REACCS in the Chemical Development Environment. 3. Graphically Nonequivalent Representations of Molecules and Reactions

JOHN E. MILLS* and BARBARA BAUGHMAN

The R. W. Johnson Pharmaceutical Research Institute, Welsh and McKean Roads, Spring House, Pennsylvania 19477-0776

Received August 1, 1990

The utility of a chemical database management system can be dependent upon the consistent use of molecular and reaction representations within the database. Graphically nonequivalent molecular representations of equivalent chemicals or mixtures of isomers may be registered using REACCS. It is also possible to register graphically nonequivalent reaction representations using REACCS. The registration of nonequivalent representations is allowed even when duplicate registration of the same molecule is not allowed. Differences between common stereochemical conventions and the conventions used by REACCS are discussed. Methods that may be used to minimize registration of data under nonequivalent representations of both molecules and reactions are presented.

INTRODUCTION

The necessity to consider the end-user's needs when determining the datatypes to be selected for inclusion in a REACCS database, as well as methods used to facilitate data storage, retrieval, and display, have been presented in previous papers.[1,2] Just as a consistent application of guidelines for the storage of data in a given datatype can enhance the effectiveness of end-user searches, the consistent application of stereochemical designations to molecules and reactions can improve the information content in REACCS databases or in any other chemical database that is intended to be searched primarily through a graphics interface.

STEREOCHEMICAL CONSIDERATIONS

REACCS uses the same stereochemical designations[3] which are familiar to organic chemists for the representation of bonds which are up, down, or undetermined (Figure 1). Some of the problems associated with stereochemical designations within REACCS are due to the organic chemists' reliance upon contextual constraints in the interpretation of structural diagrams. In addition, subtle differences between the way that the program uses stereochemical information and the way that chemists interpret structural diagrams may lead to problems.

For example, in Figure 1, most chemists would probably agree that molecules A and B are racemic and that molecules C and D may represent enantiomers. REACCS, however, recognizes molecules A and B as representing a mixture of two compounds (racemate) while molecules C and D are perceived as being identical with one another yet different from A and B. To circumvent this problem, molecules containing a bond bearing either an up or a down stereochemical designation may be labeled with a CHIRAL flag (Figure 2). Once the CHIRAL flag has been applied, molecules E and F can be distinguished from one another, while either molecule C or molecule D may still be registered. Thus, if duplicate registration of molecules is not allowed, REACCS will allow the registration of up to four molecules for 2-bromobutane (or any other molecule containing one stereogenic center) as shown in Figure 2.

The above distinction between chemical intuition and the algorithms used within REACCS at first appear to be inconsequential. These inconsistencies can be useful or detrimental depending upon the forethought used in choosing the stereochemical representation of molecules within REACCS prior to database construction. In fact, these inconsistencies are frequently overlooked during the initial phase of database construction without immediate consequence.

In molecules containing two stereogenic centers, the distinction between chemical intuition and the algorithms used by REACCS becomes more obvious. Figure 3 shows, in general form, a number of different representations of molecules containing two stereogenic centers that may be registered within REACCS. For clarity, only those structures have been drawn in which the stereochemistry is designated by using the bond between the carbon atom and the X or R substituent. If one allows the stereochemistry to be designated by using the bond to the terminal methyl group on the left-hand side of the molecule, the bond to the ethyl substituent on the right-hand side of the molecule (both valid representations within REACCS), or the bond between the adjacent stereocenters (an invalid representation within REACCS), then the number of potential representations for the same molecule becomes proportionately greater.

Of the 24 representations shown, the REACCS software recognizes 13 as being different entities. Structures **1–4** are all equivalent, and most chemists would probably agree that they represent racemic mixtures of diastereomers. Once it is accepted that REACCS does not recognize molecules as being resolved unless the CHIRAL flag has been applied, most would conclude that structures **21–24** each represent a single pure enantiomer and that structures **17–20** each represent a different mixture of epimers in which the absolute configuration of one of the stereogenic centers is known. The remainder of the structures (**5–16**), however, represent chemicals or mixtures of chemicals which are sometimes counterintuitive.

Within REACCS, structures **13** and **14** are equivalent representations of racemates of known *R*,R** relative stereochemistry. Structures **15** and **16** are also equivalent representations of racemates of known *R*,S** relative stereochemistry. All of these assignments are contrary to what is taught in introductory organic chemistry courses. Common usage is that structures **13** and **14** are enantiomeric pairs, as are structures **15** and **16**.

Perhaps the most disconcerting equivalencies for structural representations within REACCS are those depicted by structures **5–8** and **9–12**. It could be argued that since structure **5** represents the same chemical as structure **7**, then **5–8** are all representations of the same mixture of racemic diastereomers. The same argument can be used to compare structures **9–12**, with the same conclusion. However, it has already been stated that structures **1–4** are racemic mixtures of diastereomers. Consequently, REACCS allows three different representations for the same mixture of isomers when

0095-2338/90/1630-0431$02.50/0

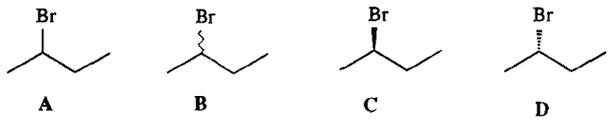

Figure 1. Stereochemical designations.

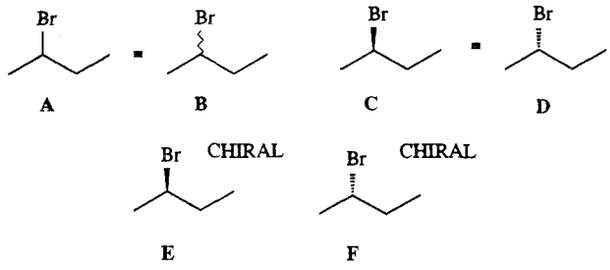

Figure 2. Stereochemical designations with REACCS.

two stereogenic centers are present in a molecule. The above considerations apply whenever there are two stereogenic centers in the molecule and are independent of the relative position of those two centers (i.e., the centers need not be contiguous).

Once a third stereogenic center is added, the issue of representation of a pure compound or mixtures of compounds becomes even more complex. With three potential stereogenic centers and the use of the chiral label, there are potentially 128 different ways to represent the individual pure isomers and mixtures of those isomers. Of those 128 different representations, REACCS allows for the registration of a maximum of 40 when duplicate registration is not allowed. Of these 40 representations, four can be shown to be chemically equivalent but graphically nonequivalent representations.

The above reasoning may be extended to mixtures of compounds containing more than three stereogenic centers. However, for most practical work, when more than four structural isomers of a compound may result from a chemical reaction, there is frequently some physical separation of isomers formed in the synthesis, or the reaction under consideration is run using a single pure isomer. Extension of the concepts elaborated above will therefore not be pursued further.

DOUBLE-BOND CONSIDERATIONS

The possibility of valence tautomers and double-bond isomers can play a role in database construction by controlling the number of representations of a given molecule that can exist either in tautomeric forms or as geometric isomers. This effect may be demonstrated most easily by examining the different nonequivalent representations allowed by REACCS in the registration of a guanidine in which each nitrogen atom bears a single substituent, but in which each substituent is different, e.g., *N*-methyl-*N'*-ethyl-*N''*-propylguanidine. Figure 4 shows the six nonequivalent representations allowed for tautomeric and isomeric forms of a single guanidine. The pair of isomers labeled A and B represent a pair of *E* and *Z* isomers

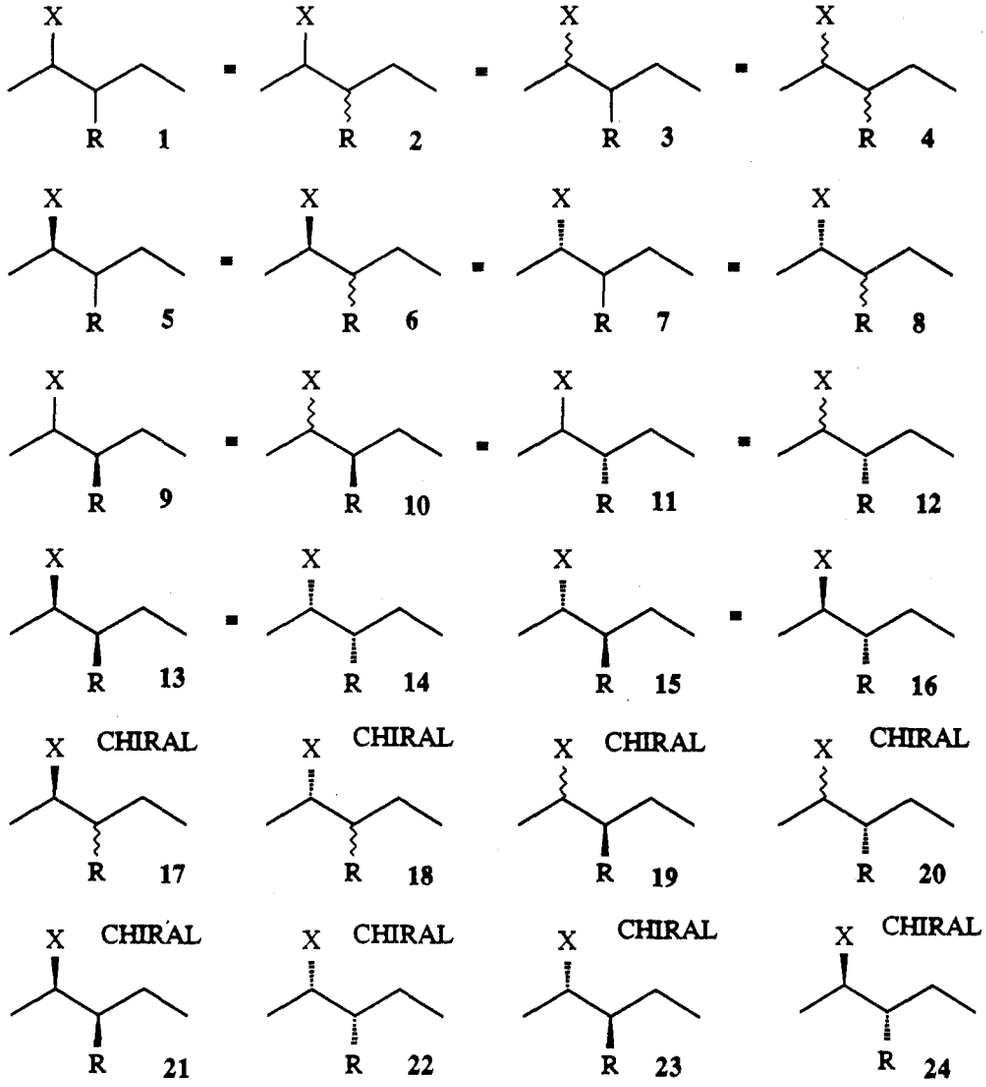

Figure 3. Partial listing of possible REACCS representations of molecules with two stereogenic centers.

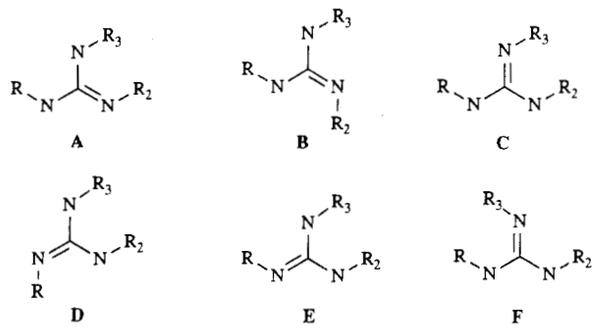

Figure 4. Nonequivalent guanidine structures allowed by REACCS.

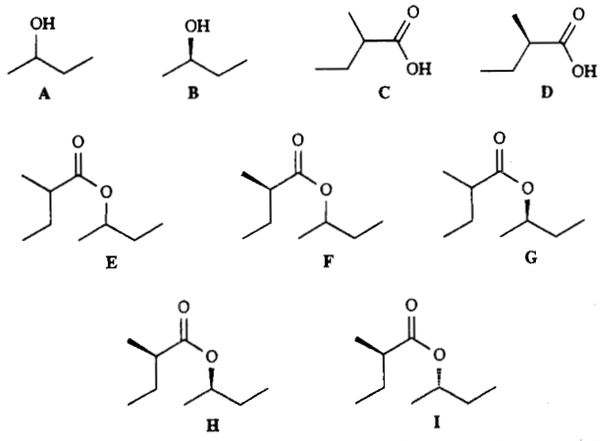

Figure 5. Nonequivalent representations of 2-butanol (A and B), 2-methylbutanoic acid (C and D), and 2-butyl 2-methylbutanoate (E–I) possible within REACCS.

which can convert easily via inversion of the lone pair of electrons on the imine nitrogen. Similarly, C and F as well as D and E represent *E,Z* isomers which can convert through inversion at nitrogen. In addition, when all three substituents are sterically small alkyl groups, valence tautomerization is typically a facile process. Consequently, in many cases, although there is the real possibility of preparing six different chemicals, the potential energy differences among the different compounds are so small that one cannot isolate or even distinguish among the species present in solution. Even when care is exercised, multiple registrations of a single chemical entity as different molecules are probable.

Other functional groups which pose similar potential problems concerning valence and geometric isomers are aminals, *S*-alkylthiouronium salts, N,S-disubstituted thioimidates, and N,O-disubstituted imidates. Oximes, hydrazines, hydrazones, and oxamates could be treated similarly; however, in these cases the potential energy barrier to inversion is frequently high enough that individual isomers may be isolated and identified. Without the capability to register *E* and *Z* isomers of imine derivatives, it would not be possible to distinguish the configuration of these chemicals.

REACTION INDEXING

The problems associated with nonequivalent representations of the same molecule or mixtures of molecules become most apparent when a reaction database is being constructed for use within the context of chemical development. In these circumstances, a number of chemists may be working on the same or similar reactions. Every chemist may submit his/her own preferred representation for a specific compound or mixture of compounds for inclusion in the corporate database. After entry is complete, numerous reactions which should actually be variations of a single reaction may be registered in the database.

Table I. Possible Representations in REACCS of Reactions in Which Two Reactants both Contain One Stereogenic Center

group	representation			
1	A + C → E	A + D → E	B + C → E	B + D → E
	A + C → F	A + D → F	B + C → F	B + D → F
	A + C → G	A + D → G	B + C → G	B + D → G
2	A + C → H	A + D → H	B + C → H	B + D → H
3	A + C → I	A + D → I	B + C → I	B + D → I
4	A + C → H+I	A + D → H+I	B + C → H+I	B + D → H+I

Figure 5 shows potential representations of the molecules in the esterification of 2-methylbutanoic acid (C and D) with 2-butanol (A and B). For simplification only one of each of the equivalent representations possible for each molecule has been displayed, and only racemic starting materials are considered.

REACCS defines reactions only in terms of reactants and products. Contrary to the way in which most chemists represent reaction diagrams, the order of those reactants or products in the REACCS diagram is not critical. Table I contains representations of the reaction cited above. The registration of every representation shown is allowed within REACCS. More importantly, the table contains generalized REACCS representations of all reactions in which two reactants, each containing one stereogenic center, react to yield either a single stereoisomer or a mixture of stereoisomers containing two stereogenic centers.

With the chemical equivalency of different representations for molecules developed above, it can be determined that the reaction representations shown in Table I can be grouped into nonequivalent representations of the same chemical reaction. The 12 reaction representations contained in the first three rows of Table I (group 1) are all chemically equivalent but graphically nonequivalent representations of the same reaction. The chemical reality displayed by these representations is the reaction of two racemic starting materials to yield an equal mixture of all four possible stereoisomers. The four reactions presented in row four (group 2) of the table are all representations of the same stereospecific reaction in which a racemic mixture of known relative stereochemistry is isolated. The four representations presented in the fifth row (group 3) of the table are also stereospecific reactions; however, the racemic product in these representations has a different relative stereochemistry than that in group 2.

The final row in Table I (group 4) contains four representations that are both the most specific and potentially the most useful for this type of reaction. These representations all indicate that two racemic diastereomers are produced in the reaction. With the appropriate data on yields, any reaction contained in groups 1–3 may be stored under one of the representations in group 4. For example, when the yield of H is equal to the yield of I, this representation is chemically equivalent to the reaction representations of group 1. If the yield of I or H is zero, then the representation is chemically equivalent to that displayed in either group 2 or group 3, respectively. The representations in group 4 allow for a more complete description than that allowed in any of the remaining groups, since group 4 allows for reactions which are only partially selective, e.g., a mixture which is 40% H and 60% I. The relationships between reactions represented by each of the groups is shown in Figure 6.

Similar problems in multiple representations of a single reaction arise when multiple representations of a single chemical are unintentionally registered in a REACCS database and subsequently used in the registration of reactions. For example, the reaction of an *S*-methylisothiouronium salt with an amine to yield a guanidine may be registered as many as 24 times depending upon the exact substituents attached to

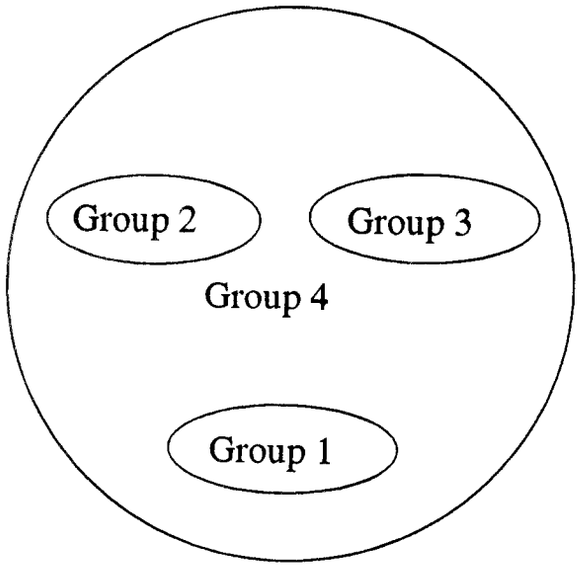

Figure 6. Sets of chemical reactions adequately described by the representations shown in Table I.

each of the nitrogen atoms. However, in this case, the precise number of chemically equivalent representations of the reaction is not readily apparent since the chemical species displayed in the reaction may or may not be distinguished.

Clearly, some process should be established to manage the potential proliferation of equivalent representations of the same reaction allowed within REACCS. Ideally, this process should allow for distinction between reactions depending upon the amount and quality of quantitative analytical data that has been collected. For example, early work on chemical process development may result in mixtures of chemicals which are not completely characterized. As work progresses, analytical methods may be established that allow for complete characterization of all isomers present in the reaction mixture. It can be beneficial to register the reactions that were incompletely characterized under one of the representations included in group 1. Those reactions in which the product mixture has been characterized more completely could then be stored under one of the representations shown in group 4.

The process used to limit the number of nonequivalent representations of the same reaction may be highly dependent upon the probability of encountering different representations of the same molecule or reaction. It may be unnecessary to be concerned about duplicate representations in a small commercial database, the purpose of which is to present broad literature coverage. During construction of such databases, it is unlikely that the same reaction will be abstracted from two different sources. However, in a proprietary corporate database used only by chemical development personnel, elimination of duplicate representations may be crucial. In the latter case, it is probable that several chemists, each using his/her own preferred representation, may be optimizing the same reaction. Retrieval of all the work done on a reaction will then be dependent upon the care taken in consistently indexing the reaction for retrieval.

DISCUSSION

As suggested above, the problems associated with the storage and retrieval of molecular and reaction data may be addressed at several different levels. Those solutions which appear most promising all involve recognition that registration of multiple nonequivalent representations of a single molecule or reaction is allowed using REACCS. Where necessary, controls should be implemented to ensure that only one preferred molecular representation is used throughout the database. Finally, standards should be set to ensure that preferred molecular and reaction representations are used consistently for the storage of reaction information.

In order to ensure that one representation of a molecule is used consistently in the R. W. Johnson Pharmaceutical Research Institute's proprietry REACCS databases, the registration of a new molecule containing two or fewer stereogenic centers is followed immediately by the registration of all allowed chemically equivalent representations. This process is used even for the registration of any molecule that can exist as valence tautomers or as *E*,*Z* isomers of nearly equal potential energy. The preferred representation is then used for the storage of data. All chemically equivalent representations other than the preferred representation are given the name USE REGNO [X]. In this case, X represents the internal registration number of the preferred representation. Before a molecule containing more than two stereogenic centers is registered, a substructure search is performed to determine if any equivalent structures are already in the database. If none are found, the new molecule is registered. When a nonequivalent representation is found, the originating chemists are contacted to determine whether a second representation is necessary.

During registration of reactions, the molecular identifier used by REACCS is set to the molecule name field. Thus, if a molecular representation other than the preferred is used in a reaction, the message to use the preferred representation is automatically displayed.

Implementation of this procedure has eliminated multiple representations of the same reaction within any of the four groups shown in Table I. It does, however, allow for chemically nonequivalent representations of the same reaction, i.e., a single representation from each of the four groups is allowed. As discussed above, a single representation may be used to store all variations of a reaction. However, it may be advantageous to classify a specific reaction as belonging to one of the groups prior to storage. Once this is done, the implied knowledge present in the representation as well as the data associated with the specific representation is more readily available to the user.

At RWJPRI, the accepted convention is that a group 1 representation is used to register those reactions that give equal mixtures of products or for those reactions in which the individual isomers have not been quantitated. Group 2 and 3 representations are used to depict those reactions in which only one isomer is prepared or isolated. Finally, a group 4 representation is used to store those reactions in which both isomers were either individually quantitated or isolated. By maintaining a distinction between those reactions defined as being in different groups, one can easily retrieve any specific reaction via a current search. Alternatively, if the complete set of data from all representations of a single reaction is required, it can be easily retrieved through a reaction substructure search.

CONCLUSIONS

With the advent of time sharing and improved graphics capabilities, reaction indexing has become a practical application for computer systems. This application, especially using REACCS, has placed rapid reaction retrieval based upon complex queries at the fingertips of those chemists with access to the program. With improved access has come the potential to overwhelm both the capabilities of the computer system and the user with redundant information. This is especially true if limitations inherent in the software are not recognized. The problem of registration of multiple nonequivalent representations of a given molecule or reaction is not typically a problem in small literature-based REACCS databases. The builders of these commercial databases typically are not concerned with

the depth of knowledge of a specific reaction, but rather with the breadth of knowledge in order to provide the user with potential ideas for new applications. The problem, however, frequently becomes obvious when the software is being used to build and maintain a database whose primary function is use by developmental chemists. In these circumstances, a great amount of data generated by numerous chemists on a given reaction or reaction sequence is frequently encountered. Consequently, consistent application of a procedure to ensure ready access to all data is imperative.

It is important to recognize that any form of reaction indexing has its own limitations and that identification of those limitations provides opportunities both for the utilization of those limitations and for the elimination of some of the potential problems caused by the limitations.

REFERENCES AND NOTES

(1) Mills, J. E.; Maryanoff, C. A.; Sorgi, K. L.; Scott, L.; Stanzione, R. REACCS in the Chemical Development Environment. 1. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 153.
(2) Mills, J. E.; Maryanoff, C. A.; Sorgi, K. L.; Stanzione, R.; Scott, L.; Herring, L.; Spink, J.; Baughman, B.; Bullock, W. REACCS in the Chemical Development Environment. 2. Structure and Construction of Proprietry Databases. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 155.
(3) Kasparek, S. V. *Computer Graphics and Chemical Structures*; Wiley-Interscience: New York, 1990; pp 514–516, 651–658.