Interactive Searching of a Structure and Biological Activity File*

V. B. BOND, C. M. BOWMAN,** N. L. LEE, D. R. PETERSEN, and M. H. RESLOCK
Computation Research Laboratory,
The Dow Chemical Co., Midland, Mich. 48640

Received June 10, 1971

A large file which contains structural information in the form of fragments based on the Wiswesser Line Notation can be searched interactively using a time sharing system. Search can be carried out using any simple boolean expression. Retrieval can be either accession number or complete notation. The interactive nature of the searching program permits recycling the modification of the query until the desired results are obtained. Coupled with the searching program is a program which scans a file of herbicidal activity for the same compounds. Searches can be made on compound number, plant organism, concentration, or activity. The use of these two programs permits the chemist to explore a variety of hypotheses which will serve as a background for further testing or synthesis.

Over the years Dow has accumulated information on the biological activity of several thousand compounds. Although a number of tools have been developed to search the information, use of the file has not been as great as expected. In the last year, efforts have been made to make these files of information directly available to those who generate them to stimulate a more effective utilization of this valuable accumulated data.

This paper reports the development of two computer programs which permit the chemist to address the files of structural and biological information directly via teletype in a time-sharing environment. We first discuss the characteristics of the two data files, then the structure searching program and the biological searching program. Finally, we discuss briefly how these two programs are used in conjunction with one another.

DATA FILES

Both the structure and the biological data files refer to the same materials. The files contain information on about one hundred thousand compounds. The structural information is represented in the form of a Wiswesser Line Notation.[1] In addition to the notation, each compound record contains a molecular formula and a compound number. The biological files contain information obtained in the screening for herbicidal activity. The various items of information, in addition to the compound number, are the concentration in which the material is used, the organism on which it is used, a test method number, and the screening result. The original file contained all the results for all the tests ever run on that compound. For searching purposes the data need not be a complete historical record, so we eliminated duplicates as well as redundant results. That is, if a particular compound showed activity at one dose level, we did not include results for tests in which the compound showed the same activity at higher dose levels. During the process of eliminating duplications and redundancies, unusual situations were retained. For example, if the compound that showed activity at a low dosage level did not show activity at a higher dosage level, these unusual test results were retained in the searching file. Because of the size of these files they are maintained on magnetic tape and searching is done in the serial fashion using the Burroughs B5500 time-sharing system.

STRUCTURE SEARCHING

The structure searching program is an interactive program which permits the user to frame questions and obtain results in a very short period of time, thus enabling him to recycle and modify his questions repeatedly to obtain acceptable answers. To do the searching in a fast yet effective manner, 156 screens or fragments were selected which the user is allowed to use. These screens were derived from the Wiswesser Line Notation and from the molecular formula. This particular small set of fragments was selected to provide a rapid, simple and selective searching tool to be used on an interactive basis. For more routine or complex searches, we resort to the larger and more complex fragmentation and searching system reported earlier.[2]

Batch searches can be used to create a sub file which then can be used for on-line searching, thus reducing time and expense and increasing the accuracy of the search results. The particular fragments selected as screens are the symbols used in the Wiswesser Line Notation, including several locants (notation position indicators). In addition, most of the commonly occurring elements as well as unique combinations of two and three Wiswesser symbols have been allocated fragment code numbers. Thus, the VQ fragment is used to identify acids. Certain ring skeletons, and a few very specific rings, such as pyridines, are identified separately. Each fragment has been assigned a bit in one of the four computer words allocated for searching purposes. A fifth computer word contains the compound number.

*Presented before the Division of Chemical Literature, 161st Meeting, Los Angeles, California, March 31, 1971.

**To whom correspondence should be addressed.

A search is carried out by building up a series of bits representative of the fragments which are to be searched and then sequentially matching these bits against the bits in the searching field for each one of the 100,000 compounds. This process requires about two minutes of searching time.

To provide flexibility, several logical combinations are allowed. The first, and perhaps the most commonly used, is the AND option, in which each one of the fragments indicated must occur. An AND/OR combination permits linking together a series of fragments, one of which is to be present, with other series in each of which one or more fragments must also be present. We have found, however, that an indiscriminate use of the OR facility causes an extremely large number of irrelevant compounds to be retrieved thus lessening the utility of the search program. An AND/NOT combination permits the presence of certain fragments and the absence of others. The final logical combination is the NOT, which is very rarely used, and in which certain fragments are excluded in their entirety.

In running a search, the user sits at a teletype, calls out the program, and in a conversational mode, communicates with the program. He has the option of suppressing certain messages and abbreviating others if he is familiar with the program. He then builds up his logic and his fragments, which result in a series of bit masks which will then be matched against the tape. The results are stored on disk in intermediate storage. Upon the completion of the search he is given the option of obtaining the compound numbers directly on the teletype or having them printed on a line printer and mailed to him. He can also receive the notation either on the teletype or the line printer. This requires an additional lookup step. He may choose to omit the search results printout. He may at this time conclude his search, delete the results, or store the results permanently in his library; or he may reenter the program and run additional searches. We have provided the user with two intermediate disk areas. Thus he can go back and ask additional questions of the entrance disk files narrowing down the results he obtained in the first pass. This recycling process can be repeated as many times as he desires until he is satisfied with his answers. At that point he can have his results printed or he can have them saved permanently on disk for future use.

BIOLOGICAL ACTIVITY SEARCHING

The program which permits the searching for biological activity operates on a series of magnetic tapes which contain separate records for each one of the tests performed on the compounds in the file. As indicated earlier the file has been abbreviated by elimination of duplications and redundancies. Every entry represents a test result and contains a compound number, concentration of the material in terms of units and amount, the organism on which the test was conducted, the test method number, the result in terms of a percentage (100% being favorable), and a small area for codes for unusual types of behavior. The number of search parameters available is much greater than for the structure file, thus the searching of this file is much more complicated. A standard question in this instance would be retrieval of all the compounds which showed 100% activity below a certain concentration for a particular organism. Since we are now dealing with discrete numerical amounts, we have to allow for ranges of concentration and activity and combinations of organisms. The program has been written in a conversational mode so the user merely needs to answer questions that will be posed to him by the program. He does not have to worry about how complicated the searching logic is going to be. We have found by experience, however, that if we provide the user with a sample of the questions that are going to be asked of him, he can go through these ahead of time and spend less time at the console.

The user can also include information to limit his search to a particular set of compound numbers or to a compound number range. The possible input parameters to this program are so numerous that it became important to make this program an interactive system which allows the user to repeat his questions, to rephrase them to achieve the answers that he really wants. As has been found so often, the user does not really know exactly what he wants until he sees what he is going to get. The ability to recycle a question and to modify the queries adds significantly to the utility of the system.

The output from this biological screening program is the stored data which matches the search parameters for that particular test. The user sees presented on the teletype an organism number, the concentration, the activity, and the compound number.

STRUCTURE AND ACTIVITY SEARCHING

The rationale behind the design of these two programs was first of all to provide the user with something that he could manipulate himself without the intervention or interference of an information specialist. It also was meant to be used by the user who generates the information thus allowing him to ask questions and draw his own conclusions from the data which are in the file. Several of our users have combined the programs by first finding structural components in which they are interested and then using the compound numbers obtained from that output program to interrogate the biological activity files using the second program. This double use of the programs has only been started in the last few months, and we cannot report any great successes in terms of new products or new advances in technology. However, we do find that our chemists are using these programs to analyze the data which have been available to us for many years.

PERFORMANCE OF PROGRAMS

Whenever one reports the use of computer programs, it is important to report their performance also. The structure searching program is capable of searching 100,000 compounds in a matter of two to four minutes. Once the initial pass has been made and the smaller file has been deposited on the disk storage it requires only a matter of seconds to make additional searches. The amount of time required in framing the question and developing search strategy has been considerably cut down by abbreviating instructions and questions directed to the user. Such a search could be done easily in a matter of five to ten minutes. The output portion of the program of course is highly dependent on the number of results that are obtained, and whether one wants the numbers or the notations. The economics of our time-sharing system permit a cost of between $10 to $25 per typical search.

The complexity and size of the biological file is such that it is very difficult to give accurate figures. There is no such thing as a typical question. Suffice it to say that a search of the entire 100,000 compounds for herbicidal activity, would require a matter of 30 to 40 minutes. In practice, questions are modified to a much greater extent, and abbreviated so that they require 15 to 20 minutes per question.

The availability of interactive programs to search structures and to search biological activities provides a tool which permits the working chemist to address large banks of information individually. This capability leads to better and more imaginative research as well as increased performance of the user. Users recognize this, and several of our top scientists are now using these programs.

LITERATURE CITED

(1) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968.
(2) Bowman, C. M., F. A. Landee, N. W. Lee, M. H. Reslock and B. P. Smith, "A Chemically Oriented Information Storage and Retrieval System. III. Searching a Wiswesser Line Notation File," *J. Chem. Doc.* **10**, 50-4 (1970).

Evaluation of the IBM Administrative Terminal System and Magnetic Tape Selectric Typewriter for Text Processing*

Herman Skolnik† and William L. Jenkins
Hercules Incorporated
Research Center
Wilmington, Delaware 19899

Received February 3, 1971

The Administrative Terminal System (ATS), designed by IBM for use with System/360, is a remote, on-line, timesharing, input-output device which allows interaction with the Central Processing Unit (CPU) in a conversational mode and with computer peripherals. ATS operations are detailed, the use of ATS for text processing is evaluated in comparison with the IBM Magnetic Tape Selectric Typewriter (MTST), and the importance of the interaction of ATS with computer peripherals is described.

Typing is an essential operation in practically every communication system. Yet it is today what it was when the first typewriter was introduced: a completely human-controlled operation. Even though the typewriter of today is far superior to its predecessors, the productivity of the average typist is about 60 words per minute, or about the same as it was 40 years ago. In most working environments, however, the average typing output from dictation or handwritten copy is something under 30 words per minute, after corrections and changes by the writer.

Because typing is a time-consuming and costly step in the communication process, and particularly so in most phases of chemical documentation, we have tried to be alert to new methods or mechanisms that reduce typing time and costs[1, 2, 8]. We have studied and designed new posting methods, investigated and utilized the IBM 870 Document Writer[6], and evaluated TEXT 360[7].

This paper reports our evaluation of the ATS (IBM Administrative Terminal System) relative to the MTST for text processing of translations, technical reports, manuals, form letters, directories, mailing lists, etc. Our understanding of the ATS software indicated that it could be particularly suitable as a remote, time-sharing, input-output terminal for text processing[4].

ECONOMIC CONSIDERATIONS

If we assume the cost of a typist to be $5.00 per hour with overhead, then the use of an MTST or ATS must increase the productivity to more than offset the additional cost. A one-tape MTST rents for about $200 per month and a two-tape MTST for about $300 per month; an ATS plus a Data Set rent for about $130 per month. The MTST rentals add $1.20 and $1.80 per hour and the ATS terminal rental adds $0.77 per hour to the cost of the typist based on 168 hours per month. Thus, the productivity of the typist must be increased by at least 24% or 36% with the MTST, or 30% or 45%, respectively, when we add the cost of the magnetic tape cartridges ($40 per month replacement cost) to the rental of the MTST. In the case of the ATS, whereas the terminal rental requires a productivity increase of only 15%, the cost of the computer can run from $1.50 to $7.50 per hour for terminal connect time, depending upon how many terminals are time-sharing the same CPU block. Assuming a computer cost of $3.00 per hour, the ATS must increase the productivity by at least 75%.

It is quite apparent from these economic considerations that the MTST and ATS are not for all typing assignments. They are not economically feasible, for example, for typing assignments that are not subjected to revisions by the writers or that are not repetitive in part or in whole. They are most economically feasible for repetitive typing assignments, for text that needs to be revised and updated, and for text that needs to be communicated in different formats.

FEATURES OF ATS vs. MTST

Table I, which summarizes the features of ATS and MTST, shows that the ATS is a far more flexible system and provides many features not available on the MTST. The greater flexibility and additional features of the ATS are possible because it is on-line with the 360, whereas the MTST is a self-contained unit:

*Presented at the 6th Middle Atlantic Regional Meeting, Baltimore, Maryland, February 4, 1971.

†To whom correspondence should be addressed.

Hercules Research Center Contribution No. 1544.