Automated Overlap Analysis of Reaction Databases

Jan-Willem Boiten, Martin A. Ott, and Jan H. Noordik*

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

Received July 13, 1994[⊗]

A fully automated procedure has been developed to assess the overlap between two reaction databases. The procedure determines the number of entries in a given database sharing the literature reference with entries in another database. This number is related to the number of shared reactions through the analysis of answers to a series of sample queries. Additionally, an earlier overlap analysis has been repeated to investigate the trend in database overlap.

INTRODUCTION

The number of new syntheses and synthetic methods published on a yearly basis is far too large for the average chemist to deal with by memory alone. Therefore, a more systematic access to the primary literature is required to effectively find literature solutions to synthetic problems. An obvious method offering this systemization is the collection of reaction information in a database. Currently, such computerized collections of reaction data can be obtained from several commercial suppliers. Some of these databases are built to offer a complete literature coverage within certain boundary conditions; others try to compile careful selections of useful reactions with no claim of completeness whatsoever. The main representatives of the former type are CASREACT[1] and Beilstein,[2,3] the most prominent database systems of the latter type are ORAC,[4,5] REACCS,[6−8] SYNLIB,[9−11] and IRDAS.[12] The comprehensive databases can usually only be accessed through external hosts (e.g., STN[13]), while the selective databases are usually purchased or licensed for in-house use. The strength of these latter systems is that they provide a quick overview of the key literature methods employed in cases similar to the user's synthetic problem. The selectiveness of the databases is an absolute condition to fulfill this aim, since a complete database will often hide the interesting methods in an avalanche of rather obscure or repetitive answers.

Nowadays, in-house database systems are accessible in many synthetic laboratories. This widespread implementation prompted independent vendors to compile their own databases and supply them in the most common formats.[14] An expected effect of this trend is a steady increase in overlap within the total in-house database. This is an alarming development since frequently occurring duplicate references would highly decrease the effectiveness of the system. Therefore, the database overlap is an important consideration when deciding on a database purchase. Hence, in a situation like ours where we have to deal with many different databases, we felt a strong need for an automated procedure assessing the database overlap. We are aware of only a few attempts to characterize database composition from the user point of view: a few tiresome manual analyses[15,16] and, recently, a probabilistic approach.[17] The current paper presents the automated procedure we developed to determine the number of shared references between two or more databases without any manual interference. In order to relate the number of shared references to the number of shared reaction entries, we performed a series of sample analyses. The sample analyses allowed us to derive a rule of thumb which can be used to interpret future comparisons performed with our automated procedure.

METHODS

General Approach. In order to compare two reaction databases, one has to compare individual reaction entries. This requires a method to establish the equality or similarity of reaction entries of different databases. We would prefer to use the chemical structures of reactants and products as the criterion to decide on equality of database entries; however, the chemical structures are less suitable for an automatic comparison for various reasons. The representation of chemical reactions in the in-house databases is in no way standardized; in particular, there is no unanimity about what distinguishes a reactant from a reagent. Moreover, stereochemical features may or may not have been included.[18] Finally, it would be a major programming task to circumvent the 10^8 comparisons needed to compare two 10 000-reaction databases, and many databases are larger. As a consequence, we decided to use the bibliographic reference as the primary datafield for reaction comparisons. Naturally, two reaction entries containing the same literature reference are not necessarily identical. In fact, several possible situations can be distinguished:

1. The reactions are exactly the same with respect to reactant and product structures.
2. The reactions are the same but performed on different substrates, i.e., the reactions have identical reaction centers.
3. The reactions are not the same but are very similar (e.g., Grignard reactions on a ketone vs on an ester).
4. The reactions are different (often different steps in a synthetic sequence).

Since most user queries are in fact reaction center searches we will treat both cases (1) and (2) as identical reactions[19] and cases (3) and (4) as different reactions. In order to have a good impression of which part of the bibliographic overlap also involves identical reactions, we decided to check the

[⊗] Abstract published in *Advance ACS Abstracts,* November 15, 1994.

0095-2338/95/1635-0115$09.00/0

relation between bibliographic overlap and reaction overlap in two ways: (A) examining a set of randomly chosen bibliographic duplicates with respect to identical reactions and (B) examining a number of answer sets of randomly chosen reaction queries with respect to overlap in literature references.

Test Set of Databases. All databases available to us have been submitted to our automated evaluation procedure. As a bonus, this massive comparison yielded the overlap between most in-house databases currently available. Table 1 provides an overview of all databases subjected to the overlap analysis. The analyses incorporated not only all databases commercially available for ORAC but also two databases translated from REACCS using our own translation program[20] and the SYNLIB core database which demanded some extension from the journal standardization program (*vide infra*).

Reference Analysis. Both ORAC and SYNLIB contain functionalities to write the references of database entries to a file. To be able to compare bibliographic references directly and automatically, they have to be brought into a standard format. The references written by ORAC do not contain standard names for the journal, nor is there a standard order of numerical items (volume, issue, page, and year). Therefore, a program was written to reorder the numerical items in an ORAC reference as the year followed by the page (the other items are ambiguous and largely superfluous). Another program was written to convert all journal synonym names to the standard name (the so-called display name) according to ORAC's thesaurus definition (e.g., in the journal thesaurus *JACS*, *Am. Chem. Soc.*, and *J. Amer. Chem. Soc.* are all synonyms of *J. Am. Chem. Soc.*). As with the volume and issue numbers, the author names were discarded. SYNLIB's reference utility writes the references as journal code, page, and year. It was a trivial matter to convert these into the same format as used with ORAC. The journal codes were simply treated as synonyms of the corresponding journals, and the same program was used to convert these to the standard journal names.

Some loss of data occurred during export of the references by the program and their reformatting, because references were sometimes incomplete (missing journal name, page number, or year). For each database, the number of references thus obtained was counted, and the duplicate references were removed. Counting the resulting sets of references gave us a measure of the "internal" duplication within the various databases. The following procedure was used to determine the duplication between two databases or answer sets (*vide infra*): the references of both databases were standardized and sorted, and a special program counted the number of references in the "target" database that also occurred in the "comparison" database. This may seem a peculiar choice, since a paper from which five reactions have been abstracted in the target database and only one in the comparison database is still counted as five duplicates. However, we have to keep in mind that the main question to be answered by these database comparisons is "If we did not have this database, would we have found this reference also through another database?" This question can be answered positively for each of those five internal duplicates if there exists one external duplicate. So, we feel that our procedure accurately meets to the actual goal.

Inspection of Selected Answer Sets. Various queries have been selected to create the answer sets used to assess the connection between structural and reference duplication. Two of the queries, which originate from the paper by Borkent et al.,[15] have been repeated to determine the increase in database overlap over the years. The original paper studied three queries: (1) the cyclopropanation of alkenes; (2) the reduction of ketones to secondary alcohols in the presence of esters; and (3) the alkylation of secondary carbons next to a carbonyl group. The third query is very hard to define unambiguously, as many reactions give the same reaction modification statistics as the requested alkylation. Hence, we decided to drop this query. The remaining two queries were the only ones which were also performed in SYNLIB in order to enable a full comparison with the original results. Both the ORAC and the SYNLIB queries of the original paper could not be reused. The SYNLIB queries missed many of the correct answers,[21] while the ORAC queries used keywords which are not implemented in all of our test databases. Our new queries were the following:

Query 1. The cyclopropanation of an alkene was defined in ORAC as a product substructure search only, which is shown in Figure 1 with the bond specifiers used. The figure also shows the SYNLIB (product side) substructure,[10,11] which had to be further specified with constraints: the two bonds created during the reaction have been specified as strategic, and the olefin has been tagged as a required functional group in the reactant. This query avoids missing any correct answers but required a careful manual inspection to remove all inaccurate answers. We did restrict the query to cyclopropanations of olefins with at least one α-carbon on the olefin and two hydrogens on the added carbon atom to stay in line with the original query.

Query 2. Figure 2 shows the double substructure search to find reductions of a ketone in the presence of an ester in ORAC. The SYNLIB query was entirely constraints based. The functional groups participating in the reaction have been assigned as required: an ester and ketone in the reactant and an ester and alcohol in the product. Furthermore, reductive conditions have been specified to exclude most condensation

Table 1. Current Reaction Databases at the CAOS/CAMM Center

database	size	source	database scope
Box1-12	60000	ORAC Ltd & MDL	general, mainly 1980−1992
ACF1-2	10000	ORAC Ltd	general
Theilheimer	41783	ORAC Ltd	general, before 1980
MOS	3303	Synopsis	general, 1993
CHC	42375	MDL	heterocyclic chemistry
Hets_box_1	5000	ORAC Ltd	heterocyclic chemistry
CSM	9587	FIZ Chemie	general, 1992−1993
PG	16500	Synopsis	protective group chemistry
CLF[20]	35064	MDL	general, 1980−1991
Orgsyn	4763	MDL	checked exptl methods
ChemSynth	69703	InfoChem	general, 1975−1988
SYNLIB	81121	Distr. Chem. Graph.	general, 1890−1992

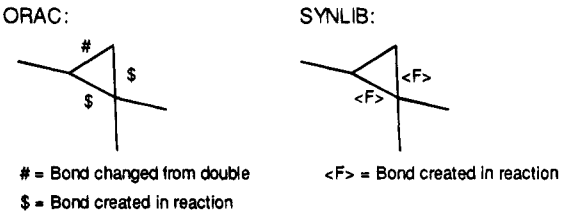

Figure 1. The ORAC and SYNLIB substructures used in the query for cyclopropanations of alkenes.

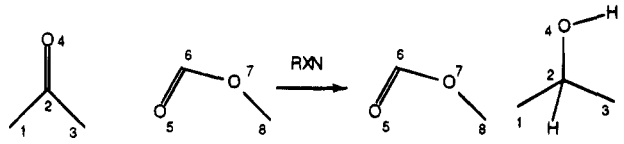

Figure 2. The ORAC substructure searches for the reduction of ketones in the presence of an ester.

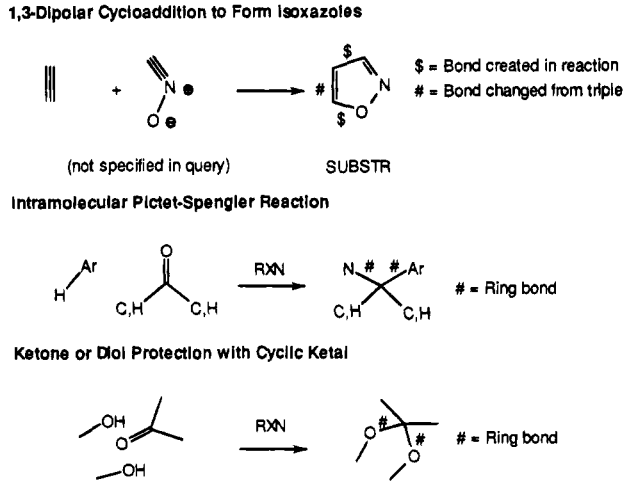

Figure 3. Three preselected ORAC queries dedicated to heterocyclic and protective group chemistry.

reactions. Still, a manual inspection was required to exclude many inaccurate answers.

Another three queries were added to our query-set to guarantee a sufficient representation of entries from the three specialized databases (CHC, HETS_BOX_1, and PG) in the answer sets. Figure 3 shows these three queries.

Another ten queries were added to counter the influence of our personal preferences, by selecting these queries semirandomly from March's book *Advanced Organic Chemistry*.[22] The selection process was not entirely random, since we proceeded through the book in an attempt to maximize variation in reaction types. The ten reactions thus selected are depicted with their ORAC queries in Figure 4.

Manual Inspection of Randomly Selected Duplicates. The references shared by two or more databases ("external duplicates") were identified by a simple procedure using the files with unique (i.e., no internal duplicates) and uniformly formatted references, as obtained from our reference analysis. These files were concatenated and sorted, after which the duplicates could be found on subsequent lines in the sorted file. The entire ORAC database has been submitted to this procedure, and sample duplicates have been taken from the output file at regular intervals, resulting in a set of 100 external duplicates to be inspected manually. These 100 sets of external duplicates have been supplemented with their internal duplicates resulting in sets with two to eleven entries. From each of the resulting sets one datacard has been randomly selected. These "key" entries have been compared with all other entries in the same set, with the exception of the internal duplicates of the key entry. Each comparison yielded a qualification according to the categories outlined before: "identical reaction", "identical reaction center", "similar reaction center", and "different" reaction. If more than one comparison had to be performed for a set of duplicates, then we only assigned the "most identical" qualification to that set of duplicates.

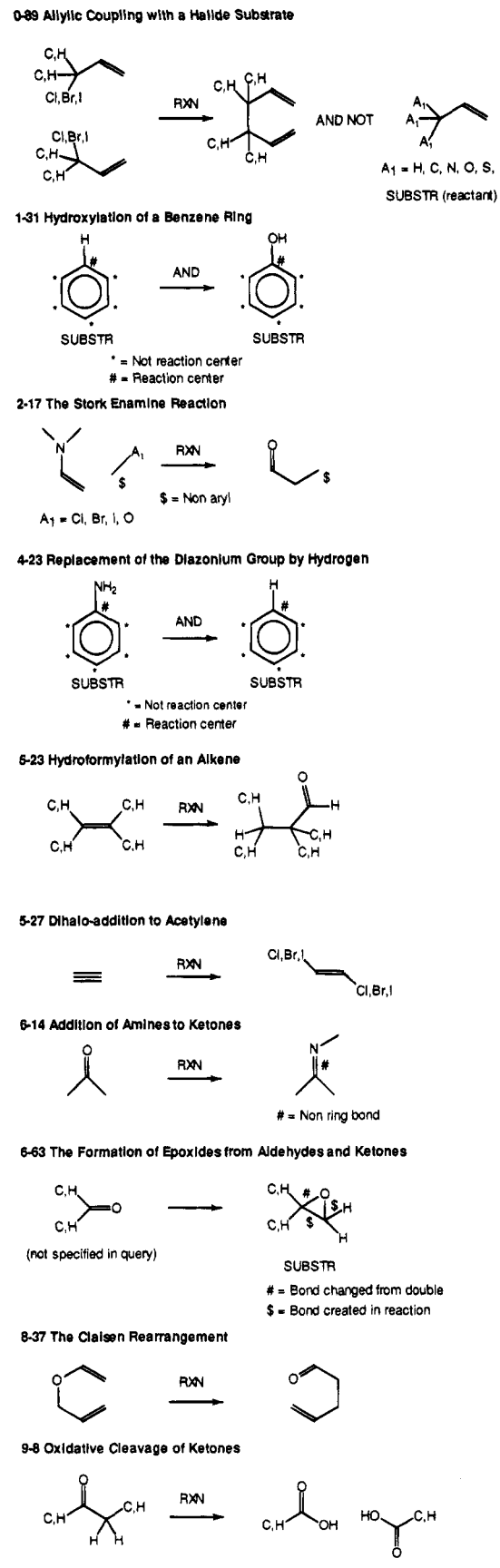

Figure 4. The 10 ORAC queries for reactions selected from March's book: RXN = reaction substructure search, SUBSTR = (single structure) substructure search, and AND = logical and.

RESULTS AND DISCUSSION

All databases have been submitted to the reference standardization process, which gave the following sta-

Table 2. Percentage of Data Entries in Each Database Sharing Their Reference with an Entry from Any Other Database

box	number of cards having external duplicates			
	all databases		excluding SYNLIB	
ACF	6052	(60.5%)	4320	(43.2%)
Boxes	38193	(63.7%)	28120	(46.9%)
CHC	15518	(36.6%)	13583	(32.1%)
ChemSynth	35618	(51.6%)	29874	(43.3%)
CLF	25280	(72.9%)	21745	(62.7%)
CSM	3414	(26.8%)	2927	(23.0%)
Hets	2886	(57.8%)	2660	(53.2%)
MOS	1610	(41.8%)	1605	(41.7%)
Orgsyn	2234	(48.6%)	619	(13.5%)
PG	9701	(59.2%)	8864	(54.1%)
Theilheimer	15220	(36.4%)	9419	(22.5%)
Synlib	53186	(67.4%)		
total	208912	(55.1%)	123736	(41.2%)

tistics for the ORAC databases:

total number of datacards	301 989
number of references written out	301 212
correctly formatted references	300 505
idem, no internal duplicates	153 078
idem, no external duplicates	121 326

Note that some references have not been written out by the reference utility, which is due to either a missing journal or an empty author field. Most of the references rejected in the standardization stage were references without a page number (549 datacards) or long references incompletely written out by ORAC (118 datacards). The remaining discarded references had miscellaneous problems, which were all very rare. We found many rejections during the standardization of SYNLIB references, which were caused by incomplete output of the program's reference utility, which assumes that each reference is formatted according to strict rules with respect to punctuation. In practice, many references disobey these rules, and, as a consequence, we had to discard 2236 references (out of 81 121).

After standardization of the references of most currently available in-house databases, it was an easy job to perform a full comparison of these data collections. The results have been condensed in a full cross-table (Table 2). Each entry in this table shows the percentage of references shared between two databases. More interesting than these individual overlaps is the overlap of each database with all other databases together, which is summarized in Table 3. The figures in this table indicate to what extent each database is redundant with the others with respect to the literature references. For example, if the ACF database would not be available, still 60.5% of the ACF references could be found elsewhere. Of course, the actual overlap in the reaction structures will be considerably smaller, as will be shown by our sample analyses.

The analyses of sample queries have been collected in Table 4. From the data in this table, we derived the average percentage of duplicates in the ORAC answer sets (18%). This number can be interpreted as the average chance of finding the same reaction (from the same reference) in another database. Table 3 showed that the average chance of finding the same reference in another database is 41% over all ORAC databases. Combining these two results shows that approximately 44% of the duplication in references consists of duplicate reactions (or actually reaction centers). This multiplication factor is important to allow better practical use of our automated database analysis method, but we are aware of several uncertainties related to this value. The value is obtained through the analysis of the answers to selected queries from which the representa-

Table 3. Overlap between All Databases Included in Our Analyses[a]

	ACF	Boxes	CHC	Chems	CLF	CSM	Hets	MOS	Orgsyn	PG	Theil	Synlib
ACF		3.2	0.33	20.9	23.4	0.03	0.20	0	0	4.0	0.13	38.9
Boxes	0.50		6.5	21.1	15.0	0.43	0.25	0	0.51	3.6	8.1	38.4
CHC	0.08	9.6		11.3	0.39	0	3.9	0	0.08	0.98	12.3	12.7
Chems	5.0	26.4	6.9		10.8	0	1.3	0	0	1.2	2.4	26.6
CLF	9.9	42.6	0.62	24.0		0.25	0.47	0	0	3.4	0.69	40.6
CSM	0.05	2.6	0	0	0.93		0	19.0	0	1.6	0	4.7
Hets	0.24	3.3	32.3	15.5	2.7	0		0	0	0.66	14.4	11.9
MOS	0	0	0	0	0	40.1	0		0	1.2	0	0.31
Orgsyn	0	11.2	1.4	0	0	0	0	0		1.0	0	41.9
PG	6.4	32.9	2.5	10.4	10.3	2.7	0.24	0.80	0.21		11.2	29.7
Theil	0.05	10.6	9.8	2.4	0.26	0	1.4	0	0	1.8		23.0
Synlib	7.3	36.3	5.6	18.8	14.5	0.61	0.68	0.03	1.5	3.2	14.2	

[a] The overlap is expressed as a percentage of the correct references in the databases listed horizontally.

Table 4. Summary of All Test Queries Manually Analyzed

answer set	no. of answers	no. of duplicates		answer set	no. of answers	no. of duplicates	
March 0.89	62	15	(24%)	het.1	114	24	(21%)
March 1.31	115	14	(12%)	het.2	150	15	(10%)
March 2.17	123	19	(19%)	PG	381	91	(24%)
March 4.23	71	6	(8%)	total	645	130	(20%)
March 5.23	68	12	(8%)				
March 5.27	46	10	(22%)				
March 6.14	165	25	(15%)	query 1	225	43	(19%)
March 6.63	93	10	(11%)	query 2	399	71	(18%)
March 8.37	121	21	(17%)	total	624	114	(18%)
March 9.8	21	7	(33%)				
total	885	139	(16%)	grand total	2154	383	(18%)

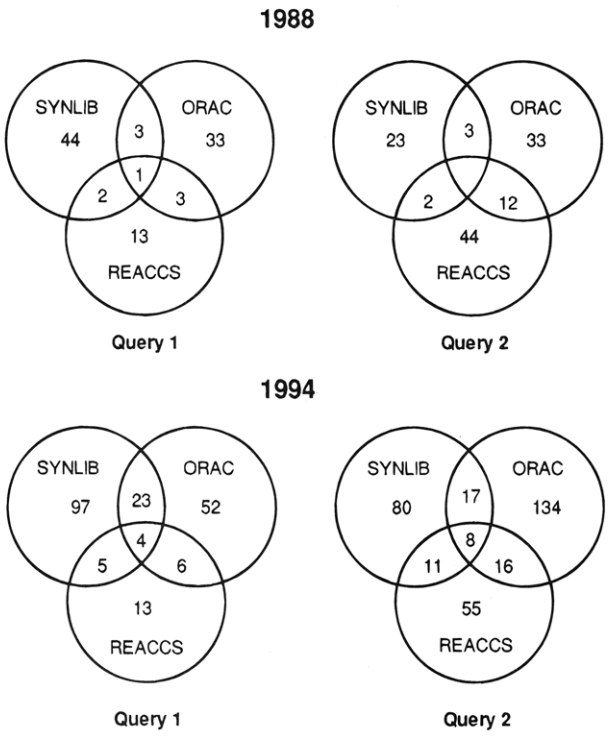

Figure 5. Number of unique references and their overlap for the two queries that were repeated from the 1988 paper.

tivity can always be disputed. However, the relatively moderate scatter in the duplication percentages is an indication that the connection between structural and bibliographic duplication is not heavily dependent on the reaction type. Nevertheless, a significant preference for some reaction types among the database compilers (which would result in more duplication) cannot be ruled out. This holds especially for some reactions strongly represented in the ORAC databases which have not been included in our sample queries (e.g., Diels–Alder, Aldol, etc.), as they would fully dominate the result. A second concern involves the potential differences between the databases. The real relationship between reference and reaction duplication will vary at least slightly among the databases. We did not attempt to assess the influence of this factor, but we have no reasons to expect major fluctuations. In view of these uncertainties in our multiplication factor (0.44), we would prefer to present it as a rule of thumb: Approximately half of the bibliographical duplicates are also reaction duplicates. A verification of this rule of thumb could be obtained from our manual analysis of the 100 sample sets of bibliographic duplicates. This analysis gave the following statistics:

sample reaction structures fully identical to one of its duplicates	24 duplicates
sample reaction center identical to that of one of its duplicates	31(+8) duplicates
sample reaction center similar to that of one of its duplicates	17 duplicates
sample reaction different from all of its duplicates	20 duplicates

We encountered a problem with the classification of eight cases which were in fact reaction center duplicates, but one of the datacards contained more reaction steps than the other. The additional step(s) caused differences in the reaction center. Therefore, these eight cases had to be classified as

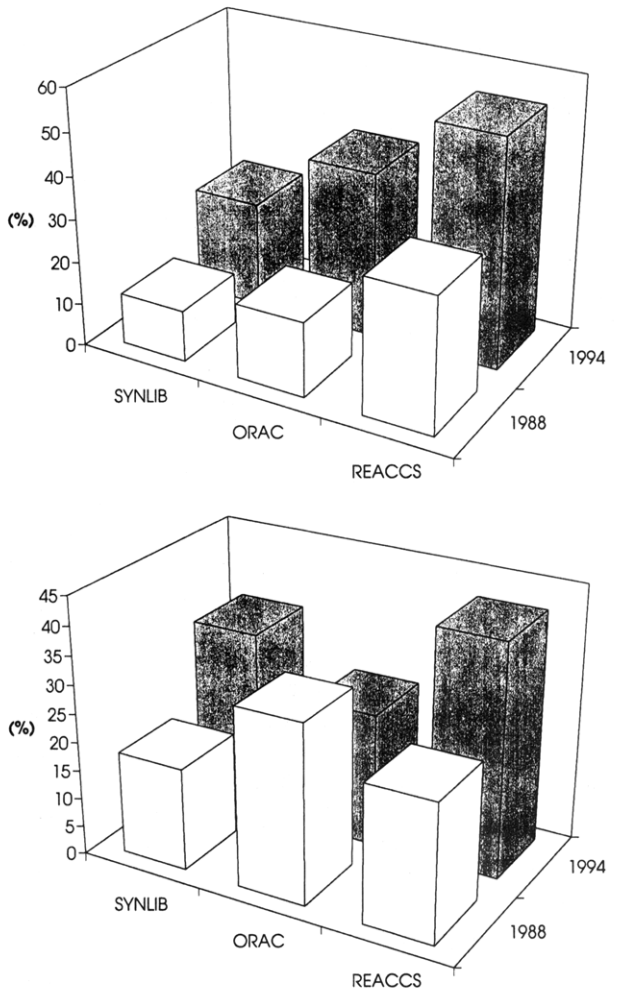

Figure 6. The database overlap (in bibliographic data) in current literature compared with the 1988 study.[15] The bars show the percentage of the references which also show up in another database. Cyclopropanation of an olefin (top). Query 2: Reduction of a ketone in presence of an ester (bottom).

different when comparing them with the results of the sample queries. This left us with the conclusion that 55% of the reference duplicates are real reaction (center) duplicates. This figure confirms the value of 44% found from the sample queries analysis sufficiently to increase the confidence in our rule of thumb.

The two independent manual analyses we performed yielded values of 44% and 55%, respectively, for the relation between bibliographic and reaction center overlap. Both methods are based on random samples from the databases, which naturally creates a certain statistical scatter. Therefore, both percentages are well within experimental error from each other, which allowed us to phrase the result as a rule of thumb: *half of the database entries with identical literature references will have the same reaction centers*. This rule suffices for our present goal: the automated assessment of the overlap caused by the addition of a new database to an existing in-house database. Excessive overlap would impair a selective database's objective, being the rapid generation of different key references to a synthetic problem. A (minor) form of database overlap missed through our analysis occurs as a result of republication of reaction data in a full paper following a preliminary communication. We do not think that this type of overlap occurs frequently enough to be a serious threat for the database efficiency.

This efficiency of the selective in-house database was our primary concern to start the overlap analyses, and we feel that we have created a useful instrument to assess this problem. An actual reduction of the overlap through the detection and elimination of individual duplicates or through selective reaction registration can obviously not be realized with our approach.

The comparison with the Borkent et al. overlap study[15] was limited to the (then) current literature files of ORAC, REACCS, and CLF, so, we had to do the same with our rerun of this work to allow direct comparison. We decided not to include MOS and CSM in the comparison, since these are more or less independent databases not directly belonging to any of the database systems. As a consequence, Boxes and ACF were used as the current literature for ORAC, CLF for REACCS (in its ORAC translation), and the entire SYNLIB database for SYNLIB. Both the results of our current study and those of the 1988 paper have been summarized in the Venn diagrams shown in Figure 5. A better insight is obtained from a graphical depiction of these results, as can be seen in Figure 6. The general trend is obviously toward increasing overlap; on average the six queries had 22% shared references in 1988 compared to 35% in our present study. Remarkably, one query showed a decreasing overlap. The overlap we find is still not outrageous (maximum 50%), given that only half of the bibliographic overlap concerns reaction duplicates according to our rule of thumb. Nevertheless, the trend toward increasing database overlap supports our concern for excessive database overlap, which demands for database comparison utilities such as we present in this paper.

CONCLUSIONS

Our results show that reference analyses can be used to assess database overlap. We are able to relate the reference overlap with the reaction overlap through our rule of thumb that half of the overlapping references have overlapping reaction centers. This paper shows the application of this method to ORAC and SYNLIB databases, but implementation of this analysis for other in-house reaction database programs should be straightforward, provided that the program is able to write references in an external (ASCII) format.

ACKNOWLEDGMENT

The use of the services and facilities of the Dutch CAOS/CAMM Center, under grant Numbers SON 326−052 and STW NCH99.1751, is gratefully acknowledged.

REFERENCES AND NOTES

(1) Blake, J. E.; Dana, R. C. CASREACT: More than a Million Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 394−399.
(2) Database accessible through external hosts, e.g., STN, based on the Beilstein Handbuch der Organischen Chemie, Beilstein Informationssysteme GmbH, Frankfurt am Main, Germany.
(3) The Beilstein database is a compound-oriented database, but its preparation field provides it with reaction database features. See, also: Jochum, C. The Beilstein Information System Is Not a Reaction Database, or Is It? *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 71−73.
(4) ORAC (Organic Reactions Accessed by Computer); Molecular Design Ltd.: San Leandro, CA, version 7.9.
(5) Johnson, A. P. Computer aids to synthesis planning. *Chem. Br.* **1985**, *21*, 59−67.
(6) REACCS (REaction ACCess System); Molecular Design Ltd.: San Leandro, CA.
(7) Wipke, W. T.; Dill, J.; Hounshell, D.; Moock, T.; Grier, D. Exploring reactions with REACCS. In *Modern Approaches to Chemical Reaction Searching*; Willett, P., Ed.; Aldershot: Gower, 1986; pp 92−117.
(8) Moock, T. E.; Nourse, J. G.; Grier, D.; Hounshell, W. D. The Implementation of Atom-Atom Mapping and Related Features in the Reaction Access System (REACCS). In *Chemical Structures: The International Language of Chemistry*; Warr, W. A., Ed.; Berlin/Heidelberg: Springer-Verlag, 1988; pp 303−313.
(9) SYNLIB (SYNthesis LIBrary); Distributed Chemical Graphics, Inc.: Meadowbrook, PA, version 3.22.
(10) Chodosh, D. F. SYNthesis LIBrary. In *Modern Approaches to Chemical Reaction Searching*; Willett, P., Ed.; pp 118−145. Aldershot: Gower, 1986.
(11) Chodosh, D. F.; Hill, J.; Shpilsky, L.; Mendelson, W. L. SYNthesis LIBrary, an expert system for chemical-reaction knowledge-base management. *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 247−254.
(12) IRDAS (ISIS Reaction Database Access System); Molecular Design Ltd.: San Leandro, CA.
(13) Barth, A. Status and Future Developments of Reaction Databases and Online Retrieval Systems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 384−393.
(14) Examples of independent vendors are Synopsis Ltd. (Leeds, England), which produces databases for protective group chemistry (PG) and current literature (MOS), and InfoChem GmbH (Gröbenzell, Germany), which distributes selections from a massive Eastern European database.
(15) Borkent, J. H.; Oukes, F.; Noordik, J. H. Chemical Reaction Searching Compared in REACCS, SYNLIB, and ORAC. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 148−150.
(16) Zass, E. A User's View of Chemical Reaction Information Sources. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 360−372.
(17) Rohde, B. Reaction type informetrics of chemical reaction databases: how "large" is chemistry? Royal Society of Special Publications [Montreux '93 Conference]; Collier, H., Ed.; 1994; Vol. 142, pp 109−127.
(18) Mills, J. E.; Baughman, B. REACCS in the Chemical Development Environment. 3. Graphically Nonequivalent Representations of Molecules and Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 431−435.
(19) In this treatment, differences in yield and reaction conditions are essentially ignored.
(20) Miller, T. M.; Boiten, J.-W.; Ott, M. A.; Noordik, J. H. Organic Reaction Database Translation from REACCS to ORAC. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 653−660.
(21) This was due to a known bug in SYNLIB version 3.22.
(22) March, J. *Advanced Organic Chemistry*, 3rd ed.; John Wiley & Sons: New York, 1985.

CI9403447