Experience with the Mechanized Chemical and Biological Information Retrieval System

D. P. JACOBUS, D. E. DAVIDSON, A. P. FELDMAN, AND J. A. SCHAFER
Division of Medical Chemistry, Walter Reed Army Institute of Research,
Walter Reed Army Medical Center, Washington, D.C. 20012

Received July 2, 1969

New computer methods have been developed in association with the drug development programs of the Walter Reed Army Institute of Research. Experiences with these systems are recounted. Special input devices and computer programming have been developed for the input and retrieval of conventional chemical structural diagrams. The costs, operation, and the advantages of this system are discussed. Associated files of biological properties and inventory control information have been created, which are searchable. The methods used in creating consolidated listings of selected chemical compounds and associated biological data are discussed.

A mechanized chemical and biological information retrieval system has been developed for monitoring drug development programs sponsored by the Walter Reed Army Institute of Research.

The system retrieves chemical structures according to either their chemical or their biological properties. These two requirements are met in separate ways. The criteria for chemical retrieval are structural identity, inclusion of a specified fragment, or the presence of given screens. Specialized programs have been developed for retrieval by these chemical criteria. Searches from the biological point of view are similarly guided by their own criteria, such as the kind of test, the type and degree of activity, the origin of compounds, etc. These biological and administrative criteria are handled by more conventional computer methods, as are the correlation of files by means of accession numbers.

Based on our own early experience, the experience of Bello,[1] and a general impression founded upon the "Survey of the Chemical Notation Systems,"[2] we estimated that the cost of handling chemical information would amount to 5% of the total drug development program. Our later experience has shown that this figure is accurate or reasonable only if the figure is averaged over a four-year span. The programming costs amount to about 10% of the first year's effort for the initial phase of the program, followed by 5% the second year, and then by 4% and 2% for the next two years. If this initial heavy commitment in information retrieval is not made, then test systems tend to become separated from the retrieval programs, and the retrieval programs are essentially a mere recording of summarized information. We have also found that a detailed inventory-control program must be coordinated with the chemical and biological programs to insure proper management of the entire drug development effort.

In a conglomerate system of this nature, the utility of each component system depends on its own performance, as well as on the smoothness of its integration with the other programs. In this context, the five basic problem areas, in our opinion, are the following:

I. Inexpensive input
II. Interpretation of input and file maintenance
III. Synonym files
IV. Questions
V. Output

INEXPENSIVE INPUT

Although chemical input is discussed in greater detail below, the input method for biological and conventional data deserves brief mention. The methodology we use is somewhat different from that reported by Waldo,[3] who keypunches laboratory results directly from data sheets. Our philosophy, however, is the same as his. Our original investigations and observations are recorded in a format suitable for direct input. The objective is to minimize human error, to reduce cost, and to provide the original investigator with summarized data which he can evaluate for accuracy before the information is transferred to the central files. The direct recording of all observations also makes it possible to send the submitter of the chemical a report of all the details of the data, rather than just a summarization.

Inexpensive chemical input has been achieved by the use of the chemical typewriter.[4] This machine, by virtue of recording coordinates, provides discrete locations for the characters typed, and therefore digitizes a chemical structure. Important to its operation is the production of hard copy (sheets or 3- × 5-inch cards) for the typist to inspect and to make corrections on. This hard copy is also used by the person responsible for the input, for checking errors discovered in machine-processing. Such compounds are then retyped. Since some incorrectly drawn structures require the staff to consult with the submitter of the chemical, we cannot imagine a system operating without the production of hard copy at the time of input. The typewriters, manufactured by the Mergenthaler Co., have been relatively satisfactory. Our typists produce approximately 1000 correct structures per 40-hour week, or 50,000 per machine per year per shift, at the cost of approximately 10 cents per entry. Overhead figures

vary according to the scheme of operation and raise the cost to 15–25 cents per structure. Of the total input typed, approximately 15% of the structures are rejected because of typing errors. One and one-half per cent of the total material is rejected by the system because of incorrect chemistry. These rejection rates and the machine downtime are not considered in the over-all input capability of 1000 structures per week.

For a typist to type a structure, the chemical diagram must first be made available. If only names are available, a manual conversion to a chemical diagram is required. Considering overhead rates, the conversion of a name to a chemical diagram has cost between 50 cents and one dollar over a broad range of compounds.

We have developed rules for the typing of chemical structures. While we believe most of these rules represent elementary precepts of chemistry, some of these rules may be peculiar to our system. It is our opinion that differences in the interpretation of conventionally-used chemical diagrams will cause much more serious problems of compatibility between systems than differences due to the nature of the various systems. This is because many conventional diagrams are, in a wider context, ambiguous. We feel, therefore, that these rules represent mostly "good" chemistry, and need to be understood and agreed upon. With experience and advanced programming, the number of rules for proper machine analysis is decreasing, although some rules are maintained to keep the hard copy looking "attractive." Examples of typing rules include the prohibition of typing fractions in the molecular formula or structural formula, the requirement of one bond per substituent to show attachments to an atom, and the typing of hydrogen atoms when necessary to satisfy valency requirements for hetero atoms in a ring system.

INTERPRETATION OF INPUT AND FILE MAINTENANCE

The tape from the chemical typewriter must be interpreted in order to produce a conventional connection-table suitable for machine processing (Figure 1). This interpretation problem exists not only for typewriters, but also for the interpretation of input from optical scanners or from cathode ray consoles. We are not aware of any other production system interpreting chemical diagrams which are written in a conventional manner. We have emphasized conventional diagrams because we feel that the value of a chemist's time is such that we should put a premium upon easily recognized output.

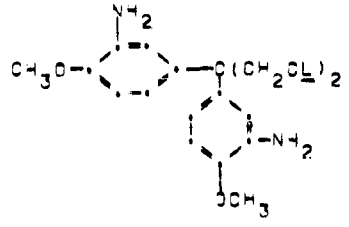
Chemists do not generally write chemical structures in which each atom and each bond is shown explicitly. Instead, there is an abundance of shorthand conventions, such as —COOH, —C_6H_5, —C_6H_{11}, —$C(CH_3)_2(CH_2)_3$, which must be correctly interpreted by the computer. For the last four years we have been involved with two philosophies for interpreting such conventions. The first philosophy, described by Gould, Gasser, and Ryan,[5] interprets the chemical diagrams by looking up each abbreviation in a table in order to find explicit bond-attachments. Programming for these chemical abbreviations (or glyphs) is time-consuming and requires a large computer core, but has the advantage that the interpretation of each glyph can be specified in advance. Since table of look-up may be easily changed and enlarged, this was a natural beginning point for our efforts in interpreting conventional chemical diagrams. Confusion in the interpretation of glyphs may result, however, from their juxtaposition. The program utilizing these glyphs has been working since 1962, and although we are not using it now, the availability of third-generation hardware with large memories permits reconsideration of glyphs as a way of handling chemical diagrams. The second philosophy for handling abbreviations does this by means of rules. Rules have the advantage of high-speed processing and small core requirements; the disadvantage is that they are not easily changed, since a change in one rule may change the interpretation of other chemical structures. These rules, like chemical glyphs, also need to be understood and agreed upon in order to determine if they, in fact, represent "good" chemistry. For example, we allow $CH_3SO_2CH_3$ to be typed on input, but the rules interpret it only as dimethylsulfone. While it is not necessary that chemists learn these rules, it is necessary that they agree on the interpretation of the diagrams. Since chemistry is a diagrammatic language, there are a fair number of special cases usually associated with multivalent atoms, many of which we have taken care of in our existing programs.

The cost of running a compound through this phase of the program is approximately 13 cents if the operation is carried out at commercial daytime rates, or 1.3 cent if the cost is calculated using government equipment. The output of this first phase is a connection table similar to those used as input in other systems. When we used such tables as input, we found them time-consuming to prepare and to check, and consequently expensive. We feel that on grounds of cost alone, we have established the advantage of typewriter input.

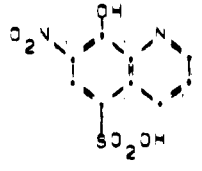
The connection table generated from the structure might be left unchanged, but most other systems process it. This processing, however, produces additional levels of incompatibility between systems. To our knowledge, our method is the only one which does not attempt to produce a condensed connection table; nor do we attempt to generate a unique canonical notation for each structure. We feel that a critical problem in a canonical notation is that, in the absence of a rigorous mathematical proof, the algorithm used may establish false identities, which would be extremely difficult to discover. We felt, therefore, that in renouncing canonical notations, we would keep the system clean. More important, when searching for fragments, canonical notations are useless. For establishing the identity of actual compounds, our program contains all applicable screens and the molecular formula. Only if these match will it trace through the connection table, atom-by-atom, to determine identity.

To reduce the time-consuming atom-by-atom matching, we compute screen bits which are assigned to each molecule. A screen bit is a simple yes-no flag reflecting the presence or absence of a particular chemical fragment—e.g., a carboxyl group. Our screens, however, can be relatively exotic—e.g., NH_2 separated by any four atoms from OH without regard to whether the intervening atoms are aromatic, aliphatic, etc. Screen bits are generated automatically, become a permanent part of the record, and are used as an extension of the molecular formula on file.

```
◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ CND   BNAV77890                    ACCEPTED
◇◇BN-AV-77890
C17H20O2N2O2
```



```
◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇◇ CND   BNAV78299                    ACCEPTED
◇◇BN-AV-78299
C9H6N2O5S
```



```
◇◇WARNING - OXYGEN PRECEDES VARIABLE VALENCE ELEMENT          X=11, Y=03
◇◇WARNING - OXYGEN GROUP PROCESSED                            X=11, Y=03
◇◇WARNING - OXYGEN FOLLOWS VARIABLE VALENCE ELEMENT           X=17, Y=12
◇◇WARNING - OXYGEN GROUP PROCESSED                            X=17, Y=12
◇◇WARNING - PEROXIDE MAY HAVE BEEN CREATED OR LOST            X=17, Y=12
◇◇◇◇◇◇◇◇◇WARNING MESSAGES ONLY - COMPOUND ACCEPTED

STRUC
01    O          D03
02    O          D03
03    N          D01,D02,S04
04    C          S03,D05,S14
05    C          D04,S06,S07
06    O          SH,S05
07    C          S03,S08,D12
08    N          S07,D09
09    C          D09,SH,S10
10    C          S09,SH,D11
11    C          D10,SH,S12
12    C          S11,D07,S13
13    C          S12,D14,S15
14    C          D13,S04,SH
15    S          D15,D17,S13,S18
16    O          D15
17    O          D15
18    O          SH,S15
```

Figure 1. Connection table generated by computer for chemical typewriter input

We have emphasized screens because of our interest in searches for inclusion. We felt that complex inclusion-questions otherwise would require multiple backtracking on large portions of the file. The screens also represent an attempt to reduce the recomputation associated with multiple questions. Of necessity we arrived at our pattern of screen bits in an arbitrary manner; we are now able to evaluate these from the point of view of information theory, since we have now accumulated both a large data base and a large series of questions. At present, our guess is that we are carrying more screen bits (500) than would result if they were selected in a systematic manner. However, since screen production is only a one-time computation, and since screens occupy relatively little space on tape, we have chosen to err on the side of safety. The cost of elaborating the simple as well as the complex screen bits would be 80 cents per compound if we did it on commercial machines during the daytime shift; the cost to us is 8 cents per compound.

Updating and sorting files is part of the interpretation of input. A test for identity determines whether a compound will be added to the master file or be treated as a duplicate. To do this operation efficiently, the chemical file is kept in some kind of chemical order, and indeed, all existing chemical systems maintain their files in some order. Our file keeps the compounds in molecular formula order, followed by their screen-bit configurations. The order in which the biological files usually are maintained

is accession number order, thus necessitating sorting whenever chemical and biological files are correlated. Even in random access systems, corresponding indexes will have to be sorted. With large files, large sorts are a difficult and expensive operation.

SYNONYM FILES

Synonyms—not necessarily names, but synonymous numbers—arise because submitters of chemicals, testing laboratories, etc., insist on retaining their own identification numbers, and also because in our system a permanent accession number cannot be assigned to a compound when it is received, but only after its record has proceeded to a certain stage in our system.

The permanent accession number cannot be assigned to a compound by our system until it is determined that the new entry does not duplicate a previous entry on the file. All processing up to this final step carries the uncertainty that the incoming compound may be a duplicate. The assignment of an accession number means that all the previously available but less favored numbers must be changed, or else that a synonym file must be maintained so that the preferred accession number can be coupled with records using older identification numbers.

The main problem up to this point is keeping the system clean—i.e., eliminating bad input, being certain that compounds which are rejected are properly re-entered, correcting incorrect chemistry, and correcting typing errors which are not chemical errors and which therefore are accepted by the system. The main problem from this point on is the maintenance of the synonym file and the production of output in a format and sequence suitable for integration with the synonym, biology, and inventory files. We have just indicated how late in the system a permanent accession number is assigned. In both our older manually operated systems and in our computer-operated system, we would prefer a single master or registry number, but we have found it impossible to operate without synonyms. There are at least five major sources of synonyms: 1) old test systems have different accession numbers, such as the serial numbers of the World War II malarial program; 2) companies and private organizations sending in compounds use their own accession numbers for inventory control; 3) our biological testers assign their own serial number to find their data. While testers often keep track of our number, their own accession number improves the operation of their system; 4) blind accession numbers are assigned to check on the reproducibility of biological tests; 5) multiple samples of the same chemical from the same source requires lot control, which of necessity means at least a subdivision of the previous accession number. We feel this problem can only get worse, especially for systems dealing with multiple unpublished private files.

QUESTIONS

The format of our chemical questions, and some of the chemical answers, are shown in Figures 2 and 3. The chemical question for identity is asked by typing the actual structure. Similarly, a substructure search is specified by typing the actual fragment. The letter *Z*,

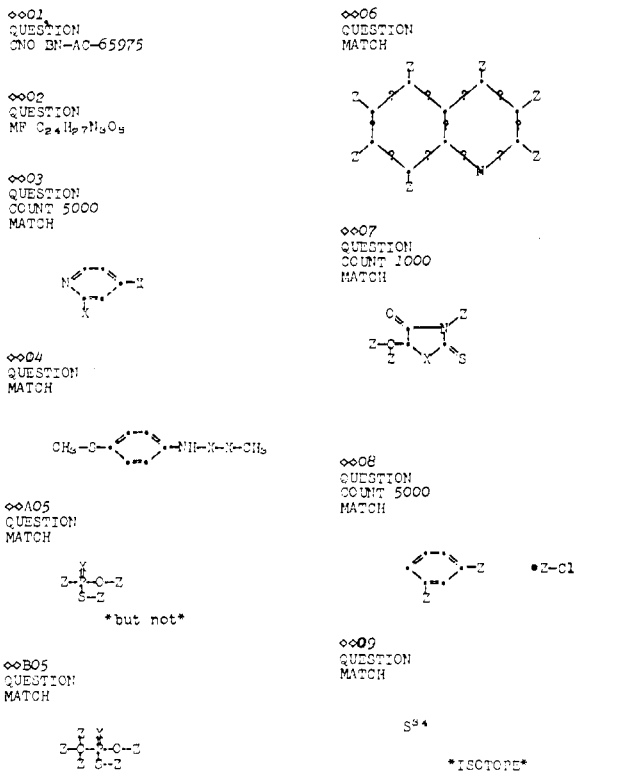


Figure 2. Examples of queries

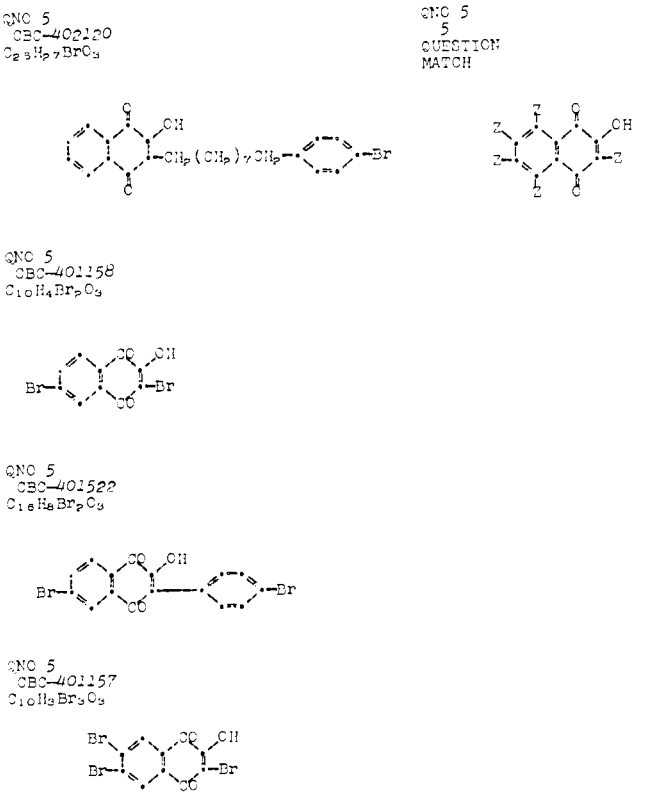


Figure 3. A chemical question and some answers

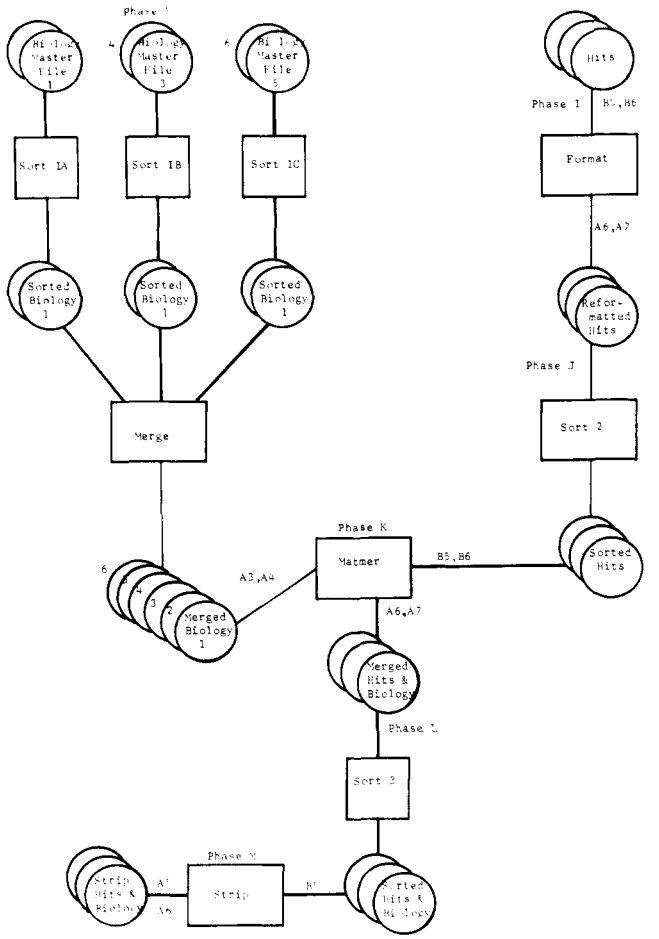


Figure 4. Merged biology and chemstructs text data

inserted at appropriate locations on the fragment, indicates locations where substituents are allowed. Following the Gould convention,[5] the letter *X*, typed similarly, indicates that a nonhydrogen substituent is required at that location. Questions including "but not" conditions can be specified by typing first the fragment desired, then the fragment to be excluded.

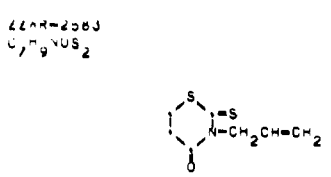
Except as a tour-de-force, we have found that chemical output without associated data is of little interest. To obtain the chemical diagrams independently from one file and the appropriate biological information from another is impractical. With small files, this has been done for many years; with large mechanical systems, the necessary manual handling becomes an impossible task. It is the machine which must present the desired chemical, biological, and associated data integrally combined.

Because of the synonym problem and because the chemical order is different from the accession number order, much sorting and merging is necessary to achieve this combined printout. These data files may always outgrow on-line memory systems. As implied by Figure 4, the conventional data processing problems are large.

INEXPENSIVE OUTPUT

The last critical area is the development of inexpensive output. At present, the cheapest output is produced by the high-speed printers associated with digital computers. We use a Data Products printer off line. We originally chose this machine because its drum was easily engraved with the characters constituting a chemical font. This printer has since proved very reliable in production use.

Our machine prints 600 half lines per minute when printing chemical diagrams or 600 full lines per minute

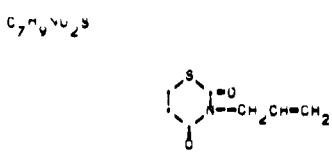


RODENT RADIATION MORTALITY DATA

NR 0002563A BN 9999999 DRUM PRINTER SOURCE 0039 SUBMITTER N=173

3-ALLYL-1,3-THIAZANE-2-THIONE-4-ONE

N	JULIAN DATE	PERCENT SOL.	PH	VEH	CHEM LD50	ROUTE ADM.	DRUG DOSE	RADIATION TYPE	DOSE	TIME	NO. MICE	MORTALITY BY DAYS	PER SURV	PI
	63/175	0.00	0.0	CMCTH		IP	000.00	G	1000	30	10	000000/00010/23121	0	000.00
	63/175	3.00	7.7	CMCTH	0800	IP	300.00	G	1000	30	15	000000/01000/14301/01001/00010/00100	7	002.65
	63/175	3.00	7.7	CMCTH	0800	IP	600.00	G	1000	30	15	041000/00022/01002/00000/00000/10100	7	001.42



RODENT RADIATION MORTALITY DATA

NR 0002564A BN 9999999 DRUM PRINTER SOURCE 0039 SUBMITTER N=226

3-ALLYL-1,3-THIAZANE-2,4-DIONE

N	JULIAN DATE	PERCENT SOL.	PH	VEH	CHEM LD50	ROUTE ADM.	DRUG DOSE	RADIATION TYPE	DOSE	TIME	NO. MICE	MORTALITY BY DAYS	PER SURV	PI
	63/176	0.00	0.0	CMCTH		IP	000.00	G	1000	15	10	000000/00012/01201/3	0	000.00
	63/176	2.00	6.2	CMCTH	0600	IP	400.00	G	1000	15	15	000000/00120/27201	0	000.00
	63/176	2.00	6.2	CMCTH	0600	IP	200.00	G	1000	15	15	000000/00212/21101/10000/00000/00000	27	003.61

Figure 5. Printout of biological data and associated chemical structures

when printing biological data (Figure 5); a newer version is available which operates at 1200 lines per minute.

From the description of the above system, which handles large volumes of data, it is obvious that a number of problems had to be solved, independently of the ones presented by chemistry. Problems were encountered in the areas of systems analysis, programming, electrical and mechanical engineering, topology, information theory, and human-factor analysis. We believe that the growth and effectiveness of our system is due in large part to our attention to these areas and by our use of experts and consultants therein. The growth of the system was stimulated by the need to depend upon it for the control of a large medicinal chemistry program.

LITERATURE CITED

(1) Bello, F., "How to Cope With Information," *Fortune*, pp. 2-10 (Sept. 1960).

(2) "Survey of Chemical Notation Systems," National Academy of Sciences, National Research Council, Publication **1150**, 1964.

(3) Waldo, W. H., "Searching Two-Dimensional Structures By Computer," J. CHEM. DOC. **2**, 1-2 (1962).

(4) Feldman, A., D. B. Holland and D. P. Jacobus, "The Automatic Encoding of Chemical Structures," J. CHEM. DOC. **3**, 187 (1963).

(5) Gould, D., E. B. Gasser and J. F. Ryan, "Chemical Search—An Operating Computer System For Retrieving Chemicals Selected For Equal, Analogous or Related Character," J. CHEM. DOC. **5**, 24 (1965).