks or dictionaries on organic chemistry.[9]

General Assignment of Keywords. The molecule keywords represent a set of atoms in a restricted environment. The number of spheres around a central point is kept as small as possible. Keywords that would require a large-scale examination along a molecule chain, such as "aliphatic" or "peptide", are not assigned according to this restriction.

0095-2338/93/1633-0128$04.00/0

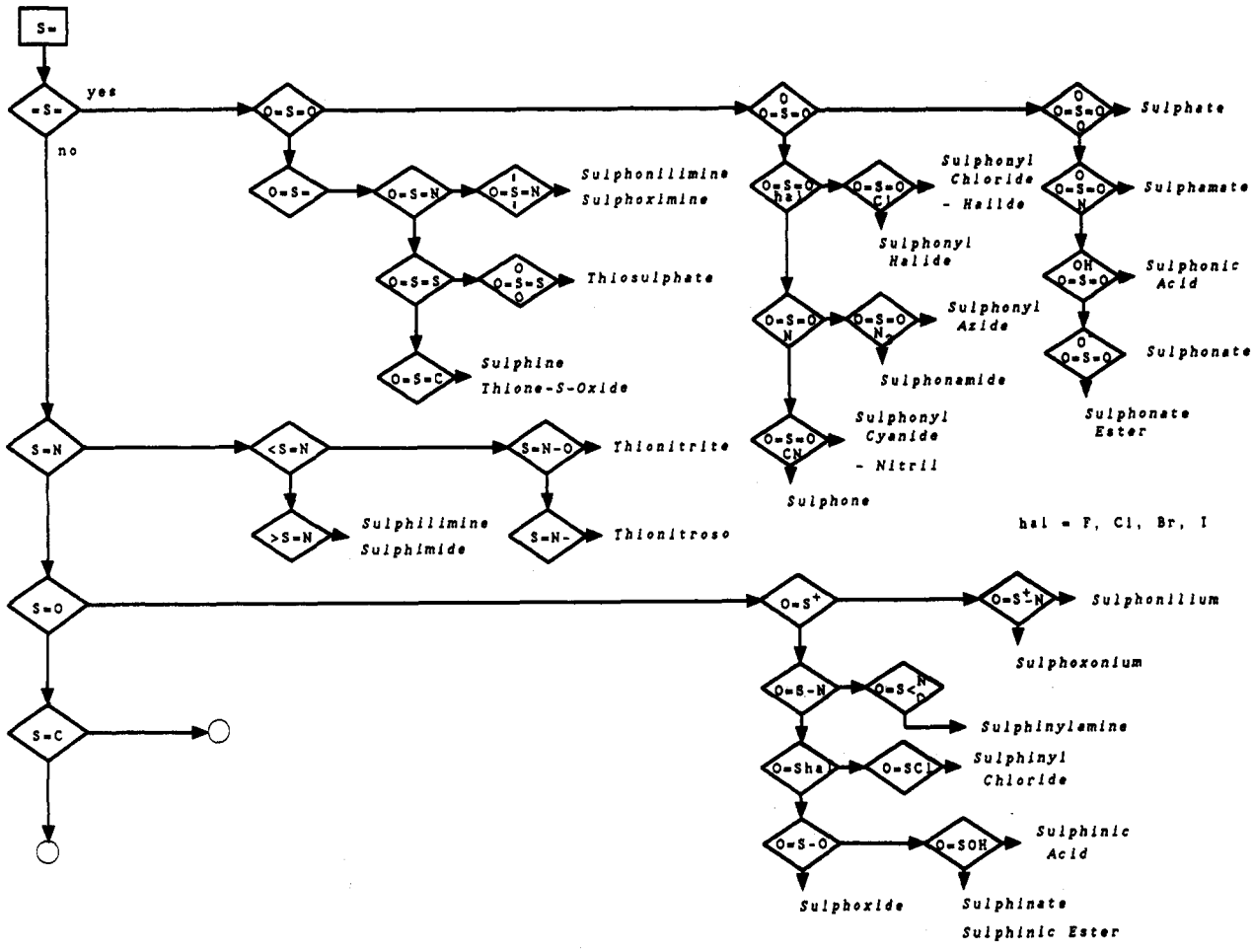

Figure 2. Keyword branch for a double-bonded sulfur atom.

The limitation to a few spheres around a central atom or group is due to the short computation time required.

The formal and automatic approach used throughout this program module ensures that all molecules are treated equally. The resulting keywords are not revised manually and are thus identical for the same groups in other molecules.

Molecule keywords for certain molecules have been assigned in a special SETUP procedure. These molecules have been stored as a standard set of molecules in the database.

All molecule keywords are based on characteristic structural elements. The assignment is performed in a dedicated module during the registration of the molecules. The registration of molecules also includes the assignment of a unique registry number for this molecule and the actual database storage.

On the next occurrence of the same molecule in a reaction, the keywords are read from the database. The calculations only have to be done once for any molecule.

Difference between Molecule Name and Keywords. The name of a molecule should be an additional *unique* descriptor beside the structure. But in real databases, different names are frequently used for the same compound.

Trivial names, systematic or IUPAC names, acronyms, and abbreviations of names are common. Nevertheless, each name identifies exactly one compound.

The name is used for the registration in the ChemInform RX ADABAS database, if there is no structure available for a named molecule.

Keywords (except for category 1) identify classes of compounds (e.g., "xylene"). In contrast to names, keywords are standardized however.

AUTOMATIC ASSIGNMENT OF MOLECULE KEYWORDS

The module for the automatic evaluation of keywords uses only the connection table (CT) of the molecules. No other information is used during the procedure.

The CT represents a data structure covering all information on the atoms and bonds of a molecule, as there are atom numbers from the periodic table of elements, isotope markers, charges, numbers of free electrons, bond orders, and bonded atoms.

Hierarchy of Entries. The keyword routine works through the CT in several steps:

In the first step, each molecule is examined for rings.[10] The appropriate keywords will be assigned by comparing the "isolated rings" with those ring systems that can be assigned (category 2, see above). The isolated rings are produced by removing all substituents that do not belong to the actual ring system. A ring system in this context is given by all rings with at least one common bond between each other. The course of this ring procedure is in accordance with Beilstein's method for reducing cyclic molecules to the registry compound.[11]

In the following steps, the other keywords (category 3, see above) are assigned according to a hierarchy of the functional groups. The procedure determines the root elements for the functionality trees. Each function is connected to one or more branches, which end in the real keywords to be assigned.

This functionality tree was set up during program development to give a systematic reduction of the functional groups to their characteristic substructure elements and the corresponding root elements.

This tree of functional groups is linked by the hierarchy of the root elements and the appropriate keyword branches.

The root groups for the treatment are shown in Figure 1. The actual assignment of the keywords is performed in each branch entered, as demonstrated below.

As a last step, the "rest" module assigns keywords that did not fit in the previous scheme of root groups, such as "carbanion" or "carbene", and finally all redundant keywords are eliminated.

Working through a Branch. On entering a certain root module (see Figure 1), the examination of the specified group is extended to adjacent characteristic bonds or neighboring atoms (first sphere). The procedure is continued by examining the next sphere around the atoms found before. In most cases, it is enough to examine these two spheres around a starting atom.

Only for a limited number of keywords will more than two spheres have to be examined. Examples are "hydroxyketone (2-)", "-(3-)", or "tosylate".

A sphere of an atom represents the entire neighborhood in a certain level:

The first sphere describes an atom and all the atoms fixed to it by their bonds; the second sphere represents, in addition, the atoms bonded to the outer atoms of the first sphere.

The branching process leading to the keywords is drawn for a molecule containing a double-bonded sulfur atom, as shown in Figure 2.

The first differentiation is given by a second double bond from the root sulfur atom, which leads to the branches for sulfates, sulfones, and related characteristic groups. In the case of only one double bond, the neighboring atoms are examined for specific atoms connected to the root sulfur atom.

If one of the atom types (nitrogen, oxygen, or carbon) could be found, they span further subtrees. Only the N- and O-branches, which result in the thionitrites or sulfimides and sulfinates of sulfoxides, are shown completely in Figure 2. In the C-branch, which is not shown, all "thio" compounds would be assigned.

IMPLEMENTATION

All parts of the molecule keyword module are implemented in FORTRAN 77, supported by either a VMS (DEC VAX) or a UNIX (CADMUS) environment.

ACKNOWLEDGMENT

I thank the German Ministry of Research and Technology (Bundesministerium für Forschung und Technologie) and the Fachinformationszentrum Chemie for their support of the development of the keyword addition system. I also thank my colleagues at CHEMODATA for valuable discussions during the project. I would like to give special thanks to Dr. A. P. Johnson of ORAC Ltd. for supplying us with the keyword thesauri.

REFERENCES AND NOTES

(1) Roden, G.; Weiske, C., Eds. *ChemInform, Selected Abstracts in Chemistry*, VCH Verlagsgesellschaft: Weinheim, FRG, published weekly.
(2) ADABAS is a database management system supplied by Software AG, W-6100 Darmstadt, FRG.
(3) (a) Glock, B. CIC in Freiberg. *Nachr. Chem. Tech. Lab.* **1992**, *40*, 239–243. (b) Mitteilungen aus dem Fachinformationszentrum CHEMIE, Berlin. *FIZ CHEMIE Aktuell* **1991**, *23/24*, 3. (c) Parlow, A.; Weiske, C.; Gasteiger, J. ChemInform: An Integrated Information System on Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 400–402.
(4) REACCS (= Reaction Access System) is a product of Molecular Design Limited, San Leandro, CA.
(5) ORAC is a product of ORAC Ltd., Leeds, U.K.
(6) Finch, A.F. The Chemical Reactions Documentation Service: Ten Years On. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 17–22.
(7) Gasteiger, J. and Johnson, P. private communication, 1988.
(8) *Nomenclature of Organic Chemistry, Sections A, B, C, D, E, F, and H.* Pergamon Press: New York, 1979.
(9) (a) March, J. *Advanced Organic Chemistry, Reactions, Mechanisms, and Structure*, 3rd ed.; J. Wiley & Sons, Inc.: New York, 1985. (b) Streitwieser, A., Jr.; Heathcock, C. H. *Organische Chemie*; Verlag Chemie: Weinheim, 1980. (c) Neumüller, O. A. *Römmps Chemielexikon*, 8th Auflage; Franckh'sche Verlagshandlung: Stuttgart.
(10) Gasteiger, J.; Jochum, C. An Algorithm for the Perception of Synthetically Important Rings. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43–48.
(11) Kennen Sie Beilstein?: Springer Verlag: Heidelberg.