

Integrated Chemical-Biological-Spectroscopy-Inventory-Reactions Preclinical Database¹

SANDOR BARCZA,* HENRY W. MAH, MIRINISA H. MYERS, and SIEGFRIED S. WAHRMAN

Sandoz Research Institute, East Hanover, New Jersey 07936

Received March 13, 1986

An integrated in-house preclinical research and development database (DB) was created containing ~28 000 chemical structures and ~256 000 administrative, physicochemical, chemical, and biological filled datafields, and it is growing, using the MACCS system. The design converts the "flat" DB into a multidimensional hierarchy. A ¹³C NMR retrieval module correlates canonical atom numbering with chemical shifts. An administrative datatype (DT) for material quantities submitted forms the basis of a compact drug inventory system, where subsequent transactions can be recorded. Biodata are organized into a six-level hierarchy of zoned data fields. Chemical data entered allow simple reaction storage and retrieval, via reagents and reaction conditions stored as searchable text and by linkage of starting materials, intermediates, and products via their registry numbers and structures. Self-documenting is incorporated in datatype definitions under the respective datatypes. Some literature references are stored in a compact form. Text, numeric and formatted data, and structures are searchable for exact matches, range of values, substructures, etc. Consolidating the structures and various kinds of data into the same DB has multiple advantages: (1) all data are stored in one DB; (2) coherence and intercorrelation are achieved; (3) all data are correlatable with structures; and (4) structures are stored only once.

INTRODUCTION

The storage, retrieval, and correlation of chemical structures and associated data are essential for organizations that research and develop biologically active substances. Sandoz made a thoughtful choice of storing these together. The MACCS^{2,4} database (DB) management system was chosen and installed in 1981 for this purpose. This paper is *not* a review of the MACCS system,^{4,5} but of nonobvious extensions and valuable practical applications. We are especially prompted to disclose these extensions because it appears that the program is underutilized, relative to its potential, at other installations. This review should facilitate enhancements and should provide synergistic feedback between users.

Our aim was to create as much of a "one-stop DB" for preclinical research and development as possible. This required compromises between opposing criteria. The chosen configuration is what we believe to be the optimized balance for current and near-future needs.

The components of this decision are (1) Specialized DBs exist, for example, REACCS⁷ for chemical reaction retrieval; spectroscopic files,^{8,9} as in the NIH-EPA CIS (Chemical Information System);⁹ inventory systems; bibliographic DBs; and many other systems with rapidly searchable inverted files. (2) Most, if not all of these systems are at extra cost. (3) Interfaces would have to be developed. (4) Redundant storage and overhead would occur for correlators, cross-references, or pointers. Especially sizable duplication of storage would occur for structures in separate reaction and spectroscopic DBs and for administrative data in a separate biological DB.

Our decision was to minimize overhead, artificial barriers, incompatibility, storage, and expense. Therefore (1) Should multiple DBs be needed, MACCS should be used. (2) Information used together should be stored together, especially structures, biological activities, synthesis, and spectra. (3) Within the DB, as much data should be put into a datatype (DT) as is feasible in order to increase the coherence and the "AND" searchability of pieces of multilayered information. The number of compounds at Sandoz and the volume of data are such that the consolidation of most preclinical chemical-biological data into a single MACCS DB is feasible and effective.

MATERIALS AND METHODS

Hardware used is a Prime 550 superminicomputer.

Software MACCS, Molecular Access System,^{2,4-6} and

DATAACS, Data Acquisition System.¹⁰ Structures are formula-, substructure-, and stereoisomer-searchable by MACCS.

DTs are either "fixed" or "flexible". The former are created at the time of DB creation, are compact (1 line, 1-20 characters), and fast-searching, and are displayed with the structure. The latter can be created any time and may hold 300 lines × 120 characters per compound per DT. DTs can be text, formatted, or numeric with text comment.

Hardcopy reporting from a DB may occur via a data file, called "datfile", via captioned structure plots, or with DATAACS forms, which can be custom configured and edited/updated. Display, transfer, and report forms designed with DATAACS deliver structure(s) and data in various flexible layouts; they can also be used for prompted (menu-driven) data input.

Vocabulary. The contents of DTs in the DB range from strictly controlled format to free text, depending on a variety of factors, and include the following examples.

1. Registry number: The form of the external registry number is strictly controlled; this information is frequently searched, and no false positives or negatives must exist.

2. Requested biological testing: The abbreviations are in-house standards for routine tests; correctness is enforced by the compound registrars. An occasional unusual test name or comment is accepted.

3. Reagents and reaction conditions for synthesis of compound: This contains only loosely controlled chemist's slang. It is less frequently searched; the searcher has to think of synonyms. For example, unless the reagent is registered by structure in the DB, the searcher has to think of searching for the string "benzyl", Bz, Bnz, PhCH₂ to locate compounds synthesized with benzyl halide reagents.

4. "Note": The vocabulary for comments is unrestricted.

RESULTS AND DISCUSSION

Spectroscopy Data Storage, ¹³C NMR DB. General provision was made to accept spectroscopic information, and, specifically, a numeric DT was created to store ¹³C NMR shifts.

After registration of a structure into the DB, MACCS provides a canonical atom numbering according to the Sema algorithm.¹¹ This numbering, once correctly generated, has been guaranteed by MDL (Molecular Design Ltd.) not to change, is unique and unambiguous, and can be optionally displayed.

The ^{13}C NMR shift DT contains the Sema atom numbers in the first column and the corresponding ^{13}C chemical shifts in the second column, followed by optional comments. The comment may show solvent, instrumental conditions, coupling constant, and uncertainties of assignment. Interchangeable carbons receive the same letter as comment; e.g., one pair of carbon atoms with similar shifts is labeled A, another pair or group, B, etc.

Typical data for the ^{13}C NMR shift are shown below:

```
> 12295 (SAH-053563) (CMR.SHIFTS) DT0016
20.0-130.1
18.0-129.2
16.0-126.7
15.0-129.2
10.0-161.7
5.0-99.4
2.0-175.7
1.0-43.1
4.0-25.9 A
9.0-22.9 B
12.0-23.3 B
7.0-27.5 A
3.0-36.0
8.0-174.6
```

Input. The registered structure is displayed with atom numbering. The numbering is noted on the spectrum (structure or peaks) or other laboratory record, and then the pairs of atom numbers and chemical shifts (plus comment) are entered for the compound. Currently about 2000 spectra are stored in the main DB, with nearly all carbons assigned.

Auxiliary information entered is sample number, spectrum number according to an overall lettering scheme for extension to other spectral analytical results, and NMR literature reference.

Searching. Typical searches are of two types, shift-originated and structure-originated.

To find compounds having a peak at a certain chemical shift (window), DT CMR.shifts is searched for the string /limit 1-limit 2.

To hit compounds having each of several shifts ("AND" search), the active list of hits from the first range search is made the reference list for the next one, and so on, thus providing an *intersection* of sets.

"OR" searches for chemical shift ranges are done by repeated range searches using the *same* reference list of compounds and then merging the hitlists, thus creating a *union*.

To search for shifts of a single structure, the user draws the structure, does a "find current", and inspects or reports the data.

To find chemical shifts for a certain substructure or carbon type, the user draws a substructure, specifying all the neighbor and bond requirements, atom lists, etc. (ring/chain), does a substructure search, and then inspects/reports the shifts for the compounds hit.

Various combination searches are easily performed, including searches of the other DTs. For example, the structure and shifts can be found for the compound that (1) was taken from the literature (has NMR ref), (2) has a peak in a certain shift window, (3) is an indole alkaloid (substructure = indole), (4) was authored by Wenkert, (5) is in journal JACSAT, and (6) is in the year range 1979-1986.

Reporting Results. The atom-numbered structure and the atom number-chemical shift data can be reported with MACCS by plotting (an) atom-numbered structure(s) and printing a data file (datfile) out of the DB.

More conveniently, a DATACCS form reports the atom-numbered structure and the adjacent shift data either as a

hardcopy plot or on the terminal (Figure 1). Thus, a spectroscopic DB within a main DB was realized, essentially utilizing only a single DT and no duplicate storage of structures.

Integration of the spectroscopic data into the main DB has synergistic benefits: the users find not only the chemical shifts but also chemist's names, solubility, melting point, warnings, and other information on the compound, as needed.

An additional value of the integration of spectroscopic, structural, and biological information is that they can be most conveniently submitted to QSAR (quantitative structure-activity relation) analysis. ^{13}C NMR shifts reflect atomic charges, which may correlate with bioactivities.

Synthesis (Reaction) Information Storage and Retrieval. Reaction documentation and retrieval is a rapidly growing field, and programs,^{7,12-14} e.g., REACCS,⁷ exist, specifically designed for the purpose.

We created within the main MACCS DB a useful reaction DB. This, with very economical means, offers many of the capabilities of a specifically designed reaction search system. The key element is a one-way, directional *link* between compounds transformed.

Implementation. A numeric DT was created called "Made.from.Regno" (RN). For each compound, the internal registry numbers of its one to two first-generation precursors, i.e., mother, and father if there is one, are entered as the data into DT made.from.RN. For starting materials a RN of 0 is entered. As much as feasible, compounds are registered into the DB in the order of synthetic sequence. Simultaneously, there is a commitment—easily fulfilled—not to change the position of compounds in the DB.

Reagents and conditions of the synthesis are entered into the comment portion of the DT made.from.RN as text. Substeps are labeled A, B, C, etc. Starting materials have the comment "Start". Some examples of made.from data are shown below:

```
> 20010 (MADE.FROM.REGNO) DT0025
0.0 START
> 20011 (MADE.FROM.REGNO) DT0025
20010.0 Me3SiCl Et3N
PENTANE-ETHER
> 20012 (MADE.FROM.REGNO) DT0025
0.0 START
> 20013 (SAH-061874) DT0025
(MADE.FROM.REGNO)
20011.0-20012.0 A) cat NaBr Et3N
CH3CN reflux 16 hr B) H2O
```

Although a chain of made.from linkages specifies even a long multistep synthesis completely, a convenient review of multistep syntheses is made possible by expressing the *summary of synthesis* in a text DT. These data consist of registry numbers, arrows, and "+" signs and parentheses as necessary, and they list the synthesis path from starting material to end product.

Since both the made.from and sum.synth DTs accept 300 lines of data, multiple paths to a compound are easily entered. Entering synthetic sequences into the DB is much less work than registering the same number of unrelated compounds, since the structure diagrams can be derived from the predecessors by small modifications.

Synthesis literature reference DT was created at the request of users, but the chemists generally do not supply this information.

Searching. Although data are entered only as "from" information, searches in the *forward and retrosynthetic* sense are both feasible, because of the vectorial nature of the information. The following are some typical searches (Figure 2):

INT.REG.NO 20880	SAH.NO	SALT CODE	CHEM.NO	UNIT	C M R . S H I F T S COMMENTS: J'S, INTERCHANGES, SOLV. FIELD, COND'S 18.0 - 22.7 16.0 - 172.3 7.0 - 52.9 11.0 - 169.9 3.0 - 27.3 1.0 - 109.4 4.0 - 122.8 A 5.0 - 136.1 9.0 - 111.3 13.0 - 121.9 A 10.0 - 119.4 6.0 - 118.2 2.0 - 127.5
MOL NAME CMR		CMR SPLE			
C14 H16 N2 O3		SPECTRUM.NO 000094-C	INIT	CMR PYTCAS-018-001869-79 WENKERT	
M W	N O T E CMR DATA LIT	DATE MAY 02, 1985			

Figure 1. ^{13}C NMR shift report. The essence is the canonically numbered structure and real number pairs: each atom number and its chemical shift. Various auxiliary data are accommodated by the form, notably a compact literature reference at the lower right.

What was compound X made from? Simply find DT made.from for compound X.

What compounds were made from X? Search the DT made.from for all compounds for the string [RN of X]/ and then for /[RN of X]. [Number/ searches in first column (=“mother”), and /number searches in second column (=father). Most compounds in our DB have one parent only. Such entries are hit by searching in either column.

Was compound X ever made from Y? Find the made.from data for X and see whether the RN of Y is listed.

In generic searches, searches for the synthesis of *types of products*, a substructure search is done, the made.from data are inspected, and then, if necessary, also the structures for the compounds whose RNs (RN) appeared in the data, are also searched. Searches of this type are aided by converting a list of *RN data* to an ASCII RN listfile. This was accomplished by using DATACCS and global editing. Work has succeeded in automating such processes using command files.

A converse search is preceded by a substructure search for the generic precursor.

Generic reaction search is feasible via substructure search for generic product, conversion of the RN in the found made.from data to a precursor listfile, and substructure search of compounds in this listfile. This search will include false positives because the product substructure and precursor substructure may be both incidentally present in product and precursor, respectively, but not converted from one to the other in the reaction. Work is in progress to test the use of “reaction centers” in MACCS to refine this aspect. However, the existing setup *will not miss any cases sought* (no false negatives).

Reagents and conditions used can be searched or inspected for any of the cases hit. Searching for all reactions that use a particular reagent is done by a string search of the text

```

DMODE: Search datatype(s) = MADE#
Search string: 20212/
> 20213 (SAH-061955) <MADE FROM REGNO> DT0025
20212.0 PYRIDINE, THEN ACETYSALICYLIC ACID CHLORIDE IN THF

> 20222 (SAH-061620) <MADE FROM REGNO> DT0025
20212.0 (A)DMF, IMIDAZOLE, T-BU-DIME-CLISILANE (B)DMF, 2NH, 2-BENZYLBR

> 20224 (SAH-061961) <MADE FROM REGNO> DT0025
20212.0 , THF, ET3N, 2 PHENYLACETYL CHLORIDE THEN REFLUX

<INTERRUPT>
DMODE: Search datatype(s) = 26
Search string: 8202128
> 20213 (SAH-061955) <SUM. SYNTH> DT0026
20212-->20213

> 20223 (SAH-061960) <SUM. SYNTH> DT0026
20212-->20222-->20223

> 20224 (SAH-061961) <SUM. SYNTH> DT0026
20212-->20224

> 20679 (SAH-062131) <SUM. SYNTH> DT0026
20212-->20590-->20584-->20678-->20679

<INTERRUPT>
DMODE: Search datatype(s) = 25
Search string: SIL
> 20182 (SAH-061946) <MADE FROM REGNO> DT0025
1963.0 REACT WITH SILA ACID CHLORIDE

> 20222 (SAH-061620) <MADE FROM REGNO> DT0025
20212.0 (A)DMF, IMIDAZOLE, T-BU-DIME-CLISILANE (B)DMF, 2NH, 2-BENZYLBR

> 20461 (SAH-062055) <MADE FROM REGNO> DT0025
20473.0 DIMETHYLDIMETHYLDICYSILANE, BULI, THF, N2, ODE9

> 20485 (SAH-062071) <MADE FROM REGNO> DT0025
20344.0 , TRIMETHYLSILYL BROMIDE, R. T., 3HR., DIISOPROPYLAMINE

<INTERRUPT>

```

Figure 2. Examples of searching the synthesis (reaction) data. In the first example, compounds are found that were made from compound 20212 in one step (20212 is the mother). The second search finds the compounds that originate from ancestor 20212 in any number of recorded steps. The last example is a search for silicon-containing reagents by text searching the reagents and conditions comment of the Made.from.RN DT for the string “SIL”. (Each search was truncated for brevity.)

comment for the name of reagent and likely synonyms.

The summary of synthesis data can be searched in one step for a registry number to find all compounds that were synthesized from the query compound in any recorded *single or multistep* reaction sequence. This capability surpasses even

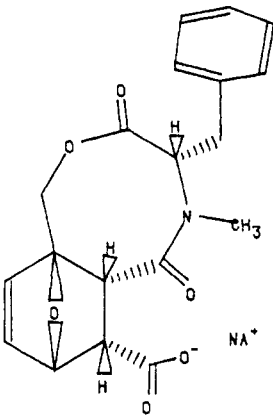
INT. REG. NO	SAH. NO	SALT CODE	CHEM. NO	SUBMITTED	UNIT	CHEMISTS	DISCL. NO	
25000	EXM-300001	NA	9999-888-44	06-28-84	BAR	BARCZA THIEDE	999-82	
<div style="text-align: center;">CHIRAL</div> 						KNOWN?	LAD. NO	
						N		22999.0
						MP		82.0 - 84.0 SINTERS AT 81
						BP		185.0 - 191.0 EVAPORATIVE
						PRESSURE		0.08 - 0.07
						OTHER. PHYS. DATA		
						ALPHA-D -86DEG. 5X. PY1		
						SUBLIMES 70-75DEG. 0.05MM		
						SOL. CODE		C. >20/E
						DETAILS		FOR PL. MAKE 10-X SOLUTION IN ACETONE. ADD TO 80-X WATER 20-X DMSO THEN STRIP OFF ACETONE IN VACUO. <40DEG.
SCREENS								
AO HG PL (PAF) TC*								
GHI AM/AV TR								
NOTES SEE LONGNOTE								
ABS CONF								
PLEASE REFRIGERATE. APPROX. 5X OF AN ISOMER MAY BE PRESENT.								
PLEASE INFORM BERNE OF AVAILABILITY OF THE COMPOUND.								
AMOUNTS. MG						COMPARE WITH		
5500.0						*300002		
40.0 - 5500.0 BERN. DR. FAUSTUS								
C19 H18 N NA O6		MW	FICTITIOUS, EXEMPLARY COMPOUND, ABS CONF					
		379.348						
MADE FROM. REGNO		24999.0 A) IN D1-1PROPYLBENZENE 160DEG OVERNITE. B) NaOH. C) HCL. TOLUENE. D) CHROM. E) NAHCO3						
SUM. SYNTH		24995+24996-->24997+24998-->24999-->25000						

Figure 3. "Chemical Information Sheet", the most important nonbiological report, has administrative (top), physicochemical (upper right), chemical (lower left), bio-test (lower middle right), and inventory (lower right) regions.

the current commercial reaction DBs.

Reporting synthesis information can be done by (1) printing a data file of the made.from and/or sum.synth data. (2) The official "chemical information sheet", as generated by DATAACCS, shows the structure, the made.from.RNS with reagents and conditions for the last step, and the summary of the multistep synthesis for the compound submitted for biological testing (Figure 3). (3) A special small DATAACCS form shows the product structure, an arrow leading from the precursor RN(S) to the product, and the reagents and conditions are given under the arrow (plus auxiliary information). (4) These forms can be stacked for all the compounds in the multistep summary of synthesis to display a linearized synthetic scheme. Registering compounds in the same order as the synthesis is advantageous since DATAACCS (just as MACCS) sorts compounds for display by registry number. The stacked plot will automatically be correctly sequenced. If the synthesis is branched, the branch compounds are "squeezed" into the linear sequence according to their RN position. An example of the DATAACCS synthesis report is shown in Figure 4. (5) For displaying an isolated reaction, DATAACCS is best used to show both the precursor and product structures, connected by the arrow and the made.from information. More elaborate (branched) displays can also be designed. The one shown is used in our regular production of synthesis sheets.

Recent work enhanced the searching and reporting of reactions via command files: (1) Given a listfile of compounds (products), a single line command by a user will cause the creation of a listfile of precursors (parents). (2) Given a listfile of products, a single command will cause the plotting of synthesis sheets for all the products with structures of precursor(s) (parents) and product, made.from synthesis infor-

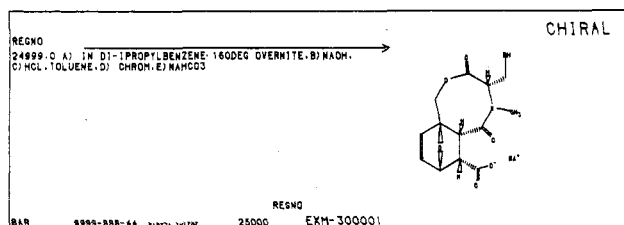


Figure 4. Synthesis step report form shows structure, etc. of product, registry numbers of 1-2 parents, reagents, and reaction conditions under arrow. Such reports stacked vertically outline a linearized synthesis pathway.

mation, and a summary of synthesis and other data (Figure 5). (If no father exists, the structure of the mother is repeated.)

The value of the foregoing is detailed in the following: (1) With no additional investment in software and only the creation of 1-3 DTs, we realized with ~20% effort ~80% of the benefit of a full-fledged specialized reaction retrieval system, and in some respects surpassed it. (2) There are about 5000 reaction entries currently in our DB; these provide a rich resource of starting materials and intermediates (existence of ..., who made them? ..., etc.). (3) Much synthetic knowledge is documented and can be followed up for details via "chemist" and "notebook number" DTs (also on report form). (4) Spectroscopic models can be located among the intermediates. (5) New ideas may reveal the need to submit some intermediates to new biotests. (6) Potential exists to use the accumulated information to feed an eventual full-fledged reaction DB, if deemed worthwhile, and to supply information to enhance computer-assisted synthesis. (7) Combined with

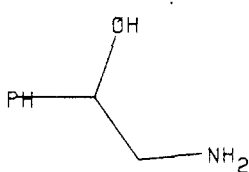
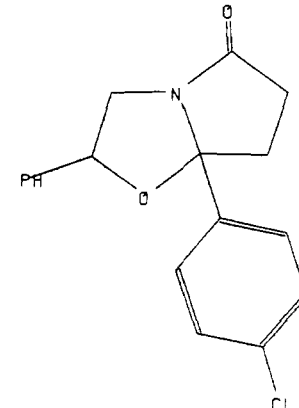
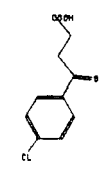
PRECURSOR 1 	LAB UNIT HOU	CHEMISTS HOULIHAN PARRINO	INT. REGNO 26945	EXT. REGNO SAH-063839	SALT	LAB BOOK REF 1088-201-30
MODE OF SYNTHESIS REPORT. 1-2 PRECURSORS ON THE LEFT FORMING PRODUCT ON THE RIGHT WITH REAGENTS, CONDITIONS IN CENTER			PRODUCT 			
22815 REAGENTS, CONDITIONS			FORM: BARCZA, JAN. 8, 86			
22815.0 - 26944.0 XYLENE REFLUX 7 HRS.						
SUMMARY OF SYNTHESIS OF PRODUCT 22815+26944---->26945						
PRECURSOR 2 	REG. DATE 11-06-85	REPORT DATE 01/13/86				
26944	DISCL. NO 364-85	L&D. NO	AMOUNTS 2340.0			

Figure 5. Complete synthesis step report, showing structures of not only the product but also of precursors (mother and father) of compound, reagents and conditions, summary of synthesis (multistep, if that be the case), and auxiliary information, e.g., in which lab notebook to find the procedure, which chemists to ask.

"Chiral" and "Abs Conf" data, the Made.from data provide us with a chiral synthon DB,¹⁵ (i.e., a DB of building blocks for synthesizing chiral molecules).

Stereochemistry, Absolute Configuration. MACCS stores stereochemical designations, employing the SEMA (stereochemically extended Morgan algorithm).^{11,16} This was augmented in the Sandoz Preclinical DB. A compound with a single stereocenter is implied to be a racemate, one with several stereocenters, a racemate of a specific diastereomer or meso, i.e., the *relative* stereochemistry is unambiguously specified at this level.

The Chiral software button can be activated for the molecule, in which case the fact is registered with the structure that the compound is optically active and represents one enantiomer only. (A regular DT stores optical rotation.)

Searches can locate racemic or chiral matches or stereoisomers. Chiral compounds are either of (1) known absolute configuration, as registered into the DB, or (2) unknown absolute configuration, in which case one enantiomer is arbitrarily drawn and registered.

A request to MDL to develop an Abs Conf software button has not yet been satisfied. Therefore, the ambiguity was removed in the Sandoz DB in the following way: when a stereostructure is known to represent the absolute configuration, the term Abs Conf is registered into a DT for notes and is also appended to the "Molname". (The Molname and ordinary DTs have slightly different and complementary search and display properties.) Thus, by searching for Abs Conf, all compounds with known absolute configuration can be quickly identified.

Benefits. (1) Since *biological activity* is (absolute) configuration dependent, *promising candidates* can be found if a sufficient lead structure or a relevant (e.g., complementary)

chiral natural molecule body constituent is known, e.g., an enzyme site. This should be expanded in the future to three-dimensional structure matching. (2) A *chiral synthon pool*¹⁵ is available within the Sandoz Preclinical DB via registration of Abs Conf, and starting materials, and intermediates, in addition to the final products. Thus the Abs Conf storage and simple reaction retrieval in the same DB synergistically reinforce their usefulness.

Literature References. A maximally filled, compact, formatted, single DT was created to store and search the literature references in a coherent manner in the form @@@@%-000-000000-00 comment, where @ is an alphabetic character, % is any character, 0 is a numeric character, and the optional comment can fill the line to 120 characters with text. These are used as follows.

CODEN-Volume Number-Page or Abstract Number-Year.

For increased user friendliness, the most frequently occurring CODEN abbreviations were listed with their expanded journal names in the self-documenting part of the DB (see later). The comment is usually filled with the names of the most prominent authors, keywords, and/or the abbreviated title. Typically, only one line per citation is used; however, a full list of authors, title, and keywords may be accommodated in multiple lines if the formatted portion is repeated at the beginning of every line. Typical data are

> 21437 (NMRREF)	DT0031
JACSAT-101-000191-79 R. L. Smith, D. W. Cochran, P. Gund, E. J. Crague	
> 21512 (NMRREF)	DT0031
JACSAT-090-000697-68 R. J. Pugmire, D. M. Grant, pyrimidines purines	
> 22501 (NMRREF)	DT0031
OMRRBD-012-000379-79 STEFANIAK	

The literature references stored in this format are eminently searchable by using the column search feature of the MACCS program. They are range searchable (most valuable for the Year field). A Boolean logical "AND" search within any line of data is provided by including several search strings and the appropriate column ranges within the same search query. "OR" logic is provided by consecutive searches.

Literature references are entered under NMR.REF. for known compounds with ^{13}C NMR shifts. Currently about 1000 compounds have these data.

An identically configured synthesis reference DT was created.

Compound Inventory. A simple substance inventory was set up within the main DB by devoting only a single DT for this purpose. A flexible numeric DT receives one line of entry per transaction per compound. The first number is the amount transacted (or requested), and the second number is the amount resulting (or remaining) in the drug room as a result of this transaction. The comment identifies the recipient and purpose. The most frequent case, sending the sample to the drug room, is the default and is without comment. Samples sent to special/outside tests simply preserve the drug room amount in the second column.

A portions of this operation which is fully implemented is that the compound registrars enter the amounts and applicable comments for all samples submitted to in-house and outside biological testing. These data are (range) searchable: the first column for amounts submitted; the second for the amount in the drug room; and the comment is text-string searchable for the investigator's name, purpose, etc. Sample data are shown below.

```
> 25000 (EXM-300001) (AMOUNTS) DT0017
5500.0
40.0-5500.0 BERN, DR. FAUSTUS
150.0-5500.0 SFI AM/AV
```

By the design described, provision has been made to enter further transactions: requests made and requests filled (with negative sign in the first column and the drug room amount being decremented. (MACCS has no arithmetic capability.) This phase is not yet being practiced, but feasibility has been shown. Sample data are shown below.

```
> 22000 (DOC-022000) (AMOUNTS) DT0017
5000.0
50.0-5000.0 Scallen
200.0-5000.0 SFI AM/AV
400.0-5000.0 BASLE Karobath CNS 7.15.83
150.0-5000.0 Req Engstrom 7.16.83 HL
-150.0-4850.0 Filled Engstrom 7.17.83 HL
600.0-4850.0 Req Aranda 7.18.83 AO
-600.0-4250.0 Filled Aranda 7.19.83 AO
3000.0-7250.0 Bottle 2 7.19.83
-500.0-6750.0 NCI 7.20.83
```

A planned third phase is to take signals from an electronic balance in the drug room, perform the necessary calculation, and automatically update the DB.

A fourth phase is to use bar-code labeling to identify samples and read them for recording transactions. This is awaiting adoption of bar-code techniques by all laboratories.

The approaches outlined are applicable to most inventory situations elsewhere as well and are partly practiced for laboratory chemicals DBs.

A DATACCS form was designed that plots bottle (vial) labels, one label for each of the first three lines of entry under the DT "Amounts". This can be done automatically out of the DB. A sample is shown in Figure 6.

Biological Data. It is the very essence of preclinical drug research to evaluate chemical structures and biological ac-

EXM-300001 NA	EXM-300001 NA	EXM-300001 NA
REGNO 25000	REGNO 25000	REGNO 25000
CHEM# 9999-888-44	CHEM# 9999-888-44	CHEM# 9999-888-44
MW 379.348	MW 379.348	MW 379.348
5500.0 MG.	40.0 MG. BERN, DR. FAUSTUS	150.0 MG. SFI AM/AV
DATE 06-28-84 LAB BAR	DATE 06-28-84 LAB BAR	DATE 06-28-84 LAB BAR

Figure 6. Bottle labels are automatically plotted with data from the DB. Each label receives the weight for a different bottle, for a different destination, and the applicable comment. Other information is internal and external registry numbers, salt code, notebook number, date of submission, molecular weight, and lab unit code. DATACCS offers much flexibility in customizing such plots. Storing the information, especially the "amounts", in the same DB helps.

tivities in each other's context, and therefore we made and realized a total commitment to this need.

Biological data are stored in (currently about 180) DTs organized in a hierarchical fashion: the most significant, leftmost, two letters of each DT name represent the major disease goal, the next the subgoal, the remainder the test family. Within most DTs, the hierarchy continues in the form of zones and subzones of entry. Each line of data is a test result, and its description is organized into zones and subzones for dose, percent response, rating, species, duration, date, investigator, comments, etc. These can be searched in a coherent fashion, by "AND" logic within the line, using the column search feature of MACCS. Sample data are shown below. Details have been published.³

```
> 9042 (SAH-050283) (DIHG.MOUSE HYPOGLY-
CEMIC) DT0094
```

```
200.000 MK 52 - PO MOUSE 02-13-74 ... RSH 0453-044
20.000 MD 29 - PO MOUSE 02-04-75 ... RSH
40.000 MM 34 - PO MOUSE 02-04-75 ... RSH
80.000 MK 36 - PO MOUSE 02-04-75 ... RSH
14.500 MK 25 - PO RAT 06-17-76 ... RSH 0526-061
30.000 MK 40 - PO RAT 06-17-76 ... RSH 0526-061
15.000 MK 30 - PO RAT 06-17-76 ... RSH 0526-061
```

Discussion of the biological data is restricted in this paper to the context in the overall structure-activity DB. Thus, the most typical searches that capitalize on having the structures and biodata in the same DB are (1) substructure searches for a type of structure. With the resultant hitlist in the reference file, the user range searches for specific levels and types of biological activity (optionally combined with species or other restriction) or simply browses the biological activities for this structure class. And (2) searches for specific levels and types of bioactivities over all compounds. The user then browses and/or substructure searches these active compounds for common structural elements. This can be followed up by examining whether a correlation exists between spectral (^{13}C NMR) shift values of certain atoms and the activity or by checking for other activities, etc.

Many other useful search combinations exist, which are made possible by the unified DB.

Self-Documenting Database. A tier or layer of DB-specific help was implemented *within the very DB* it serves, superposed on the general help, in the following way.

One "compound" with a memorable registry number is the depository of information about the format, structure, and meaning of data. For each DT, this information is registered under this compound, into the DT itself, which it is designed to document, format permitting. "Fixed" DTs, having the capacity of a single line (maximum 20 characters), are documented in those "flexible" DTs (maximum 300 lines \times 120 characters) that are most closely related. For example, the fixed DT "Solubility.code" is explained in the flexible DT "Solubility.details". As much as length and format allow, a cross-reference points to the site of explanation. Similarly, for those flexible biology DTs in which, for the purpose of consistency and easy learning, similarly zoned data format is

used, cross-reference is made to a single DT where the format is described.³ This mode of self-documenting a DB was adopted at our recommendation by MDL in the Fine Chemicals Directory DB.

The types of information placed into the DT documentation include format of the data, especially for "zoned" and "subzoned" DTs,³ and cross-references to similar data (types/formats).

Visibility of the DT Documentation. Placing the DT documentation into the DB itself is ideal since its visibility coincides with the uses of the DB. Thus, any of the following modes of access can be exploited.

(1) Within a MACCS session accessing the DB, DT definition/documentation can be inspected by "Find DT XXX" for the documentation and exemplification compound.

(2) A stand-alone file ("Datfile") can be written of the DT documentation and subsequently printed out and used as hardcopy reference. The most frequently used biology data zone formatting was posted in this manner at all terminals; also, laminated plastic cards were given to many users ("card-carrying MACCSers") for pocket reference.

(3) The same information may be called into a DATACCS display form on the terminal screen (Box for the data can be temporarily expanded if necessary) in exactly the same manner as the real data it explains.

(4) A self-documenting DATACCS form can be plotted.

(5) DATACCS tables can be made containing the documentation information.

(6) Each of the hardcopy outputs (2, 4, and 5) can be prepended or appended to a series of reports.

(7) "Docu" data can be used as a header.

It is most convenient that in the absence of any deliberate changes read-access by the same users is allowed to the data and its definition automatically, in both MACCS and DATACCS.

An illustrative example of a DT definition is shown below.

) 22000 (COC-022000) (SUM.SYNTH) DT0026
TEXT DATATYPE, LISTING THE (MULTI-
STEP) SYNTHESIS OF THE COMPOUND.

REGNO1 → REGNO2 → REGNO3 + REGNO4
→ THISREGNO, FOR EXAMPLE.

PARENTHESES MAY BE USED FOR COM-
PLICATED, BRANCHED SYNTHESIS TREES.

TO INSPECT THE SYNTHESIS PATHWAY
FOR A CPD, FIND THIS DT FOR CPD.

TO FIND DESCENDANTS OF A CPD, SEARCH
FOR ITS REGNO IN THIS DT.

@THISREGNO@ IS THE FORM FOR
SEARCHING FOR A NUMBER EMBEDDED IN
TEXT.

THE SYNTHESIS SEQUENCE MAY BE DIS-
PLAYED WITH STRUCTURES BY DATACCS,
USING THE FORM TMPL)SYNBX5.

SEE ALSO DT 25, MADE.FROM.REGNO.

MADE.FROM.REGNO AND SUM. SYNTH
ARE POWERFUL TOOLS FOR VERY LITTLE
COST.

July 2, 85, S. Barcza.

Prospect. The next planned major phase of development is to close the gap between laboratory instruments and the DB, employing laboratory automation, data reduction, and online input and avoiding data transcription by humans.

CONCLUSIONS

This work illustrates that careful analysis of needs and the capabilities of a set of tools combined with experimentation and some ingenuity gives rise to a far better and more cost-

effective satisfaction of needs than the standard application of those tools would suggest. The alternative, choosing separate specialized tools for the purpose at this stage of development, would mean some incompatibilities, wasteful storage, and other disadvantages. The essence of preclinical drug research and development information is (synthesis → spectra) → structure ↔ activities. Our work embodies this maximally with a single coherent system.

We summarize the approximate vital statistics of the DB: 28K compounds (structures) plus 256K filled [DT × RN] combinations including 86K biological activities, 5.5K synthesis steps, 2.3K summaries of synthesis, 2K ¹³C NMR spectra, and 2.7K drug inventory input in ~210 DTs occupying ~40M words.

ACKNOWLEDGMENT

We thank numerous scientists for contributions, help, and discussions: K. Mensler, S. Detar, D. del Rey, J. D. Dill, and W. D. Hounshell of Molecular Design Ltd. The contribution of supplying connection tables of our first 19 000 compounds by U. Hegi and H.-K. Kaundl of Sandoz Ltd., Switzerland, was very valuable. M. J. Shapiro supplied and S. N. DiCataldo and M. T. Schafer input most of the ¹³C NMR information. L. A. Kelly and L. Roberts translated older records of administrative information for several thousand compounds and made other valuable contributions along with R. E. Kirschenbaum and H. R. Lukas. J. L. Cooke, B. McKay, and S. Peacock performed much of the installation of the computer and the MACCS system along with G. Lawler, who provided superb user education.

REFERENCES AND NOTES

- Presented in part at the 190th National Meeting of the American Chemical Society, Chicago, September 1985; paper CHINF 14.
- MACCS, DATACCS, and REACCS are programs and trademarks of Molecular Design Ltd., Inc., 2132 Farallon Drive, San Leandro, CA 94577.
- Barcza, S.; Kelly, L. A.; Wahrman, S. S.; Kirschenbaum, R. E. "Structured Biological Data in the Molecular Access System". *J. Chem. Inf. Comput. Sci.* **1985**, 25, 55.
- Wipke, W. T.; Dill, J. D.; Peacock, S.; Hounshell, W. D.; Marson, S. "Search and Retrieval Using an Automated Molecular Access System". Presented at the 182nd National Meeting of the American Chemical Society, New York, August 1981.
- Anderson, S. "Graphical Representation of Molecules and Substructure—Search Queries in MACCS". *J. Mol. Graphics* **1984**, 2, 83.
- Adamson, G. W.; Bird, J. M.; Palmer, G.; Warr, W. A. "The Use of MACCS within ICI". *J. Chem. Inf. Comput. Sci.* **1985**, 25, 90.
- French, S. E. "REACCS Applied to a Corporate Database". *Abstracts of Papers*, 187th National Meeting of the American Chemical Society, St. Louis, MO, April 1984; CHINF 45.
- Bremser, W. "Expectation Ranges of C-13 NMR Chemical Shifts". *Magn. Reson. Chem.* **1985**, 23, 271 and references contained therein.
- Heller, S. R. "The Chemical Information System and Spectral Databases". *J. Chem. Inf. Comput. Sci.* **1985**, 25, 224.
- Dill, J. D. "DATACCS—An Interface from MACCS to Other Software Systems". *Abstracts of Papers*, 187th National Meeting of the American Chemical Society, St. Louis, MO April 1984; CHINF 46.
- Wipke, W. T.; Dyott, T. M. "Stereochemically Unique Naming Algorithm". *J. Am. Chem. Soc.* **1974**, 96, 4834 (see also 4825).
- ORAC (Organic Reactions Accessed by Computer). A. Peter Johnson, A. P., University of Leeds, UK.
- SYNLIB (Automated Organic Synthesis Library). Chodosh, D. F.; Mendelson, W. L. "A Graphics Approach to Reaction Retrieval". *Pharm. Technol.* **1983**, 7(3), 90. Chodosh, D. F. *Drug. Inf. J.* **1983**, 17, 231.
- Schubert, W. "ASSOR—Allgemeines Simulations—System Organischer Reaktionen". *MATCH* **1979**, 6, 213.
- Hanessian, S. "CHIRON, an Interactive Computer Graphics Program for the Analysis of Stereochemical Features and for the Selection of Chiral Precursors in Organic Synthesis". Presented at the Symposium on Computers in Organic Synthesis, 191st National Meeting of the American Chemical Society, New York, April 1986. Hanessian, S. *Total Synthesis of Natural Products: The Chiron Approach*; Pergamon: New York, 1983.
- Wipke, W. T.; Dyott, T. M. "Simulation and Evaluation of Chemical Synthesis. Computer Representation and Manipulation of Stereochemistry". *J. Am. Chem. Soc.* **1974**, 96, 4825.