should achieve the same results as those now described, if the query substructure is properly formulated. However, these systems are not yet in general use for searching privately held files.

The few string searches so far done on the output of the present suite of programs have thrown up a sufficient number of suggestive heterocyclic–aliphatic relationships to justify pursuing this approach more thoroughly. Tree-path tapes (see Figure 7) could be prepared on a large computer from WLN as issued by *Index Chemicus* (or from private files). These could then be searched at leisure on a small computer (with rapid input/output), to follow out ideas as they cropped up, by any chemist familiar with the CROSSBOW bonded-atom symbols and capable of writing very simple string-search programs.

ACKNOWLEDGMENT

I am grateful to Prof. P. M. Stocker and his staff at The Computing Centre of this University (and specially to Dr. P. Anstey and R. A. Jenyon) for much constructive advice about using their facilities, to Prof. M. F. Lynch and his colleagues at The Department of Information Science of Sheffield University (particularly Drs. J. M. Barnard and P. Willett), Pamela A. Chubb, and Dr. Wendy A. Warr for varied and valued advice, and to the Deans and staff of this School for office space and much day-to-day help.

REFERENCES AND NOTES

(1) Robinson, R. "The Structural Relations of Natural Products"; Clarendon Press: Oxford, 1975. Todd, Lord; Cornforth, J. W. *Biogr. Mem. Fellows R. Soc.* **1976**, *22*, 415–527.
(2) Prager, B.; Jacobson, P., Eds. "Beilsteins Handbuch der organischen Chemie," 4th ed.; Springer: Berlin, 1918; Vol 1, p xvi.
(3) Randić, M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 23–31, 31–37.
(4) Wilson, R. J. "Introduction to Graph Theory"; Longman: London, 1972.
(5) Ash, J. E. In "Chemical Information Systems"; Ash, J. E.; Hyde, E., Eds.; Horwood: Chichester, England, 1975; Chapter 11, pp 156–176.
(6) Gasteiger, J.; Jochum, C. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43–48.
(7) The marketing and development rights for the CROSSBOW suite of programs are held by Fraser Williams (Scientific Systems) Ltd., Poynton, Cheshire, England.
(8) Lynch, M. F. *J. Chem. Doc.* **1968**, *8*, 130–133.
(9) Bond, V. B.; Bowman, C. M.; Davison, L. C.; Roush, P. F.; Young, L. F. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 103–105. Warr, W. A. "Proceedings of the CNA (UK) Seminar on Chemical Structure Searching of the Published Literature", March 17–19, 1980, Daresbury, Warrington; Chemical Structure Association: London, 1983; pp 165–80.
(10) Page, E. S.; Wilson, L. B. "Information Representation and Manipulation in a Computer"; Cambridge University Press: Cambridge, England, 1973; p 124, algorithm B.
(11) Windholz, M.; Budavari, S.; Stroumtsos, L. Y.; Fertig, M. N., Eds. "The Merck Index—An Encyclopedia of Chemicals and Drugs", 9th ed.; Merck: Rahway, NJ, 1976.
(12) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.
(13) Attias, R. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102–108.

# Structured Biological Data in the Molecular Access System

SANDOR BARCZA,* LAWRENCE A. KELLY, SIEGFRIED S. WAHRMAN, and RICHARD E. KIRSCHENBAUM

Preclinical Research, Sandoz Research Institute, East Hanover, New Jersey 07936

Received June 11, 1984

Chemical, administrative, and biological information at Sandoz Inc. Research and Development was put into a database created with the MACCS program.[1,2] The configuration of the database and of the "datatypes" in it was done in a way that made the essentially "flat" original design of the database hierarchically structured and searchable. This was accomplished by two devices: (1) The biological activity datatypes were given structured names. The characters went from left (broadest category) to right (most specific category), expressing the major disease goal, then the subgoal, and finally the actual test name. (2) The data within the datatypes were structured into zones and subzones of columns, corresponding to species, dose, effect, direction, date, etc., for each line, while the rows of entry were successive instances of testing. This additional organization of the data offered significant advantages in economy of storage, coherence (interrelatedness) of data, searching, user comprehension, and overview. The orderly entry of data into this system was assured through a data entry interface to the MACCS program. It is the purpose of this paper to describe the innovative adaptation of MACCS to the handling of pharmacologic data, as well as some associated problems and solutions.

## INTRODUCTION

Every organization that makes decisions on the basis of data obtained from biologically active substances is confronted with the problem of organizing and retrieving a multiplicity of data elements on a large number of compounds. Sandoz Pharmaceuticals of New Jersey selected MACCS[2] (the Molecular Access System), a data management program based on molecular structures, to manage its large chemical and biological data base. Developed by Molecular Design Limited (MDL), MACCS was chosen because it offered the best commercially available system capable of storing, searching, and retrieving both molecular structure information and associated data.

Sandoz's data management system had to accommodate information derived from over 300 tests on approximately 20 000 compounds—nearly 300 000 lines of information 94 characters wide (27 megabytes). Commitment was made to store chemical and biological information together, in accordance with modern drug research needs. Storage of both chemical and biological information had to be open-ended, allowing Sandoz to add both compounds and data fields as needed. Sandoz developed a chemical and biological information system with MACCS that is graphical, interactive, and

0095-2338/85/1625-0055$01.50/0 

user friendly. The system provides storage, searchability, and prompt retrieval of both molecular information and administrative and biological data.

This paper describes the configuration of the biological part of the database as developed at Sandoz. We believe this is the first use of MACCS for *combined* chemical and biological retrieval in a sizeable database. This system was implemented after extensive experimentation in a rather unique fashion. The structure-search and substructure-search capabilities of MACCS are well worked out and already widely applied.

## METHODS

The chemical–biological information retrieval system described here is organized as a typical MACCS database: a matrix of compounds vs. data fields. The data fields contain information relative to each compound. The list of compounds and the list of data felds are each sequential and extendable. Insofar as the data fields appear simply in sequence, the database that they form is "flat"; i.e., the data base lacks hierarchical organization.

MACCS provides three different types of user-defined data fields (datatypes) for storing nonstructural information: (1) Numeric datatype contains one or two real numbers plus a text comment. Numbers and comments are searchable separately. (2) Formated datatype permits data items of up to 20 characters to be formated in one to five fields. The remaining characters in the line can be used as a text comment. Formated fields and comments are searchable together. (3) Text datatype may contain any text characters and requires no structural format. The capacity of each datatype (numeric, formated, and text) for each compound is 118 columns in width and 300 lines in depth. While full usage of this capacity exceeds the current bounds of the report generator (MDL's DATACCS program), it is possible to output selected subsets of the data into multiple reports, so that data storage is not really limited by these restrictions.

Datatypes may be searched for a string of characters. A search string must match within the same line of data to produce a hit, e.g.

| effect | species |
|---|---|
| 80 | rat |
| 70 | dog |

Searching by string (columns 1) 75 to 100 (columns 2) rat will produce a correct hit, but searching for 75 to 100 and then species dog would allow this compound to survive as a (false positive) hit. Each line is examined in succession. Consecutive searches (using different strings) may result in search strings matching across different lines. These amount to OR searches and, if the objective is an AND search, would include false positive hits.

When MDL added column-searching capabilities to MACCS, however, we found that we could now reliably match several search strings within a single line. We could search for several strings (or ranges of strings) within specified column limits. [As an example of ranges of a (nonnumeric) string, "MK to MM" will find compounds with "marked" as well as "moderate to marked" activaties.] In effect, we could produce Boolean AND logic searches within the line proper.

## RESULTS AND DISCUSSION

**Structuring Data within Datatypes.** MACCS's new column-searching capabilities created the opportunity to introduce a degree of hierarchy into the database. Accordingly, we initiated "zoned datatypes". The available columns in each datatype are divided into zones. Each zone (Zn) is designated for a qualitatively different entry; e.g., zone 1 = columns 1–8 = dose; zone 2 = columns 10 and 11 = rating of the compound,

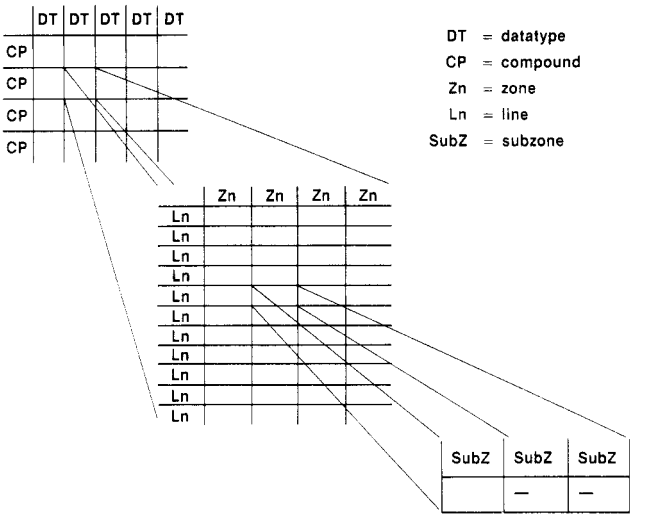


**Figure 1.** Database as a matrix of compounds (rows) and datatypes (columns). The subsets show hierarchy resulting from structuring the data into zones and subzones within the datatype.

etc. Blank spaces function as delimiters and separate the different zones (for exemplification and further definition please see Figures 3 and 4).

The zoned datatypes create a new data substructure within the database: an ordered, organized set of data under each compound and each datatype. Each datatype in a database could, of course, be zoned differently; however, uniform structuring of datatypes provides more efficient searching capabilities. Each zone (range of columns) in a datatype contains data entries of a similar kind. Each line in a datatype contains data items entered into each of the different zones. The entries on each line are internally related: e.g., all were obtained at the same time; all were obtained within the same measurement; etc. A typical datatype might describe the effect of a compound on the blood sugar of a test animal. Typical zones in this datatype would describe the dose of the compound administered to the animal, the magnitude of the effect, and the species of the animal tested.

The introduction of subzones (subZ) adds a further degree of hierarchy within the database. Subzones are used to further define zones; i.e., the zone "reference" is subdivided into the subzones "book" and "page", the zone "date" is subdivided into the subzones "month", "day", and "year" (MM-DD-YY). Different subzone entries are separated with hyphens and are functionally distinguished in the searches but not of themselves in the data. A conceptual description of the zoned and subzoned database is presented in Figure 1.

The new datatypes were implemented in MACCS's "text" format, chosen because it provides no format restrictions and allows the entire datatype to be searched at one time. Use of the text format involves some trade-offs, however. When this work began, searching of the text data was based on the ASCII value. Consequently, range searches of numbers with decimal points and/or leading blanks as part of the data provided initial pitfalls. MACCS was changed recently to provide correct range searching of decimal (floating point) numbers embedded in text and to ignore leading and trailing blanks.

Data entered into zones and subzones provide neat, easily read tables when printed. Data structured in this manner store and deliver more information than would be reviewable without this organization.

**Structuring of Datatypes.** Further hierarchy is introduced with the structuring of datatype names. Typically, the first two characters of the name are abbreviations for a disease group. The next two characters (following a delimiting

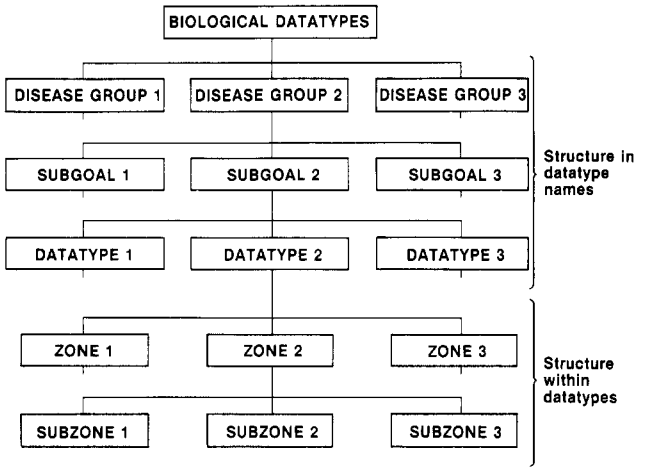


**Figure 2.** Hierarchical structure of biological portion of database.

character) represented a subgoal in the disease group. The remainder of the name is unique and identifies the tests[3] stored in that datatype. Thirty characters are available for a datatype name. Four characters are used to describe the disease group and subgoal, and two characters function as delimiters. Consequently, 24 characters remain to describe the test—almost by name. Each datatype name within a disease group starts with the same pair of characters, and each datatype name within a disease subgoal has the fourth and fifth characters in common as well. The disease group and subgoal abbreviations and the first two characters of the test description are compatible with a similar system employed by Sandoz Ltd., Switzerland.[4] The hierarchical structure of the biological portion of the database is shown in Figure 2.

**Entry of Data into Zoned Datatypes.** Column searching of text data is a relatively recent addition to the MACCS system. MACCS will assure that the user enters data in a proper manner in both the "numeric" and the "formated" datatypes. At present, however, MACCS offers no method of data entry into the "text" datatype that will assure entry of data into the proper zones. Because the success of the zoned data procedure depends entirely on a highly reproducible, easy-to-use method of entering the data into the respective zones, a means of data entry outside of MACCS had to be found. Several systems were briefly considered: (1) SAS (Statistical Analysis System);[5] (2) PROPHET—a laboratory data handling system;[6] (3) RS/1—a subset of PROPHET; (4) software resident on the Prime computer.

SAS, PROPHET, and RS/1 provide great statistical computing power for data manipulation before entry into MACCS (and may be used in the future). These systems would be functionally suitable but would have meant overhead and delays. Further, since these systems were not currently available on our Prime computer, we developed our own data entry routines.

The first attempt to control the data entry function was a CPL routine (CPL is the Command Procedure Language that runs under the Prime operating system). The routine, called "Enter Tab Data", allows data to be entered under control of the standard system editor. Execution of this routine creates a file that can be registered directly into MACCS with a "transfer datatype" command. The CPL routine prompts for the MACCS internal or external registry number (the compound identifier) and for the name of the file that stores both the data to be entered and the MACCS commands that will actually transfer the data.

Several commands are preset in the editor to aid in the data entry process. Tab stops for the zoned data are set so the user does not have to remember or type in the correct column numbers. (With the tab feature, data can be entered in the proper columns without counting spaces and perhaps misaligning the data fields.) We defined a new editor command that prints a column ruler and allows the user to double check that data are being entered in the proper columns. In addition, we set commands that cause the editor to display a prompt character and print line numbers.

The CPL routine automatically enters the compound and datatype identifiers into the first line of the file in the format required by the transfer datatype command of MACCS. The editor then prints the column ruler, changes to "input" mode, and allows the user to enter data. Once the data are entered, the user typically enters "edit" mode and makes any necessary corrections. When the data file is complete, the user may either send it to a printer or register it into a MACCS database.

While this procedure eliminates many sources of error, there are still limitations in using the standard editor. The editor will prompt for the correct first line and provide formating for one compound only. It is the user's responsibility to separate the entries for subsequent compounds with a blank line and the compound identifier and datatype name or number, formated as required by MACCS. Since the Enter Tab Data routine provides no error checking, the user can enter meaningless data or enter data into the wrong column unknowingly. We decided, therefore, to reserve this data entry method for experienced users of the system.

We then developed a procedure using the BASIC language on the Prime computer. This routine contains a table of all the biological datatype names and numbers. The user simply enters his test number or abbreviated test name and the routine handles the rest. The user is prompted for the compound identifying number, dose, two-letter rating (MK, MM, MD, WM, WK, NA, or ED for marked, marked to moderate, moderate, weak to moderate, weak, not active, or effective dose, respectively), effect, route, species, duration, date, initials of the researcher, reference, and comment. The format of the entire field of 80 characters and the column justification within that field are governed by the BASIC "PRINT USING" formating command. The routine examines entries for length and numeric value and returns error messages where appropriate. Help pages currently only contain the phone number of the author but will be written to explain the most commonly encountered problems. The output of this routine is a data file properly formated for entry into a MACCS database. The file may be printed for review by the user if desired. One program module edits the data file if necessary, and another enters the quality-assured data into MACCS. The interfacing and connecting programs are written with CPL with Prime command files. The zoning of most biological datatypes is shown in Figure 3.

**Searches and Retrievals.** At the highest level of the hierarchy, all data belonging to all datatypes within a disease group can be retrieved or searched by specifying only the first two characters of the datatype name (the abbreviation for the disease group) suffixed with the wild-card character, "@". At the second highest level, all data belonging to a disease subgoal can be retrieved or searched by specifying the first five characters (including one delimiter) of the datatype name followed by the wild-card character.

The names of all currently used tests become unique with the first character after the subgoal delimiter. Therefore, any biological test can be individually retrieved by simply entering the first seven characters (five specifiers and two delimiters) of datatype name plus the wild-card. We refer to these seven characters as the abbreviated test name. Zoned data are retrieved and displayed with the entire datatype, but they can be individually searched with the column-searching technique. MACCS will search the entire datatype (or datatypes) for a match to the search string. A "hit list" will be generated for all compounds with datatypes containing the search string. In

```
          1         2         3         4         5         6         7
1234567890123456789012345678901234567890123456789012345678901234567890123456789
DDDD.DDD RT EEEE - RRRR SSSSS DD-MM-YY DURATION INT BBBB-PPP COMMENT HERE.....
 > Dose  RT Effect Dir Route Species Date Dur   Init Ref       Comment

      Dose              < 1- 8>    9 SP  0000.000 TO 9999.999 Range
      Rating(RT)        <10-11>   12 SP  MK, MM, MD, WM, WK  ED for ED50
      Effect            <13-16>   17 SP   % Change from Control
      Direction         < 18  >   19 SP    + or -
      Route             <20-23>   24 SP  PO, IV, IN FEED, etc.
      Species           <25-29>   30 SP  RAT, MOUSE, etc.

   The date is in Sub-Zones
      Total Date        <31-38>
            Month       <31-32>   33 -
            Day         <34-35>   36 -
            Year        <37-38>   39 SP

      Duration          <40-47>   48 SP Format is 1HR, 1DAY etc.
      Initials          <49-51>   52 SP Responsible Biologist.
      Reference         <53-60>   61 SP Lab Notebook BOOK-PAGE.
      Comment           <62-80>   19 Characters, free format.
```

**Figure 3.** Tabulation of zones and subzones of the most commonly used biodatatype. The first two rows contains column tabulations, the third row contains symbolic character representations of the data fields, and the fourth row is a comment header. The remaining lines list the zones and subzones, the columns in which they appear, and the form in which they are entered.

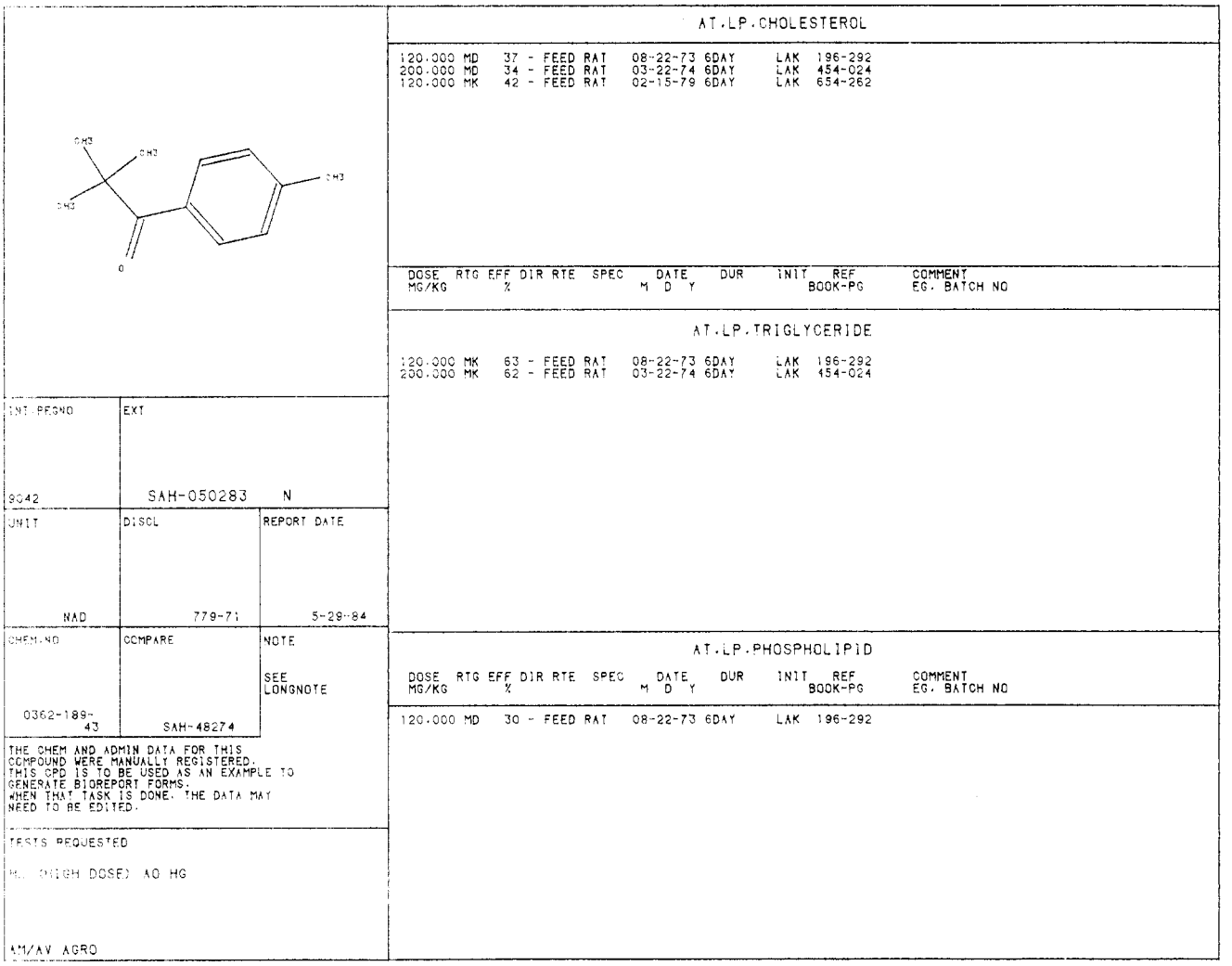


**Figure 4.** Example of data report from MARGEN.

the case of a range search, a "hit" is recorded whenever the ASCII code value of a character sequence falls within the limits of the ASCII code values of the range of characters specified in the search query. (As mentioned above, recent changes in MACCS allow for proper searching of real number ranges.) If column numbers are specified, however, MACCS will only search (either for an exact match or for a character sequence within the specified range) within those columns. A column search query can contain several search strings and the columns (zones) in which they may be found. Multiply queries (within a single search) are treated with AND logic, i.e.; all parts of the query must be satisfied within the same

line before a hit is returned. Therefore, we can serach for several entries within the same biological test measurement at the same time. We could specify the correct zones and search for a compound that effected a 50% change in a parameter with a dose of 200 mg/kg. We can search subzones by simply narrowing the column range. We could, for example, specify those columns in which we store the test year, enter the characters "81", and further limit our search to compounds tested in 1981.

**Reporting.** Hard-copy reports can be printed out in the datatype format as it is stored in MACCS. The desired datatypes for the compounds in the hit list are transferred to a data file and subsequently sent to a printer. We can print more flexible and pleasing reports, however, with MDL's MARGEN (Molecular Access Report Generator) or DATACCS (Data Access System) program by using a preformed report template and data retrieved from the database. Each page contains those datatypes (for one or more compounds on the hit list) specified when the template was designed. MARGEN will truncate the elements of a datatype but cannot further subdivide the datatype elements. With DATACCS, however, we can selectivity print out any line or lines of the desired datatype. An example of a biological data report is shown in Figure 4.

Header-type information is not stored with every compound; rather, header and unit information are designed into the MARGEN or DATACCS report templates as captions. Consequently, they need be keyed in only once, when the template is designed. The header information itself can be hierarchical since the caption can contain multiple lines as well as different type sizes.

## CONCLUSIONS

The MACCS program can be used to store and retrieve both biological information and molecular structure information on a practical, sizeable database. Within a typical MACCS database, we were able to create an additional and hierarchical ordering of data. Our experience in setting up an ordered, efficient method of biological data storage (in conjunction with molecular structure storage) is applicable to other organizations that currently maintain separate chemical and biological databases and routinely must store, search, and evaluate biologically active compounds.

Our principal MACCS database designed with hierarchically arranged data currently contains about 24 000 compounds and 200 hierarchically ordered data fields. We have data for about 40 000 compound/data field combinations, many with multiple-line entries. We are continually expanding this database.

## ACKNOWLEDGMENT

Valuable help and contributions of Kathleen Mensler, Michael Weinshelbaum, Susan Anderson, and Michael Savage of Molecular Design Limited, Inc., are gratefully acknowledged.

## REFERENCES AND NOTES

(1) Dill, J. D.; Hounshell, W. D.; Marson, S.; Peacock, S.; Wipke, W. T. "Search and Retrieval Using an Automated Molecular Access System". Presented at the Symposium on Molecular Substructure Searching, 182nd National Meeting of the American Chemical Society, New York, Aug 23–28, 1981.
(2) MACCS, MARGEN, and DATACCS are registered trademarks of Molecular Design Limited, Inc.
(3) Tests for different species, duration, and doses are stored in the same datatype.
(4) Kaindl, H.; Hegi, U.; Hummel, R., Sandoz, Ltd., Basle, Switzerland, private communication.
(5) Barr, A. J.; Goodnight, J. H.; Sall, J. P.; Helwig, J. T. "A User's Guide to SAS"; SAS Institute: Raleigh, NC, 1979.
(6) "Prophet Molecules. A User Guide to the Molecule Facilities of the Prophet System"; Rindone, W. P.; Kush, T., Eds.; *NIH Publication* **1980**, *July*, No. 80-2168.