Duplications among Reaction Databases

James B. Hendrickson* and Ling Zhang

Department of Chemistry, Brandeis University, Waltham, Massachusetts 02454-9110

Received August 11, 1999

We have examined the extent of duplication in 16 reaction databases from four commercial sources, totaling nearly a million reactions, by converting them to a common format for comparison. The overlap is surprisingly small, less than 3%.

INTRODUCTION

A number of large databases of organic reactions have been compiled by different groups of abstractors over the past 2 decades. The procedure has generally been to assign current journals to a number of individual chemists who then enter into the database their choice of the most interesting or valuable reactions in each journal issue. This naturally raises the question, to what extent is there duplication among the choices made in the several different databases?

A previous effort by the CAOS/CAMM group in The Netherlands[1] approached this problem by offering three specific reaction queries to each of the three databases then available: REACCS, ORAC, and SYNLIB. For two of the queries, cyclopropanation and enolate alkylation, there was less than 10% overlap between one database and another; for the third, ketone reduction, less than 15%.

The authors observed that this relatively small overlap indicates the considerable difference in the interests and expertise of the abstractors in the three systems. They concluded that it is therefore preferable to use all three together for a complete archive. They added: "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."

In this spirit a decade later, with many more and larger databases now available, we undertook to tabulate their extent of duplication. If we find extensive duplication, we may have more confidence in using just one major database for reaction retrieval, but if the overlap among databases is small, it becomes important to use all the databases for serious retrieval of reaction precedents. We were also interested in establishing a single master archive from the whole collection without duplication.

Rather than just comparing an arbitrary set of particular reaction queries, we elected to compare the entire collection entry by entry for a more accurate tabulation. This has only now become possible by using the COGNOS system for indexing and retrieving reactions.[2,3] This system allows us first to convert all the entries from the several databases into the same COGNOS format so that they may be directly compared.

We had available 16 databases from four commercial sources, totaling nearly a million reaction entries and summarized in Table 1. The four databases from InfoChem[3] in Germany are ChemSynth (CS) and three ChemReact (CR) databases. The collection from ISI[5] includes Current Chemical Reactions (ccr) from 1986 to 1993 and separate annual collections from 1994 to 1997. The major collection from MDL[6] is RefLib, which contains Theilheimer and some current literature files (CLF) closed in 1991. The other MDL databases are a set of current synthetic methods (csm), a set of current heterocyclic chemistry (chc), the *Journal of Synthetic Methods* (jsm), and the *Organic Syntheses* set (OrgSyn). The fourth collection, from Synopsys[7] in England, includes biocatalytic reactions (bcrx) and a set of protecting group reactions (pgroup).

METHOD

To appreciate the comparisons, it is important to understand how the COGNOS system identifies reactions. Unlike the common practice in other systems, reactions here are not identified by the structures of substrate and product. Rather the identification here is dynamically based on the net bonding *change* at the reaction center, i.e., just the atoms which change their bonding in the reaction. Reactions are indexed taxonomically in five nested levels of generalization, so that matching precedents can be sought at various levels of similarity. These levels are as follows, in increasing refinement of detail.

(1) Thirteen broad reaction classes are defined first, to distinguish the following: single and multiple constructions; refunctionalizations at one carbon, more carbons, or heteroatoms; fragmentations; rearrangements; and several minor special reaction types.

(2) Bonding to carbons is generalized[2] to distinguish four types of bonds: σ- and π-bonds to other carbons and to more and to less electronegative atoms. This affords a simple digital description of each carbon so that the reaction change is then identified by the numerical product−substrate difference in bonding type at those carbons that change. Generalized in this way[2] all possible reaction changes on up to eight linked carbons are uniquely characterized by a simple binary number.[8] This number is the reaction type.

(3) The immediate environment of the changing carbons in the substrate further distinguishes the entries within each reaction type. These are also treated in the same generalized terms—number of attachments of each reacting carbon to other carbons, heteroatoms, π-bonds, or electron-withdrawing

10.1021/ci990100s CCC: $19.00
Published on Web 01/15/2000

Table 1. Distribution of Database Entries

	InfoChem				ISI					MDL					synopsys		
	CS	CR1	CR2	CR3	ccr86-93	ccr94	ccr95	ccr96	ccr97	RefLib	csm	chc	jsm	OrgSyn	bcrx	pgroup	sums
<1974	201	745	1	2	0	0	0	0	0	46 945	0	22 820	7	2994	2 319	4 168	80 202
1975	1639	6 452	0	0	0	0	0	0	0	3 148	0	1 963	1	36	255	429	13 923
1976	1480	5 593	0	0	0	0	0	0	0	2 771	0	2 093	2	43	170	518	12 680
1977	1645	6 373	23	0	0	0	0	0	0	3 078	0	1 894	1	27	151	613	13 805
1978	2016	8 306	28	0	0	0	0	0	0	3 977	0	2 051	41	39	147	668	17 273
1979	3832	14 156	40	0	0	0	0	0	0	2 927	0	2 226	472	20	167	892	24 732
1980	4899	17 009	100	0	0	0	0	0	0	3 848	0	2 060	3 292	20	194	731	32 153
1981	5170	13 747	3 808	2	0	0	0	0	0	5 038	0	1 850	2 930	50	269	953	33 817
1982	5377	1	17 972	0	0	0	0	0	0	4 293	0	540	3 430	19	254	839	32 725
1983	6364	0	20 636	5	0	0	0	0	0	6 160	0	53	3 026	7	255	878	37 384
1984	6840	0	20 445	567	0	0	0	0	0	10 564	0	0	3 247	39	322	864	42 888
1985	6623	2	7 322	13 052	8 921	0	0	0	0	6 557	0	2	2 910	3	373	859	46 624
1986	6127	0	7	18 916	22 493	0	0	0	0	8 292	0	9	2 556	49	419	1 011	59 869
1987	7461	0	0	23 886	18 434	0	0	0	0	8 217	0	15	2 927	56	566	1 019	62 581
1988	9300	0	7	12 894	15 748	0	0	0	0	9 534	0	0	2 848	263	610	1 038	52 242
1989	9395	0	4	3 011	16 643	0	0	0	0	9 076	0	1	2 953	67	662	1 051	42 863
1990	8078	0	0	1 243	17 871	0	0	0	0	6 893	200	0	2 631	60	863	886	38 725
1991	2211	0	0	0	16 607	0	0	0	0	4 064	2 600	0	2 677	85	1 114	1 069	30 427
1992	0	0	0	0	18 536	1	22	0	0	0	5 062	0	2 837	58	1 309	996	28 821
1993	0	0	0	0	6 547	4 752	5 560	0	0	0	5 276	0	2 875	0	1 291	1 153	27 454
1994	0	0	0	0	0	14 709	5 481	2	0	0	7 414	0	2 552	0	1 242	1 216	32 616
1995	0	0	0	0	0	0	16 163	4 968	104	0	5 629	0	2 847	0	1 418	1 166	32 295
1996	0	0	0	0	0	0	0	17 353	4 877	0	5 547	0	1 969	0	1 630	1 228	32 604
1997	0	0	0	0	0	0	0	0	5 874	0	3 832	0	0	0	475	638	10 819
None	6	57	0	0	45	0	7	7	0	0	0	0	0	0	0	0	122
sums	88 664	72 441	70 393	73 578	14 1845	19 462	27 233	22 330	10 855	145 382	35 560	37 577	49 031	3935	16 475	24 883	
totals	305 076				221 725					271 485					41 358		839 644

groups, all characterized by a digital code, which extends the binary number for the reaction type.

(4) The nature of the groups lost or gained in the reaction is also digitally described and is itself capable of several levels of refinement. These first four levels of similarity concern only the reaction center, i.e., those carbons that change their bonding.

(5) The generalized nature of groups and features elsewhere in the molecule, and unchanged in the reaction, is coded last. A simple list of 1024 features is used to code the unchanging parts of the reacting molecules. These include not only obvious functional groups, such as ester, halide, and nitro, but also general features, such as unsubstituted or variously substituted aromatic rings or heterocycles.

This taxonomic classification of reactions allows any reaction entry in any database to be defined by five digital codes for increasing similarity: reaction class; reaction type; environment of the reaction center; groups lost or gained; and features unchanged in the reaction. Any two entries with all five codes the same will be extremely similar but not necessarily identical, since homologous series and simple regiochemical differences, for example, are not distinguished. Nevertheless, in the search for reasonable precedents for a query reaction, these final differences are relatively unimportant.

A special advantage of the COGNOS retrieval system in actual use is the possibility of tuning the extent of similarity between the query and the hits to afford either more or less closely refined pruning. Thus the controls for using COGNOS allow the user to specify the desired detail of matching at each of the five levels. The response of the program to these choices is essentially instantaneous because the index simply groups the reactions together by their digital codes. Therefore, for the user tuning to a manageable number of hits is fast and facile, and this is also provided automatically by the program.

To identify duplicates fully between any two database entries, we must find not only a complete match of all five digital codes for the chemistry but also the same text reference. While the format for the chemistry codes which define the reaction is the same for all the databases once they are classified by COGNOS, the text formats for the references are not uniform among the several databases. After some exploration of their various formats we were able to establish five text fields accessible in all databases: author, journal, year, volume, and page. Some of these reference items are presented differently in different databases, however, and cause problems for comparison.

Some databases quoted all authors, others no more than three, and others only the first author. The author names were variously found with first names or first initials only, and these either before or after the last name. Accordingly, for uniformity we compared only the last name of the first author. The journal names were sometimes abbreviated in different ways, so only the first initials of the journal name were compared. Even this may be ambiguous, however, since a paper may be quoted as *Ber.* in one and *Chem. Ber.* in another, or *Ann.* against *Liebigs Annalen.*

The volume number was sometimes ambiguous since some individual entries recorded an issue number instead, and in journals for which the publication year suffices for the volume number, the page number was sometimes entered erroneously in the format location for volume number. Much of the latter problem was avoided by accepting only volume numbers smaller than 600.

Ultimately, only three text fields are quite unambiguous: last name of the first author; year; and page. We made the pairwise comparisons both with the five reference text fields and with just these three, but the differences between the two modes were quite small, as discussed below.

Finally, a significant number of entries quoted more than one reference; the maximum number we found was 22. In

Table 2. Pairwise Duplication in the 16 Databases

	InfoChem				ISI					MDL					Synopsys	
	CS	CR1	CR2	CR3	ccr86-93	ccr94	ccr95	ccr96	ccr97	RefLib	csm	chc	jsm	OrgSyn	bcrx	pgroup
CS	−	57	50	58	0	0	0	0	0	2882	12	468	1970	0	8	182
CR1		−	0	0	0	0	0	0	0	558	0	765	320	1	0	33
CR2			−	0	2	0	0	0	0	785	0	73	503	0	1	43
CR3				−	970	0	0	0	0	1062	0	5	374	0	6	36
ccr86−93					−	0	0	0	0	570	94	0	218	0	113	120
ccr94						−	0	0	0	0	38	0	10	0	26	9
ccr95							−	0	0	0	62	0	31	0	47	29
ccr96								−	0	0	23	0	13	0	47	18
ccr97									−	0	26	0	1	0	23	19
RefLib										−	45	1509	4056	62	189	731
csm											−	0	1757	0	127	512
chc												−	191	5	0	42
Jsm													−	3	90	536
OrgSyn														−	0	2
bcrx															−	146
pgroup																−

Summary[a]

	InfoChem	ISI	MDL	Synopsys		InfoChem	ISI	MDL	Synopsys
InfoChem	165 (165)	972 (896)	9777 (8826)	309 (303)	MDL			7655 (6369)	2229 (2015)
ISI		0 (0)	1086 (971)	451 (373)	Synopsys				146 (134)
$\Sigma = 22\,790\ (20\,052)$									

[a] Comparison by three reference text fields (parentheses for five fields).

such cases *with the same chemistry* it was necessary to compare all quoted references in one entry with all in another; then if any two are identical the entries are labeled as duplicates.

Patents constitute a special kind of reference, not covered by the text fields chosen above, and not comparable with publication entries. The patent number is given, but rarely the year, and in a number of cases the same reaction is presented by the same authors in more than one patent. Fortunately, the proportion of patents in the whole collection is quite small, only 18 273 entries (2.1%), so we elected to compare them separately and not include them in the overall database comparisons.

The procedure for comparison has two parts. First, each COGNOS entry in the database is stripped to just its five chemistry codes above and its five reference text fields. The entries are then sorted taxonomically by the five digital codes for the chemistry, in effect creating a new list of entries, reduced and ordered. This stripping and sorting procedure prior to the comparisons makes the pairwise comparison of the database lists much faster.

In comparing two databases the comparison is made first with reaction class, then with the next three chemistry codes as a unit, and last with the unchanged features code, aborting the matching as soon as one is different. Only if alike in all five chemistry codes are the entries further compared for identity of the reference text fields.

RESULTS AND DISCUSSION

The original 16 databases examined have a total of 1 075 484 reaction entries. However, the number classifiable for COGNOS is only 857 917. Commonly the missing reactions cannot be classified because of incorrect mapping of the atoms between substrate and product. This mapping was often obtained originally by an automatic program instead of input by the abstractor. Frequently a simple reactant, like methyl iodide, for example, is only annotated on the reaction arrow and not mapped into the product. In unmapped cases such as this, COGNOS classification cannot parse the overall reaction change.

When the patent entries are also deleted, the total number of database entries to be compared becomes 839 644. These are sorted by their publication year in Table 1. In 122 entries no publication year was recorded ("None" row).

In the COGNOS system the whole rationale of the five nested generalizations of chemical identity above was to facilitate matching literature precedents with a reaction query at several levels of similarity. The closest possible match, at all five levels, is used here to identify "duplicates", but two such reactions may still not be exactly the same. Even when the reaction center is the same in both entries, there may be variations in the rest of the molecule, which remains unchanged in the reaction. The identity codes for the unchanged features of the molecule (code 5 above) will distinguish entries with substituent halogens, methoxy, etc., as nonidentical reactions, but will not distinguish isomers or homologues.

When one reference presents a number of variations on a single reaction type, the abstracter may just make a single database entry from it, or he may make a separate entry for each example. In cases for which the only variations are undistinguished homology, an entry in one database will be found to be "identical" with several in the compared database. In these instances *only one such pairing* is counted as a duplication; the others are recorded as different. This is illustrated by the reaction[10] in eq 1, in which the ChemReact3 entry no. 15069 for $n = 3$ is recorded as duplicated by all six ccr86−93 entries (the odd numbers from nos. 9391−9401) for values of $n = 3{-}8$.

$$2Bu_3SnCl + HS(CH_2)_nSH \rightarrow Bu_3SnS(CH_2)_nSSnBu_3 \quad (1)$$

The results of pairwise duplication among the 16 databases are displayed in Table 2. It will be apparent from Table 1 that a number of cells in the Table 2 array must be zero since there is no overlap of their reference years. Of the four sources[4−7] the ISI databases have no internal duplication since they were established as a consistent abstraction of current literature by year. By contrast, the MDL databases are a disparate collection and exhibit much overlap of entries among themselves.

The number of duplicate pairs shown in Table 2 is surprisingly small considering the size of the combined databases. The total of 22 790 pairwise duplicates is only 2.7% of the 839 644 entries overall. When the pairwise duplications are assessed separately by reference year, i.e., Table 1, the percentage of duplications varies from 1.02% (1997) to 4.33% (1980), averaging at 2.7%.

These duplicates are identical in both their chemistry codes and the three best reference fields discussed above, i.e., without journal name and volume number. When all five reference text fields are included in the comparisons, the number of duplicate pairs is only a little less at 20 052. The duplication among the four abstracting sources[4−7] is summarized at the bottom of Table 2 for just three reference text fields. The tighter comparisons with five text fields are included in parentheses.

The very low proportion of duplicates, even less than earlier estimated by the Dutch group,[1] strongly supports their conclusion that any broad search for precedents for a query reaction should cover all available databases.

If the four abstracting groups showed so little overlap, each was implicitly passing over the choices of the others. This in turn suggests that there are many reactions not abstracted at all from the major literature. However, it is not clear that there is any reasonably accurate way to estimate the total number of reactions reported in any given year. If we could know this number, we could assess just how extensive these abstracting operations really are. The very low overlap among four groups, however, suggests that the true overall coverage of the literature here is not very complete, even with all the databases taken together.

Comparison among just the 18 273 patents revealed 2249 pairs which duplicate only the chemistry code. In fact patent entries were only found in the ISI databases and the bcrx from Synopsys. At 12.3% duplication this represents a considerably higher overlap for patents than for the publications in Table 2; however, there are only chemistry comparisons here and no reference texts. The only reference text field available for comparison in the patents is the first author's last name. When this criterion is added, the number of duplicate pairs is reduced to just 377, and this is 2.1%, the same overlap as found in the publications in Table 2.

It would be advantageous to combine these databases into a master reaction archive without duplicates, but this would entail satisfying the proprietary concerns of the four commercial sources.[4−7] We found that such an archive would consist of 817 931 unique reactions, mainly over the period 1975−1997. Subtraction of the 22 790 pairwise duplicates from the total of 839 644 classifiable reaction entries would leave only 816 854.

The total of 817 931 unique reactions is somewhat different from this subtraction of pairwise duplicates owing to the effect of some instances with three or more identical entries. Thus, if three entries are the same, this counts as three pairwise duplicates but only two would be subtracted from the total in assembling the master archive. The difference between the two totals suggests that there are about a thousand instances of these multiple identities, composed mostly of three entries alike.

CONCLUSION

We have examined a collection of 16 reaction databases compiled by four different agencies[4−7] in order to find the extent of duplication among them. The main and surprising result is that so few common entries were abstracted by the four groups: there were only 22 790 duplications among 839 644 reaction entries overall, i.e., only 2.7%. Of these, only about a thousand instances represent multiple identities more than just duplication.

The two fold implication of this result is that any search for reaction precedents will very likely be seriously incomplete if conducted with only one database, and conversely a search with all of these databases will probably fall substantially short of the total literature for the period.

ACKNOWLEDGMENT

We are grateful to the Environmental Protection Agency for support of this work through EPA Grant No. R825329. We also wish to express our thanks to Dr. Camden Parks for guidance and much helpful discussion.

REFERENCES AND NOTES

[1] Borkent, J. H.; Oukes, F.; Noordik, J. H. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 148−150.

[2] Hendrickson, J. B.; Sander, T. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 251−260; *Eur. J. Chem.* **1995**, *1*, 449−453.

[3] The name COGNOS was introduced in the original paper[2] and is used here, but cannot be used commercially because of its prior registration as a proprietary name for another software product. It is currently on the web, somewhat modified, as *www.WebReactions.net*.

[4] InfoChem GmbH: Munich, Germany (*www.springer.de/newmedia/chemist/infochem*).

[5] Institute of Scientific Information: Philadelphia, PA (*www.isinet.com*).

[6] Molecular Design Ltd.: San Leandro, CA (*www.mdli.com*).

[7] Synopsys: Leeds, England (*www.synopsys.co.uk*).

[8] Reactions are sharply defined as "unit reactions", which have only a single exchange of one bond for one other at each changing atom. The binary identification number is unique for all unit reactions and for all composite reactions of two overlapped unit reactions.[2] Over 90% of all database reactions are either unit reactions or composites of two over the same site.[9]

[9] Hendrickson, J. B.; Miller, T. M. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 403−408.

[10] Harpp; et al. *Tetrahedron Lett.* **1986**, 441.

CI990100S