Correlative Searching of Bibliographic and Chemical Structure Data Bases of Chemical-Biological Activities*

H. J. HAMILTON and M. K. PARK
Information Science Unit, Computer Center, University of Georgia, Athens, Georgia 30601

Received February 16, 1970

A method of interfacing the text and chemical structure data bases for *Chemical-Biological Activities* was studied. Recall and precision of text searches for chemical structure-biological activity correlations using nomenclature to represent structural concepts were compared with the result of searches made after augmentation of the search request by a substructure search. Statistical tests showed no significant difference in recall or precision between the two methods. However, an interaction between the search method using nomenclature versus Registry Numbers and the quality of nomenclature available to express the structural concept was observed.

This study is the first of a series aimed at evaluating the various ways of interfacing two computer files of information related to *Chemical-Biological Activities* (CBAC). CBAC, produced by CAS, covers the journal literature which reports the effects of organic chemical compounds on biological systems.[1] One of the two files is the CBAC text search file, a computer-readable version of the printed CBAC publication. It contains the bibliographic citation, including the article title and authors' names; an abstract or digest of the article; and Registry Numbers and molecular formulas for specific chemical compounds cited in the digest. The second file of CBAC-related information is a subset of the compound structure records from the CAS Registry System.[2-4] All compounds cited in the first eight volumes of CBAC are represented. The structure file contains an atom-bond computer-generated notation of each structure, recorded in such a manner that compounds can be retrieved on the basis of structural characteristics. The common link between the two files is the Registry Number, a unique identification number for each different compound structure in the Registry System.

There are several ways of using these files, both singly and in conjunction with each other. A text search system is available for retrieving bibliographic citations from the text file by posing questions in terms of words or word phrases, authors' names, journal Codens, Registry Numbers, and molecular formulas.[5] Another search system, called Substructure Search (SSS), has been developed by CAS for retrieval of chemical structure data. Search results for this system are a list of Registry Numbers for compounds satisfying the structural requirements. This study was designed to determine if bibliographic retrieval could be improved by supplementing text search with a substructure search. The relative retrieval capability of nomenclature versus structural notations, in the form of Registry Numbers, for retrieving documents from CBAC was the principal investigation. The data bases used were volumes 1 through 8 of CBAC, a total of 51,453 documents.

Twenty-two (22) questions pertinent to the CBAC subject area were obtained from university research staff. Questions used had to fulfill two basic criteria. First, the question had to contain two concepts, one of which was chemical structure-related and the other biological. A few questions in which the biological concept was all inclusive were also included due to the selection criteria for CBAC, i.e., the effect of organic compounds on biological systems. In other words, the question requested any information reported in CBAC for a given class or classes of compounds. Secondly, no more than two classes of compound structures could be represented in the question. This kept the number of search results to a manageable size for analysis.

The majority of the search profiles, 18 of the 22, were selected from the set of profiles run routinely in a CBAC current awareness search service. The search results of these profiles had been reviewed, and the profiles revised as necessary over an extended period of time. The remaining four profiles were obtained especially for this study during a seminar held on the campus. These four profiles were coded and reviewed by two experienced information specialists, but were not searched on this data base prior to the runs made for this study. Since profile coding for free text retrieval systems such as CBAC is more art than science, the quality of the profiles is inevitably argumentative. The quality of the profiles in terms of where their retrieval capabilities fall on the relevance-recall curve is, of course, of interest in evaluating the performance of the profiles, per se. However, it does not appear to have affected the comparisons of the two search methods in this study.

Two search routes were used. The first, which will be referred to as the Nomenclature method, utilized only the CBAC text files. The questions or profiles, were coded using trivial and/or systematic nomenclature to express the compound class. Text books, standard reference books,

* Presented in part before the Chemical Literature Division, 158th Meeting, ACS, New York, September 9, 1969.

Nomenclature Profile		Registry Number Profile	
blood	*dicumarol*	blood	66762
fibrin	*acadyl*	fibrin	548005
coagul	BHC	*coagul*	1821198
anticlot*	*dicumarin*	anticlot*	2964229
anti-clot*	melitoxin	anti-clot*	4006966
thrombosis		thrombosis	4263438

Figure 1. Representative search profiles

Table I. Profile Characteristics

	Nomenclature	Registry Numbers
Av. terms/compound parameter	10.7	90.6
Av. terms/biological parameter	7.6	7.6
Av. terms/profile	18.3	98.2

CBAC printed issues, and the CDB Desktop Analysis Tools[6] were used as vocabulary sources. The coding conventions were those specified for use with the CAS text search system.[7] In the second method, the Registry Number method, the compound concept in each question was first run as a substructure search. The Registry Numbers obtained as answers replaced the nomenclature used in the text profiles, and the resulting profiles were then searched against the same CBAC text file. The terms of the biological concept were constant and common to the profiles of both methods. Thus, a given set of questions (profiles) was searched twice against the CBAC text file, once using nomenclature to describe the compounds and the other using Registry Numbers obtained via a substructure search of the corresponding structure file.

Representative terms for a typical example are shown in Figure 1. The question called for information on the use of dicumarol and its derivatives in the treatment of post-surgical blood clot formation and coronary thrombosis. There were 10 biological terms in each profile, six of which are shown. Asterisks are used to indicate the truncation form used and are not part of the search term. The nomenclature profile had 11 nomenclature-related terms while the Registry Number parameter contained a total of 23 Registry Numbers obtained by substructure search. Substructure search questions were coded for both fragment and iterative search using the conventions specified by CAS.[8]

Characteristics of the profiles for the two search methods are shown in Table I. Numbers reported are for the 16 questions which were processed by both methods. In the compound parameter, there was an average of 10.7 terms in the Nomenclature search and 90.6 Registry Numbers replacing the text terms in the Registry Number search. The biological parameter is, of course, constant in both search methods at an average of 7.6 terms per parameter. The line shows the average number of terms per profile.

ANALYSIS OF SEARCH RESULTS

Search results for the questions are shown in Table II. In the Nomenclature searches, two of the profiles retrieved no answers from the eight volumes of CBAC; three profiles retrieved only answers which were judged irrelevant to the question. In the Registry Number searches, five of the substructure search questions retrieved no Registry Numbers; consequently, the text searches utilizing Registry Numbers in lieu of nomenclature terms were not run. Two of these five questions corresponded to Nomenclature questions which also retrieved no relevant documents. One Registry Number question retrieved only irrelevant answers. Thus, for both Nomenclature

Table II. Text Search Results

Profile Number	Nomenclature			Registry Numbers		
	No. Citations Retrieved	No. Relevant Citations	Precision (Percent)	No. Citations Retrieved	No. Relevant Citations	Precision (Percent)
45	27	17	63	21	16	76
48	0	0	indeter.	...[a]	...	...
50	66	48	73	36	35	97
51	315	73	23	...[b]	...	...
53	253	147	58	297	170	68
54	48	13	27	32	7	22
56	1	0	0	1	1	100
57	0	0	indeter.	7	0	0
58	3	1	33	...[a]	...	...
60	56	13	23	57	13	23
62	1759	757	43	1631	726	45
63	1	0	0	163	56	34
64	637	420	66	18	12	67
65	123	3	2	...[a]	...	...
66	456	154	34	406	364	92
67	823	355	43	...[c]	...	...
68	35	15	43	24	13	54
69	34	33	97	29	29	100
70	22	0	0	...[a]	...	...
71	75	52	69	18	7	39
73	81	76	94	31	31	100
74	19	13	68	32	26	81

[a] These profiles not searched since no Registry Numbers were obtained from Substructure Search. [b] 2497 Registry Numbers were obtained for this question: text search not run.

Table III. Summary of Search Results

	Nomenclature	Registry Number
No. of Answers Retrieved	3548	2803
No. of Relevant Docs.	1758	1506
No. of Unique Relevant Docs.	718	466

and Registry Number search, there were five questions for which no relevant documents were retrieved. In addition, one Registry Number question was omitted due to the excessive number of answers.

Analysis of the relative retrieval capability was based on the number of relevant documents retrieved by each method. Documents retrieved were tabulated as the total number of documents retrieved, the number of relevant documents obtained by each method, and the number of relevant documents unique to each method of search. The total number of relevant citations present on the data base for a given question (i.e., the basis for recall calculations) was taken as the sum of the relevant documents retrieved by both methods minus the number of relevant documents retrieved by both searches. A summary of the raw search results for the 16 profiles processed by both search methods is shown in Table III. A total of 2224 relevant answers was obtained for the 16 questions searched by both methods.

The relevance of the documents was determined in terms of relevance to the question. The CBAC digests were used to judge the relevance. Documents which showed false coordination of concepts were designated as irrelevant. Also, studies in which the subject compounds were used solely as reference standards or as inducing agents for a biological condition, and were not themselves the subject of the study, were judged irrelevant. For example, an article reporting use of quinazolones to potentiate epinephrine and norepinephrine induced responses was judged an irrelevant answer to a search for articles on the sympathomimetic effects of hydroxylated beta-phenylethylamines. Although it can certainly be argued that articles of this type satisfy the search logic, they do not answer the intention of the question as posed. Precision determinations made under these criteria definitely affect measurements of computer search effectiveness, but they do not affect the comparison in this study since a consistent set of judgment criteria was used.

Early experimental observations indicated that some questions were obviously performing better than others in terms of retrieving relevant documents, apparently as a function of the nomenclature. This was tested by arbitrarily classifying the questions as to whether the nomenclature terminology available for describing the desired structures was "good" or "poor." That is, compound classes which were readily described in terms that were widely used in the literature and which accurately and, to a large degree, unambiguously defined the class of substances were termed "good." Examples of compound classes with "good" nomenclature are prophyrins, lysergic acid amides, and penicillins. Classes which were very difficult to define in terms of nomenclature were classed as "poor." Two specific types of compound descriptions became apparent in this "poor" classification. One was compound classes which were very highly specific substructures for which there was no general name. An example is the compound having a structure of a pyrazine ring ortho fused to a five-membered ring containing a double bond and a keto group. The substructure also has resonance and tautomeric forms. In terms of nomenclature, this was searched using the term "pyraz" with both right and left truncation. The substructure search question, on the other hand, was coded to retrieve only the highly specific compounds desired. The "poor" nomenclature condition also occurred for generic compound classes, such as the substituted aliphatic acids, where it was necessary to list numerous specific examples in order to search by nomenclature. It was virtually impossible to preconceive all possible compounds of interest and to name them so as to match the authors' nomenclature.

The experimental observations for both precision and recall were analyzed by standard statistical methods to determine what differences, if any, existed between the two search methods, the two types of question classification, and the interactions between these two factors. Analysis of variance and tests for statistical significance were made according to the methods presented in Snedecor.[9] All percentages were transformed to angles represented by the arcsin of the square root of the percentage.

Considering all questions, both types of search performed equally well for both precision and recall for questions retrieving some relevant documents. Questions retrieving no answers were not included in the precision and recall analysis as these are indeterminate conditions. In terms of the number of relevant documents retrieved (i.e., the precision of search), Registry Numbers yielded 53.5% relevant documents using the criteria for relevancy discussed earlier. The percentage relevancy for the Nomenclature search was slightly lower, 49.6%; however, tests of variance indicated no statistically significant difference. This is due to the large amount of variance within subclasses. Percentage relevance on individual questions ranged from 0% to 100% for questions which retrieved at least one answer. These values for precision are only slightly higher than the mean value of *c.* 40% reported for CBAC by Kent.[10]

Analysis of recall data for the two methods also showed no significant difference in the two methods. Recall for nomenclature was 72.7% and for Registry Numbers, 71.7%. The basis used for measuring recall in this study was the total number of relevant documents retrieved by both methods. The total number of relevant documents was taken as the number of relevant documents common to the two methods plus the unique relevant documents retrieved by each method. Calculations on this basis do not take into consideration relevant documents missed by both methods, and the resulting recall percentages are undoubtedly biased high. These results are in the range of the 65% and 80% recall values reported by Kent for computer and manual searches, respectively.[10]

Variance analysis on documents recalled also indicated that there was no significant difference in recall of relevant documents, depending on the classification of the question, when all questions were considered. "Good" nomenclature had a mean recall percentage of 82.7 and "poor," 52.6. The analysis did indicate an interaction between the method of search and the classification of nomenclature, however. The percentages are shown in Table IV. Within

Table IV. Percentage of Relevant Documents Retrieved (Recall)

	Method	
	Nomenclature	Registry Numbers
"Good" Nomenclature	82.7	63.3
"Poor" Nomenclature	52.6	88.0

the "good" nomenclature class, the Nomenclature method retrieved a larger percentage of relevant documents than did Registry Numbers, primarily because the nomenclature terms retrieved documents which contained only a generic reference to the class of compounds. No specific chemical compounds, consequently no Registry Numbers, were cited in the digest. On the other hand, Registry Numbers significantly improved recall for questions containing "poor" nomenclature. This is an expected result if a major function of structural notations is in fact true. That is, the notation provides for selection of compound-related data on the basis of chemical structure without regard to nomenclature. This indicates that Registry Numbers do augment the vocabulary, hence the retrievability of CBAC documents, over simply the title and digest, provided that ready access to Registry Numbers is made available as a supplementary tool.

SEARCH TIME FACTORS

The search time required by the two methods was also examined. Times have been computed in terms of averages per profile over the entire eight volumes of CBAC. For the text search using nomenclature terms, it took an average of 8.5 minutes of CPU time per profile for 22 profiles. The text search using Registry Numbers required 11.2 CPU minutes per profile. Substructure search prior to the Registry Number text search averaged 33 seconds or 0.6 minutes per profile. This gives a total of 11.8 minutes per profile for the combined search and 8.5 minutes per profile for the nomenclature search. As shown in Table I, replacement of the nomenclature with Registry Numbers significantly increases the average number of terms per profile. Text search time is a function of the number of search terms and the size of the data base. Since the data base size is constant, the increased time per profile is due to the increased number of terms. All searches were run on an IBM 360/65 operating in MVT environment. Both text and substructure search programs were supplied by CAS.

Even though there are considerably more terms per profile by the Registry Number method, the search time is not a linear function of the number of terms. The average search time per term is relatively high for less than 200 terms per run with the CAS text programs on this size data base.[11] With more than 200 terms, the average time per term begins to level off and becomes relatively constant for larger numbers of terms. Thus, the larger the number of terms per search run the lower the CPU time per term. Consequently, the five-fold increase of terms from nomenclature to Registry Numbers does not result in a five-fold increase in time.

Another factor is the way in which text terms and Registry Numbers are searched. Registry Numbers are recorded in uniquely identifiable fields in the data base and are matched directly, while each text term in the profile must be matched against every character sequence in the text fields (i.e., title and digest). Thus, the number of Registry Numbers can be increased significantly over text terms for the same amount of search time. The average processing time per profile, and thus the cost in this study, was increased by 3.3 minutes per profile when a substructure search was run to obtain Registry Numbers prior to the text search, a factor of 39%.

CONCLUSIONS

The conclusions drawn from this study, realizing that the sample was small, are summarized as follows. When compounds can be described in terms of "good" nomenclature, that is, nomenclature or terms which are widely used in the literature and which adequately describe the class of compounds under consideration, there is no significant difference in the number of relevant documents retrieved (recall) or in the relevancy of the citations between the two types of search. Questions with "good" nomenclature do retrieve some documents not obtained by using Registry Numbers due to generic references to the class of compounds. When nomenclature and text terms are inadequate or incomplete for defining the compound classes, there is a measurable improvement in the recall of documents relevant to the question when a substructure search is used prior to text search. The Registry Numbers obtained via substructure search can be used in lieu of nomenclature in the text search. In determining the cost of the computer search for the two methods, the number of Registry Number terms can be increased substantially over text terms before the increased number of terms becomes a significant factor. The technique used to search Registry Numbers allows for a rather large increase in the number of this type of term before the search timing increases over that required to match a few text terms against every word in the CBAC titles and digests. The best retrieval, however, is a combination of "good" nomenclature terms and Registry Numbers for specific compounds. The computer time for this type of search would approach the sum of the two individual methods, however. The individual user would have to decide whether the improved retrieval justified the additional cost.

ACKNOWLEDGMENT

The authors gratefully acknowledge the consultation of James L. Carmon, Professor of Statistics, University of Georgia, on analyses of the search results. This work was partially supported by the National Science Foundation under Grant GN-851.

LITERATURE CITED

(1) Ish, C. J., and S. W. Terrant, Jr., "Chemical-Biological Activities: A Specialized Information Service in Biochemistry," *Am. Jour. of Pharmaceutical Education* **32**, 201–10 (1968).

(2) Leiter, D. P., Jr., H. L. Morgan, and R. E. Stobaugh, "Installation and Operation of a Registry for Chemical Compounds," J. CHEM. DOC. **5**, 238 (1965).

(3) Morgan, H. L., "The Generation of a Unique Machine Description for Chemical Structures," J. CHEM. DOC. 5, 107 (1965).
(4) Tate, F. A., H. L. Morgan, D. P. Leiter, and R. E. Stobaugh, "A Mechanized Registry of Chemical Compounds," presented at the 1965 Congress of International Federation for Documentation (FID), Washington, D. C., October 10–15, 1965 (unpub).
(5) "Text Searching," Chemical Abstracts Service, Columbus, Ohio, 1968.
(6) "Desktop Analysis Tool for the Common Data Base," Chemical Abstracts Service, June 1968, PB 179 900, Clearinghouse for Federal Scientific and Technical Information.
(7) "Preparation of Search Profiles," Chemical Abstracts Service, Columbus, Ohio, 1967.
(8) "Substructure Search—Background Information and Question Coding Instructions," Chemical Abstracts Service, Columbus, Ohio, 1968.
(9) Snedecor, G. W., "Statistical Methods," Iowa State College Press, Ames, Iowa, 1946.
(10) Kent, A. K., "United Kingdom Experiences in the Operation of a Retrieval and Dissemination Service Based on CAS Search Tapes," presented at the American Chemical Society Meeting, Atlantic City, N. J., September 1968.
(11) Unpublished experimental data, J. L. Carmon and M. K. Park, University of Georgia, Athens, Ga.

Implementation and Evaluation of Two Computerized Information Retrieval Systems at the University of Pittsburgh

NEALE S. GRUNSTRA and K. JEFFREY JOHNSON
Pittsburgh Chemical Information Center, University of Pittsburgh, Pittsburgh, Pa. 15213

Received April 15, 1970

This article describes the Pittsburgh Chemical Information Center data processing group's implementation and evaluation of two information retrieval systems: TEXT-PAC, an information retrieval system developed by the International Business Machines Corporation, and a system developed by the Chemical Abstracts Service (CASCON). Both systems use the Chemical Abstracts Condensates *(CA Condensates)* tape as the input data base.

A group has been organized within the chemistry department of the University of Pittsburgh as an experimental station for the computerized dissemination of information.[1,2] One of the goals of this group, the Pittsburgh Chemical Information Center (PCIC), is to evaluate new chemistry data bases and the programs that search them. Thus, this study began shortly after Chemical Abstracts Service (CAS) began publishing *Condensates* in the fall of 1968.

CA Condensates is a machine-readable magnetic tape service of CAS. One tape is issued by CAS each week corresponding to the hard-copy issues of *Chemical Abstracts* (*CA*). Each record on the tape includes an abstract number, title, authors, bibliographic citation, and keywords which amplify the content of the article. The abstract and molecular formulas included in the hard-copy of *CA* are not included on the tape. Papers from approximately 12,000 journals are included. The odd-numbered issues of *CA Condensates* cover papers in biochemistry and organic chemistry. The even-numbered issues include chemical engineering, applied, physical and analytical chemistry, and macromolecular chemistry.

CA Condensates has three features that make it more appealing for current awareness than *Chemical Titles* (*CT*)[3], another CAS data base. First, since *CT* includes only 650 journals, *CA Condensates* offers significantly greater journal coverage as well as books and patents. Second, the keyword feature is absent in *CT*. And third, the *CA Condensates* user has the option of searching even or odd issues, or both. Thus, if he desires, the user may eliminate wide areas of chemistry, thereby reducing the number of irrelevant hits and search costs.

The Pittsburgh Chemical Information Center is currently providing current-awareness service to approximately 270 users. Chemical information specialists translate user interests into computer readable search strategies. These data are then submitted to a data processing group, which is responsible for both routine production processing as well as development of new information retrieval capabilities. The computer output is returned to the information specialists who maintain statistics on the processing costs and the number of citations per user. In addition, a group of behavioral scientists are in regular contact with Pittsburgh Chemical Information Center users to ascertain in-depth information about the ways in which chemists procure, use, and communicate scientific information. This information is obtained through a variety of sources including structured and unstructured interviews and feedback cards. The latter are used in the compilation of statistics concerning relevancy of the