NAPRALERT: Computer Handling of Natural Product Research Data

W. D. LOUB,* N. R. FARNSWORTH, D. D. SOEJARTO,[†] and M. L. QUINN

Program for Collaborative Research in the Pharmaceutical Sciences, College of Pharmacy, University of Illinois at Chicago, Chicago, Illinois 60612

Received March 1982

Research in natural products, not unlike many other fields of investigation, requires access to large amounts of prior experimental data relevant to specific projects. The most efficient method of identifying and analyzing these data currently employs computer handling of the information. The NAPRALERT database has been designed to meet this need relevant to the development of natural products. It has also been designed, through analysis of existing literature and its contained data, to provide a means to predictively identify taxonomic sources most promising from specific biological activities.

INTRODUCTION

Databases generally fall into two major categories by the type and form of information they contain. The most frequently encountered is the bibliographic resource, which for the most part lists only citation information. The second major category is the so-called "source"-type file, which provides, in addition to bibliographic information, numerical and textual data contained in the communications it records. Some bibliographic databases include as part of their citation information an abstract derived from the article, which confers limited "source" properties to these files.

NAPRALERT is a database in the source category. It is a textual-numeric collection of records regarding the chemistry and pharmacology of natural products and their appropriate taxonomic data. It is presently accessible only off-line through the Program for Collaborative Research in the Pharmaceutical Sciences (PCRPS), University of Illinois at Chicago, but communications with various database vendors are under way to determine on-line capabilities.

The NAPRALERT file has been especially designed to be of value in drug development and contains information relevant to the research efforts of natural product chemists, biochemists, pharmacognosists, agricultural chemists, and others. It covers the chemistry and biological activities of extracts and/or secondary constituents isolated from or identified in plants, marine organisms, microbes, and, to some extent, animal data as well. Chemical information regarding vertebrate animals, including enzymes, proteins, amino acids, simple sugars, nucleoproteins, and lipids, with the exception of certain reptilian toxins, is not included in the database. One important design feature of this database is the recording of information capable of predicting or rank ordering organisms as to their probability of having specific biological properties if properly investigated. This feature and other possible uses that can be made of NAPRALERT's wide variety of information on natural products are discussed in the following paragraphs.

THE NAPRALERT PROGRAM

Development of the NAPRALERT database required considerable effort prior to computerization of appropriate data. A simplified flow diagram of these efforts is presented in Figure 1.

The first step, literature collection, has been carried out by using both systematic and nonsystematic sources. For example, more than 150 journals, who entries are primarily devoted to natural product research, are systematically reviewed by Ph.D.-level scientists for articles pertinent to the NAPRALERT file. In addition, various comprehensive abstracting sources are scanned, page by page, for relevant articles. From time to time, special subjects identified under projects being carried out for a user of the database or through in-house research interests have been given a systematic retrospect search with a variety of secondary sources. These sources allow the search process to be carried backward even into the 14th century. Abstracting services used are listed in Table I. Other data vital to the concept of NAPRALERT have been encountered more or less nonsystematically from books, reviews, and personal communications. These data, collected randomly, are prepared for entry as well.

Table I. Secondary Literature Indexes Used in the NAPRALERT Program

title	period covered
Index Catalog of the Surgeon General	1880–1961
Index Medicus	1897–1927; 1960 to present
Chemical Abstracts	1907 to present
Quarterly Cumulative Index Medicus	1916–1956
Biological Abstracts	1926 to present
Current List of Medical Literature	1941–1959
United States Armed Forces Medical Journal	1950–1960
National Library of Medicine Catalog	1956–1965
National Library of Medicine Current Catalog, Cumulative Listing	1966 to present
Current Contents, Life Sciences	1967 to present

After selection of appropriate articles and abstracts, each is scanned for data regarding the organism studied, chemical compounds isolated or identified therein, and pharmacological effects recorded for compounds and/or extracts of the organism. This information, along with the demographic data, e.g., author, title, journal, etc., is recorded on special data-entry forms utilized by terminal operators for computerization.

FILE DESIGN AND DATA PROCESSING

Articles dealing with studies on natural products contain four major types of data. These include the usual citation (demographic) information as well as taxonomic, chemical, and/or pharmacologic data. The NAPRALERT file has been constructed to contain these four record types in a vertical hierarchy. A horizontal hierarchy also exists in order to enter one to many records at each level as shown in Figure 2. Records are associated with one another through a common citation number assigned to the demographic data, through one of four record type (RTYPE) assignments, i.e., D, O, P, or C representing *D*emographic, *O*rganism, *P*harmacologic, or *C*ompound data, respectively, and an "occurrence" number for each record created under a single parent record type. All of these data are maintained with the Indexed Sequential Access Method (ISAM) file management for rapid retrieval. Currently, more than 43 000 citations have been entered. Associated with these demographic records are more than

[†] Research Associate (Hon.), Field Museum of Natural History (Botany), Chicago, IL 60605.

0095-2338/85/1625-0099$01.50/0

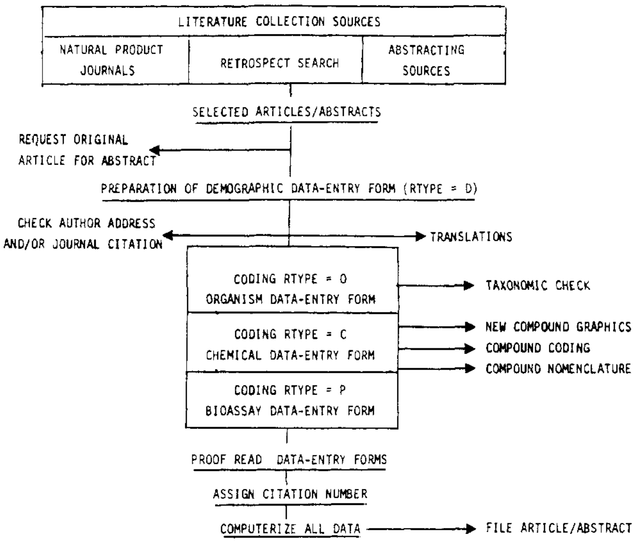
LITERATURE COLLECTION SOURCES
NATURAL PRODUCT JOURNALS
RETROSPECT SEARCH
ABSTRACTING SOURCES
SELECTED ARTICLES/ABSTRACTS
REQUEST ORIGINAL ARTICLE FOR ABSTRACT
PREPARATION OF DEMOGRAPHIC DATA-ENTRY FORM (RTYPE = D)
CHECK AUTHOR ADDRESS AND/OR JOURNAL CITATION
TRANSLATIONS
CODING RTYPE = O ORGANISM DATA-ENTRY FORM
TAXONOMIC CHECK
CODING RTYPE = C CHEMICAL DATA-ENTRY FORM
NEW COMPOUND GRAPHICS
COMPOUND CODING
COMPOUND NOMENCLATURE
CODING RTYPE = P BIOASSAY DATA-ENTRY FORM
PROOF READ DATA-ENTRY FORMS
ASSIGN CITATION NUMBER
COMPUTERIZE ALL DATA
FILE ARTICLE/ABSTRACT

Figure 1. Diagram of NAPRALERT organization.

105 000 organism names, 195 000 pharmacological results, and 190 000 compounds identified. Those numbers given above for organisms and chemical names above do not represent unique names since a given organism may have been investigated many times and the same chemical compound may exist in many sources.

Information contained in the "Demographic" record is quite similar to that found in most bibliographic files but, in addition, contains data unique to the NAPRALERT file. For example, a complete address for the senior author or the publishing organization is only infrequently encountered in other databases but is recorded in the NAPRALERT file. This greatly facilitates correspondence with the author for additional information and/or reference materials when dealing with current literature citations. Since updating this information would not be practical, it is only of value in current awareness reports. Other fields maintained on this record as well as those on other record types (O, P, C) are presented in Table II along with a brief description of the type of data recorded.

The second record type used by the NAPRALERT system, the "Organism" record, contains a full taxonomic description of the organism studied. These data include the organism class, i.e., monocot, dicot, mollusca, protozoa, etc., the family, genus, species, species author citation, subspecies and its authority, strain, cultivar, common names, geographic origin, and synonomy. Data are also recorded to indicate the organism part studied, i.e., leaf, root, culture filtrate, etc., the condition of these materials, i.e., dried, fresh, lyophilized, etc., and the amount studied. These latter data can be valuable to the researcher for several reasons. A major consideration is that the organism part and its condition can have a profound effects on constituents to be identified and/or the biological activity of an organism's extract. In addition, logistical problems are greatly simplified when, for example, the collector is faced with a large tree but knows that only the leaves or root bark are necessary for study.

Where chemical constituents have been isolated or evaluated in a given organism, these data are recorded in the "Compound" record type. Vernacular or trivial names are used for chemical constituents where possible due to their familiarity to the researcher. Many natural products are complex molecules and require highly involved systematic nomenclature. This can be readily demonstrated with a relatively simple molecule among natural products, 6-(D-5-amino-5-carboxyvaleramide)-3,3-dimethyl-7-oxo-4-thia-1-azabicyclo(3.2.0)-heptane-2-carboxylic acid, commonly known as penicillin N. Where neither name has been provided by the author, a compound is then named as a derivative of a common trivial base, for example, quercetin 3,3′-dimethyl ether, or given an IUPAC-derived name as a last resort.

A code is then assigned to the compound in a manner unique to NAPRALERT, which consists of the following. First, a binary numeric code defines the major chemical class of natural product. By virtue of this code, the NAPRALERT program can retrieve and/or display entire classes of natural products within a single sort. Approximately 80 major types of natural products are identified in this manner, i.e., flavonoes, indole alkaloids, sesquiterpenes, quinoxaline-type antibiotics, etc. Following the chemical class code is a three-digit code, which defines the carbon skeleton or substructure. This portion of the code can be used to sort compounds by unique carbon bases, but their graphic presentation requires the use of a noncomputerized base code dictionary, which associates the numerical value with a graphic representation of the molecule's carbon base. A final step in coding a compound is the use of up to 12 different binary alphanumeric designations for functional groups present. More than 200 can be identified

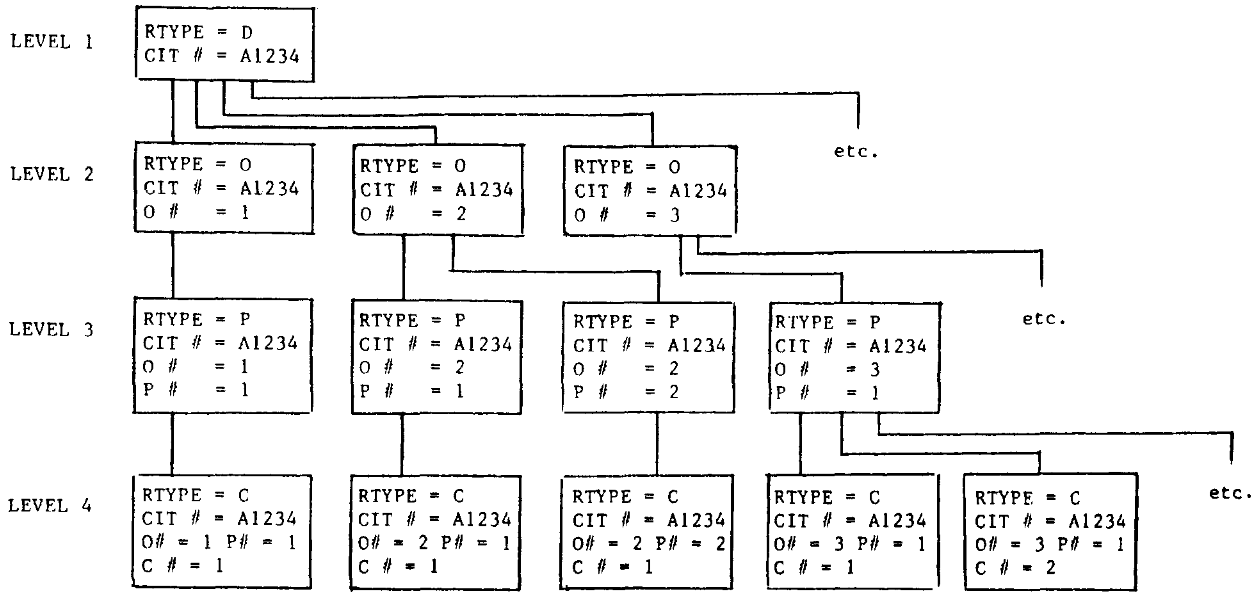

Figure 2. NAPRALERT file hierarchy. RTYPE (record type) = D (demographic), O (organism), P (pharmacology), and C (compound). CIT# = citation number assigned to each article; O#, P#, and C# represent record sequence numbers used to tie each level.

Table II. Field Names and Types of Information Computerized by NAPRALERT

field name	description
	Demographic Record Type
citation number	file number assigned sequentially at time of data entry
citation title	accommodates first 765 characters of the title
author	lists all author names associated with the article
journal	stores journal name or appropriate alphanumeric code
volume number	volume designation of the above journal entry
issue number	accommodates issue number when needed
page number	number of first page on the article
last page	number of the last page of the article
year	year of publication
language	language used in the article, i.e., German, English, etc.
article type	designates type of article, i.e., research, review, etc.
abstract	lists title of secondary reference source, i.e., CA
reference volume number	secondary source volume number
abstract number	secondary reference abstract number
paragraph number	secondary reference paragraph number
address code	alphanumeric address code if one has been assigned
department code	departmental address for senior author, if any
college address	school, college, or institute name in senior author address
university address	university institute or company name in address
city	city for senior author's address
state	state, district, or province for address
zip code	zip code if given
country	country of residence for senior author
grant agency	agency supporting reported research
grant number	granting agency identifying number
	Organism Record Type
organism	code identifying organism class, i.e., angiosperm, gymnosperm, dicot, monocot, etc.
family	botanical family name of the organism studied
genus	botanical genus name of the organism studied
species	botanical species name
species citation	species authority citation
subspecies	subspecies name, if any
subspecies citation	subspecies authority citation, if any
common name	acommodates common names of the organism
taxon synonym	synonym for the taxon and its authority
organism part	lists organism part studied
	Compound Record Type
compound isolated	amount of compound isolated, if any
organism country	geographic source of organism studied
compound name	compound name, vernacular name where possible
compound code	a binary numeric code for the chemical class, i.e., indole alkaloid, flavone, steroid, proteid, etc.
substructure code	stores a three-digit numeric code for the carbon skeleton
functional group code	stores two character codes for functional groups present
	Pharmacology Record Type
worktype	alphanumeric sorting code to designate type of work performed, i.e., in vitro, in vivo, in situ, and/or in humans
major pharmacologic activity	binary code for 16 different pharmacological classes of study, i.e., CNS, chemotherapeutic antifertility, etc.
specific pharmacologic activity	three-digit code for specific pharmacological activity studied; 1000 codes available to date
director codes	enters "PDC" code
weighting codes	enters weighting point designator (see Table II)
alert codes	enters "Alert Data" codes for sorting efficiency
experimental modifications	stores miscellaneous statements describing a special disease condition or test parameter
extract	binary code used to identify type of extract studied
mode of administration	binary code used to identify mode of administration
test species	binary code describing type of animal used, if any
sex	sex of above animal if appropriate
dose expression	identifies type and numeric amount of dose, i.e., LD_{50} 1.0, MLD 2.5, or concentration, i.e., MIC 25.0
dose unit	dose unit of above dose or concentration, i.e., mg, mcg, etc.
per unit weight	per unit weight of above dose or concentration, i.e., kg, per plate, g, person, etc.
qualitative result	qualitative expression of result, i.e., active, inactive, equivocal
quantitative result	numerical expression of result data
expression	type of quantitative result, i.e., increased life span (ILS)
pathological system	alphanumeric code for disease test organism substrate or tissue used

and were chosen for their value in either structure–elucidation or structure–activity relationships (SAR) of interest in drug development. A mechanism has also been incorporated to reveal negative data, i.e., those situations where a compound has been evaluated but was found *not* to be present. For example, a plant could be studied by gas chromatography and a number of compounds detected. The author may have been looking for a specific potentially toxic substance, e.g., safrole, but had no evidence of its presence. This then becomes an important piece of practical information relevant to the use of this plant in herbal teas. An example of a natural product and its NAPRALERT chemical code is presented in Figure 3. A final but important data element collected in this record is statement of the yield for any isolate where the information has been given by the author or can be calculated from experimental facts presented.

When biological activity, either a positive or negative effect, has been reported for a constituent or for an extract of an

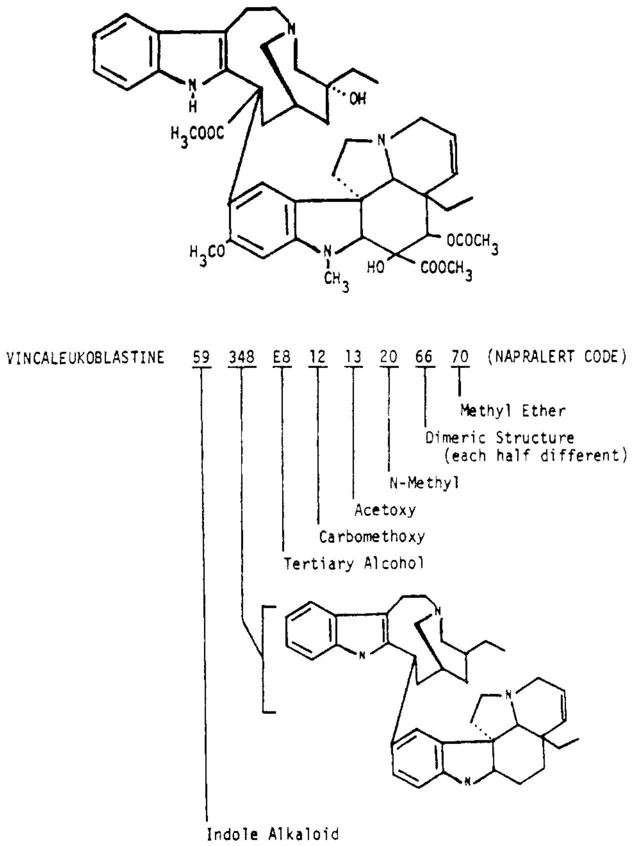

Figure 3. NAPRALERT chemical coding.

Table III. General Pharmacological Class Codes

code[a]	pharmacological category
11	central nervous system (CNS) effects
12	autonomic effects
13	hematologic effects
14	chemotherapeutic effects
15	toxic effects
16	effects on enzyme systems
17	hormonal effects
18	cardiovascular effects
19	muscle effects
20	pheromone effects
21	immunologic effects
22	effects on plants
23	miscellaneous metabolic effects
24	effects on organisms other than mammals
31	miscellaneous pharmacologic effects
66	activities effecting fertility

[a] An additional three-digit code follows any of the above general pharmacologic codes to describe a specific bioassay or biological activity.

organism, including purported folkloric or ethnomedical uses, a "Pharmacology" record type is prepared for data entry. This document uses a variety of alphanumeric work-type codes for rapid sorting purposes, differentiates the biological studies into one of 16 major pharmacologic categories (see Table III), and states the specific biological effects studied, within each major category. Currently, more than 1000 specific biological activities are identified as individual key-field values. Additional data contained in this record are defined in Table II and include type of extract used (where appropriate), mode of administration, test animal, sex, dose expression, qualitative and/or quantitative results, and pathological system or substrate evaluated. A textual area on the data-entry forms permits an unlimited number of modifying remarks to be entered that define special bioassay parameters, such as dosing schedules, and/or pertinent observations of the researcher. Other fields have also been added to this record, which subjectively evaluate conditions of the study and assign point values used in predictive programs to rank-order organisms in a variety of ways.

Much of the scientific information being computerized, e.g., organism class, family names, organism part and condition, names of biological activities, etc., is repetitive and has necessitated the implementation of alphanumeric codes to represent it. This procedure has not only dramatically reduced the number of typographical errors during data entry but has also made storage of this rapidly growing data set, which now exceeds 130 million characters, more cost efficient. Textual retrieval of these codes for report generation are afforded by appropriate driver records also stored in the NAPRALERT file.

After computerization of the information contained on the various data-entry forms, a permanent file of all computerized articles and abstracts is maintained. Thus, if a user cannot obtain copies of more pertinent articles from local library collecitons, viz., in a developing country, they can be made available through the NAPRALERT program.

DATA RETRIEVAL

The most frequently requested information from the NAPRALERT file is concerned with the purported ethnomedical (folklore) information and biological activities for extracts in vitro, in vivo, or in human studies, as well as listings of chemical constituents reported in the literature to have been isolated or identified in a given genus or species of plant. Other commonly desired data include known biological activities for natural products and identification of taxonomic sources for specific natural products, as well as their yields and geographic distribution.

As stated earlier, an important design feature of the NAPRALERT database would be its ability to rank-order organisms as to their probability of having specific useful biological activities. An example of utilizing NAPRALERT in this way has been the generation of a predictive analysis program for the Task Force on Indigenous Plants for Fertility Regulation under the W.H.O. Specieal Programme of Research Development and Research Training in Human Reproduction. In this instance, a search of existing data in the NAPRALERT file, representing literature citations from 1975 to 1978, provided the names of more than 1300 plants identified with potential fertility-regulating effects. A series of retrospect literature searches produced the names of an additional 3200 plant species with similar fertility-regulating data, either as a folkloric notation or through recorded bioassay data. Random selection of plants for laboratory investigation from a list of this size could not be considered a feasible approach, and therefore, it became purdent to utilize the NAPRALERT file construction and appropriately developed software to predictively select the most promising plants to study.

In this particular project, it was realized that a large number of pharmacological activities existed that are pertinent to fertility regulation in the male or the female. It then became important to resolve what types of fertility regulation, i.e., preimplantation vs. postimplantation effects, etc., were of interest to the W.H.O program. With this information it was possible to identify only those plants having appropriate folklore data and its repetition among noncommunicating cultures and bioassay data within specific pharmacological activities, which could further be evaluated on the basis of animal type used and dosage schedules effected, as well as other details from recorded research data. By initially assigning numerical values to details of interest and negative values to undesirable effects noted for these plants, the computer performed an unbiased summation of these values for each plant based on computerized research and ethnomedical information.

The end result was the identification of approximately 300 plants from a list of more 4500 that appeared to be the most promising for initial study. Fifty of these, chosen for their availability in the locality of a particular research center, have undergone preliminary investigation. Eight have provided confirmed desirable activity within two different laboratory animal assays. Although the true value of the NAPRALERT approach must await the successful development of clinical drugs from these plants, it would appear the utilization of such a computer-generated analysis can be an indispensible adjunct to natural products research. Certain aspects of this predictive program have been published elsewhere.[1,2]

Information contained in the NAPRALERT file has also been used by the National Cancer Institute, as well as by the herbal, pharmaceutical, and cosmetic industries in the development of new products. Future considerations for the use of the NAPRALERT-type database include the direct preparation of handbooks for the natural products researcher and the preparation of phylogenetic "density maps" for use in chemotaxonomic and biotaxonomic decisions, as well as the possible prediction of an impending endangered species through periodic examination of newly constructed maps.

These are but a few of the possibilities to which this comprehensive file of scientific data on natural products can be used. Individual needs and advanced computer technology will dictate future resource applications.

ACKNOWLEDGMENT

This work has been supported in part by the National Cancer Institute, Contract NCI-CM-97259, and the World Health Organization, Project 78135.

REFERENCES AND NOTES

(1) Soejarto, D. D.; Bingel, A. S.; Slaytor, M.; Farnsworth, N. R. "Fertility-Regulating Agents from Plants". *Bull. W. H. O.* **1978**, *56*, 343-352.

(2) Farnsworth, N. R.; Loub, W. D.; Soejarto, D. D.; Cordell, G. A.; Quinn, M. L.; Mulholland, K. "Computer Services for Research on Plants for Fertility Regulation". *Korean J. Pharmacogn.* **1981**, *12*, 98-110.

CSEARCH: A Computer Program for Identification of Organic Compounds and Fully Automated Assignment of Carbon-13 Nuclear Magnetic Resonance Spectra[†]

HERMANN KALCHHAUSER and WOLFGANG ROBIEN*

Institut für Organische Chemie der Universität Wien, A-1090 Vienna, Austria

Received July 13, 1984

A computer program for the analysis of ^{13}C NMR spectra by various search strategies, including different methods for line search, molecular formula search, and structure-oriented search, is presented. The key algorithm of the program performs a fully automated assignment of ^{13}C NMR resonances to the respective carbons of a known structure. A database of 8000 ^{13}C NMR spectra taken from the literature and from our own measurements was created, containing carbon-centered substructural environments and their corresponding chemical shifts. The assignment algorithm is based on the prediction of chemical shift ranges from these data and permits a stepwise solution of the assignment problem with chemical shift arguments up to a five-bond radius.

INTRODUCTION

During the last decade NMR instrumentation became more sophisticated, and carbon-13 NMR data are now routinely reported in many papers dealing with natural product chemistry. The interpretation of ^{13}C NMR measurements is based on multiplicities, either from SFORD or *J*-modulated spectra, and on chemical shifts, which are mainly used to determine the number of sp^2 carbons and some functionalities with narrow shift ranges like methoxy groups. The chemical shift value of a certain carbon resonance depends strongly on the environment of the corresponding carbon. This sensitive probe cannot be fully utilized by manual interpretation of carbon-13 resonance data. The number of published reference data exceeds many thousand spectra per year; therefore, computerized databases have been built up.[1-16] In this paper we describe our program package, which includes spectrum estimation, many different file search strategies, and the complete automated assignment algorithm.

DATA STORAGE AND OVERALL DESIGN

Each reference data set contains the information given in Table I. From these data, several subfiles containing special information can be created by the computer allowing efficient execution of the different search strategies. The search methods available in the program are given in Table II. Each search function can be called by a three-letter abbreviation. A second program, named C13ADD, performs all other tasks concerning addition, modification, and control of existing records and also includes routines for generation of sorted lists by name, bibliography, or molecular formula. The CSEARCH program is designed for interactive use, but every search can be performed as batch job without user interaction.

Table I. Data Stored in Each Record

(1) entry number
(2) compound name, as given in the literature, up to 160 bytes
(3) comment and experimental conditions (temperature, reference, ...)
(4) structure: atom type, connectivity matrix, and bond type, up to $C_{40}H_{99}O_{63}$ and all other elements up to 15
(5) solvent
(6) molecular formula
(7) literature
(8) chemical shifts and multiplicities
(9) assignment of the resonance lines

SEARCH STRATEGIES

ISO. This option allows the user to find all compounds in the database having a specific molecular formula. During one

[†] Dedicated to Prof. Dr. K. Schlögl on the occasion of his 60th birthday.

0095-2338/85/1625-0103$01.50/0