A Microcomputer-Based System for Chemical Information and Molecular Structure Search†

M. LEONOR CONTRERAS,* MAURICIO DELIZ, ANTONIO GALAZ, ROBERTO ROZAS, and NELSON SEPULVEDA

Department of Chemistry, University of Santiago de Chile, Casilla 5659, Santiago-2, Chile

Received January 16, 1986

ARIUSA, a personal generalized interactive microcomputer system of storage and retrieval of chemical information, is described. This system makes use of the graphic capabilities of microcomputers for working with the drawings of molecular structures. Its performance is described with examples that show the retrieval of references containing identical structures to the query molecule, substructures, superstructures, and similar structures. This system is useful in the design of organic syntheses and in the search of structure–activity studies.

INTRODUCTION

Due to the rapid growth of the scientific literature, research and development work in the area of chemical information has been carried out with extraordinary intensity. A number of general and extensive chemical information systems are now available to facilitate literature searching including CAS ONLINE, DARC, MACCS, DIALOG, and others.[1-6] As the volume of chemical information increases, though, one needs some method of storing this material in a convenient and inexpensive way. This paper presents a personal information retrieval system that, because of its design, is both much more specific and more flexible than other systems that are currently available.[7,8] Two particular attributes that make this system attractive are that the stored information may be done in the user's own language and also only information principally important to the user need be stored.

New useful retrieval algorithms for main frames have recently appeared,[3,9-11] although they have not yet been implemented on microcomputers. It is felt that it would be a useful contribution to have software that allows a microcomputer to handle information employing most of the advantages that a mainframe can offer.

This microcomputer-based personalized information system allows retrieval of information with the flexibility and efficiency similar to those offered by a mainframe. This algorithm uses both alphanumeric and graphic accesses. This system, which is called ARIUSA, is oriented to the research needs of the organic chemist especially interested in organic synthesis and structure–activity relationships. ARIUSA has been developed for an 8-bit microcomputer and allows retrieval of references related to query molecules, according to structures, substructures, superstructures, and similar structures. Input of the structures associated with the reference and their retrieval are accomplished according to the description given below. Examples are provided that illustrate the general retrieval strategies used by the program. In addition, some technical details and the mechanism of correlation of the attributes are specified.

DESCRIPTION OF THE SYSTEM

ARIUSA is a friendly menu-driven generalized system for the storage and retrieval of chemical references. Each reference is described via the following attributes: Title, Authors, Journal, Subject, Physical Location, Keywords, and Molecules (i.e., common names and structures that use a wedge representation for the stereochemistry). Each attribute is related to the others in the way shown in Figure 1.

† Presented in part at the 1st Computational Methods Users National Meeting, Santiago, Chile, July 1984.

Menu 1, the main menu, allows four choices: (1) input step, (2) search step, (3) modification step, and (4) utilities step. Each of these choices transfers the user to one of four submenus to further define the desired operation.

Input Step. This module allows the user to start a new literature file, which is done interactively. In this step, the researcher can provide only some of the attributes of the reference, allowing data entry to be completed later, even by a nonscientist. Graphic input of the targets is easily made from the keyboard as in a previous study.[12] The position and the speed of movement of the cursor are controlled by the use of some predefined keys. Hydrogen atoms are implicit. Provisions are made in the program to interactively correct any error of bond or atom. The molecular structure information is then stored in an encoded form. Four data-archives with relative organization are used to store the information. The records are arranged in fixed-length cells that are numbered from 1 to n beginning with the first record in the file. Records can be placed in or deleted from any cell and can be accessed sequentially or directly. The program keeps track of the record number.

Search Step. A search can be performed with any one of the attributes or a combination of them. Partial attributes can also be used (i.e., the year). The first part of the search is made in a sequential way, and then, the relationship with the other corresponding attributes is made with the help of pointers to the different data-archives, which is explained in the following section. The search algorithms used are a combination of a quick preliminary scan followed by a more accurate atom-by-atom search to select the final targets. To accomplish this, small manually encoded screen sets are generated by the user to define important structural features of a substance. Augmented atoms and linear sequence screens were used.[13] This is transparent to the user, who only has to answer the questions posed by the program. Structures that do not share the screens required by the query are quickly eliminated. Only those structures that satisfy these requirements are used in the atom-by-atom search. If no screens are selected, an atom-by-atom search is performed. Atom-by-atom searching is a relatively slow process compared with the Boolean logic operation of screen searching. To do the atom-by-atom search, the program uses the encoded data for the molecule to generate a connectivity table that covers its full topology. By employment of all of this information, the identity or exact match is accomplished.

On the other hand, similarity, substructure, and superstructure searches use basically the same search algorithms as described above, i.e., screen search followed by an atom-by-atom search. Examples that show the results of these searches are presented in a subsequent section.

When the program has finished the required search, the chemist can review each of the references found and can select

0095-2338/86/1626-0105$01.50/0

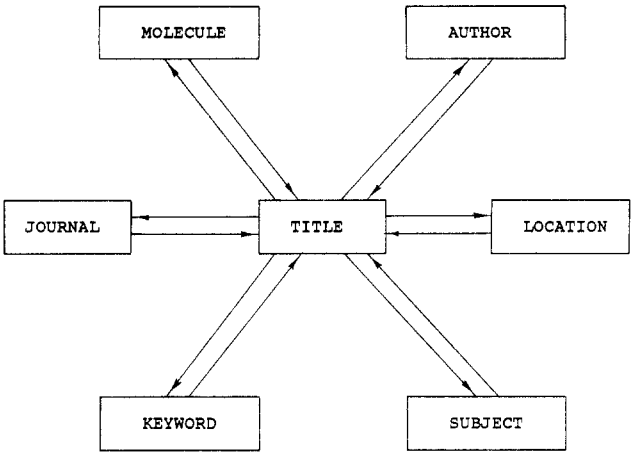


Figure 1. Relation of the attributes associated to a reference.

those that are to be printed. Each printed reference may also include a drawing of the molecule(s) of interest. Each time a new search is made, a subfile containing the first references is formed. This file can be reviewed and modified in a successive manner in order to refine the information.

Modification Step. In this step, it is possible to modify an attribute that is incomplete or erroneous. It is also possible to add or delete any of the principal subjects, keywords, molecular structures, or names.

Utilities Step. In this step the user will be able to perform two tasks: (1) The first is to format a new diskette for use with ARIUSA. The formatting includes giving a label, the date, a password (optional), and the total number of registers that will be occupied for each attribute of the references. (2) The second is to determine the free space and the extent to which each register is occupied by the different attributes of a reference. This allows the user to determine, in an empirical way, the number of real occurrences of each attribute. In this way, each user will know the necessary number of registers needed for each attribute so that the diskette may be formatted with a maximum of storage efficiency.

RELATIONSHIPS BETWEEN THE ATTRIBUTES

The structure of the archives used by ARIUSA is described in this section. A specific example is then provided that elucidates the relationships between the attributes. As previously mentioned, there are four data-archives: REFERENCE, KEYWORD, SUBJECT, and MOLECULE.

REFERENCE. Each register of the REFERENCE archive has the following fields: title, author, journal (here books, theses, patents, and other are also considered), and location (see Figure 2). In addition, each register includes three pointers, namely, the keyword entry (PUBKEY), the subject entry (PUBSUB), and the molecule entry (PUBMOL).

KEYWORD. Each register of this archive contains a keyword and a pointer to the KEYPUB archive.

SUBJECT. Each register of the SUBJECT archive contains the name of a general subject and a pointer to the SUBPUB archive.

MOLECULE. Here, each register contains the name of one molecule, its encoded structure, and a pointer to the MOLPUB archive.

So, there are six pointer-type archives: PUBKEY, PUBSUB, PUBMOL, KEYPUB, MOLPUB, and SUBPUB. Each register of the pointer-type archive has two fields. The first field has the register number of the attribute archive, and the second field has the register number of the pointer archive where the next attribute can be found.

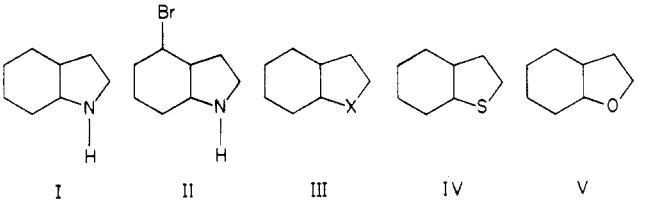
Figure 3 depicts the way in which implementation of the relationships between the data-archives is accomplished. For this purpose, a reference from the literature has been used.[14] The data-archives (REFERENCE, KEYWORD, SUBJECT, and MOLECULE) and three of the pointer-archives (PUBKEY, PUBSUB, and PUBMOL) are shown. In this figure, reference number 3 is related with the keyword entry number 7. Register number 7 of the pointer-archive PUBKEY has a 3 in the first field, which corresponds to register number 3 in the KEYWORD archive, in this case, "electroreduction". The number 1 in the second field of register number 7 in PUBKEY is a pointer to register number 1 of the same PUBKEY archive where more information about keywords related to reference 3 can be found. Register number 1 contains "21 and 5". The number 21 specifies the keyword "amides" in register number 21 of the KEYWORD archive and the "5" the next step of the chain. The first field of register number 5 in the PUBKEY archive gives the keyword "sulfur electrode". This keyword is the last one as is indicated by the second field, which is zero. This indicates that the chain of information about keywords terminates here. The same kind of chain is used for the subjects related to reference number 3. The number 2 in the subject entry specifies the subject "synthesis" and then the subject "electrochem". Finally, the relation between the reference number 3 and the "mol.1" molecule is easily seen through the pointer-archive PUBMOL.

In a similar way, to implement a correlation between KEYWORD and REFERENCE, the second field of the registers of the KEYWORD archive is used as the input for the pointer-archive KEYPUB. This archive contains the register number of the REFERENCE archive with the associated bibliographic references. The chain of information continues as previously explained. To implement the relationships between SUBJECT-REFERENCE and MOLECULE-REFERENCE, the pointer archives SUBPUB and MOLPUB respectively are used.

MOLECULAR RETRIEVAL EXAMPLES

A search can be initiated by starting with any of the attributes. The graphic capabilities of ARIUSA allow for the retrieval of references based on the molecular structures drawn on the screen. After the molecule is drawn, there are four general types of searching procedures that can be performed, which are described below.

(i) The system can retrieve all of the references that contain a molecule identical with that which was drawn on the screen. This is called an *identical molecule* search. For example, if molecule I is the query molecule and the archives contain only



I II III IV V

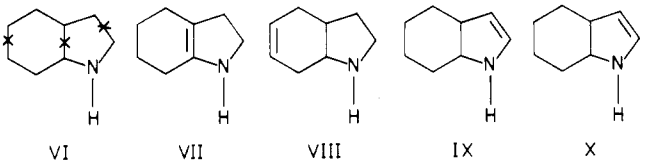
some derivatives like II, then a response of "no references found" will be displayed.

(ii) The system allows for the selection of some variable atoms in order to expand the searching to find structures similar to that which was drawn on the screen. This is called a *similar structure* search. To accomplish this, the bond order must be kept fixed. For instance, if the query molecule is structure III, where X is any bivalent atom, the system will give all of the references with similar structures like IV and V. In the same way, the system allows for the selection of bonds that may have variable bond orders. For instance, if molecule VI is the query and the bonds marked × are selected to be variable, then ARIUSA will retrieve all of the references

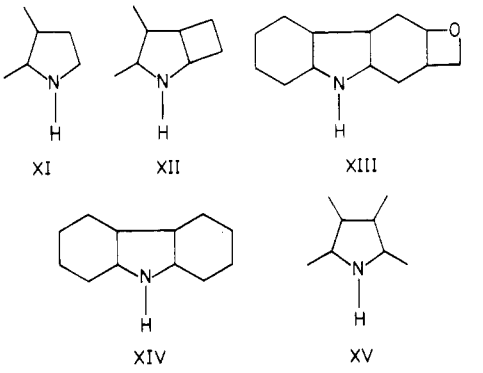
TITLE	AUTHOR	
JOURNAL	LOCATION	
KEYWORD ENTRY	SUBJECT ENTRY	MOLECULE ENTRY

Figure 2. Register fields of the REFERENCE archive.

that contain not only VI but also VII, VIII, and IX or any *similar compound* with different bonds on the × positions like molecule X:



(iii) ARIUSA can also retrieve the references of all of the structures that contain the query molecule as a fragment within them. In this case, the query is a *substructure*. For instance, if molecule XI is drawn and the system contains XI, I, XII, and XIII, then the references of all of them will be displayed. In this case, the size of the related molecules can be fixed by the chemist, who determines the number of extra atoms to be considered. The search can be extended even further by using variable atoms and bonds as described before in (ii).



(iv) ARIUSA can also retrieve references with structures less complex than the query molecule. The query is now a *superstructure*. For instance, if molecule XIV is drawn and the system contains I–XIII and XV, then all of the references related to I, VI, XI, and XV will be displayed. Additional sophistication can certainly be obtained by using variable atoms and bonds as previously described.

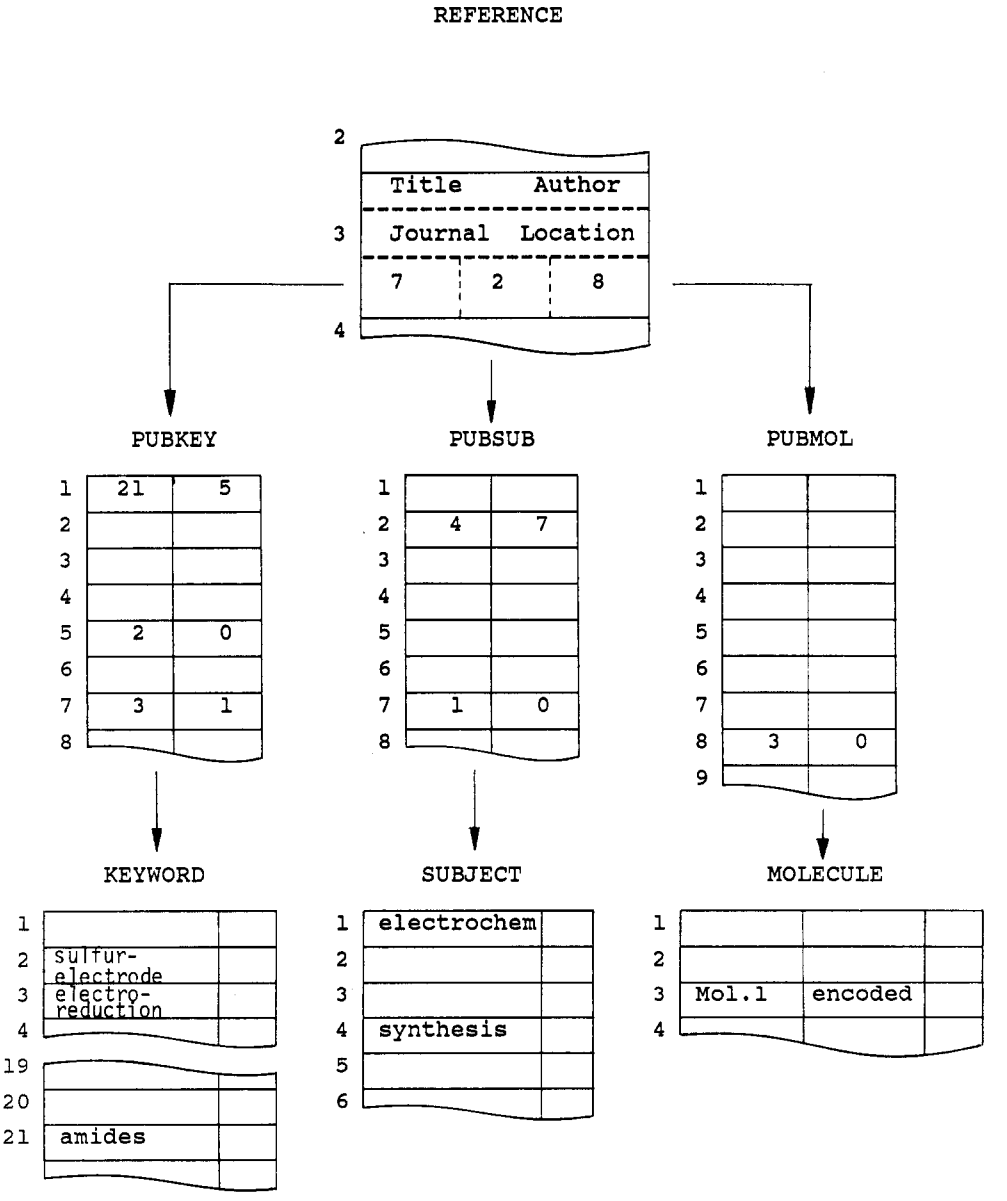


Figure 3. Implementation of the relationships between the attributes of a reference.[14]

TECHNICAL DETAILS

The software is written in PASCAL-80 (Microsoft Release 7.00, Dec 1981) and runs on a North Star Model Advantage microcomputer with 64K RAM. The system works with two floppy disk drives running under the CP/M 2.2 ver. 1.0 operating system and has a graphic resolution of 640 × 240 pixels. The microcomputer has an EPSON MX-100 printer with graphic capabilities. Two different formats are available for printing: a narrow (about 6 in. wide) and a wider format (about 10 in. wide). Each diskette can store about 500 references. The characteristics of the attributes of each reference as used here are specified as follows: title, 160 characters; authors, 80 characters; journal, 80 characters; location, 10 characters; keywords (0–10), 30 characters; subject (0–5), 30 characters; molecules (name, 0–5), 30 characters. The system can draw molecules with a maximum of 28 atoms and 28 bonds. At the moment, the program is being implemented on an IBM-PC.

DISCUSSION AND CONCLUSIONS

The system described here has been designed to help in the handling of the information needed for doing experimental and theoretical research.[15] This system is able to treat references catalogued by title, authors, subjects, keywords, year of publication, physical location of the reference, and molecular structures and names. This makes it a useful tool for the researcher, especially when dealing with interdisciplinary information.

The input and the output of ARIUSA are quite flexible. The alphanumeric or graphic input can be modified or completed at any time. The output can be printed even with molecular structures, with selected sizes and according to any of the attributes, partial attributes, or a combination of up to five of them.

This personalized system runs on a microcomputer but has the retrieval capabilities of a mainframe.[2,9,10] In addition to the normal alphanumeric searching, this system can also perform a search initiated by graphic input. This type of search enables the location of (a) references containing structures identical with the query molecule, (b) references containing substructures of the query molecule, (c) references containing superstructures of the query molecule, and (d) references containing structures similar to the query molecule, where some of the atoms or bonds (type and number) are selected according to the chemist's need.

Handling of chemical information has been accomplished by both inverted and serial file organization.[1,2] Inverted files provide a rapid response but need large amounts of disk space. Serial files provide easy updating but slow response time. ARIUSA uses a relative file structure, which allows a rapid response and easy updating. The storage space used by relative files is greater than that used by serial files but much less than that used by inverted files. In the present case, in addition to the four data-archives, six pointer-archives are used. In this data structure, the information is linked with pointers to create a chain of information. Resequencing of the information chain can be achieved by adjusting the linking pointers when new data are added and old data are deleted from the files. A normal molecular search through 500 stored molecules takes about 2–5 min according to the number of atoms of the query molecule and to the general search strategy being used. Similarity searching is the slowest of the techniques used here.

This interactive system can easily be modified according to the chemist needs. For instance, in the retrieval of a superstructure query, a fragment search can be done according to some chosen characteristics like the number and the type of the participating atoms (variable screen selection). Another adaptation that can be made is to limit the number of the character assignments for each attribute in order to get a more efficient storage. ARIUSA can also be easily modified to work on larger computers.

The capacity (about 500 references per diskette) and efficiency (requires 15–20 s for making a typical keyword or author search through a completed diskette) of the system described here are comparable with one nongraphic personal literature retrieval microcomputer system.[7] However, ARIUSA, has a greater flexibility and selectivity, due to its file design, data processing, data retrieval, and graphic capabilities. Compared with a mainframe system, though, it has the inherent limitations of speed and storage capacity. ARIUSA is much more specific because it contains only the references that are principally important to the user. For example, in order to save time in searching, each diskette could be reserved for a particular type of information.

Due to the relatively low cost of microcomputers and the associated hardware, this system can be made available to most research groups. Its graphic capabilities render it a valuable tool for the organization and retrieval of literature information in the search for chemical transformations, in the development of computer-aided synthesis design studies,[10,12] and in structure–activity studies, and it is also applicable to many other areas of chemistry.

Copies of the program and detailed operating instructions are available on request to M.L.C.

ACKNOWLEDGMENT

Financial support from DICYT of the University of Santiago de Chile is appreciated. Help with the manuscript by J. Briggs is also appreciated.

REFERENCES AND NOTES

(1) Wipke, W. T.; Heller, S. R.; Feldmann, R. J.; Hyde, E. *Computer Representation and Manipulation of Chemical Information*; Wiley-Interscience: New York, 1974.
(2) Ash, J. E.; Chubb, P. A.; Ward, S. E.; Welford, S. M.; Willet, P. *Communication, Storage and Retrieval of Chemical Information*; Ellis Horwood: Chichester, England, 1985.
(3) Rusch, P. F. "Chemical Information from DIALOG Information Services". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 192–197.
(4) Stobaugh, R. E. "Chemical Substructure Searching". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 271–275.
(5) Kao, J.; Day, V.; Watt, L. "Experience in Developing an In-House Molecular Information and Modeling System". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 129–135.
(6) Barcza, S.; Kelly, L. A.; Wahrman, S. S.; Kirschenbaum, R. E. "Structured Biological Data in the Molecular Access System". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 55–59.
(7) Ensley, H. E. "Personal Literature Retrieval System". *J. Chem. Educ.* **1983**, *7*, 571.
(8) Smith, S. F.; Jorgensen, W. L.; Fuchs, P. L. "PULSAR: A Personalized Microcomputer-Based System for Keyword Search and Retrieval of Literature Information". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 209–213.
(9) Wipke, W. T.; Rogers, D. "Artificial Intelligence in Organic Synthesis. SST: Starting Material Selection Strategies. An Application of Superstructure Search". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 71–81.
(10) Willett, P. "An Algorithm for Chemical Superstructure Searching". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 114–116.
(11) Willet, P.; Winterman, V. "Implementation of Nearest-Neighbor Searching in an Online Chemical Structure Search System". *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36–41.
(12) Barone, R.; Chanon, M.; Contreras, M. L. "Microcomputer and Organic Synthesis. Application of an Interactive Program to the Photochemical Synthesis of Pheromones". *Nouv. J. Chim.* **1984**, *8*, 311–316.
(13) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. "The CAS ONLINE Search System. 1. General System Design and Selection, Generation, and Use of Search Screens". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.
(14) Contreras, M. L.; Rivas, S.; Rozas, R. "Electrosynthesis of Methyldithiooxalamides with a Sulfur Reactive Electrode". *J. Electroanal. Chem.* **1984**, *177*, 299–302.
(15) Fugmann, R. "Peculiarities of Chemical Information from a Theoretical Viewpoint". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 174–180.