(40) "MARK IV File Management System. Reference Manual," 2nd ed., Change 4, Informatics Inc., Canoga Park, Calif., 1973.
(41) Converting all data to upper- and lower-case would provide better legibility. A program was written for such a conversion, but time constraints prevented the final test run on live data to verify its reliability.
(42) For example, if the 236 entries which are to be printed in four columns on one page of the personal name index are designated 1 to 236 in alphabetical order, then the sort sequence for printing is 1, 60, 119, 178, 2, 61, 120, 179, 3, . . . , 235, 59, 118, 177, 236.
(43) Tauber, S. J., and Elias, A. W., "Directory of Cancer Research and Control Projects and Organizations," National Cancer Institute, Bethesda, Md., 1974.
(44) *Mail* addresses were used in order to be able to use the data in mailing lists for the type of follow-up discussed under "Further Work".
(45) Tauber, S. J., and Elias, A. W., "Listing of Cancer Research and Control Organizations," National Cancer Institute, Bethesda, Md., 1974.

A Rapid Generalized Minicomputer Text Search System Incorporating Algebraic Entry of Boolean Strategies

T. L. ISENHOUR,* W. S. WOODWARD, and S. R. LOWRY

Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27514

Received November 11, 1974

This paper presents a rapid and efficient generalized minicomputer text searching system. The system has been applied to *Chemical Condensates* and enjoys search speeds comparable to services operating on large computer systems. Complete Boolean algebraic search strategy expressions may be used as direct entries, and all forms of truncation are automatically processed. Benchmark search speeds and results are presented for realistic profiles serving varied research groups in a major university chemistry department.

INTRODUCTION

The chemical literature has grown to the extent that only very narrow fields can be exhaustively surveyed by classical means with a reasonable expenditure of the investigator's time. Chemical Abstracts Service (CAS) presently adds about 400,000 new citations per year to the literature base which it began in 1907.

Starting in June 1968, CAS has recorded *Chemical Condensates,* a citation collection including titles, references, and keywords, but not actual abstracts, on computer readable magnetic tape. Several major commercial efforts have been made to provide current awareness search services utilizing the condensates files on large computer systems.[1-4] While these approaches have achieved a certain degree of success, they have the typical disadvantages of large, centralized systems: specifically, fairly high costs for other than very routine services; locations (and attitudes) often remote from those of the users; and increasing inflexibility as the size of the routine operation grows.

Wilde and Starke have reported on a literature search system oriented toward smaller machines.[5] While their system has been in operation for several years, search times are inconveniently slow, and complex profiles require a great deal of operator effort to translate the search logic to the format used.

This paper presents a rapid and efficient generalized minicomputer search system. The system has been applied to *Chemical Condensates* and enjoys search speeds comparable to those operating on machines costing one to two orders of magnitude more. Furthermore, complete Boolean algebra search strategy expressions may be used as direct entries, and all forms of truncation are automatically processed. Benchmark search speeds and results are presented for realistic profiles serving varied research groups in a major university chemistry department.

THE SYSTEM

The computer system involved is a 64K-byte, 1.0-μsec cycle time, Raytheon 704, equipped with two Peripheral Equipment Corp. 800 bpi, 25 ips, IBM compatible magnetic tape drives, a 500-cpm card reader, and a 300-lpm line printer. Total equipment investment is about $33,000.

BACKGROUND

The result of the development and testing of this system is a simple proof that multiple-profile searches can be done rapidly and economically on a small computer system. The system developed runs directly from standard issue Chemical Abstracts tapes, handles a number of profiles simultaneously (maximum 224 per run), handles elaborate profiles, and allows all possible Boolean logic and left and right truncation of search-text fragments. Specific results of test runs are given later.

It is perhaps necessary to dispel some of the popular misconceptions about minicomputers. First, the comparison of minicomputers to major computer installations is not analogous to that of research equipment comparisons such as low-resolution and high-resolution mass spectrometers. The numbers produced by minicomputer calculations are not of lesser quality than those generated by larger installations. Usually accuracy to any degree desired can be accomplished by a trade off in time of calculation. The principal difference between large, batch-oriented computer systems and smaller machines is more, and often faster, hardware; not necessarily more accurate hardware. Also, large systems tend to have more and diversified input/output devices as well as larger and more sophisticated operating systems. However, it should be realized that the currently popular minicomputers with approximately 1-μsec cycle times, memories from 8K to 100K bytes, and standard pe-

Table I

Program		Run time[a] (min:sec for 30 simultaneous profiles)	Estd av run time for 100 profiles (min:sec)
I. CARUNCH	Compress CA tape	12:05, 17:30[b]	15:00
II. WARPSET	Profile and search structure set up (with complete checking)	0:20, 0:20[c]	1:40
III. WARP-8	Multiprofile search	2:18, 3:28[d]	9:40
IV. SORPRINT	Citation output	6:30, 8:10[d]	23:20
			Total 49:40 or 30 sec per profile

[a] 30 profiles simultaneously. Values are given for two consecutive issues. [b] Time independent of number of profiles—must be done once each week. [c] Time linear with number of profiles—must only be done whenever profiles are added or changed. [d] At worst, linear with number of profiles—must be done once each week.

```
/  H. H. DEARMAN
/  RESEARCH PROFILE
KT,A1=NUCLEIC ACID*
KT,A2=NUCLEOTIDE*
KT,A3=NUCLEOSIDE*
KT,A4=PURINE*
KT,A5=PYRIMIDINE*
KT,A6=DNA
KT,A7=RNA
KT,B1=NMR
KT,B2=SPECT*
KT,B3=PHOSPHORESCENCE
KT,B4=FLUORESCENCE
KT,B5=CIRCULAR DICHROISM
KT,B6=CD
KT,B7=ORD
P1=(A1:A2:A3:A4:A5:A6:A7)
P2=(B1:B2:B3:B4:B5:B6:B7)
S1=P1&P2
KT,C1=COMPLEX*
KT,C2=DYE*
KT,C3=HYDROCARBON*
KT,C4=DRUG*
KT,C5=ANTIBIOTIC*
KT,C6=MUTAGEN*
KT,C7=CARCINOGEN*
P3=C1&(C2:C3:C4:C5:C6:C7)
S2=P1&P3
**=S1:S2
```

Figure 1.

ripheral equipment are very competitive to IBM 7090 and Control Data Corp. 1604 installations which were the major equipment at most computer centers only 10 years ago. A single laboratory installation costing on the order of $30,000 now surpasses all of the computer equipment available to the Manhattan project and the complete computer repertoire of Oak Ridge National Laboratory in 1960.

Another difference between minicomputers and large-scale batch operating systems is the ability of the large systems to simultaneously process several jobs. This does not mean, however, that any individual job is better executed. It means simply that total throughput can be much greater. However, the large center's administration and bureaucracy, because of its charge to serve the average user on a routine basis, can often be more conservative and less capable of implementing innovative ideas.

CURRENT AWARENESS CHEMICAL ABSTRACTS SEARCH SERVICE

The system developed consists of four programs written and implemented on the Raytheon 704. The purposes of the programs are as follows:

1. CARUNCH—compresses the *Chemical Condensates* tapes for canonical search purposes and to remove extraneous information. the compression accomplished allows about 12 weeks of CAS tapes to be written on one 2400-ft reel. (Note that this is the only step dependent upon the nature of the CAS data base. A corresponding front end program could be written for any other literature-access data base and used on the same system.)

2. WARPSET—compiles the search profile definitions into a set designed for sequential and canonical searching of the data base. (This program need be run only when profiles are added or changed.)

3. WARP-8—is a multiprofile search which will handle up to 224 profiles simultaneously with a tradeoff in the number of possible search key strings. Optimum operation probably lies somewhere between 50 and 150 profiles per search pass. However, there is no limit to the number of passes that can be made. This program makes all matches, solves Boolean logic expressions defining search logic, and outputs citation hits with labels according to the profiles scoring the hits.

4. SORPRINT—sorts the output tape in the order of profile numbers and prints the individual outputs.

The test data shown in Table I were based on 30 profiles prepared on Chemistry Department members by one of the authors. The profiles, with the exception of one, were very simple and were developed without consultation with the individual. They are not expected to be sophisticated profiles but were only for the purpose of testing the search system. Since the profiles averaged about to 20 hits each, they were considered adequate for test purposes. (Figure 1 is an example of a sophisticated profile currently in use. The citations selected by profile from *Chemical Condensates*, Vol. 77, issues 1 and 2, 1972, are listed as Figure 2.)

Note that certain programs need only be run at certain times. For example, CARUNCH must be run only once a week. WARPSET need be run only whenever profiles are changed. WARP-8 and SORPRINT must be run each week and run times are, at worst, proportional to the number of profiles run. The benchmark tests allow us to estimate run times for 100 profiles at under one minute per profile including program 2 which need not necessarily be run every week. As seen in Figure 1, the actual setup includes three types of cards. The first are cards starting with "/" which allow comments to whatever extent the user wishes. The second are sets of cards that define search fragments. These start off with mnemonics defining fragment type such as "KT" (standing for keyword title) or "AU" (standing for author) and assign user-generated reference labels. In all, nine mnemonics are recognized. Asterisks indicate left and right truncation where desired. Following these definition cards are expression cards which may be combined in any Boolean manner using the ampersand (&) to represent "AND", the colon (:) to represent "OR", and the minus (−) to represent "NOT". Parentheses are inserted as necessary to specify sequence of operations. The default order of operations is left to right. The final expression is defined with two asterisks.

Operation of the generalized text-search system is orga-

```
40
/ H. H. DEARMAN
/ RESEARCH PROFILE

CA07701001905A***J***ZHIZHINA, G. P.***OLEINIK, E. F.***INFRARED SPECTROSCOPY OF NUCLEIC ACIDS***USP. KHIM.***000072***41***3***474-511***REVIEW IR SPECTROSCOPY NUCLEIC ACID

CA07701001997G***J***ALLEN, F. S.***GRAY, DONALD M.***ROBERTS, GARY P.***TINOCO, IGNACIO, JR.***ULTRAVIOLET CIRCULAR DICHROISM OF SOME NATURAL DNAS AND AN ANALYSIS OF THE SPECTRA FOR SEQUENCE INFORMATION***BIOPOLYMERS***000072***11***4***853-79***UV CD DNA***POLYNUCLEOTIDE UV CD

CA07701002005U***J***ERFURTH, STEPHEN C.***KISER, ERNEST J.***PETICOLAS, WARNER L.***DETERMINATION OF THE BACKBONE STRUCTURE OF NUCLEIC ACIDS AND NUCLEIC ACID OLIGOMERS BY LASER RAMAN SCATTERING***PROC. NAT. ACAD. SCI., U. S. A.***000072***69***4***938-41***NUCLEIC ACID RAMAN SPECTROMETRY***DNA RAMAN SPECTROMETRY

CA07701002006V***J***CITTANOVA, N.***PETRISSANT, G.***PURIFICATION OF MAMMALIAN TRYPTOPHAN TRNA. FLUORESCENCE PROPERTIES OF THE FREE AND THE ENZYMICALLY ACYLATED TRYPTOPHAN TRNA***BIOCHIM. BIOPHYS. ACTA***000072***262***3***308-13***TRYPTOPHAN TRANSFER RNA PURIFN

CA07701002010S***J***JORDAN, C. F.***LERMAN, L. S.***VENABLE, J. H., JR.***STRUCTURE AND CIRCULAR DICHROISM OF DNA IN CONCENTRATED POLYMER SOLUTIONS***NATURE (LONDON), NEW BIOL.***000072***236***64***67-70***DNA CONFORMATION***CD DNA

CA07701002013V***J***STUDDERT, DAVID S.***PATRONI, MARIA***DAVIS, ROBERT C.***CIRCULAR DICHROISM OF DNA. TEMPERATURE AND SALT DEPENDENCE***BIOPOLYMERS***000072***11***4***761-79***CD DNA THYMUS

CA07701002027C***J***PHILLIPS, D. J.***BOBST, A. M.***CIRCULAR DICHROISM MELTING STUDIES ON R17 PHAGE RNA***BIOCHEM. BIOPHYS. RES. COMMUN.***000072***47***1***150-6***CD RNA BACTERIOPHAGE***PHAGE RNA CD

CA07701002322V***J***EPINATJEFF, CHRISTA***PONGS, OLAF***RIBONUCLEASE T1. SPECTROPHOTOMETRIC STUDIES OF THE INTERACTION OF THE ENZYME WITH SUBSTRATE ANALOGS***EUR. J. BIOCHEM.***000072***26***3***434-41***RNASE INTERACTION NUCLEOTIDE***GUANOSINE PHOSPHATE BINDING RNASE***DIFFERENCE SPECTRA RNASE COMPLEX***UV RNASE COMPLEX NUCLEOTIDE

CA07702011077W***J***VANDENHEUVEL, W. J. A.***SMITH, J. L.***HAUG, PAT***BECK, J. L.***MASS SPECTROMETRY OF TRIMETHYLSILYL DERIVATIVES OF DISUBSTITUTED PYRIDINES, QUINOLINES, PYRIMIDINES, AND PTERIDINES***J. HETEROCYCL. CHEM.***000072***9***2***451-5***MASS SPECTRA TRIMETHYLSILYL HETEROCYCLIC***PYRIDINE TRIMETHYLSILYL HETEROCYCLIC***QUINOLINE TRIMETHYLSILYL HETEROCYCLIC***PYRIMIDINE TRIMETHYLSILYL HETEROCYCLIC***PTERIDINE TRIMETHYLSILYL HETEROCYCLIC
```

Figure 2.

nized around so-called SEARCH STRUCTURES which comprise the internal representation of search profiles during search execution. Generation of search structures is performed by the profile compiler program WARPSET through the translation of user-generated search profiles. Actual comparison of these data structures to chemical condensates text files is the function of the search program WARP-8.

The content of the search structure includes text fragments or KEYSTRINGS provided by the users for selection of citations, and compressed versions of the user-supplied Boolean expressions or PREDICATES specifying those combinations of presences (or absences) of keystrings which characterize desired citations. Keystrings are stored as elements in threaded lists which are in turn nodes in a tree structure designed to permit rapid comparison of text elements to stored keystrings during the WARP-8 search. Predicates are entered in the structure as elements in one threaded list common to all profiles. The order of appearance of predicates in this list reflects their order of original definition so that, as predicate evaluation proceeds, the results of subpredicates will be available for the evaluation of expressions which reference them.

Compilation of user search profiles begins with the presentation of profile-defining card decks to the WARPSET program. Each profile deck is initiated by an operator-assigned job number and an arbitrary number of user-supplied "comment" cards. These cards, in combination, serve to identify the profile and associated search result output. Following the job number and comment cards are profile elements which create profile search logic, i.e., keystrings and predicate definition cards.

The format of keystring definition cards consists of three components. The keystring category mnemonic is a two-character identifier signifying the type of text data to be searched for the associated keystring. Examples of category mnemonics and the corresponding data type are "KT" (keyword-title), and "AU" (author). In all, nine keystring mnemonics are recognized. Each keystring definition also includes a one- or two-character user-assigned PRESENCE VARIABLE which allows reference to the definition in Boolean search logic expressions. The third element in the keystring definition format is, of course, the keystring itself. A keystring represents a text fragment to be detected in searched text and may consist of two or more (up to 795) characters. Any alphanumeric character, including blank, is permitted in keystring definitions.

The asterisk has special significance when encountered within keystring definitions. An asterisk appearing at the beginning of a keystring signifies so-called "left-hand truncation." An occurrence of a left-hand truncated keystring will be detected in a matching segment of search text even if it is preceeded in the text by an arbitrary sequence of characters. A non-left-hand-truncated keystring (i.e., no asterisk at the beginning) must be preceded by a blank when encountered in searched text in order to be recognized. The insertion of an asterisk at the end of a keystring indicates "right-hand truncation". As left-hand truncation allows the recognition of a keystring in the presence of an arbitrary *prefix*, *right-hand* truncation permits any *suffix* to be present without affecting detection. A typical keystring definition statement is shown below just as it might occur in a user search profile.

$$\text{KT, A} = \text{*POLYMER*}$$

The keystring category mnemonic "KT" indicates that only keyword title type text fragments are to be searched for this string. The user-assigned presence variable "A" will allow reference to the keystring in subsequent logic expressions. The actual string "POLYMER" is both right- and left-truncated and, hence, will be detected in a searched text fragment regardless of its environment.

Predicate definition cards contain two components. They begin with a user-assigned, one- or two-character, logic variable to which is assigned the result of predicate evaluation. Because a profile may incorporate more than one predicate, the logic variable permits the results of one predicate evaluation to be referenced in the evaluation of subsequent predicates. The use of the symbol "**" as the logic variable of a predicate identifies it as the TERMINAL PREDICATE for the profile. If and only if the terminal predicate of a profile evaluates as "true", then the citation under consideration is listed as a "hit" for the profile and will be included in the search output for that profile.

The second element of the predicate is a Boolean algebra expression which is evaluated in the course of processing every citation searched. The Boolean expression consists of keystring presence variables and logic variables of previously entered predicates in combination with the usual

Boolean logic connectives: "and", ampersand ("&"); "or", colon (":"); and "not", hyphen ("-"). In addition, parentheses may be freely used to denote the desired sequence of expression evaluation. The parsing of Boolean expressions by the WARPSET compiler is performed by an uncomplicated operator precedence technique with parser output consisting of "reverse-polish" code. In the present system, this code represents final parser output and is interpreted in that form by the search program.

Overall, compilation of profiles proceeds as an I/O-bound process with profile input via a 500-card-per-minute reader. Output of profile comments and the finished search structure is routed to one of the two nine-track tapes available to the system. During compilation, certain syntactic checks are performed upon processed profiles to detect such errors as unrecognizable keystring category mnemonics, references to undefined variables, unbalanced parentheses, etc. The encounter of a syntactic error during the examination of a profile results in the output of a printed message indicating the job number of the offending deck and the error type. That entire profile is then automatically deleted from the input stream and compilation resumed.

The compilation process continues until either all profiles have been processed or until the search structure being generated has grown to the limit allowed by the available memory. In the present system, 32,768 bytes are allocated to search structure storage during execution of the WARP-8 program. Because of the internal format used for data storage, a keystring N characters in length requires $4 + N$ bytes of memory. An expression containing I operators (exclusive of parentheses) and J variable references will occupy $6 + 2(I + J)$ bytes. If a typical profile is assumed to contain 25 keystring definitions[6] averaging 10 characters each, and each keystring is referenced in one expression with one operator associated with that reference, then an estimate of the number of profiles which may be accommodated in one search pass (P) may be made by equating

$$32{,}768 = P[25(4 + 10) + 6 + 2(25 + 25)] = P \cdot 456$$

yielding $P = 71.8$. Thus, around 70 profiles of this order of complexity may be accommodated in each search pass.

Actual text search is performed with text input taken from magnetic tape, the search structure being core resident, and selected citations are output to the second tape along with job numbers of the selecting profiles. Search processing is concurrent with text input to permit overlap of the time required for these functions. Heavy use is made of the tree-structure organization of keystring data to minimize time-consuming character-by-character comparisons and thereby increase search speeds. Overall search strategy involves the utilization of table look-up and chained hash-coding techniques to subset the body of keystring data. In addition, free use is made throughout the search algorithm of threaded-list scratch-memory structures to implement a wide variety of search functions. In this way, relatively powerful programming techniques are employed to compensate for the limitations of the small computer.

Of some interest are the sizes of the programs involved. WARPSET, the search structure compiler, the largest of system programs, consists of 733 assembly-language program statements and occupies 1680 bytes of memory (exclusive of I/O buffers). WARP-8 comprises 624 cards and 1358 bytes.

ECONOMICS OF THE SYSTEM

Currently the system is operating on 44 complex profiles developed by various members of the chemistry department. Total process time including all of the set averages under one minute per profile per week. Hence, dedication of the system (and one operator) to the CAS current awareness searching during only the standard 40-hour week means that around 2500 sophisticated users could be served.

FUTURE PROSPECTS

The Current Awareness Chemical Abstracts Search System is only one of three systems being developed. Basic aspects of the already developed search programming will be used in two other systems.

The second system is a Retrospective Search. The compression algorithm used in CARUNCH records between 12 and 15 weeks of *Chemical Condensates* on one standard 2400-ft reel. The Retrospective Search System is designed to compile one or more research profiles and search through the available *Chemical Condensates*. The estimated time now is an upper limit of eight hours for a single complex profile for a 5-year data base. A survey of the minicomputer market has shown that there is equipment available which, for on the order of $35,000 total installation cost, could do the retrospective search in approximately one hour.

The third system under development is an Interactive System. This system will have a CRT display and will allow searches at the rate of approximately 1.5 minutes per issue-week (this is a benchmark figure). With this system available, the researcher could design and test his profile on several weeks' tapes and modify the profile until it is optimized.

A grander scheme could involve placing an approximately $35,000 installation in a chemical or other library capable of serving as the retroactive search system and the interactive search system. This could not only serve researchers in developing profiles and making ordinary and retrospective retrieval searches, but could also act as a valuable educational device to students in the sciences. The coupling of a rapid retrieval system with CRT output that would give numbers to allow direct interaction with *Chemical Abstracts* or journals, either by the bound journal or microfilm reader, would offer an innovative and useful retrieval system to the scientist.

CONCLUSIONS

The feasibility of doing multiprofile Chemical Abstracts Current Awareness searches on a minicomputer system has been established. Details and economics have been described. Furthermore, more advanced projects in the area of retrospective searches and interactive searches are proven feasible by the basic search system's capability.

It should be noted that the entire search system is independent of data base. Only the front-end program which prepares the initial search tape is peculiar to the *Chemical Abstracts* data base. Therefore, this system can be applied to other literature bases which may be benefically searched in this fashion.

ACKNOWLEDGMENT

The authors gratefully acknowledge Mr. James deHaseth's assistance in obtaining the operating statistics and Dr. H. H. Dearman for use of his profile. This work was supported in part by the Institutional Grants Committee of the University of North Carolina and the William R. Kenan, Jr., Chemistry Department Endowment. Thomas L. Isenhour is an Alfred P. Sloan Research Fellow, 1971–75.

LITERATURE CITED

(1) Williams, M. E., and Schipma, P. B., "Design and Operation of a Computer Search Center for Chemical Information," *J. Chem. Doc.*, **10**, 158–162 (1970).

(2) Grunstra, N. S., and Johnson, K. J., "Implementation and Evaluation of Two Computerized Information Retrieval Systems at the University of Pittsburgh," *J. Chem. Doc.*, **10**, 272–7 (1970).

(3) Park, M. K., Carmon, J. L., and Stearns, R. E., "The Development of a General Model for Estimating Computer Search Time for CA Condensates," *J. Chem. Doc.*, **10**, 282–4 (1970).

(4) Roberts, A. B., Hartwell, I. O., Counts, R. W., and Davila, R. A., "Development of a Computerized Current Awareness Service Using Chemical Abstracts Condensates," *J. Chem. Doc.*, **12**, 221–3 (1972).

(5) Wilde, D. U., and Starke, A. C., "A Chemical Search System for a Small Computer," *J. Chem. Doc.*, **14**, 41–4 (1974).

(6) Schipma, P. B., "Computer Search Center Statistics on Users and Data Bases," *J. Chem. Doc.*, **14**, 25–9 (1974).

Interactive Pattern Recognition in the Chemical Laboratory

JAMES R. KOSKINEN and BRUCE R. KOWALSKI*

Laboratory for Chemometrics, Department of Chemistry, University of Washington, Seattle, Washington 98195

Received March 6, 1975

An interactive pattern recognition system has been developed for utilization in the chemical laboratory. This system has been designed so that the user need not know computer programming. Actual data analysis is done on a time-sharing host computer, and communication with the chemist is via an intelligent computer graphics terminal. The graphics terminal also displays projections of *n*-space data structure as two- or three-dimensional plots on the display screen. These plots can be manipulated by the chemist in real time to provide an approximate view of the *n*-space data structure. By examining the results of the various pattern recognition methods using the display terminal, the chemist can direct the application interactively, thereby increasing operational efficiency and allowing the gain of new insights into *n*-space ($n > 3$) data analysis applications.

Pattern recognition has been demonstrated to provide a powerful method for interpreting chemical data.[1] It has provided a general approach to solving a class of data processing problems commonly encountered in experimental chemistry. A statement of the general problem is: can an obscure property of a collection of objects (elements, compounds, mixtures, etc.) be detected and/or predicted using indirect measurements, made on the objects, that are known to be related to the property via some unknown relationship?

Obviously, the problem is not only to find and predict the property, but also to try to find the mathematical relationship that links the measurements to the property. Therefore, the problem can be considered as a mapping of objects from measurement space into property space. For pattern recognition methods, the objects, usually called patterns, can be considered as points in an n-dimensional hyper-space, where n is the number of measurements made on each object. Likeness among the objects is assumed to be reflected via the measurements as nearness of corresponding points in n-space. Thus, pattern recognition can also be described as a collection of methods that analyze n-space plots.

There are four major branches of pattern recognition which correspond to the four types of operations performed on the n-space data structure. First, the measurements can be preprocessed[2] by forming linear or nonlinear combinations of measurements to generate features. These features can be fewer in number than the measurements. This process is called feature reduction and is often a necessary step. The features are selected to yield a new information representation that is more amenable to data analysis by the other pattern recognition methods. An example will help demonstrate the need for preprocessing. Suppose a chemist submits a sample for NMR analysis and receives an intensity vs. time plot from a free induction decay experiment. The plot will have very little meaning to the chemist until it is preprocessed using a Fourier transform to change the information representation to the frequency domain. There are several preprocessing methods available to the chemist that do scaling operations, weighting, etc. Choice of a particular method is dependent upon the application.

The second type of operation comes from the scientist's excellent ability to recognize patterns in the familiar two- or three-dimensional space. Since the scientist cannot view n-space when n is greater than 3, the data structure (relative position of points) in n-space can be mapped to two-space or three-space by display methods[3] that seek to minimize information loss.

The last two branches of pattern recognition are called supervised learning and unsupervised learning[2] and are divided on the basis of what must be learned from the objects via the measurements. Supervised learning assumes that the sought-for property is known for some of the objects. Those objects are "tagged" as knowns (collectively called the training set) and are used to develop a classification rule that can be used to predict the property for unknown objects. Developing classification rules to separate known active anticancer drugs from inactive drugs and then classifying drugs that have not been tested in biological screening systems is an example of such an application.[4] If no training set exists and the goal of the study is to discover a useful property of the objects, unsupervised learning methods are used. The discovery of the periodicity of the elements from properties of the elements is an excellent example of unsupervised learning even though it was done before the advent of computers.

There are many different pattern recognition methods described in the literature. Yet most of the applications of