Strategy of Data Retrieval and Analysis from Large Biological and Chemical Files*

R. O. PICK,** E. H. ECKERMANN,*** J. A. SCHAFER, and J. F. WATERS
Division of Medicinal Chemistry, Walter Reed Army Institute of Research, Washington, D. C. 20012

Received October 25, 1971

Modern research programs in chemotherapy and drug development have generated huge masses of data. The method of attack used by an investigator to find information in his files depends on the size and content of those files and his means of analysis. An example approach to the analysis of data from the chemistry, biology, and inventory files of the antimalaria drug development program at the Walter Reed Army Institute of Research is presented. A general approach to the analysis of data from large files based on the organization of those files is discussed.

Modern research programs in drug development and chemotherapy have generated huge masses of data. The problems involved in searching and retrieving relevant information from these large files become comparable to those involved in searching the literature itself. The method of attack used by an investigator to find information in his files depends on the content and size of those files, his desires, and the means of analysis available.

An "investigator" in a drug development program may be an administrator, biologist, chemist, physician, pharmacologist, or a member of any other profession involved in biomedical research. The data base generated by such a program must be used by all of these people, and each will have a different viewpoint as to what is of interest. We shall use the data base generated by the Army Antimalarial Drug Development Program in examples of several analyses of data from large biological and chemical files.

The technical objectives of the Army Antimalarial Drug Development Program are to carry out research on the prevention, treatment, and cure of malaria. This research involves chemical, biological, pharmacological, toxicological and clinical studies, and the development of new approaches for study. Work is directed toward (1) the chemical synthesis of potential antimalarial compounds; (2) testing of all compounds synthesized and of thousands of off-the-shelf compounds for antimalarial properties; (3) the study of pharmacokinetics and toxicity of compounds which have shown antimalarial activity; (4) continued study of candidate drugs by administering them to volunteers, accompanied by complete pharmacologic, toxicologic, and therapeutic investigations; and (5) the study of the various mechanisms of drug resistance.

The data base consists of interfaced inventory data files, chemical data files, and biological data files which are described in detail elsewhere. The inventory file[1] contains such information as: compound identification number, location, source, quantity on hand, and a history of shipments made from that sample. Over 300,000 physical samples are represented, with nearly 500,000 shipments. A generalized search program is used to extract records relevant to user queries.

The main chemical data file[1] searchable by number, chemical identity, and chemical substructure,[2] contains over 200,000 structures. The biological files,[1] which include all raw data, contain over 2,000,000 data records encompassing fourteen antimalarial test systems.

This large data base is somewhat unwieldy, and therefore, the first step in data analyses is to determine what parameter is of basic interest to the investigator and to set limits on that parameter. This can be considered as the "primary question." Once a subfile consisting only of answers to this question is created, further file manipulations are simplified due to the decrease in file size. An example of such manipulations is shown in Figure 1. In this case, the primary question was a nonchemistry question, such as "What compounds show activity in test system X?". The answers to this question, asked of the large bio-

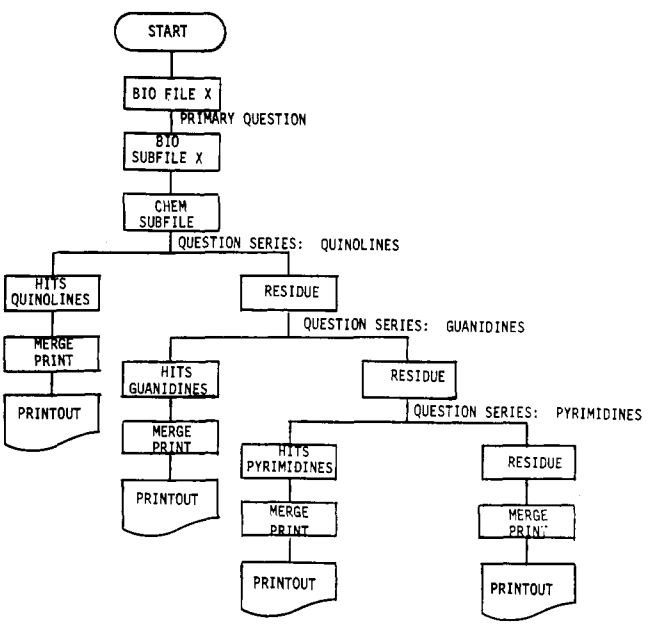


Figure 1

* Contribution No. 985 to the Army Research Program on Malaria. Presented at the 162nd Meeting, ACS, Washington, D. C., September 12-17, 1971.

** Present address: Department of Medical Research & Development, William Beaumont General Hospital, El Paso, Texas 79920.

*** Present address: Office of the Assistant for Veterinary Affairs, Forrestal Bldg., Washington, D. C. 20314. To whom correspondence should be addressed.

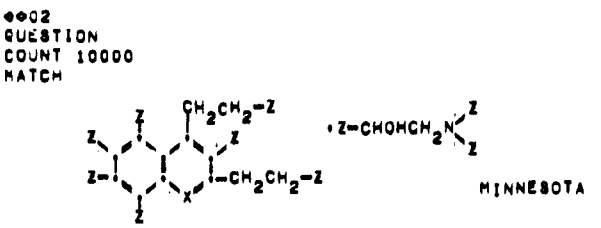
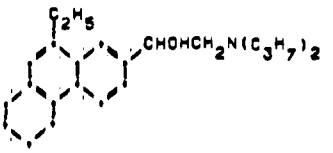


Figure 2. Illustration of output.

The question is a naphthalene system allowing any substitution where there are "Z's" and allowing one ring atom to be any non-H-atom as indicated to the computer by "X". Other groups are demanded using the .Z- to indicate that somewhere else in the molecule these groups are present. The first answer is then printed, followed by biology data. In the second answer, the inclusive parts of the query are indicated by circling the fragments demanded. The only difference between the two answers is the alkylamino side chain. Biological data in this example are from the WWII program.

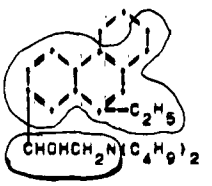
CBC-511271
$C_{24}H_{35}NO$



HEISLOGLE ANTIMALARIAL SCREENING DATA

COMPOUND NO.	TEST	STANDARD DRUG	DRUG FACTOR	PAGE REFERENCE
SN 10218	84	0	000.1500	0301

CBC-511322
$C_{26}H_{39}NO$



logical file X, formed a biological subfile X, which was used to extract a chemical subfile from the large chemical structures file utilizing a number match. This chemical SUBFILE contains the structures of all samples found to be active in test system X.

The chemical substructure search program which we use compares each file structure simultaneously to several substructure questions. Substructures are input via the chemical typewriter. Substructure matching is done first on an at least basis—that is, the file compound must contain at least those screens[1] contained in the query. Once this criteria is met, atom-by-atom comparison is done to determine inclusion. A file structure can, of course, satisfy more than one question and be output more than once. Also, an option in the search program allows the lack of a match or hit to be considered as the last question and all the nonhits to be output at the end of the hits *or* on a separate file.

In this example (Figure 1), three series of questions were asked in separate runs (quinolines, guanidines, and pyrimidines), and the nonhits or RESIDUE were output as a separate file. In each series of hits, a given structure can appear more than once, but a structure can appear in only one series—for example, 3-amino-4-(1-hydroxyethyl) quinoline would appear as both an aminoquinoline and as a quinolinemethanol, but any guanidino pyrimidines would appear as guanidines and not as pyrimidines. The structures are sorted by questions and printed, either alone or after merging with other data (Figure 2). Note, using this method, an investigator must know what he wants or duplicate his output by asking both general and specific questions. Note that any structure, including quinolines, guanidines, and pyrimidines *not* covered in the question series will appear in the residue. This can be avoided when the SUBFILE is large (in our case it contained about 5000 structures) by subdividing the SUBFILE.

The QUINOLINE SUB-SUBFILE in Figure 3 can be created by asking only one general question: Quinolines, instead of a series of more specific quinoline questions. This file can then be queried with a series of questions and residue left on as the last question. This yields a grouping of active quinolines including those types for which the investigator did not specifically ask. These structures can then be printed or merged with biological subfile X (all of file X is not necessary) and the biological and chemical data printed together.

Figure 4 illustrates versatility in the addition of data. Let test system Y be some specific secondary test, either for activity or perhaps for a suspected toxic effect of guanidines. It would be of interest, therefore, to merge this small file with the test system X data and the structures of the active guanidines. Absence of secondary data might indicate that a compound should be sent to test system Y, so inventory data can also be added to the merge. In this way, the investigator knows whether or not the compound has already been shipped and results are pending, or if there is sufficient quantity to send, or if he needs additional sample for test system Y and any other tests he may want done. He would also find out if the active guanidine had been sent to other test systems.

The use of the inventory file on this small file illustrates another point. The inventory file is a large one, and if

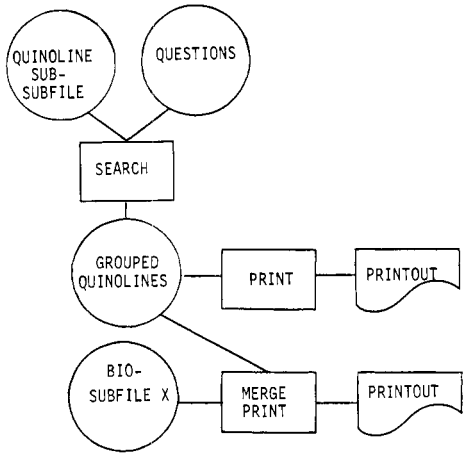


Figure 3

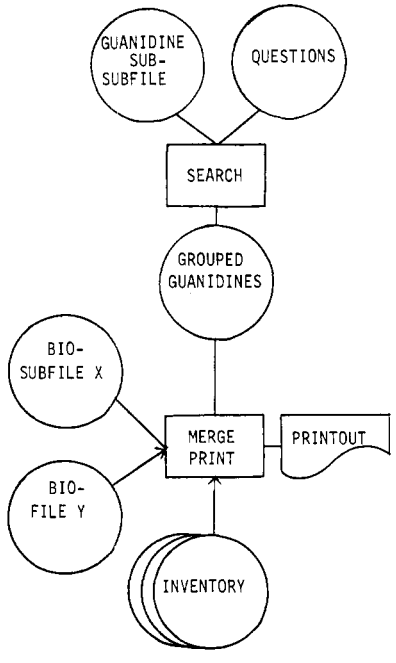


Figure 4

inventory information or information from other large files is needed at more than one point in the subfile manipulation, it would be more economical to create subfiles of these large ones from the answers to the primary question in the same way as the original chemistry subfile (Figure 1) was produced.

The above examples illustrate how a biological parameter was used as the primary question and the resultant subfile was further analyzed using the chemical structure as a parameter. Of course, any parameter on any of the three major files can be used, such as source or quantity from the inventory file, toxicity, dose, or sex of test animal from the biological data file, or chemical structure or substructure from the chemical data file. In this way, we hope to be able to accommodate any investigator involved in the program, no matter what his specialty. Actual uses of the files include: (1) guidance in the chemical synthesis program, both with respect to target compounds and availability of intermediates, (2) patent applications, (3) using the "residue" portion to find new series of compounds, (4) comparison of large structure files, (5) structure activity studies, (6) to aid in publication preparation (7), and to aid in evaluating chemical synthesis proposals.

The information system for the Antimalarial Drug Development Program is only one of many systems which have been developed in government and industry. The review of chemical structure information handling published by the National Academy of Sciences describes many of those in use up to 1968.[3] Of course, several journals describe many more. Thus, the investigator in a new program has a great deal of background work available.

The analysis of data from large files is dependent upon the content and structure of those files. The design of these files should be done considering all of the potentially different types of uses the files will have. It must also be kept in mind that basic changes in a file are not always easy to make without encountering a large conversion process for the older parts of the file. Thus, versatility and user orientation are very important in systems and file design when large files and many different disciplines are involved.

User education or orientation is a must. Even when investigators are working in a program from its beginning and mature with it, they must be periodically educated as to changes in data carried and capabilities added or deleted. New investigators must be educated as to the information and methods available for retrieval and analysis. When this is accomplished, a drug development program will not only be creating huge masses of data, but the investigators within such a program will be able to use it.

LITERATURE CITED

(1) Eckermann, E. H., Waters, J. F., Pick, R. O., Schafer, J. A., "Processing Data from a Large Drug Development Program," *J. Chem. Doc.* **12,** 38-40 (1972).

(2) Jacobus, D. P., Davidson, D. E., Feldman, A. P., and Schafer, J. A., "Experience with the Mechanized Chemical and Biological Information Retrieval System," *J. Chem. Doc.* **10,** 135-40 (1970).

(3) National Academy of Sciences, Publication No. **1733** "Chemical Structure Information Handling: A Review of the Literature: 1962-1968," Washington, D. C. 1968.