Computer Interpretation of Biological and Chemical Data*

B. B. MORPHIS, N. TORBET, W. W. HUNTER, and F. K. BROOME
Science Information Department, Abbott Laboratories, North Chicago, Illinois.

Received September 15, 1965

Processing of biological and chemical data by computer programs is briefly described. Application is made in up-dating drug records and searching organic chemical structure and/or biological data files, with complete interpretation of coded biological information into human language and printout of two-dimensional organic chemical structures.

The data referred to in this paper are those generated by our primary screening programs in the Research Division.

Successful exploitation of the accumulated data from many years of research on the pharmacological application of chemicals requires the means for relating two or more elements of the data. The most important of these elements are: species used, test performed, test results, and chemical structure of the compound. Since 1957–58, our means of relating various elements of these data have been machine-readable tab cards. We designed our card system on the principles that data should be recorded once, as near the source as possible, and in a machine-processible form. From the beginning, it was planned that as these files became too large for convenient handling, they would be put on magnetic tape and be manipulated by means of computer programs. In late 1963, we initiated the feasibility study for computerizing our primary screening data. At that time the total number of cards in our files was about 180,000.

Computer programs were written to perform file analysis to facilitate the feasibility study. We needed to know the minimum, average, and maximum number of cards the files contained for each compound. All the various card formats had to be considered. All elements necessary to determine the layout of the tape record of a compound had to be elucidated.

Because of the heavy demands on our Corporate Systems department, on which we have to depend for major system design and programming support, the creation of our magnetic tape system has not progressed as rapidly as we had anticipated. We do have our master tape file created. The necessary file maintenance programs are functioning. We also have the print-edit program working, which interprets and formats the data for the printed reports. We do not have our search program completed at this time. The search program will be a large general-type program which can handle a number of query cards at one time.

* Presented before the Division of Chemical Literature, 149th National Meeting of the American Chemical Society, Detroit, Mich., April 6, 1965

OBJECTIVES

The objective in placing our organic chemical-structure and biological-data files into a computer program were, first of all, to have the ability to search the entire file for specified chemical structures and/or biological information. A complete interpretation and printout of the chemical structures and biological information for the investigator was desired. This printed record should remain up to date as new data are accumulated. It should be in a language familiar to the investigator. Up-to-date reports on a compound should replace older ones to maintain records that would be brief as well as current, thus eliminating the job of hand posting screening data. The program should be able to generate screening reports and annual reports along with relevant statistics. We feel the program should serve both the research investigator and research management.

Organic Chemical Structure Cards. In 1956, the late Dr. Marvin Spielman and Nancy Torbet began to use what is now called the Spielman code. This is a multiple-punch code used for representing Abbott chemical structures in tab cards for machine searching. The code is a fragmentation one, similar to many of those now in use. It describes chemical compounds according to ring systems, carbon chains, and functional groups present. Structures are also hand drawn on the tab cards into which they are coded. The cards are searched on an IBM 101 Electronic Statistical Machine, resulting in a selected deck of cards through which the scientist can thumb to eliminate those structures not desired. This has caused considerable wear and tear on the cards, and no permanent record of the search, other than a list of Abbott compound numbers, has been readily available.

This multiple-punch code is acceptable by the IBM 101 sorting machine for searching, but it cannot be converted directly to magnetic tape. Since we did not have a column binary card reader, special computer programs were required to convert these cards, in stepwise fashion, to the tape record for computer searching. With the chemical-structure file on magnetic tape rather than on punch cards, we will be able to do several chemical-

structure searches at one time without the present concern about possible overlap of searches.

Organic chemicals are most conveniently expressed and most widely understood in terms of two-dimensional structural formulas. Therefore, we decided to use the "Hecsagon" chemical structure display developed by Horowitz and Crane (*1*) at Eastman Kodak for our computer printout of chemical structure. The structure of each compound is drawn on quadrille-ruled paper, edited, and key-punched, using one card per line of structure. One person working about three days per week drew some 20,000 "Hecsagon" structures in seven months. We expect the job to be completed in three to four months. A computer program was written to compact the "Hecsagon" structures onto magnetic tape and to print them out from the tape with the Burroughs 280 computer. We have been putting the structures onto the tape as they are keypunched, thus making them available for printed records of current data. The investigator will receive a printed structure of each compound that meets the requirements of his search request.

Still another card file had to be generated for input to the tape record, giving for each compound the molecular formula, source, notation of any clinical work, and date coded. This card is used with a computer program to compute the molecular weight for each compound. A new card is generated in this process, containing the above information as well as the molecular weight. This is the first card read onto the tape record of a compound.

Biological Data Files. Our biological data are contained in two major card files: toxicity and symptomatology data cards (30,000); and screening data cards (125,000).

In 1958 Dr. J. D. Taylor and B. B. Morphis began coding Abbott biological screening data using codes patterned after the CBCC (*2*) system. These card files were used to produce current screening reports and retrospective searches. By sorting and collation techniques the chemical structure cards and the biological data cards could be used for structure-activity correlation studies. These card files are now too large to conveniently handle in this fashion.

THE TOXICITY AND SYMPTOMATOLOGY DATA are generated from primary mouse toxicology studies done by our neuropharmacology section. Most of our organic compounds receive this primary screening before being submitted for other testing. The data obtained from the toxicity and symptomatology screening program provide the pharmacologist and chemist with a pharmacological profile of a compound. Figure 1 shows the card layout for these data. We use numerical codes to record these data into fixed fields. Each symptom is assigned a specific column, and a number indicating the dose level at which the symptom occurred is punched into that column. These cards were converted directly to the tape record. The computer program to interpret and print these data is a very significant asset to the automated system.

Figure 2 shows a card record and computer-interpreted output. The investigator no longer needs a guide or key to interpret the printout. A dictionary was set up in the computer program to tell what each column and the punch within it represent. The advantage of having these types of data in a flexible magnetic-tape file for computer searching is readily seen. The investigator can specify the profile of desirable compounds for a particular screening program and let the computer find them.

THE BIOLOGICAL SCREENING DATA obtained from past

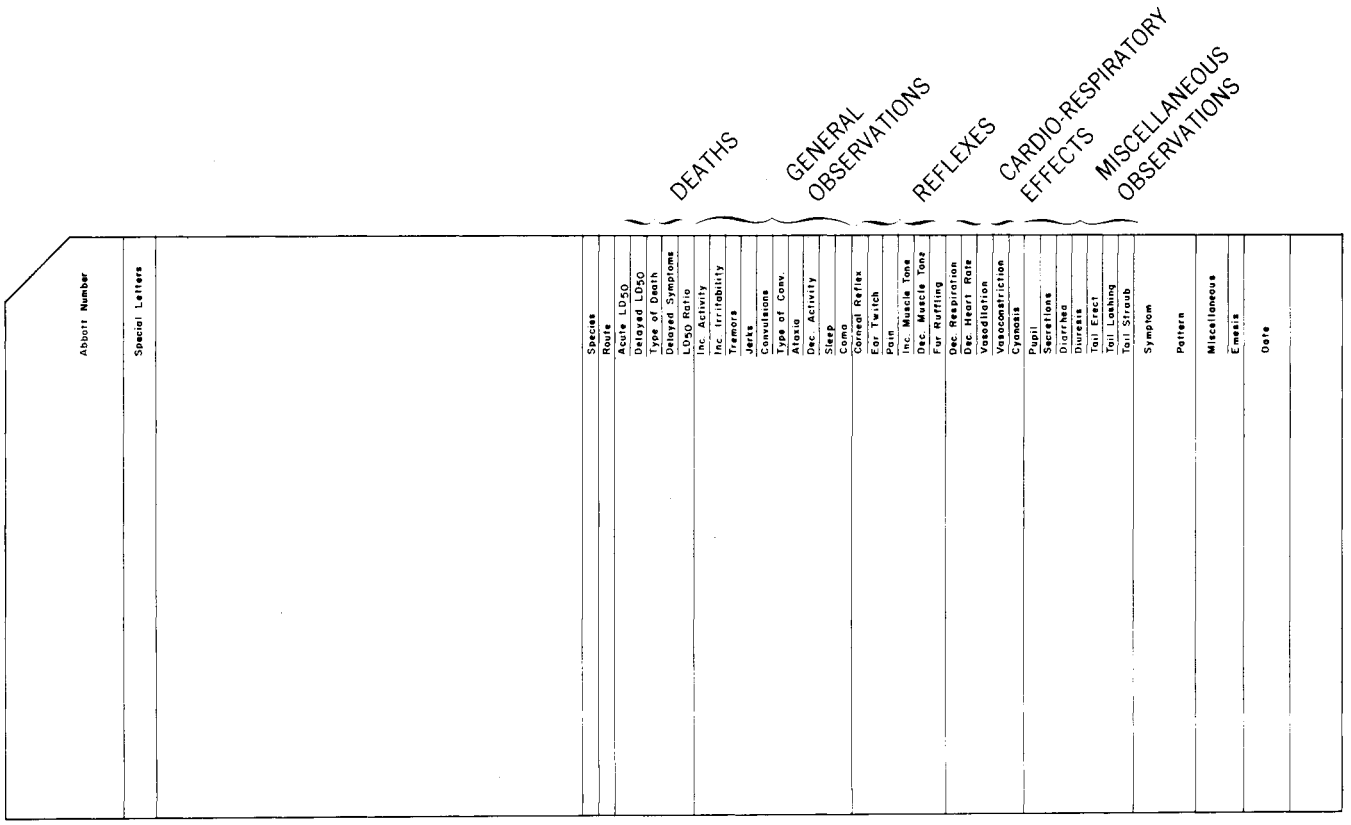


Figure 1. Card layout for toxicity and symptomatology data.

```
7207                026      3 44 43                        561          <-- Card record

                                                                         Computer interpretation
A NUMBER  7207                                              10-22-64

TOXICITY AND SYMPTOMATOLOGY  ROUTE IP  MOUSE LD50= 500   MG/KG   ALL DOSES ARE APPROXIMATE MG/KG   DATE  5-61
    INC IRR  100 , JERKS   200 ,CONVUL   200 ,ATAXIA   200 ,DEC ACT  100 ,
```

Figure 2. Toxicity and symptomatology data on compound 7207.

and present screening programs indicate more specific types of activity exhibited by the compounds. Leads to active compounds may be obtained from one screening program for testing in a completely different program. For instance, one of our *in vitro* enzyme programs may supply leads to active antihypertensive agents. In order to facilitate data handling on large-volume screening programs, we modified protocol forms and laboratory notebooks where needed, and arranged to keypunch data directly from them. We keypunch the standard alphabetic abbreviations used by the investigators to designate the screening program and test results. The keypunched cards are used to prepare the screening data report for the investigator. We are using an IBM 870 Document Writing System for this purpose.

The screening data cards contain fixed fields which are the same for all screening programs (Figure 3). These fields contain the Abbott compound number (which is the accession number to all our records), screening program, test number, species, frequency of dosage, amount of dose, route of administration, etc. All other columns in the card are available for recording variable information, which is different for each program.

One computer program was written to convert the fixed fields of all screening data cards to tape, and separate subroutines were written to convert the variable information from each screening program. The program for interpreting and printing the data also required a separate subroutine for each screening program.

SYSTEM OPERATION

Figure 4 is a flow chart of the system.

Figure 5 shows a computer-produced Drug Record Sheet for a compound. Our tape record of a compound contains the following information:

- Abbott number
- Molecular weight
- Source of the compound
- Notation of any clinical work
- Accession date
- Molecular formula
- Two-dimensional "Hecsagon" structural formula
- Chemical-structure search code
- Toxicity and symptomatology information
- Data from screening programs

We will be able to search and print out any part of the record as desired, with the exception of the "Hecsagon" part,which is used for printouts only. Each part of the record has a fixed length and location, with the exception of the screening data. These data, of necessity, are variable in length. The screening tests per compound may vary from none to many.

In a search, the "hits" will be written onto a hit tape, which is then used with the print-edit programs to interpret and arrange the output into any order desired. The hit tape will be retained until the user assures us that the output has been arranged in as many classifications as he needs.

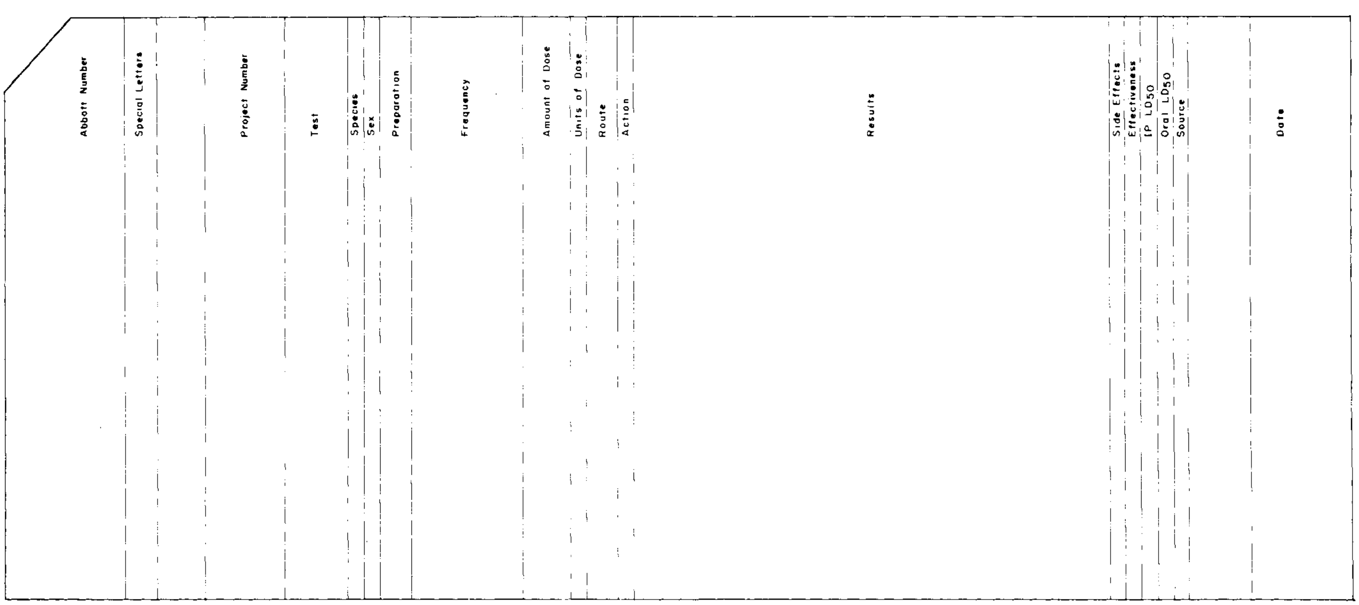


Figure 3. Screening data card.

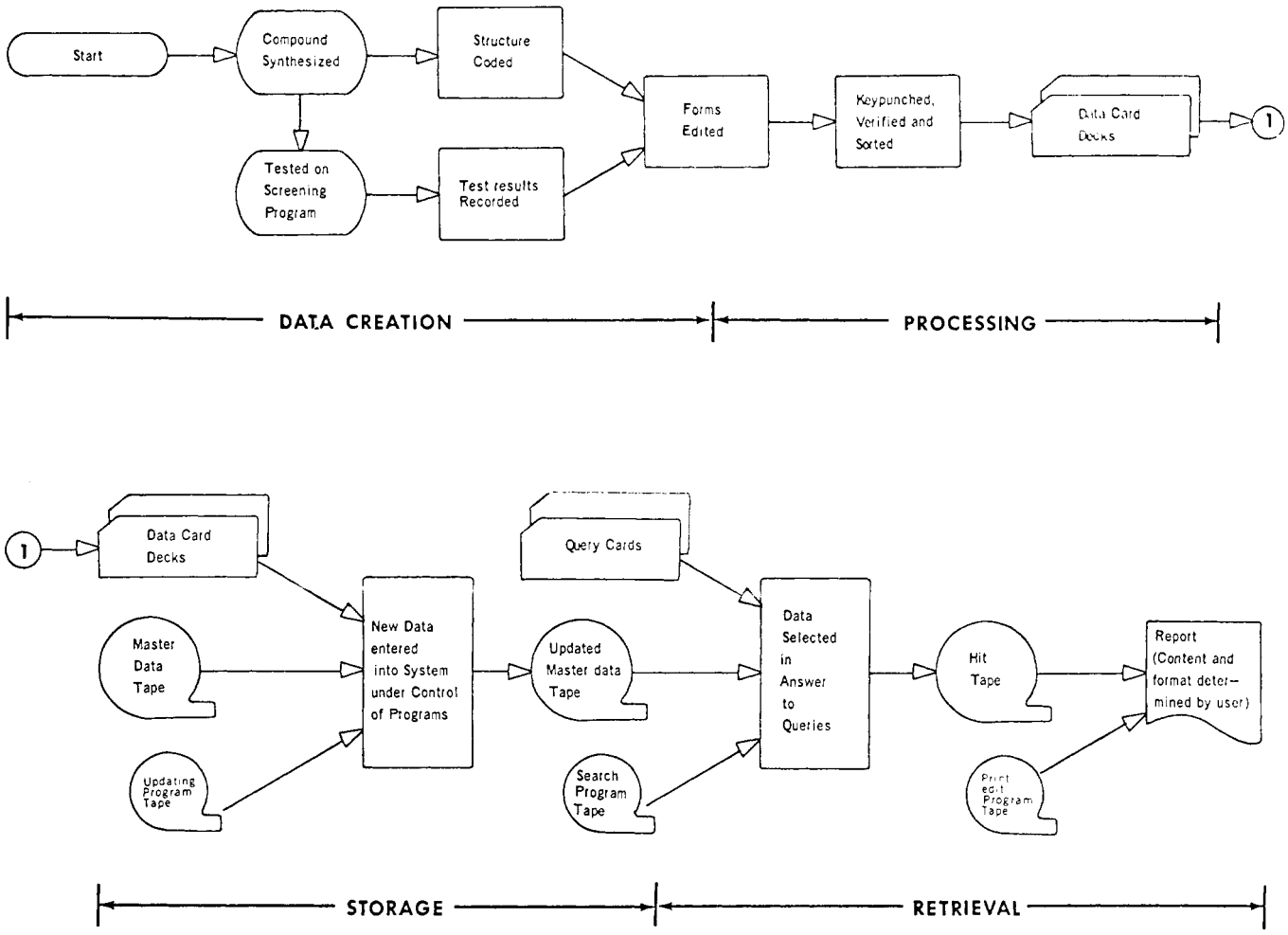
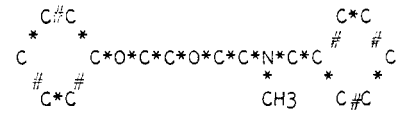


Figure 4. System flow chart.

A-NUMBER 5578 MASTER DATA RECORD AS OF 2-13-65

MW 285 39 SOURCE ABBOTT. NO CLINIC DATA, DATE 6-46. FORMULA C18 H23 N1 O2



TOXICITY AND SYMPTOMATOLOGY ROUTE IP MOUSE LD50= 750 MG/KG ALL DOSES ARE APPROXIMATE MG/KG DATE 3-61
INC IRR 100 , JERKS 500 ,DEC ACT 750 ,DEC RESP 750 ,VASODIL 100 ,DIURESIS 100 ,
DIARRHEA 500 .

AH-4 INEFFECTIVE VERSUS EGG WHITE-INDUCED BRONCHOSPASM IN GUINEA PIGS @ 40.00 MG/KG IP

AL-3 CAT BLOOD PRESSURE 10.00 MG/KG IV + 14 MMHG 5 MIN. CAROTID OCCL INH.,EPI POT., SYS P+ 18 DIA P+ 10

AP-1 60 TIMES LESS POTENT THAN ATROPINE ON ISOLATED RABBIT ILEUM

LF-3 @ 10.00 MG/KG OR IN RATS, NO EFFECT ON FOOD INTAKE

Figure 5. Computer-produced drug record sheet.

There is provision in the computer programs to delete or correct data when an occasional error in original recording may be found.

SUMMARY

Having our data on magnetic tape makes prescreening of compounds by computer possible. The computer can select those compounds which have desirable characteristics for testing in a specific screening program.

When a compound shows activity in some screening test we can easily pull from the tape file all similar compounds which have not been tested.

Complete records of compounds selected for screening may be printed out for review by an investigator. These records aid the investigator in planning his testing.

It is the ability of the computer to store, process, compare, search, and produce the completely interpreted record of the data that will make this system so useful to our Research Division.

LITERATURE CITED

(1) Horowitz, P., Crane, E. M., "HECSAGON: A System for Computer Storage and Retrieval of Chemical Structure," Technical Memorandum, Kodak Research Laboratories, Rochester, N. Y., 1961.

(2) Seitner, Philip G., Livingston, George A., Williams, Ann S., Eds., CBCC: Chemical-Biological Coordination Center Biology Code. A system for coding results, procedures, and conditions of tests for biological responses to chemicals, Publication 790 National Academy of Sciences, National Research Council, Washington 25, D. C. 1960.

A New Information System for Organic Reactions*

HERBERT J. ZIEGLER
13, Avenue Cardinal Micara, Brussels 16, Belgium

Received September 15, 1965

An information system for organic reactions on slotted punched cards for the use of the chemist in his laboratory is described. Using no new symbolism, it enables reactions to be found in the same simple manner as are chemical compounds, and provides information concerning structure and functions of initial and final products, reacting groups, reaction conditions, possibilities of synthesis, and side reactions.

It is a truism to say that it has become impossible for an organic chemist to remember, assimilate, or even read every article which describes a new method of preparation in organic chemistry. That information—deriving from these sources—should be readily available, is nonetheless a pressing necessity. To acquire facts concerning organic reactions, we need a reference tool, from which it should be possible to ascertain directly the appropriate method to solve any particular problem of synthesis whenever the information is required.

What form should this information tool take?

A collection of information on chemical compounds presents no difficulties of principle. Each molecule is defined by its elementary composition and can therefore be traced either through its formula or its name. This latter method of identification may sometimes cause difficulty, but the problems involved are negligible compared with those encountered when the information required comprises chemical reactions and methods of preparation, since a reaction cannot be defined as easily as a molecule. A chemical reaction is no static entity, but a transformation of one compound into another. Often, when a working method is being sought, it is not a question of merely reproducing the synthesis of another chemist, but rather one of finding the method which may be applied by analogy, with the greatest likelihood of success, to the synthesis of a new compound. Thus, the question is not one of recall, but of discovery.

If such data are sought by hunting for descriptions of synthesis of similar molecules—for example, by turning to the Formula Index of *Chemical Abstracts*—we run the risk of just missing the most promising methods, since, unfortunately, the analogous compound may be poorly selected; the publication which might have indicated the desired solution for a particular problem deals, for instance, with a compound which differs from the selected analogous compound only by the presence of a chlorine atom or a methyl group. Differences of this kind, which are small enough in themselves, constitute insuperable barriers to effective documentation, that are all the more dangerous because they are invisible.

* Presented before the Division of Chemical Literature, 149th National Meeting of the American Chemical Society, Detroit, Mich., April 6, 1965.

* As a result of the appreciated suggestion of Professor S. Hünig, the working methods borrowed from the periodical chemical literature will in future be quoted in unabridged form.