Chemical Reaction Searching Compared in REACCS, SYNLIB, and ORAC

J. H. BORKENT,* F. OUKES, and J. H. NOORDIK

CAOS/CAMM Center, Faculty of Science, University of Nijmegen, 6525 ED Nijmegen, The Netherlands

Received January 8, 1988

Over the past two years, the data-base contents of the three commercially available reaction retrieval systems REACCS, ORAC, and SYNLIB have been developed to a level that makes a "chemical" comparison meaningful. The results of such a comparison, based on functional group transformation queries relevant to the bench chemist, show a remarkably small overlap in retrieved references.

INTRODUCTION

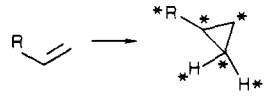
About two years ago, a comparison study of the reaction retrieval systems REACCS,[1] SYNLIB,[2] and ORAC[3] was published.[4] This comparison concentrated on the functionality of the systems rather than on the "chemical contents". The latter was omitted because at that time the data bases of the three systems were thought by the authors to be not representative. Since then the situation has changed considerably. The *Current Literature File* of REACCS (release 6.1) used in the present study contained 18 848 reactions,[5] ORAC (release 6.5) contained 25 000 reactions, and SYNLIB (release 2.2) comprised 39 174 reactions. With data bases this size, similar abstracting procedures—mainly academic chemists abstracting the primary literature, giving special attention to their own area of expertise—and the same literature being abstracted, the intriguing question arises whether the systems provide a chemist with the same answers for a particular query.

METHOD

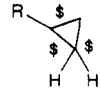
To perform the comparison, three different functional group transformations were selected to construct a query: query 1, the cyclopropanation of an alkene; query 2, the reduction of a carbonyl group to a secondary alcohol in the presence of an ester group; and query 3, the alkylation of a secondary carbon next to a carbonyl group. This choice, although rather arbitrary, is illustrative in that it makes use of the different options in the three systems needed to construct similar queries. The queries yielded a fair number of hits (in the range 20–50), which makes them realistic and representative.

To construct a similar query in each of the three systems for the transformations at hand, the following procedures were followed.

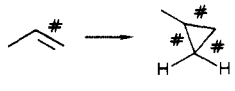
Query 1. SYNLIB: Reactant and product were drawn, with the product "substructured" (0 SUBS), and entered as a reaction (2 REQR in the RXN mode).



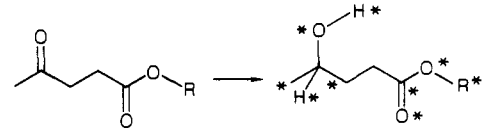
ORAC: The cyclopropane substructure was drawn (substructure mode, product), and the pertinent bond order changes during the reaction were indicated.



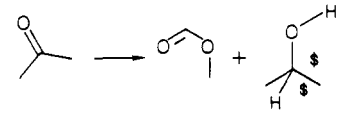
REACCS: Reactant and product were drawn, and the reaction centers were indicated in both the reactant and the product.



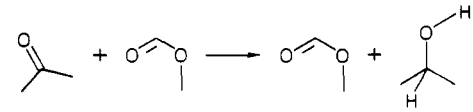
Query 2. SYNLIB: Reactant and product were drawn in the RXN mode; atoms in the product were substructured except for the carbons between the functional groups. The reaction was REQR'ed with no further constraints used.



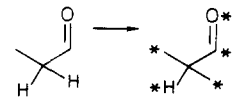
ORAC: A first selection was made by using the keyword REDUCTION. Within this set a search was performed with a keto group as REACTANT and a secondary alcohol and an ester group as PRODUCT. Also, bond order changes were indicated in the alcohol.



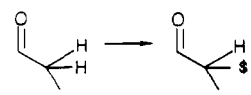
REACCS: The keyword REDUCTION was applied first, and then the ketone and ester were entered as REACTANT, the secondary alcohol and ester as PRODUCT.



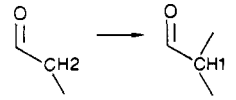
Query 3. SYNLIB: In the RXN mode, reactant and product were specified, with the product fully substructured. Constraints "WEAK BASE" (REQUIRED) and "ELECTROPHILIC" (FORBIDDEN) were added, and the reaction was entered with 2-REQR.



ORAC: The keywords BASE and NUCLEOPHILIC were used first (with AND logic). Substructures for reactant and product were ANDed. The keywords ELECTROPHILIC, PHOTOCHEMISTRY, ALDOL, and MICHAEL were used with AND NOT logic to exclude unwanted hits.



REACCS: Substructures of reactant and product were used, with "2 HYDROGENS" at the α-carbon in the reactant and "1 HYDROGEN" in the product. The keyword ALKYLATION was not used, as it resulted in loss of relevant hits.



0095-2338/88/1628-0148$01.50/0

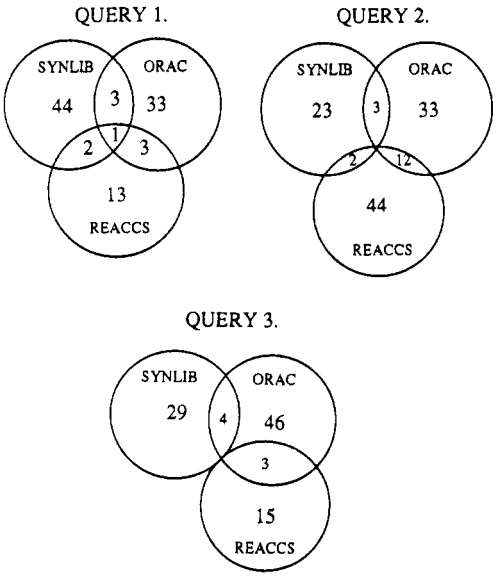


Figure 1. Number of references and overlap.

All queries were constructed in such a way that the maximum (as far as could be determined) number of relevant hits was found. Under these conditions the appearance of some "noise" (not fully relevant hits) turned out to be inevitable, in particular in SYNLIB, where bond transformations cannot be specified in such detail as in the other two systems. An example is, in query 2, the formation of a secondary alcohol by an aldol condensation instead of a reduction.

The references obtained from the relevant hits were collected and compared for the three systems. There was no apparent difference as to the chemical quality of the abstracts. It is difficult to develop criteria for a comparison on this point, but when the systems were used over a period of time, no pattern emerged that could be used to make a distinction. Also, the amount of information shown in an abstract is more or less the same.

RESULTS AND DISCUSSION

The result of the searches, in terms of the number of relevant literature references found for the same query in each of the three systems, is summarized in the Venn diagrams in Figure 1. Figure 1 shows the overlap, i.e., the number of references appearing in more than one list. For query 1, only one reference appeared in all three of the lists, for queries 2 and 3 none. The number of references retrieved for each of the queries in each of the systems is in the range 20–50, which is just about the number digestible for browsing, although in general the number of hits will be considerably larger than the number of different references. This is not only because more than one hit will point at the same reference but also because of the retrieval of irrelevant (or rather not exactly relevant) hits. In our opinion it is more important not to lose relevant hits, and when this policy was maintained throughout the study, it appeared inevitable that some extra ones were included. For that matter it should be noted that in browsing systems like the ones investigated these hits can be interesting in terms of idea generators. However, they have not been included in the figures presented here.

The figures for the overlap in the results are remarkably small: For queries 1 and 3 less than 10% of the total number of references occurs in more than one list, for query 2 less than 15%. These figures indicate that, although the same current literature is being abstracted for all three systems (roughly, the common organic chemistry journals over the past 15 years), the interest and expertise of the (academic) abstractors determine which article will be considered for abstracting and which particular reaction(s) from that article will be included in the data base. That both considerations play a role can be seen from the figures in Table I. For each list of references resulting from a query in a particular system, it was checked whether these references occur in the other two systems (irrespective of the direct overlap in Figure 1). For instance, of the 46 references found in SYNLIB and not in ORAC, for query 1, 4 were present in the ORAC data base, albeit with a different reaction. The ratios in Table I range from 0 to more than 60%, illustrating the variety of identical references with different reactions in the data bases. The higher ratios are found for cross-checking the references from ORAC and REACCS in SYNLIB, which can be explained by the significantly larger size of the data base of the latter. The low ratios in Table I reflect the variety in the three data bases. Two examples will illustrate this point.

Table I. Cross-Checking Reference Lists[a]

	query 1	query 2	query 3
SYNLIB ref in ORAC	4/46	7/25	4/29
SYNLIB ref in REACCS	3/47	4/26	0/33
ORAC ref in SYNLIB	13/26	20/45	25/49
ORAC ref in REACCS	2/27	4/36	9/50
REACCS ref in SYNLIB	2/15	18/56	11/18
REACCS ref in ORAC	5/16	13/46	11/15

[a] See text for an explanation in figures.

One of the references found for query 2 in REACCS is to a paper by Larson and Danishefsky.[6] In total there are seven entries with this reference in REACCS, of which one is retrieved by query 2. It turns out that ORAC and SYNLIB each contain four entries with the same reference, but none of these fulfill the requirements of query 2. One entry in SYNLIB mentions the reduction searched for but as part of a multistep synthesis, without the alcohol appearing as a product. Also, a comparison of the other 14 entries reveals that there is little agreement among the (at least three) abstractors who interpreted the contents of this two-page paper. On the other hand, a typical synthetic paper on the selective reduction of keto groups in steroids, published by Tal et al.[7] in 1984, appeared to be present in only one of the three systems.

CONCLUDING REMARKS

From this study, the following conclusions can be drawn:

A query in any of the three systems results in a more or less random sample of references from the current literature.

The reactions abstracted from these references reflect the interest and expertise of the abstractors, making these systems expertise files rather than literature files.

At the moment the three systems can be used next to each other (as we do at the CAOS/CAMM Center), without any serious danger of duplication, let alone triplication, of results. Actually, such a usage is preferable if possible.

Although the system distributors do not claim anything contrary to our findings (regarding the partial coverage of the literature), we feel it is interesting and useful to make occasional quantitative comparisons to see how these systems cope with the growing stream of primary literature and to check whether they approach each other in contents or not.

REFERENCES AND NOTES

(1) REACCS: REaction ACCess System, Molecular Design Ltd., San Leandro, CA.
(2) SYNLIB: SYNthesis LIBrary, Distributed Chemical Graphics Inc., Philadelphia, PA.
(3) ORAC: Organic Reactions Accessed by Computer, Wolfson Unit for Computer Aided Design, Leeds, U.K. The 1988 release of ORAC gives also access to a data base derived from Theilheimer's *Synthetic Methods of Organic Chemistry*. This data base was not included in the

present study, which focuses on the current literature.
(4) Zass, E.; Mueller, S. *Chimia* **1986**, *40*, 38-50.
(5) REACCS gives access to two other reaction data bases, derived from Theilheimer's *Synthetic Methods of Organic Chemistry* and *Organic Synthesis*, respectively, not included in the present study, which focuses on the current literature.
(6) Larson, E. R.; Danishefsky, S. *J. Am. Chem. Soc.* **1983**, *105*, 6715.
(7) Tal, D. M.; Frisch, G. D.; Elliott, W. H. *Tetrahedron* **1984**, *40*, 851.

ACS Committee on Nomenclature: Annual Report for 1987

KURT L. LOENING
Chemical Abstracts Service, Columbus, Ohio 43210

Received May 20, 1988

Nomenclature committees, both national and international, were very active in 1987, resulting in substantial progress in many different fields. A summary of the more important meetings and accomplishments follows.

The ACS Committee on Nomenclature continues to be active. Editors of ACS journals are ex officio members of the Committee. The Committee held its annual meeting at Chemical Abstracts Service (CAS) in November. The Committee continues its efforts to communicate with the ACS membership as well as all other groups who have an interest in chemical nomenclature. As part of this effort, open meetings were held again at the ACS National Meetings in Denver and New Orleans, and the Committee continues to look into simultaneous divisional nomenclature programming at a future national meeting. Good communication with high school chemistry teachers is being maintained through A. Saturnelli and the Bureau of Science Education of the New York State Education Department. Cooperation with the American Society for Testing Materials (ASTM) Committee on Medical Terminology has been established on an ongoing basis. Close liaison with other ACS bodies such as the Committees on Education and Science as well as various Divisions is being pursued. The promotion of and input into International Union of Pure and Applied Chemistry (IUPAC) recommendations is, as always, a primary objective of the Committee. A subcommittee is investigating problems relating to the nomenclature of biotechnology by focusing on the nomenclature of altered proteins. The subcommittee on Chemical Pronunciation continues to be active. The Chairman made an oral report on the activities of the Committee to the ACS Board of Directors at its December meeting.

The IUPAC Interdivisional Committee on Nomenclature and Symbols (IDCNS) continued to function effectively this year. It held its annual meeting in Boston in August. In addition to the IUPAC publications listed in the Appendix, specific documents in process and thus not yet recorded in this Appendix deal with the following topics: steroids; mass spectroscopy; X-ray spectroscopy; representation and symbolism of reaction mechanisms; and organic chemical transformations. The final chapters of the revised Red Book from the Commission on Nomenclature of Inorganic Chemistry have undergone the review procedure. The *Compendium of IUPAC Terminology* has been published, and work on an expanded second edition has been initiated. Publication is expected in 1988. The final manuscript of the revised manual "Quantities, Units, and Symbols in Physical Chemistry" has been completed, and its publication is imminent. Kurt Loening's chairmanship of IDCNS terminated in 1987 at the Boston meeting. The new chairman is Professor N. Sheppard of the U.K.

The IUPAC Commission on the Nomenclature of Inorganic Chemistry met in August in Boston. All 11 chapters of the revised Red Book have been completed and approved and are receiving final editing. A document on rings and chains has been sent to referees, and recommendations on polyanions have been published (see Appendix). The Commission again reaffirmed its position on the periodic table and on the systematic names of the elements of atomic numbers greater than 100. Other topics under study are organometallic compounds, advanced stereochemical topics, metal clusters, and abbreviations.

The IUPAC Organic Nomenclature Commission met in Boston in August. The Commission continued its study of the reorganization and revision of the present edition of the IUPAC organic rules and the development of new techniques for longer range consideration. In connection with the latter, projects dealing with nomenclature for cyclophanes, oxo acids, and nodal numbering are continuing to develop. Recommendations for generating numerical prefixes beyond 200 have been published (see Appendix). A convention for describing rings and ring systems with cumulative double bonds was approved for publication. Comprehensive documentation on classical ions and radicals, natural products, and fusion nomenclature is well advanced. A glossary of class names and terms has been compiled. In addition, projects on revision of Section E (Stereochemistry), indicated hydrogen, and numbering priorities are under study.

The IUPAC Commission on Macromolecular Nomenclature met in Boston in August. The Commission completed its work on a report offering about 75 definitions of terms dealing with crystalline polymers, agreed on a revision of the report on stereochemistry, and appointed a working party to coordinate the publication of the *Compendium on Macromolecular Nomenclature*. The Commission discussed and made progress on documents dealing with (a) the classification of polymerization reactions (mechanism and stoichiometry), (b) the nomenclature of cross-linked and nonlinear (branched, star, etc.) polymers and of polymer networks, (c) the structure-based nomenclature of irregular polymers, (d) the conventions for structural formulas of polymers, (e) the terminology for static and dynamic mechanical properties of polymers in bulk state, and (f) the nomenclautre of ladder polymers. Work was initiated on the document dealing with the terminology of liquid crystals. In 1987, recommendations on the use of abbreviations were published (see Appendix), and two reports on classification of polymers and on solutions were completed and submitted for publication.

In biochemical nomenclature both the Joint Commission on Biochemical Nomenclature (JCBN) and the Nomenclature Committee of the International Union of Biochemistry (NC-IUB) met jointly in Szeged, Hungary, in May. Recommendations dealing with the nomenclature of folic acid, prenols, and tetrapyrroles have been published (see Appendix).

0095-2338/88/1628-0150$01.50/0