The Computer-Based Chemical Structure Information System of Merck Sharp and Dohme Research Laboratories

HORACE D. BROWN, MARIANNE COSTLOW, FRANK A. CUTLER, JR.*, ALBERT N. DEMOTT, WALTER B. GALL, DAVID P. JACOBUS, and CHARLES J. MILLER

Merck & Co., Inc., Rahway, New Jersey 07065

Received November 3, 1975

The computer-based chemical structure system now in use for in-house compounds is described. Distinctive features include structural formula input for both file updating and substructure searching, multilevel Markush-type capability for structure definitions in questions as well as in file compounds, Boolean logic for complex questions, and atom-by-atom matching. Data from other files, particularly of biological data, are associated logically in the printed output, providing chemists and biologists with a resource for structure–activity studies.

OVERVIEW

This paper describes elements of the computer-based system established at Merck for handling the chemical structures of research compounds. This system was developed not only to meet the need for automated registration and substructure searching of in-house compounds but also to interface with other computer-based files, notably sample location and biological data.

The technology for processing chemical structural information has recently been summarized by Holm,[1] Wipke,[2] and Ash.[3] Our own surveys of the field led us to base our system on the approach used at the Walter Reed Army Institute of Research (WRAIR), since this permitted the input and output of structural diagrams and handled benzenoid double bond resonance in a straightforward manner.

The principal features of the WRAIR system have appeared in a number of publications.[4-9] We have made major changes in the WRAIR system to adapt it to our needs and hardware, and in particular to broaden its substructure searching capabilities.

In the construction of our file, the structures are simply typed on a magnetic tape/Selectric typewriter equipped with a typing element bearing appropriate bonding characters. Descriptive information is also entered according to format (accession number, molecular formula, source, stereochemistry) or in free text (explanatory comments). The magnetic tape produced from typing a group of structure records serves as input to the computer-based system.

For query of the file, questions are typed in a similar manner. Structures, part structures, molecular formulas, years of registration, sources, and accession numbers can all be used to retrieve in printed form the structure records which answer the question, together with information drawn from other files. The substructure searching capability is especially powerful and is the initial focus of this paper. Certain features of the search mode are also important to the processing of input records for tautomers and indefinite structures, as discussed subsequently.

The Merck Chemical Structure Information System is now in operation on a file of over 120,000 compounds. It has been well received by research chemists and biologists and has proved to be relatively simple and straightforward in operation and economical of computer resources.

A moderately efficient bit screen and a very fast atom-by-atom matching routine make it practical to provide the well-known advantages of using a connection table for the computer representation of structures.

* To whom correspondence should be addressed.

SUBSTRUCTURE SEARCHING

General. While the system was designed so that questions could be phrased by the ultimate user, in practice there is consultation with a chemical information specialist who is familiar with the capabilities of the system, and who supervises the input of groups of questions to the computer, operating in batch-mode. Within 24 hours the user receives a printout listing the structures of compounds meeting the criteria, together with associated information that may have been requested.

The question facilities are highly flexible and were designed so that no valid answer would be missed. Virtually any chemically valid structure diagram can be used as a question. As described subsequently, special conventions are provided in order to allow indefinite substituents, alternative positions of substitution, rings and chains of variable size, specification of atoms as being ring or acyclic, and logic.

Variable Substituent. The letter "X" with an identifying subscript can be used as an indefinite atom, defined separately so as to allow any of a limited number of alternatives. This facility closely resembles the conventional use of "R" to represent a radical group in Markush-type patent claims. Figure 1 gives an example of the use of X's. (The capital delta, standing for "definition", is a flag required by the computer program.) As indicated in the example, alternatives specified for the "X" do not necessarily have to be single atoms. Moreover, unspecified substitution may be present on the nitrogen atom attached to the ring.

Indefinite Position. The letter "Z" is used in a manner rather similar to the use of "X." The Z, however, indicates that each alternative in its definition must occur at one *and only one* of the positions indicated by Z's in the main structure. This facility resembles the conventional use of a slashed ring, but is a great deal more flexible. Figure 2 shows a simple use of Z's. The question requires one and only one chlorine substituent on the ring and one and only one hydroxyl on the carbon chain, but does not specify the exact position of either. No other substituents of any kind are allowed, except on the nitrogen, which may connect the question structure to a far larger structure in compounds retrieved.

Multilevel Z and X. The power of the Z and X facility is substantially increased by allowing Z's and X's to be used in the definitions of other Z's and X's in any combination and to any depth. The only limitation is the obvious requirement that no Z or X may be defined, directly or indirectly, in terms of itself. (Questions containing such recursive definitions are rejected by the edit programs to guard against possible loops in atom-by-atom matching.) Figure 3 shows a way of requiring at least one chlorine substituent, while allowing any

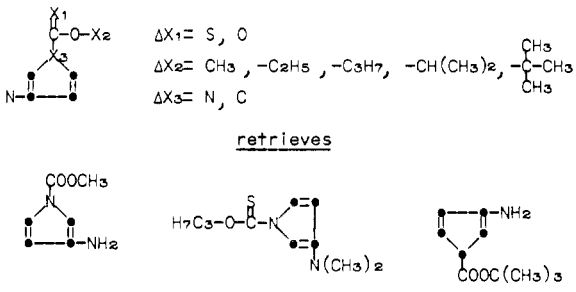

Figure 1.

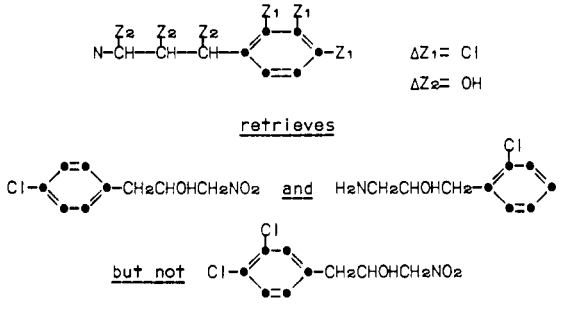

Figure 2.

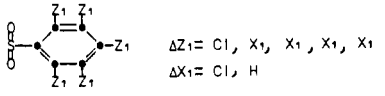

Figure 3.

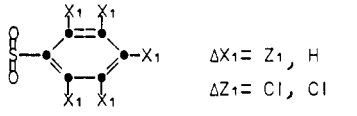

Figure 4.

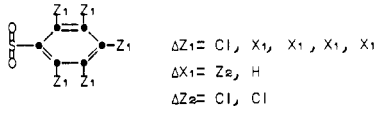

Figure 5.

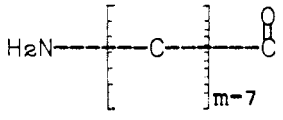

Figure 6.

number of additional chlorines. (By using the "ANY" facility described later, the restriction on the nature of the additional substituents could be eliminated.) Conversely, Figure 4 allows a maximum of two chlorine substituents, but does not require any. Figure 5, finally, requires at least one but no more than three chlorines.

Indefinite Size. To allow for rings and chains of indefinite size, the Merck system provides for repeating units. Figure 6 gives an example in which compounds retrieved are required to have a carbonyl and an amino group connected by a chain of cyclic and/or acyclic carbon atoms; the length of the chain may vary in different compounds between one and seven carbons. In Figure 7, a repeating unit is used to specify a ring of variable size; this question will retrieve all cycloalkanes from cyclopropane through cyclooctane, provided the ring carries an amino substituent. Repeating units may be of any size, and there are no restrictions on the maximum number of repetitions which may be specified, although some restraint is recommended. The only limitation on the contents of a repeating unit is that it must not contain another, subordinate, repeating unit.

Logic. Questions may consist of several independent

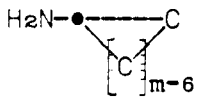

Figure 7.

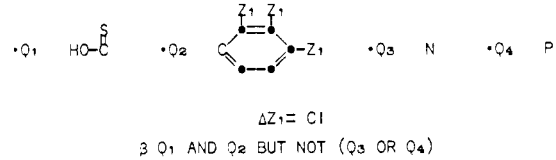

Figure 8.

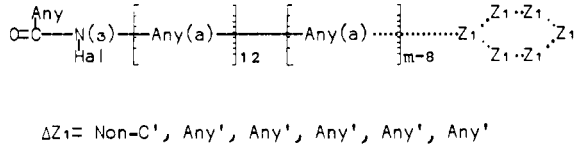

Figure 9.

structures to which Boolean logic (AND, inclusive OR, and NOT) is to be applied. Figure 8 gives an example. This question will retrieve all thioacids containing a monochlorinated phenyl, but excludes any compounds which contain either nitrogen or phosphorus. Parentheses may be nested to any depth. A single-term statement ("NOT Q1") is permissible and occasionally useful. The only limitation on the Boolean statement is that enough parentheses must be used to make the statement unambiguous. The statement "Q1 AND Q2 or Q3", for instance, would be rejected.

Other Search Features. A number of other search facilities is available. The letters "r" (ring) and "a" (acyclic) may be written in parentheses after an element symbol to specify that the atom must be a ring or a non-ring atom. (The letter "r" is unnecessary after atoms actually shown within a ring, since the program automatically identifies and flags such atoms.) An Arabic numeral in parentheses can be used to specify a particular valence of a variable-valence element. The abbreviations "Any", "Non-C", and "Hal" can be used to specify an atom of, respectively, any element other than hydrogen, any element other than hydrogen or carbon, or any halogen. An apostrophe (prime sign) can be used on an element symbol to specify that the atom must not be connected to any atoms other than hydrogen except those to which it is connected in the question. Prohibition of unwanted substitution on carbon is handled by use of the carbon dot rather than C'. A dotted bond line (used in file compounds to indicate stereo) may be used in questions to show that the nature of the bond between two atoms can be of any type; single, double, and triple bonds are equally acceptable. Figure 9 shows the use of a number of these facilities. The question would require a six-membered, unsubstituted, hetero ring, which may be saturated or unsaturated. The number and position of the hetero atoms are unspecified. This ring must be connected by a chain of from 13 to 20 non-ring atoms to the group shown at the left.

A special feature of the system permits the retrieval of peptides by specifying the classical three-letter codes together with indication of ring or acyclic if desired, and bonding to ordinary atoms or groups.

In addition to structural parameters, questions may specify a molecular formula (exclusive of hydrogen), year of registration, and/or the name of the chemist who synthesized the compound. Any or all of these criteria may be combined in a single question. The molecular formula may be specified as "at least" or as exact match. The latter case thus will retrieve all compounds with a given molecular formula, without the need for structural specifications.

Compounds can also be retrieved by laboratory number. Where a single compound had been assigned more than one

laboratory number in the original manual file, those synonym numbers have been entered in a single master file record. A compound can be retrieved by any one of its numbers, and in printed reports a compound is always identified by all its numbers. This has avoided the need for any special synonym file.

Computer run times for structure searching can vary greatly, depending on the number and nature of the questions in the run. When a run exceeds the estimated time, compounds answering each question prior to cutoff will be reported, together with indication of the proportion of the file searched for the run.

To avoid flooding the user with too many answers, retrieval for each question is normally restricted to 500 compounds. When retrieval is cut off because the count limit is reached, an estimate can easily be made of the number of answers not printed. A count limit greater or less than 500 may be specified whenever desired.

The searching procedure operates through bit-matching and then atom-by-atom comparison. It is possible to restrict the search to bit-matching only, a feature rarely used. It does permit "browsing" for unexpected compounds.

Structure searches are invariably associated with requests for other information, notably biological data and sample availability. Under the broadest option, the user receives a printout which lists, in addition to the structure diagram and ancillary information provided in the input record, also the chemical name, the shelf location of any available sample, and all computer-stored biological testing data. [The biological data system with which the structure system interfaces operates under the direction of Mrs. Idamarie R. Eggers. The structure system with its related files operates under the direction of one of the authors (W.B.G.).] The user may restrict the printing of data to a particular test or group of tests, and even to specified levels of activity. By special arrangement, it is possible to carry out searches involving logic across both chemistry and biology and/or availability. Such searches are important in structure–activity correlations.

The system is designed so that structures can be printed whenever data indexed by laboratory number are printed. This option is especially valuable when biological data are being presented in reports.

MASTER FILE

General Characteristics. The master file consists of five data sets: a structure file, a picture file, a screen index, a number index, and a hold file. The structure file has one record per compound and includes all searchable data. The picture file contains the original input records (structural diagrams and ancillary information) in a format suitable for printing when a compound is retrieved. The screen index indexes the structure file by bit screen, while the number index indexes both the structure and picture files by laboratory number. Entries in either index can be located by applying a hashing algorithm to the bit screen or laboratory number. The indexes are designed so that in the present file practically all overflow entries are stored in the same block as the prime entry (the location determined by the hashing algorithm). It is rarely necessary to perform more than one direct access in order to obtain the address of a particular structure or picture record. The hold file contains problem compounds reported for manual resolution (see below), and its function is simply to avoid the need to retype records which prove to be acceptable.

Updating. Updates are run in batch mode. Each run is a single job and is normally run overnight. Before a new compound record is added to the file, the program compares the structure to the structures of compounds already in the file which have the same bit screen. Since about 90% of the new compounds being added to the file have unique screens, the number of such atom-by-atom comparisons is small. If duplication is not found, the system assigns a new laboratory number or validates the assigned number. If duplication is found, the compound is reported for manual resolution and the structure records for both compounds are printed in the report from the run. The input record is added to the hold file to await a chemist's decision as to proper disposition.

Update runs may include orders releasing (or discarding) compounds from the hold file. The release order may simply add the compound to the file as a new compound. A release order may, instead, declare that the compound in the new record is identical with a compound already in the file, identifying the latter by its laboratory number. The laboratory number in the new record is then added as a synonym to the existing master file records for the old compound, and the new compound record is discarded.

Delete orders may be included in update runs to remove records from the master file. It is possible to concurrently delete an old number and add it as a synonym to a record being released from the hold file. This feature was especially valuable during conversion of the original file.

INPUT RECORD

Description. The input record for the master file is comprised of the following elements: the unique laboratory number, the molecular formula, stereo descriptor, reference, the chemical structure, and explanatory comments. The year of registration is entered automatically by the computer, but does not appear in the picture record.

The laboratory number is composed of a six-digit portion which is characteristic of the generic structure, a two-digit suffix which signifies the different salts or solvates, and finally a check letter determined by computation over the eight preceding digits and used to detect transcription errors. Omission of the laboratory number in the input record will cause the system to assign and enter a number if the compound is new.

When appropriate, the molecular formula and the structural formula are faceted into the generic portion and the salt or solvate portion. To provide for ease of computer recognition and registration, the generic portion of an acid salt is described in the protonated form in the first facet and the salt forming portion in the second facet. Amine salts are faceted into the free base and acid portions. Quaternary salts are faceted into ionic pairs. Molecular complexes are treated as single molecules with nonbonded parts. (The computer calculates and displays a faceted molecular weight corresponding to the faceted molecular formula during output.)

The reference field shows the source of the compound, generally the name and notebook reference of the scientist who prepared the first batch of the material or in the case of compounds from external sources, an appropriate code. (Samples of compounds bear, in addition to the laboratory number, a suffix corresponding to the batch number; data obtained on a sample also bear the batch number.)

The chemical structures in the input record appear in the form familiar to chemists, even to "crossed bonds". Certain conventions (typing rules) are followed to assure unambiguous interpretation by the computer. The structure segment of the input record may contain comments in free-text form offset by asterisks. An example of an input record is shown in Figure 10.

Incompletely Defined Structures. For certain types of incompletely defined structures, advantage may be taken of the Markush capability employing Z's and X's, as well as the element symbols "Any" and "Non-C" as has been illustrated earlier. Entry of such structures into the master file then

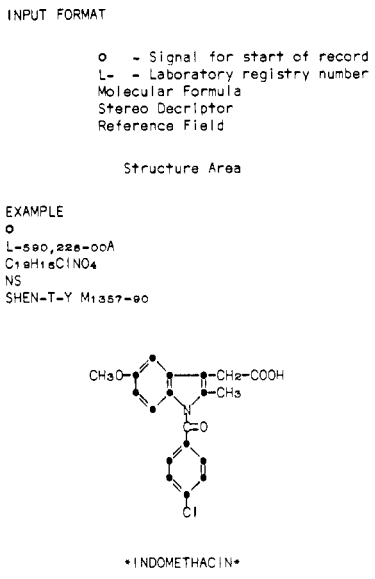

Figure 10.

permits retrieval by questions specifying any one or more of the specific structures they cover. In principle, whole classes of compounds could be defined in this way and would be very useful for special files. In practice, use of the capability has so far been largely limited to structures where the position of ring substitution is unknown.

In situations where the structure is less definite, a second capability is used. In this instance, the fragments of the molecule whose structures are known are shown as bonded to a molecular formula grouping flagged by a percent sign, as in CH_3-%$C_8H_{12}O$-NH_2. The record is retrievable by the molecular formula method previously given or by matching on the known fragments.

Tautomers. The Markush facility is used to handle the problem of tautomers. A set of tautomeric patterns is defined for the system. When one of the patterns is found in a compound, the edit program generates X's for the terminal atoms sharing the hydrogen atom. The program allows for the fact that a particular pattern may occur in several places in a single compound, and that a particular atom may participate in two or more tautomeric patterns as in guanidines. Duplicate compounds are thus properly recognized in updates when the input is in other forms. In substructure searches, retrieval occurs when a tautomer is expressed in an alternate protonated form.

The tautomeric patterns recognized include the atom strings of the form A=B–CH where A and C are oxygen or sulfur and B is sulfur, phosphorus, nitrogen, or carbon. The atom strings –NH–N=N– and –NH–C=N– are also recognized with no restriction as to whether the carbon or nitrogen atoms are cyclic or acyclic. The latter string thus allows tautomerism in amidines, imidazoles, and α-imino N-heterocycles. A special pattern is defined to make cyclic amides and cyclic thioamides equivalent to the hydroxy and mercapto forms. While this convention implies tautomerism in compounds such as caprolactam, the widespread occurrence of the functionality in pyridines, pyrimidines, purines, and the like militated for its easy recognition in either form. Additional patterns have been defined to make 4-hydroxypyridines equivalent to 4-pyridinones, as well as to recognize 1,2-prototropy in pyrazoles.

Simple keto–enol tautomerism is not recognized in compounds input to the file; questions involving keto vs. enol must be appropriately phrased. Cyclic systems of alternating double and single bonds are recognized inherently by the way the records of atoms comprising them are stored in the connection table.

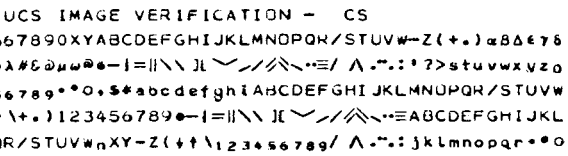

Figure 11.

Polymers. Polymers are handled by special conventions. Polymers are entered as the monomer, enclosed in brackets with a subscript "p". The end groups, which may be hydrogen, are shown on either side, outside the brackets. A comment set off within asterisks in the picture gives the specific details of the polymer. For copolymers, the monomers are shown separately and the comment includes information about proportions. Polymers are retrieved by any questions which would retrieve the monomer, with or without its end groups.

Peptides. In peptides of three or fewer amino acid units, the structures are input in atom-by-atom detail. In larger peptides the units are abbreviated according to the classical three-letter symbols treated by the system as nondirectional pseudo atoms. The units may be bonded together to form chains, branches, and rings and may also be joined to ordinary atoms and groups. The system does not recognize the equivalence of a three-letter code with its full molecular structure. Therefore questions requiring retrieval in both instances must be phrased with the three-letter symbols and multiatom fragments in an OR-type question.

Isotopes. The presence of isotopes is indicated in the comment portion of the input record, and thus is not a searchable parameter. In the case of hydrogen isotopes, "D" and "T" are acceptable in the structure diagram itself but are treated in the atom description table as if they had been expressed as "H".

Stereoisomers. The connection tables do not distinguish between stereoisomers, although structure diagrams may use dotted bond lines to make a visual distinction. For this reason structure records contain stereo descriptors chosen from a fixed set. When an atom-by-atom match on connection tables indicates that two compounds are duplicates, the stereo descriptors are compared, and when different, the latest compound is entered into the file as new but with a listing in the hold file report to warn of any stereoisomer found. If subsequent review shows it to be a duplicate, it is then deleted.

In a substantial number of cases this procedure will not give an adequate or reliable unique identification. In such cases, a free-form stereo descriptor, suitably flagged, is used to require a chemist's decision as to whether an isomer is different.

Input to System. All input, whether a file compound or a question, is typed on a standard IBM magnetic tape/Selectric typewriter, Model IV, with reverse search and reverse index. The input of questions follows a somewhat simpler format than that for a file compound. To facilitate atom-by-atom searching, structures in questions are oriented or even distorted to put the most unusual atom at the upper left. The typing element is that developed at the Smith Kline and French Laboratories[10] featuring "octobliques" and commercially available from Camwil, Inc., Honolulu, Hawaii. The magnetic tape output is converted to nine-track computer tape on a Digi-Data System 30 converter.

OUTPUT

Output is printed on an IBM 1403 impact printer, adapted for 10 lines to the inch vertical spacing, and operating at about 400 lines per minute, either on-line or off-line according to the operational convenience of the computer area. A special print train is used, having the character set shown in Figure 11. The lack of full size numerals in the typing element is

compensated for by input edit programs which test the context of numerals to determine which should be converted to full size for storage and output. The incomplete set of lower case letters in the typing element is filled out by a special typing convention which flags letters which should be converted to lower case.

OPERATIONAL DETAILS

Processing. The computer program which converts structural diagrams to bit screens and connection tables allows virtually all the condensations and abbreviations normally used by chemists. Parentheses and brackets can be nested to any depth. Structures are extensively checked for validity and ambiguity, and invalid or ambiguous structures are rejected. Warning messages are issued to call attention to doubtful points. The system will reject an otherwise valid structure if the atom counts do not match the molecular formula which was input. All accepted structures are printed for the run, and these are normally reviewed against original copy to assure that proper interpretation, particularly of isomers, was made by the typist.

Bit Screen. The bit screen consists of 288 bit screen positions, each precisely defined and falling into several categories: elements, bonded atom types, ring counts, and multi-atom fragments.

Element bits, of which there are 168, are of two kinds depending on frequency of occurrence of the elements. The common elements C, O, N, S, P, and the halogens are each assigned sets of bits based on atom count, so that all bits up to and including that count are turned on by the atom counts in the formula. The less common elements are assigned single bit positions, while the unusual elements are grouped together as a single bit position. The amino acid "elements" are each assigned bit positions. Hydrogen content is ignored completely in the element section of the bit screen.

The bits for bonded-atom types are derived from the atom description in the connection table. A total of 21 bonded-atom types are recognized. The more common fragments such as a carbon bearing a double bond and two single bonds (the benzenoid carbon) are distributed over several bits according to count. A total of 52 bits is reserved for bonded-atom types.

The remaining two classes of bits require processing of the structure as a whole and are thus essentially built-in search routines. Five bits are assigned corresponding to the count of rings. Naphthalene turns on the bits corresponding to one ring and two rings.

The set of fragment bits totals 63 for one or more occurrences of 57 fragment types. Most of these were taken from the WRAIR system and selected on the basis of statistical information supplied by WRAIR. Others were derived from our own experience.

Connection Table. The computer-stored connection table consists of two parts: an atom description list and a set of neighbor lists. The atom description list contains one entry for each nonhydrogen atom in the structure to a limit of 255 atoms and/or pseudoatoms. The neighbor lists provide information as to which atoms are bonded to which atoms but not by what type of bond. There is no attempt to make the connection table canonical, although some sorting is performed to facilitate matching.

Each atom description contains codes specifying the element, the valence, whether or not the atom is in a ring, the size of the charge (if any) on the atom, and counts of single bonds, double bonds, triple bonds, and hydrogen atoms. (The count of single bonds includes bonds to hydrogen as well as bonds to nonhydrogen atoms.) The atom description also contains a pointer to the neighbor list for that atom and provides for flagging atoms (such as X or Z) which need special treatment during atom-by-atom matching. The atom description codes are designed to facilitate comparison on an "at least" basis.

Use of Computer Resources. The system was written for the IBM S/360, Model 65, to be run under OS/MFT. It was subsequently run on a 370, Model 158, under HASP, and currently runs on SVS 1.7. The main system consists of an input conversion program, two edit programs, an update program, and a search program, to be run as successive job steps within a single job. Input can consist of search questions, new compounds, delete orders, and orders to release compounds from the hold file, all intermixed in any combination or any order. Compounds can be deleted and replaced in a single run. In practice, however, combining search and update runs has not proved operationally desirable. Instead they are submitted as two separate jobs at one time.

The programs require a 100K partition when run under OS/MFT. The present master file of 120,000 compounds requires (exclusive of the hold file) about 65 million bytes of disk space. Structure records average 309 bytes, picture records 124 bytes. The present file is contained on two 3330 disk packs (with the hold file included).

In operation, the program normally uses two disk drives (for the master file) and one tape drive (for input). An additional tape drive is frequently used so that reports are written on tape for printing off line.

Run times for adding compounds to the file have been averaging about half an hour per thousand compounds, clock time, with a fairly wide variation depending on what other jobs are being run simultaneously. CPU times show less variation and range from 7 to 8 min per 1000 compounds.

Search times vary greatly, depending on the number and nature of the questions in the batch. On six recent runs, chosen at random, the number of questions ranged from one to eight, and the clock times ranged from 10 to 23 min. CPU times ran from 53 sec to 2 min 16 sec. The matches on screens ranged from 600 to 15000 per run, and the number of compounds retrieved varied from 99 to over 1300. The averages for the six runs were 14 min clock time and 1 min 21 sec CPU time.

We have no way of getting accurate figures on the speed of atom-by-atom matching on real searches. However, dividing the total CPU time for the search phase of a run by the number of matches on screen (i.e., the number of atom-by-atom matches performed during the search) will give an upper limit, since the time used in atom-by-atom matching must be something less than the total CPU time used in searching. Applying this approach to the six searches described above, atom-by-atom matches are averaging something less than 11 msec per match. How much less can only be a guess, but our feeling is that the true average is in the neighborhood of 7 or 8 msec.

ACKNOWLEDGMENT

We wish to thank Walter Reed Army Institute of Research and, in particular, Mrs. June A. Page for making available its Chemical Structures Storage and Retrieval System. In addition to the program for converting structure diagrams to connection tables, the Merck system has taken over the format WRAIR uses for connection tables, its system of atom descriptions in connection table entries, its method of automatically handling aromatic bonds, and its approach to generating bit screens. The WRAIR approach to atom-by-atom (iterative) matching has also been adopted, with major modifications to handle the Markush facilities which do not exist in the WRAIR system. WRAIR also supplied statistics from its file of some 250,000 compounds (including frequency of occurrence of bits in its bit screens) which were of great value in designing the Merck bit screens.

We also wish to thank Chemical Abstracts Service for providing extensive information on their registry system which was most helpful in the design of the Merck system, particularly those aspects dealing with stereochemistry and tautomerism. The suggestions and support given us by Herner and Co., Washington, D.C., our contractor for keyboarding the original file, are gratefully acknowledged.

Finally we note our gratitude to numerous staff members of Merck and Co., Inc., for their roles in the design and development of the system. Particular thanks go to Dr. K. H. Cram, Dr. D. R. Hoff, Mr. W. L. Henckler, Dr. A. H. Land, Mr. H. B. Lewis, Dr. M. Leyzorek, Mr. M. G. Ly (deceased), Mr. R. T. Palmer, and Mr. W. A. Pater.

REFERENCES AND NOTES

(1) B. E. Holm, M. G. Howell, H. E. Kennedy, J. H. Kuney, and J. E. Rush, "The Status of Chemical Information", *J. Chem. Doc.*, **13**, 171–183 (1973).
(2) W. T. Wipke, S. R. Heller, R. J. Feldmann, and E. Hyde, "Computer Representation and Manipulation of Chemical Information", Wiley New York, N.Y., 1974.
(3) J. E. Ash and E. Hyde, "Chemical Information Systems", Wiley New York, N.Y., 1975.
(4) A. Feldman, D. B. Holland, and D. P. Jacobus, "The Automatic Encoding of Chemical Structures", *J. Chem. Doc.*, **3**, 187–189 (1963).
(5) D. P. Jacobus, D. E. Davidson, A. P. Feldman, and J. A. Schafer, "Experience with the Mechanized Chemical and Biological Information Retrieval System", *J. Chem. Doc.*, **10** 135–140 (1970).
(6) R. O. Pick, E. H. Eckermann, J. A. Schafer, and J. F. Waters, "Strategy of Data Retrieval and Analysis from Large Biological and Chemical Files", *J. Chem. Doc.*, **12**, 35–37 (1972).
(7) E. H. Eckermann, J. F. Waters, R. O. Pick, and J. A. Schafer, "Processing Data from a Large Drug Development Program", *J. Chem. Doc.*, **12**, 38–40 (1972).
(8) A. N. DeMott, "Interpretation of Organic Chemical Formulas by Computer", Spring Joint Computer Conference, 1968, p 61.
(9) A. N. DeMott, "Computer Processing of Non-Linear Text", Proceedings of the Second Hawaii International Conference on System Sciences, University of Hawaii, Jan 1969.
(10) R. Gottardi, "A Modified Dot-Bond Structural Formula Font with Improved Stereochemical Notation Abilities", *J. Chem. Doc.*, **10**, 75–81 (1970).

Information Activities in Support of the EPA Pesticide Program†

WILLIAM C. GROSSE

Technical Services Division, Office of Pesticide Programs, U.S. Environmental Protection Agency, Washington, D.C. 20460

Received November 5, 1975

The Environmental Protection Agency (EPA) is responsible for regulating the supply and use of pesticides. The EPA Office of Pesticide Programs (OPP) is involved in (1) supply control via product registration, (2) use control, (3) monitoring and hazard evaluation, and (4) research and economic studies. Major information activities in support of these four activities are reviewed.

The Environmental Protection Agency (EPA) is a regulatory agency. Among its familiar programs, such as those directed toward cleaner air and purer water, is another, Pesticides, which constitutes somewhat of a special interest in connection with the first two, and has potentially far-reaching implications for the world's agricultural community as well. Indeed, pesticides have brought substantial benefits to man. Yet, these benefits to our health, welfare, and comfort, which arise from using chemical compounds to control undesired forms of animal, plant, and microorganism species, are sometimes offset by the adverse effects that they also can have on man and the environment.

With the passage in 1972 of an amendment to the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) of 1947, EPA was given broad new powers and responsibilities to regulate the supply and use of pesticides so that, with the desire for ever more benefits, and the urge of the marketplace to provide them, sufficient attention is given to minimizing the negative consequences as well.

FIFRA, with the 1972 amendments, is a comprehensive new law. It provides for Federal regulation of all pesticide products that are marketed and used within the United States, while preserving the right of the States to regulate products within their own boundaries. It assures a more complete scientific review of pesticide products, so that the circumstances of their use are designed to minimize the potential for creating undesired adverse effects. It provides for denying use when the expected or observed hazards of such use far outweigh the benefits, and it incorporates penalties for misuse. It provides for training those who intend to use highly toxic products and those who apply pesticides as a commercial service. It seeks more descriptive and understandable product labeling, so that information on proper care and use may be communicated more effectively. It also provides for specific responses to special or emergency needs, by permitting use of certain chemicals under highly controlled conditions.

While the foregoing list of provisions is by no means complete, it does serve to highlight most of the major ones and provide some background for a discussion of EPA's Pesticide Program and the information activities that are emerging to support it. A diagram of the process of product registration and regulation is shown in Figure 1. Applicable sections of the FIFRA are shown for the major activities.

The EPA component responsible for conducting the program to carry out FIFRA is the Office of Pesticide Programs (OPP). Toward this end, OPP has developed a four-point strategy through which the program is divided into major thrusts and under which it is actually conducted. These four points are:

1. Register and classify all pesticide products to identify more precisely those that can be safely and effectively used.
2. Provide more fully for the safety of pesticide use and facilitate improved product labeling, packaging, and education programs to reduce unwarranted adverse effects due to misuse.
3. Improve hazard evaluation activities and the monitoring programs that are necessary to support them.

† Presented in the symposium on "Information Requirements Resulting from Environmental Impact Laws", Division of Chemical Information, 170th National Meeting of the American Chemical Society, Chicago, Ill., Aug 27, 1975.