DOCUMENT RESUME

ED 025 413

SE 004 965

By- Scholtz, R. G., And Others
Feasibility Study of the Development of a Specialized Computer System of Organic Chemical Signatures of Spectral Data.
Illinois Inst. of Tech., Chicago.
Spons Agency- National Science Foundation, Washington, D.C.
Report No- IITRI-C6104-4
Pub Date Apr 68
Contract- NSF-C514
Note- 186p.
EDRS Price MF-$0.75 HC-$9.40
Descriptors- *Chemistry, *Information Retrieval, *Information Science, *Organic Chemistry
Identifiers- National Science Foundation

This final report of a feasibility study describes the research performed in assessing the requirements for a chemical signature file and search scheme for organic compound identification and information retrieval. The research performed to determined feasibility of identifying an unknown compound involved screening the compound against a file of chemical signatures of known compounds. The chemical signatures were obtained from (1) infrared spectrometry (IR), (2) nuclear magnetic resonance (NMR) spectrometry, (3) mass spectrometry (MS), (4) gas chromatography (GC), and (5) ultraviolet (UV) spectrophotometry. A physical state and element search preceded the signature search. The basic system contained data for 500 representative pure organic compounds. The discriminating ability of each signature, from most to least discriminating, follow the order (1) IR, (2) MS, (3) GC, (4) NMR, and (5) UV. The logic and search procedures of the experimental system were designed for ready conversion to a computer system. Appendixes contain a listing of the 500 compounds in the data base and other experiment details. (DH)

Report No. IITRI-C6104-4
(Final Report)

FEASIBILITY STUDY OF THE DEVELOPMENT OF A SPECIALIZED COMPUTER SYSTEM OF ORGANIC CHEMICAL SIGNATURES OF SPECTRAL DATA

National Science Foundation

Report No. IITRI-C6104-4
(Final Report)

FEASIBILITY STUDY OF THE DEVELOPMENT OF A SPECIALIZED COMPUTER SYSTEM OF ORGANIC CHEMICAL SIGNATURES OF SPECTRAL DATA

April 17, 1967, through April 16, 1968

Contract No. NSF-C514
IITRI Project C6104

Prepared by

R. G. Scholz, E. S. Schwartz,
and M. E. Williams

of

IIT RESEARCH INSTITUTE
Technology Center
Chicago, Illinois 60616

for

National Science Foundation
1800 G Street
Washington, D. C.

Attention: Mr. T. Quigley

Copy No. 95

April 29, 1968

FOREWORD

This is Report No. IITRI-C6104-4 (Final Report) on IITRI Project C6104, entitled "Feasibility Study of the Development of a Specialized Computer System of Organic Chemical Signatures of Spectral Data." This program is being performed for the National Science Foundation under Contract No. NSF-C514. The work reported herein was performed from April 17 1967, through April 16, 1968.

The project leader is Dr. Robert G. Scholz, Research Chemist, and the project is under the administrative guidance of Dr. Warner M. Linfield, Manager, Organic Chemistry Research. Major contributions to this report were made by Eugene S. Schwartz, Senior Scientist, and Miss Martha E. Williams, Manager, Technical Information Research.

The authors acknowledge the excellent work of the following persons who contributed to this study: M. M. Bogolin, P. A. Demink, B. E. Edwards, J. J. Finn, P. A. Llewellen, B. L. Mankus, H. J. O'Neill and A. J. Starshak.

Respectfully submitted,

IIT RESEARCH INSTITUTE

R. G. Scholz

R. G. Scholz
Research Chemist
Organic Chemistry Research

Approved by:

W. M. Linfield

W. M. Linfield
Manager
Organic Chemistry Research

RGS:dfp

FEASIBILITY STUDY OF THE DEVELOPMENT OF A SPECIALIZED COMPUTER SYSTEM OF ORGANIC CHEMICAL SIGNATURES OF SPECTRAL DATA

ABSTRACT

This final report describes a study of the feasibility of a proposed information file for organic chemical signatures. The purpose of the study was to assess the requirements for a file and a search scheme which could be used for compound identification and information retrieval. Use of the system could be made by chemists from industry, government and educational institutions. Users would be able to identify compounds and obtain references to the primary data sources by supplying either analytical data or a compound sample from which data could be derived. These data would then be submitted to the computer for screening against the master file. Each compound in the file will bear the Chemical Abstracts registry number and source references so that the chemical signature file will be tied in with the existing National Chemical Information System and original data sources.

The research performed determined the feasibility of identifying an unknown compound by screening it against a file of chemical signatures for known compounds. The chemical signatures were obtained from Infrared (IR) spectrometry, nuclear magnetic resonance (NMR) spectrometry, mass spectrometry (MS), gas chromatography (GC), and ultraviolet (UV) spectrophotometry. A physical state and element search preceded the searches on the five signatures. Data for 500 representative pure organic compounds were used to develop the basic system. Data were obtained from existing sources, such as ASTM, Sadtler, Varian, and other Government and private collections. Some additional data were generated.

The work encompassed selecting the representative compounds; acquiring source data; analyzing, indexing and cataloging the data; devising a file structure and search strategy; and demonstrating feasibility by operating the data-matching and -screening system.

A series of experimental searches was conducted using the 500 representative compounds as the data base. In these searches 100 compounds all having five signatures were treated as unknowns and screened against the file compounds. Also, eleven compounds of the 100 were analyzed by the five techniques and the data obtained were screened against the file. An inverted file with optical coincidence (Termatrex system) was used to demonstrate the feasibility of the identification system.

A candidate had to fulfill the following criteria:

1. Same physical state at standard conditions

2. Elements

3. Match on two of the three most intense IR bands, within tolerances, in any order

4. Match on two of the four most intense MS peaks in any order

5. Match on relative retention time, within tolerances, on any one of 2 selected GC columns

6. Match on the most intense peak, within tolerances, in NMR

7. Match on the most intense band position, within tolerances, in UV.

Of the 100 "unknowns" screened using these original matc ι criteria, 68 came through as unique candidates while 32 came through with multiple candidates. The large number of multiple candidates is attributed primarily to the lack of data (defaults).

A search made for the 11 "unknowns" which were analyzed produced 9 unique candidates, one lost to suspected erroneous data in the file and one with 4 candidates, attributed to lack of data.

A test on the 32 compounds having multiple candidates using modified, more rigid match criteria resulted in 24 coming through as unique candidates.

The discriminating ability of each signature, from most to least discriminating, follow the order: IR, MS, GC, NMR and UV.

The logic and the search procedures of the experimental system were designed for ready conversion to a computerized system.

To evaluate the results of these searches a weighting scheme based on chemical criteria, i.e., relative usefulness of the types of analytical data, was established. Exact, tolerance, and default matches were also considered in the weight assignments. When data for more than one compound survived the search, the differences in ratings clearly identified the better candidate.

The study has demonstrated that the composite chemical signature method for screening and matching data is feasible and provides a reliable base for compound identification.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES (cont.)

LIST OF FIGURES

FEASIBILITY STUDY OF THE DEVELOPMENT OF A SPECIALIZED COMPUTER SYSTEM OF ORGANIC CHEMICAL SIGNATURES OF SPECTRAL DATA

I. INTRODUCTION

In the past year IIT Research Institute has been engaged in a program to study the feasibility of an analytical data information file for organic compounds. The program was carried out under the sponsorship of the National Science Foundation.

Although there is a wealth of analytical data available, the analytical chemist does not know where it is located nor does he have ready access to it. This information would be of considerable help particularly for identification of unknown compounds. In the area of organic chemistry, the number of compounds is steadily increasing and research in the areas of organic synthesis as well as instrumental analysis is becoming more diversified. The volume and complexity of available analytical data are growing and even well-trained analytical and organic chemists find it difficult to keep pace with this growth. Thus, there is a need on the part of chemists to have a place where they can go for help with their analytical problems.

The usual procedure for analytical and organic chemists trying to confirm the identity of a compound is to obtain spectral data by as many analytical techniques as are available to compare these data with reference data and identify the unknown compound. All too often the chemist is limited by inadequate files. A search would be facilitated by a centralized system which can use either spectral data or can provide a more complete array of spectral data to search for and identify the unknown compound and direct the user to additional analytical, chemical and physical data. To date, there is no such system. There are available various collections of spectral data from industrial and government organizations such as Sadtler Research Laboratories, Varian Associates, the American Society for Testing and Materials, the American Petroleum Institute and the National Bureau of Standards. Most of these collections are limited in scope and, more important, are not correlated with each other. At present a researcher will have to spend considerable time and money in an effort to locate a source that has the analytical data on file and which will help identify an unknown compound and reference the original data. A well-planned and ultimately computerized central data file would provide a single location to which a scientist could submit his data or bring compounds from which the spectral data could be obtained. Such a file containing all published standardized data could be screened to identify an unknown compound. If the data are not available he need not go elsewhere if the file incorporated data from other collections. If the data are contained in the file he will be able to identify the compound, learn the original source of the data and be directed to additional chemical and physical data pertaining to the compound. The objective of the research study reported herein was to determine the capability of a composite signature to identify an unknown compound. The composite signature is made up of a minimum number of selected data points obtained from 5 analytical techniques. Identification would be accomplished by matching the composite signature of the unknown against a file of composite signatures. In a sequential search, candidate compounds would be selected by obtaining the logical intersection of the individual signature matches.

We have prepared a data file and developed the logical procedures for searching the file. The work involved acquiring data for a reference file of 500 pure organic compounds, indexing and tabulating the data, defining the search strategy and finally testing the system for its ability to select the proper compounds. This final report on the project describes the work and the results.

Recommendations based on the study are provided. A comprehensive list showing data sources, type of data available, costs and other information are also included.

II. SELECTION OF COMPOUNDS

Our initial efforts were directed towards selecting 500 compounds for which we could obtain 2, 3, or 4 sets of data derived from infrared spectrometry, nuclear magnetic resonance, mass spectrometry and gas chromatography. We wanted the compounds to represent many compound types and, in some cases, compounds whose composite signatures were similar to test the selectivity of the sequential search to select the proper compound. To aid in the selection of compounds the SOCMA Handbook for Commercial Organic Chemical Names was used to identify the compounds by their IUPAC name and CAS registry number. When a compound was identified by our system the CAS registry number would accompany each compound allowing the user to obtain additional chemical and physical information about the compound. Ultimately, 838 different organic compounds were considered for use. These represented hydrocarbons, terpenes, heterocyclics, polyfunctionals, aromatic, amino acids, steroids, alkaloids and others. Also, we made the decision to include ultraviolet as a fifth signature. From the 838 compounds, we hoped to select as many as possible (up to 500) that had 5 signatures. In the 500 selected compounds, 23 of the original 24 compound types were represented with only carbohydrates being excluded.

III. ACQUISITION OF DATA

A. General

Our plan was to extract and index data from various literature and commercial sources to build our data base. Initially, since we had not yet determined the minimum amount of data necessary for identification we were concerned with getting more data than we thought we might need for the final system.

We tried to find all signatures for all 838 compounds. This proved to be a major problem and more difficult than had been anticipated. The difficulty lay not only in the lack of data but in our method of compound selection - we were limiting the compounds to several hundred and restricting ourselves to finding as many signatures as we could for just these compounds. However, this approach was reasonable at this stage of the program inasmuch as we were trying to maintain the diversity of compound types.

Of the 838 compounds considered for use, 357 had NMR data, 233 had UV data, 478 had MS data, 476 had IR data and 308 had GC data. Of these, 500 were selected for use in our file. The following is a breakdown of the data distribution for these 500 compounds:

100 had 5 signatures each

174 had 4 signatures each

195 had 3 signatures each

31 had 2 signatures each

Table 1 lists the number of compounds available in 5-, 4-, 3-, and 2-signature combinations. Table 2 summarizes the number of signatures available in the 23 chemical groups used in the experiment. Tables 3, 4 and 5 list the combinations of signatures available in the 23 chemical groups in 4-, 3-, and 2-signature combinations.

Table 1

COMPOUNDS AVAILABLE IN MULTISIGNATURE COMBINATIONS

SIGNATURES	NUMBER
UV/MS/GC/IR/NMR	100
MS/GC/IR/NMR	89
UV/GC/IR/NMR	2
UV/MS/IR/NMR	50
UV/MS/GC/NMR	0
UV/MS/GC/IR	33
UV/MS/GC	3
UV/MS/IR	21
UV/MS/NMR	7
UV/GC/IR	2
UV/GC/NMR	1
UV/IR/NMR	7
MS/GC/IR	62
MS/GC/NMR	4
MS/IR/NMR	85
GC/IR/NMR	3
UV/MS	6
UV/GC	0
UV/IR	1
UV/NMR	0
MS/GC	6
MS/IR	9
MS/NMR	3
GC/IR	3
GC/NMR	0
IR/NMR	3
TOTAL	500

Table 2

DISTRIBUTION OF MULTISIGNATURES BY FUNCTIONAL GROUP

FUNCTIONAL GROUP	NUMBER OF COMPOUNDS IN GROUP	NUMBER OF SIGNATURES AVAILABLE FIVE	FOUR	THREE	TWO
ALCOHOLS	50	19	20	10	1
ALDEHYDES	21	8	5	7	1
ALKALOIDS	3	0	0	2	1
AMIDES/IMIDES	8	0	1	4	3
AMINES	18	0	4	13	1
AMINO ACIDS	4	0	1	2	1
CARBOXYLIC ACIDS	24	0	8	15	1
DIOLS	14	4	9	1	0
ESTERS	56	4	23	26	3
ETHERS	22	2	8	9	3
HALOCARBONS	57	5	22	30	0
HETEROCYCLICS	22	0	11	11	0
HYDROCARBONS	70	23	22	22	3
KETONES	27	10	9	7	1
NITRILES	18	3	7	8	0
NITRO	9	6	1	1	1
POLYFUNCTIONALS	48	10	14	18	6
STEROIDS	2	0	0	1	1
SULFIDES	8	5	2	1	0
SULFONES	1	0	0	0	1
SULFOXIDES	1	0	1	0	0
TERPENES	11	0	4	5	2
THIOLS	6	1	2	2	1
TOTALS	500	100	174	195	31

Table 3

4-SIGNATURE COMBINATIONS BY FUNCTIONAL GROUP

FUNCTIONAL GROUP	MS/GC/IR/NMR	UV/GC/IR/NMR	UV/MS/IR/NMR	UV/MS/GC/NMR	UV/MS/GC/IR
ALCOHOLS	3	1	3	0	13
ALDEHYDES	4	0	0	0	1
ALKALOIDS	0	0	0	0	0
AMIDES/IMIDES	0	0	1	0	0
AMINES	1	0	3	0	0
AMINO ACIDS	0	1	0	0	0
CARBOXYLIC ACIDS	0	0	8	0	0
DIOLS	0	0	3	0	6
ESTERS	19	0	3	0	1
ETHERS	6	0	2	0	0
HALOCARBONS	19	0	2	0	1
HETEROCYCLICS	6	0	5	0	0
HYDROCARBONS	7	0	10	0	5
KETONES	4	0	2	0	3
NITRILES	7	0	0	0	0
NITRO	1	0	0	0	0
POLYFUNCTIONALS	6	0	7	0	1
STEROIDS	0	0	0	0	0
SULFIDES	0	0	1	0	1
SULFONES	0	0	0	0	0
SULFOXIDES	1	0	0	0	0
TERPENES	3	0	0	0	1
THIOLS	2	0	0	0	0
TOTALS	89	2	50	0	33

Table 4

3-SIGNATURE COMBINATIONS BY FUNCTIONAL GROUP

FUNCTIONAL GROUP	UV/MS/GC	UV/MS/IR	UV/MS/NMR	UV/GC/IR	UV/GC/NMR	UV/IR/NMR	MS/GC/IR	MS/GC/NMR	MS/IR/NMR	GC/IR/NMR
ALCOHOLS	1	3	1	2	0	0	2	0	1	0
ALDEHYDES	0	0	0	0	0	0	6	1	0	0
ALKALOIDS	0	0	0	0	1	0	0	1	0	0
AMIDES/IMIDES	0	1	0	0	0	0	0	0	3	0
AMINES	0	2	0	0	0	3	0	0	8	0
AMINO ACIDS	0	0	0	0	0	0	0	0	0	2
CARBOXYLIC ACIDS	0	2	0	0	0	0	0	0	13	0
DIOLS	0	0	0	0	0	0	1	0	0	0
ESTERS	0	0	0	0	0	0	23	0	3	0
ETHERS	0	1	0	0	0	0	8	0	0	0
HALOCARBONS	0	0	0	0	0	0	4	0	26	0
HETEROCYCLICS	0	1	2	0	0	1	1	0	6	0
HYDROCARBONS	2	8	4	0	0	0	1	1	6	0
KETONES	0	0	0	0	0	2	5	0	0	0
NITRILES	0	0	0	0	0	0	2	1	5	0
NITRO	0	0	0	0	0	1	0	0	0	0
POLYFUNCTIONALS	0	2	0	0	0	0	6	0	9	1
STEROIDS	0	0	0	0	0	0	0	0	1	0
SULFIDES	0	0	0	0	0	0	0	0	1	0
SULFONES	0	0	0	0	0	0	0	0	0	0
SULFOXIDES	0	0	0	0	0	0	0	0	0	0
TERPENES	0	1	0	0	0	0	3	0	1	0
THIOLS	0	0	0	0	0	0	0	0	2	0
TOTALS	3	21	7	2	1	7	62	4	85	3

Table 5

2-SIGNATURE COMBINATIONS BY FUNCTIONAL GROUP

FUNCTIONAL GROUP	UV/MS	UV/GC	UV/IR	UV/NMR	MS/GC	MS/IR	MS/NMR	GC/IR	GC/NMR	IR/NMR
ALCOHOLS	1	0	0	0	0	0	0	0	0	0
ALDEHYDES	1	0	0	0	0	0	0	0	0	0
ALKALOIDS	0	0	0	0	1	0	0	0	0	0
AMIDES/IMIDES	0	0	0	0	0	1	1	0	0	1
AMINES	0	0	0	0	0	1	0	0	0	0
AMINO ACIDS	0	0	0	0	0	0	0	1	0	0
CARBOXYLIC ACIDS	0	0	0	0	0	0	1	0	0	0
DIOLS	0	0	0	0	0	0	0	0	0	0
ESTERS	0	0	0	0	1	0	1	1	0	0
ETHERS	1	0	0	0	1	0	0	1	0	0
HALOCARBONS	0	0	0	0	0	0	0	0	0	0
HETEROCYCLICS	0	0	0	0	0	0	0	0	0	0
HYDROCARBONS	2	0	0	0	0	1	0	0	0	0
KETONES	0	0	0	0	1	0	0	0	0	0
NITRILES	0	0	0	0	0	0	0	0	0	0
NITRO	0	0	1	0	0	0	0	0	0	0
POLYFUNCTIONALS	1	0	0	0	2	2	0	0	0	1
STEROIDS	0	0	0	0	0	1	0	0	0	0
SULFIDES	0	0	0	0	0	0	0	0	0	0
SULFONES	0	0	0	0	0	1	0	0	0	0
SULFOXIDES	0	0	0	0	0	0	0	0	0	1
TERPENES	0	0	0	0	0	1	0	0	0	0
THIOLS	0	0	0	0	0	1	0	0	0	0
TOTALS	6	0	1	0	6	9	3	3	0	3

The list of the 500 compounds in the data base shown in Appendix A contains, in addition to the IITRI number assigned to a compound, common names and the name designated by Chemical Abstracts, and the data availability for each compound.

B. Gas Chromatography

The analytical data of qualitative value in gas chromatography is retention time. Worker's have been trying to find the best retention designation that will be accurate, universally applicable and acceptable. The main difficulty is the lack of universal applicability; each worker has a private collection of columns for analysis and their retention data are unrelatable to other workers' data. Most commonly used data are relative retention values and a recently introduced value called the Kovats' Retention Index.[1] We decided to use the relative retention values because of their wide use and ease of measurement. We also chose toluene as the internal standard. We believe that for any future work with this system we would select another internal standard because under the operating conditions we had selected, the retention distance of toluene is somewhat short, possibly resulting in relatively large errors in measurement. A compound with a longer retention distance, such as hexyl or heptyl acetate, would be better suited for this purpose.

The operating conditions chosen for our data base are shown in Table 6. Two columns, Carbowax 20M and Silicone SE-30

[1] Kovats, E., Z. Anal. Chem., 181, 351 (1961).

thermostatted at 160°C were selected. These columns are frequently used by gas chromatographers and offer excellent thermal stability and high-temperature operation if needed. The liquid phases also are readily obtainable from a multitude of distributors.

Table 6

OPERATING CONDITIONS FOR GAS CHROMATOGRAPHIC RETENTION DATA

Condition A	Condition B
Column: 20% SE-30 (Gen. Elec. Co.) 0.5% Polytergent J-300 60/70 mesh Celite 545(J-M Co.) 3/8 in. x 12' Aluminum 160°C	Column: 21.9% Carbowax 20M (Carbide & Carbon Chem. Co.) 0.5% Polytergent J-300 60/70 mesh Celite 545(J-M Co.) 3/8 in. x 12' Aluminum 160°C
Solvent: toluene & chloroform	Solvent: methylene chloride
Conditioned: 48 hr @ 180°C 18 hr @ 160°C	Conditioned: 40 hr @ 170°C
Carrier gas: Helium	Carrier gas: Helium
V_g toluene: 14.6 ml	V_g toluene: 15.9 ml

We obtained gas chromatographic data from two sources:

1. W. O. McReynolds' published collection[2]
2. Data generated in our laboratory.

McReynolds' compilation includes relative retention data and Kovats' Retention Indeces (I_x) for about 350 compounds on 80 different substrates at two different temperatures. We were able to use about 180 compounds out of his compilation. Relative times (R_{V_g}) were calculated from the following equation.

$$R_{V_g} = \text{Relative Retention} = \frac{R_x - R_{air}}{R_{tol} - R_{air}}$$

[2] McReynolds, W. O., "Gas Chromatographic Retention Data, Preston Technical Abstracts Co., Chicago, Ill., 1966.

where R_x, R_{tol} and R_{air} are measured retention distances for the compound, toluene and air, respectively. Although Kovats' retention indices are included in our data file they were not used in the screening. Hence, no detail will be given concerning its computation except to say it is a number that relates the retention time of the unknown to the retention times for the normal hydrocarbons which bracket the unknown.

In order to increase available gas chromatographic data we found it necessary to generate data. At first, an attempt was made to assess the usefulness of literature data. ASTM published a gas chromatography data compilation (ASTM DS25A) which compiles data from the literature through 1965. Unfortunately, as Table 7 illustrates, relative retention distances obtained under similar operating conditions can vary widely making such data unusable for our purposes. Two conclusions drawn from the literature survey are:

1. It is not feasible to use literature sources for standard gas chromatographic data
2. We were (and would be) restricted to the GC data we have on hand or can generate under the conditions used by McReynolds.

Consequently, we requested and received permission from the National Science Foundation to generate gas chromatographic data for 100 to 150 compounds. Mr. W. O. McReynolds supplied us with the Carbowax 20M and Silicone SE-30 columns he used and we generated data for about 125 compounds.

Table 7

VARIABILITY OF GC RETENTION DATA

Compound	Relative Retention Data[a]					
	On 20M Relative to Acetone					
Acetoldehyde	0.68	0.53	0.37			
Propionaldehyde	0.80	0.86	0.87	0.90		
Butyraldehyde	1.40	1.35	1.35	1.34		
Ethanol	2.20	1.27	1.40	1.48		
Isovaleraldehyde	3.70	2.6	2.06			
Formaldehyde	0.60	0.39				
	On 20M Relative to Toluene					
Dipropylether	0.20	0.26				
Diethylether	0.68	0.11				
Methanol	0.35	0.35				
n-Amylacetate	2.08	1.82				
n-Heptanol	5.73	4.7				
n-Hexane	0.09	0.10				
n-Decane	0.65	0.70				
Acetone	0.29	0.31	0.21			
2-Butanone	0.44	0.47				
Benzene	0.62	0.64				
Ethylbenzene	1.52	1.53				
Methylformate	0.19	0.20				
n-Butylacetate	0.99	0.94				
Acetaldehyde	0.16	0.16	1.2			
Butyraldehyde	0.38	0.42				
	On 20M Relative to Benzene					
Butylacetate	2.40	2.01				
Cyclohexane	0.277	0.228				
1-Butanol	3.58	3.08				
Ethanol	0.74	1.29				
	On SE30 Relative to Cholestane					
Cholesterol	1.78	1.95	1.94	1.96	1.74	1.2
	On SE30 Relative to Pyrene					
Chrysene	2.78	2.81	2.81	2.27	2.08	1.83
	On SE30 Relative to Pentane					
Dimethylbutane	1.35	1.24				

[a] Each column represents a different literature source or different column temperature. Some groups have only one entry per column; e.g., cholesterol shows 6 different sources and/or temperatures.

The actual time spent in generating these data was approximately 20 man days. We estimate that, on the average, we can generate GC data for 10 compounds during an 8-hour working day per instrument.

Our final data file for gas chromatography contained data for 308 compounds of which 36 had data on Silicone SE-30 only, 7 had data on Carbowax 20M only and 265 had data on both columns.

Although we are unhappy about the lack of GC data we have decided to retain it in the system because of the excellence of gas chromatography as a qualitative tool. Since there are little or no additional gas chromatographic data available which we could incorporate into our file, any future data acquisition would require generation. In order to include gas chromatography in the final working system, it will be necessary to standardize the operating conditions such that all users of the data retrieval system would obtain data on columns specially supplied for that purpose. This is a reasonable solution to the perplexing problem now existing among gas chromatographers regarding variability of "standard" data. The logic underlying GC's qualitative ability forecasts greater importance, increased emphasis, and improved data and techniques for qualitative analysis by GC. Thus it should be relatively easy to ask the scientific community to generate and contribute data by using a prescribed set of conditions. Such a proposition would probably be accepted because scientists have already accepted such systems in the areas of NMR, MS, X-ray diffraction, IR, etc.; universally accepted "standard" data that have been compiled are accepted by all for comparison with individual data.

ASTM is approaching the problem of standard GC data with their ASTM DS25A compilation, but, as evidenced by the data in Table 7, unique data for each compound are not available and won't be until a set of standard conditions is agreed upon by the workers in this area.

C. Nuclear Magnetic Resonance

Probably the most difficult area for indexing data thus far encountered is the area of NMR. A number of systems have been developed by various workers none of which is either completely satisfactory or widely accepted. The difficulties are inherent in the nature of NMR spectra.

In theory, and frequently in practice, a complete NMR spectrum for a given chemical structure may be generated by a computer. However, the reverse process, generation of a structure from a spectrum has been too complex for satisfactory handling by computer methods. A skilled practioner may analyze the patterns present in a spectrum almost intuitively, but the rules by which such analyses are made become too complex for practical translation into computer language.

The purpose for which we are indexing these and all other data must be kept in mind, however. Our purpose is not to derive chemical structures from the data but to use the data as a "map" to help identify unknown compounds and to direct us to the complete spectra or original data source. The methods of indexing must be simple (free of interpretive quality) and amenable to computerization. With these goals in mind we can employ indexing methods which are of little value in terms of relating to structural characteristics but are significant in terms of defining, in an abbreviated fashion, the uniqueness of the total data.

The problems related to various NMR indexing systems may best be discussed with reference to particular examples. The NMR spectrum of diethyl phthalate (Figure 1) is a typical example of a relatively simple spectral pattern in the format supplied by the Sadtler Research Laboratories, Inc. The molecule is symmetrical and has only four chemically distinct kinds of hydrogen (indicated by a, b, c, and d). The pattern of a triplet at 1.35 ppm and a quadruplet at 4.29 ppm with integral ratio of 3 to 2 is immediately recognizable as characteristic of an OC_2H_5 group. The pattern in the aromatic region (7.47 and 7.63 ppm) is typical of adjacent hydrogens on an aromatic ring, further coupled to more distant aromatic ring hydrogens. It should be noted here that the Sadtler assignments are chemical shift values arising from interpretation of the spectrum, and in the cases of b, c and d do not correspond to any actual peak positions on the spectrum. They represent the centers of multiplet patterns.

PHTHALIC ACID, DIETHYL ESTER

$C_{12}H_{14}O_4$ Mol. Wt. 222.33 B. P. 298-299°C/735 mm

Source: Eastman Chemical Products, Inc., Kingsport, Tenn.

Filter Bandwidth: 4 cps
Sweep time: 250 sec
Sweep width: 500/250 cps
Sweep offset: -/225 cps
Spectrum amp: 10
Integral amp: 25
Conc. ~100mg/0.5ml CCl_4

ASSIGNMENTS
a 1.35
b 4.29
c 7.47
d 7.63
e
f
g
h
i
j

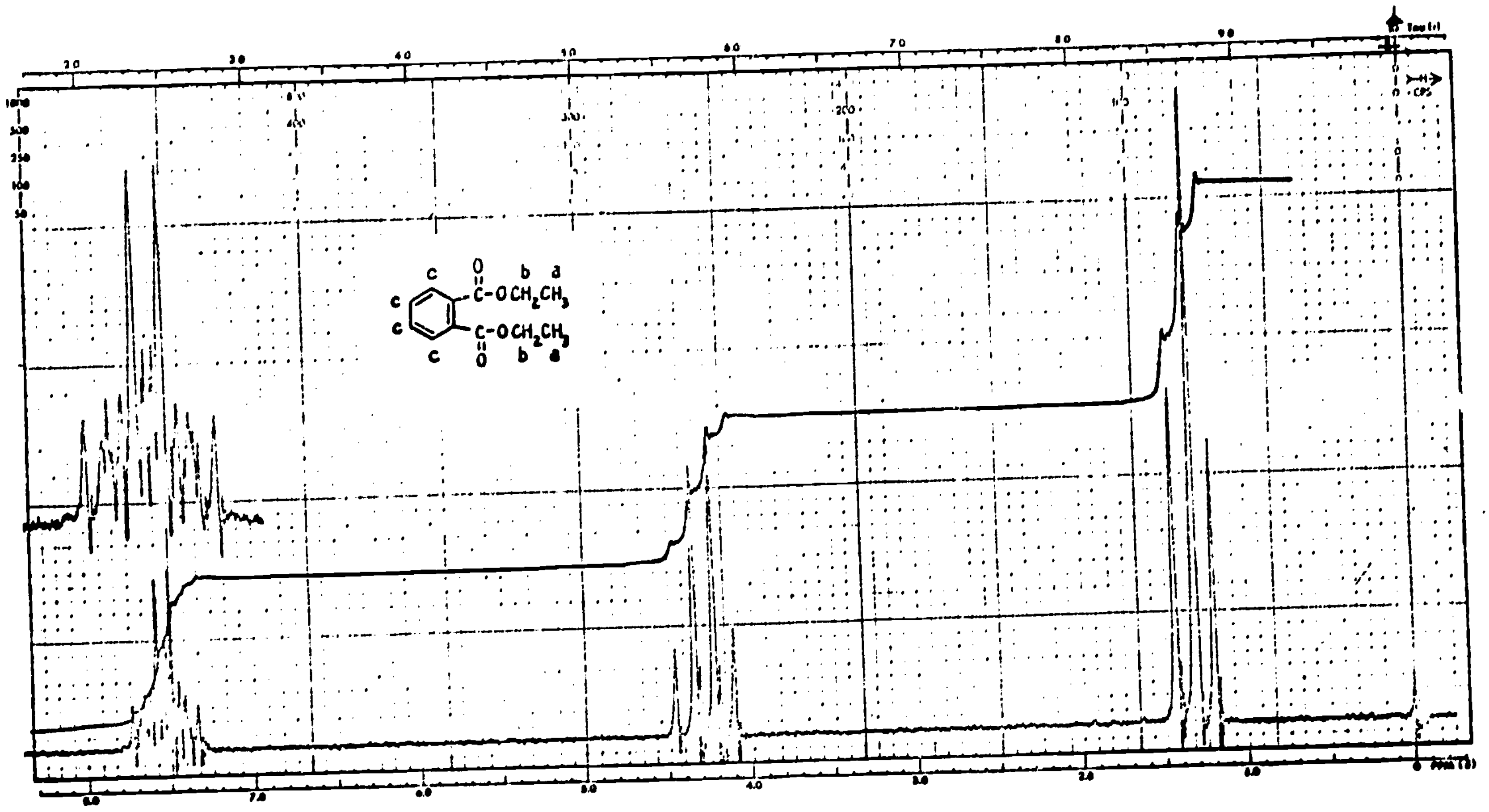


Figure 1

NMR SPECTRA OF DIETHYL PHTHALATE

To further complicate the situation, the instrumentation of NMR is undergoing a continuing evolution towards higher magnetic fields. The most widely used instruments at the present time operate at 60 Mhz, but 100 Mhz instruments are commercially available and experimental studies at 220 Mhz have been reported.

The significance of this evolution to our program may be illustrated by reference to the spectrum of diethyl phthalate at 60 Mhz and at 100 Mhz. Two parameters determine the observed patterns in the NMR, the chemical shift δ , expressed in ppm from trimethylsilane (TMS) and the spin-coupling constants, J in hz. The observed value of δ is dependent only on the magnetic field, and when expressed in the dimensionless terms of ppm is independent of the field strength of the instrument. The value of J is constant in hz, and is independent of the magnetic field. The quartet (b) in diethyl phthalate has a δ of 4.29 ppm. The observed peak positions at 60 Mhz are shown in Table 8. Simple calculations give the values at which they would appear in a spectrum run at 100 Mhz. Note that the <u>only</u> number which is the same in both spectra is the chemical shift δ , which is a value obtained by interpretation of the spectra and does not represent a peak position. In more complicated cases where multiplet patterns overlap, the difficulties obviously increase rapidly.

Table 8

NMR PEAK POSITIONS FOR DIETHYL PHTHALATE

	60 Mhz		100 Mhz	
	ppm	hz	ppm	hz
δ	4.29	257.5	4.29	429
Peak 1	4.12	247	4.19	419
2	4.23	254	4.26	426
3	4.35	261	4.33	433
4	4.47	268	4.40	440

δ = Chemical shift of multiplet.

A simple version of indexing observed band positions is the Sadtler Specfinder system, a method apparently derived from IR experience. In this system the most intense peak in each 1 ppm region is recorded, as well as the most intense peak in the whole spectrum. The data for diethyl phthalate is presented in Table 9 below. The range from 0-14 ppm is required to insure inclusion of all possible compounds although peaks are very rarely found beyond 9 ppm. While this system is far from ideal, from a practical point of view it appears to be best for the purposes of this study. The resulting series of digits represent a set of data which is characteristic of a relatively small number of compounds, and which may nevertheless be readily obtained from raw published data. The problem of peak position shift with changes in operating frequency outlined in the previous section is still present, but is minimized by the fact that the strongest peak in a multiplet will be near its center, and the shift with frequency will therefore be small.

Table 9

NMR PEAK POSITIONS FOR DIETHYL PHTHALATE
ACCORDING TO SADTLER

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Strongest Band
-	35	-	-	23	-	-	50	-	-	-	-	-	-	-	1.35

A more refined system, developed by Varian Associates, consists of recording the relative integrated intensity of the peaks in each 1 ppm range. For our example of diethyl phthalate, these data would have the form shown in Table 10. While this form of data recording contains a little more structural information (i.e., the ratios of the various chemically different kinds of protons in the molecule) it is less sensitive for distinguishing between structural isomers since diethyl isophthalate, diethyl terephthalate, and probably any ethyl esters of trisubstituted benzoic acids would give the same profiles. This is because the profile simply represents a ratio of 2 aromatic protons to one ethyl ester group. The shape of the aromatic signal would change markedly, but it would not be shifted outside of the 7-8 ppm region. An advantage of the system is that the data may be transferred directly from the spectrometer to computer storage, but this would be useful only in a program of data generation. Also, for our present purposes, a good deal of the requisite data would be unavailable from the literature except via calculations from theoretical integrals since very little integral data is published. We investigated this method through discussions with Dr. LeRoy Johnson of Varian Associates so as to have a complete appraisal of the method on hand. This method of indexing did not appear to be practical for our purposes.

Table 10

NMR PEAK POSITIONS FOR DIETHYL PHTHALATE
ACCORDING TO VARIAN

ppm	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
%	-	43	-	-	29	-	-	29	-	-	-	-	-	-	-

An experimental Termatrex index of NMR data is being evaluated in an ASTM project. In this system, compounds were indexed by the number, position, area and number of lines per "cluster" in their spectra, as well as by the presence or absence of certain functional groups or elements. While the system seems fairly effective on the scale tried, it is too cumbersome for a project of the magnitude we propose, and the method of indexing is open to many of the objections mentioned above in connection with other techniques.

Other workers have devised indices based on chemical shifts and spin-multiplicities, but since those techniques all require considerable interpretation of the spectra prior to storage of data, we can eliminate these possibilities without further discussion.

Therefore, the best compromise of reasonably definitive yet brief descriptions of the experimental data which are easily handled by computer techniques is the Specfinder system.

We have used this approach in recording our data but used the first most intense peak for searching. Some tests, which will be discussed later, were conducted using the first and second most intense peaks.

A system based on this approach, particularly where the instrument is connected directly to the computer, would require a standard condition in frequency (60, 100 or 220 Mhz). Data obtained at different frequencies would have to be interpreted before presentation to the data file for screening. Although somewhat inconvenient, this is what would prevail with any file that exists now. Such a compromise would not detract from the usefulness of the system considering its qualitative application and purpose.

Our main sources of data were the Sadtler and Varian spectra collections. In addition to these, we have searched several other sources in order to increase the number of compounds on our selected list for which we have all five types of data. We have used the "Formula Index to NMR Literature Data," Volumes I and II, to locate the spectra of a number of compounds. In addition to the MCA and API collections, which are handled by the Data Distribution Office of the Thermodynamics Research Center at the Texas A&M Research Foundation (and which we have on hand), we have obtained copies of the Tiers' Tables, a collection of chemical shift values privately circulated by G. Tiers of 3M in 1958, and the Humble Collection, circulated by N. F. Chamberlain of Humble Oil Company in 1959. The Jeolco Collection of spectra has also been obtained from Sadtler.

Although the current practice is to report chemical shifts in ppm from TMS, the data in the older literature are reported in a variety of forms. This situation has provided some interesting exercises in calculation, but so far the quality of the data appears satisfactory after conversion to the ppm scale. For instance, the Humble spectra are calibrated in mgauss, relative to benzene. These values are converted to ppm relative to TMS by dividing by the magnetic field strength and adding a constant to change the reference from benzene to TMS. Many of the spectra are run at 30 or 40 Mhz, which magnifies the spin coupling constants relative to the chemical shifts. Further, a variety of solvents have been used, and solvent shifts must therefore be considered. However, qualitative inspection of the data and comparison of data for a single compound from several sources indicate that the constancy of the values recorded in our system is adequate, although it leaves something to be desired in absolute reproducibility.

NMR data derived from literature sources comprised 6 percent of our total NMR data input. The majority were obtained from Sadtler and Varian compilations.

D. Infrared

Two basic approaches to indexing IR data were considered. These are the methods used by Sadtler and ASTM. We decided to use ASTM's approach with the exception that we listed the peaks in order of decreasing intensity. ASTM's method is to list, in order of wavelength to the nearest 0.1 micron, the most intense band plus all bands whose intensities are 10 percent or greater of the most intense band. Sadtler, on the other hand, lists the most intense band within each 1 micron region from 0 to 15 microns. Using a modified ASTM approach we indexed the three most intense bands (except 3.4μ which was so prevalent as to be relatively non-discriminating) in order of decreasing intensity.

Indexed IR data for about 700 compounds from our list were obtained from ASTM through Dr. Kuentzel. About 1,800 punched IBM cards were provided by Dr. Kuentzel and represented, in many cases, multiple listings of IR data for a given compound.

Because the 1,800 ASTM IR punched cards were not interpreted, printouts of the deck were prepared to facilitate data extraction from the cards. The ASTM cards record all significant IR peaks, but they do not indicate intensities. Because our search routine which will be discussed later requires data for the three most intense peaks, it was necessary to consult the original spectra (mainly Sadtler) to determine intensities. The intensities were recorded on the printout from the ASTM deck, and the printout with intensities is now a more complete source document for the project than the ASTM cards.

The entires representing 542 spectra from Sadtler Research Laboratories were all characterized by punches 4 and 3 in column 79 and y and 1 in column 80. In columns 74 through 78 the ASTM number gave these in numerical order; thus it was a much simpler task to check the value of the Sadtler spectrum and the accuracy of the ASTM listing.

Fully 10 percent of the Sadtler spectra (1962 edition) reviewed were poor enough to raise doubts concerning some of the absorption peaks. Usually this lack of acceptability can be attributed to concentration control (the samples were too thick or too thin), to impure samples, or to poor mulling techniques. It is understandable why Sadtler has published a number of these again as "series B." Furthermore, in almost 5 percent of the spectra viewed, ASTM had made an error either in interpreting the spectrum or in failing to transcribe and keypunch the data correctly. In either case, strong absorptions (not attributable to the solvent or mulling medium) were not listed, or medium absorptions were listed incorrectly. Medium strength absorptions that had not been listed were not counted as a mistake in the above error analysis. This emphasizes the importance of having trained personnel verify any data placed in the data bank of the retrieval system.

While the intensities of the 3 most intense bands were accorded the intensities, per se, were not used in the screening process. The intensities were used only to rank the three peaks in order of decreasing intensity.

E. Mass Spectrometry

Indexing mass spectrometral data proved relatively easy. The ten most intense peaks were indexed along with their relative intensities. The only standard criteria restricting the selection of the peaks is that they be obtained with electron bombarding energies of 70 electronvolts, a widely accepted practice. The most intense peak (base peak) was given an intensity value of 1000 and the remainder were given values in percents of 1000. In our actual screening only the first 4 peaks (in order of decreasing intensity) were used. The intensity, per se, was not used in the screening process.

All of the mass spectral data were taken from two sources:

1. ASTM Publication No. 356, "Index of Mass Spectral Data, 1st Edition, ASTM, Philadelphia, Pa., 1963.

2. Cornu, A., Massot, R., "Compilation of Mass Spectral Data," Heyden & Son, Ltd., London, 1966.

The ASTM publication contains data for about 3,200 compounds while Cornu and Massot's has data for about 5,000 compounds. Most of the compounds in the ASTM compilation are also found in the Cornu-Massot compilation.

F. Ultraviolet Spectrophotometry

About two-thirds of the way through the program we became concerned about the lack of gas chromatographic data. We decided to incorporate ultraviolet (UV) spectral data into our file to provide a broader basis for identification of compounds. Accordingly, spectral data for the 838 compounds were sought using the following three data sources:

1. Sadlter's compilation located at Abbott Laboratories, North Chicago, Illinois

2. "Ultraviolet Spectra of Aromatic Compounds," R. A. Friedel and M. Orchin, John Wiley & Sons, Inc., New York, 1957

3. "Ultraviolet and Visible Absorption Spectra Index for 1930-1953," H. N. Hershenson, Academic Press, New York, 1956.

The third source is a literature index that provided references to the original data. Sadtler's file was the major source of data. Because ASTM's UV spectral files were not readily available, we did not use that source for UV data. The ASTM file has data for almost 26,000 compounds.

The majority of UV spectra have been obtained in solvents over a wavelength range of 200 to 400 millimicron ($m\mu$). The most common solvent was methanol. Hence, data for compounds run in methanol were selected. Our data were reported in the range of 200 to 400 $m\mu$. We recorded only the strongest peak and excluded extinction coefficients. The strongest absorption band was located to the nearest $m\mu$. A tolerance of ± 2 $m\mu$ was used in the search. Selection of this tolerance was based on experience in reproducing wavelength readings that are subject to both instrumental and interpretational variations. A sample that is transparent in the UV region is so indicated in the data file.

G. Survey of Signature Availability

A survey was made by information specialists between June, 1967 and March, 1968 to determine the overall availability of signature data from the major data sources in the country.

The American Society for Testing and Materials (ASTM) is responsible for the identification and indexing of the significant spectral collections throughout the United States. The ASTM index identifies virtually all spectra available from Sadtler Research Laboratories, National Bureau of Standards (NBS), Documentation Molecular Spectroscopy (DMS), Thermodynamics Research Center (TRC) (which includes the American Petroleum

Institute (API) and Manufacturing Chemists Association (MCA) collections), the Coblentz Society, the Infrared Documentation Center, Japan (IRDC-Japan), and the open literature, domestic and foreign.

As of March, 1968, ASTM had in its files information concerning 100,000 IR, 3,200 MS, and 25,749 UV spectra. A subcommittee is working on the indexing scheme for NMR. A trial NMR indexing scheme has been prepared by Jonker Business Machines, Inc. This scheme, organized for use of the Termatrex system, is now under review. ASTM has also prepared an updated GC Compilation DS25A (approximately 4,000 compounds) which was published in late 1967.

ASTM expects to add 15,000-20,000 IR spectra yearly to its index. Some 9,000 MS spectra will be added this June from European collections.

Sadtler Research Laboratories have 32,000 IR spectra, 4,000 NMR spectra, no GC, no MS, and 22,000 UV spectra on organic compounds. All Sadtler spectra are identified in the ASTM indices. It is anticipated that 2,000 signatures in IR and NMR will be added annually.

Dr. Zwolinski, director of the Chemical Thermodynamics Research Center (TRC), states that there are presently 13,000 complete spectra available at TRC. As of December 31, 1967, this collection consisted of 3,950 IR, 1,246 UV, 2,651 MS, and 1,260 NMR spectra. All TRC spectra are identified and indexed through ASTM. However, TRC expects to have its own index available in the Summer of 1968.

The Coblentz Society, which also has its spectra identified and indexed through ASTM, has 5,000 complete IR spectra available. These are sold by Sadtler Research Laboratories.

A tabulation of major spectra sources appears in Appendix B. Each table lists the major sources, number of spectra available, the form of the spectra, the cost, and the data sources. The tables are accompanied by a listing of other data compilations.

Cost figures for data acquisition, cannot be estimated adequately on the basis of the procedures used in this research inasmuch as the procedures were experimental, varied and unique for the method of compound selection used. Methods for data acquisition will be different if future research is conducted.

In general, data acquisition costs which will account for a large share of costs in an operating system fall into three categories:

1. Data obtained from published sources and collections

    Cost of collection

    Cost of transcription

2. Data measured from published sources

    Cost of measurement

    Cost of transcription

3. Data obtained from laboratory measurements

    Cost of compounds

    Cost of manual processing

    Cost of data reduction and transcription, or

    Cost of automatic processing.

The data placed in the file were obtained from a combination of the three categories except that automatic processing of laboratory measurements was not employed in this phase of the study.

H. Cataloguing Data and Data Master Card Design

For every compound selected for inclusion into the data file there is a master data sheet on file for that compound. A search mechanism presently has as its purpose the selection of one to, at most, a few of such master sheets which represent the compound(s) which most closely match the input data. The data contained on the 8 1/2 x 11 form include the indexed data for the analytical techniques involved plus the original sources of data.

A sample of the master sheet used in the project is found in Appendix C. A sheet was prepared for each compound used in the file. The information on the sheet includes:

Compound name - The compound will be listed by IUPAC name, and trivial or trade names are included for cross reference

CAS registry number - The CAS registry number, obtained from the SOCMA Handbook also aids in identifying the compound

GC data - The gas chromatographic data is listed for Carbowax^{R} 20M and SE-30 columns. Both the relative specific retention volume and the Kovats retention index appear

MS data - The mass-to-charge ratio is listed above the relative intensities in decreasing order

IR data - The three most intense infrared peaks (excluding 3.4μ) are listed in order of decreasing peak transmittance

NMR data - The NMR is listed as it appears in the Sadtler Specfinder Index. The strongest peak and solvent are indicated

UV data - The most intense band is listed to the nearest millimicron (mμ) along with the solvent used.

Structure, empirical formula, melting point, boiling point, physical state and the original data source are also indicated on the master data sheet.

IV. TESTING SIGNATURE CONCEPT

A. Experimental Design

1. Objectives

A search experiment was designed to determine the feasibility of identifying chemical compounds by matching their signatures with those of known compounds. The experiment tested the hypothesis that an unknown compound can be identified by sequentially searching a data file consisting of a minimum number of significant data points in each designated signature and obtaining the logical intersection of the data matches. By careful selection of the data points, the tolerances, and the matching criteria, it was anticipated that the intersection of the matches would contain a small set of candidate compounds, one of which would be the desired compound.

The experiment was also designed to provide data for establishing search procedures, formulating match criteria, and investigating the discrimination power of the five signatures in the data base. These auxiliary objectives of the experiment are listed below.

(1) To investigate the following match criteria in the search of each signature:

(a) Data parameters

Number of peaks
Magnitude of peaks
Ranking of peaks

(b) Data tolerances

(c) Conditions of data acquisition

Solvents (when applicable)
Standard measurement conditions

(2) To determine the discrimination capabilities of each signature for individual compounds, chemical groups, and all the compounds together

(3) To measure the selectivity of sequential search by:

(a) Combining searches

In 10 combinations taken two at a time
In 10 combinations taken three at a time
In 5 combinations taken four at a time

(b) Determining the minimum number of signatures required to achieve identification of an unknown

(4) To develop a measure of search effectiveness that incorporates:

(a) Rapid screening

(b) The least number of search operations

(c) The maximum selectivity

(5) To investigate a weighting system for evaluating and ranking compounds.

2. Test Inputs

The experiments consisted of a set of searches with three categories of inputs:

Input 1: Compounds having all five signatures available in the data base were used as "unknowns."

Input 2: Compounds having different combinations of signatures were used as unknowns; these data were not included in the data base.

Input 3: Experimental data derived from in-house laboratory measurements were matched against stored data.

Input 1 data were used to determine the ability of the search to make positive identifications, given that the compounds are stored in the data base. Input 2 data were used to test the ability of the search to discriminate against closely related compounds which had data stored in the data base. Input 3 data were used to test the selectivity of search under non-controlled input conditions that may be encountered in practice.

Input 1 and 2 compounds served essentially the same discrimination-testing function. By eliminating an Input 1 compound from a candidate list, an Input 2 condition was established. The remaining candidates are those that would be obtained if the input data were not available in the data base.

The Input 3 compounds were used to test the tolerance limits that would be required in an operational system in which data describing unknown compounds to be identified would be derived from laboratory measurements made externally or at IITRI.

3. Tests

All "unknown" compounds from inputs 1, 2, or 3 were processed in the same manner. Input data were inserted in designated locations on the data sheet shown in Appendix D. The data base, stored in an inverted-term-Termatrex system (described in Report No. IITRI-C6104-2) was searched, one signature at a time, and the numbers of compounds whose stored data matched the input data accoruing to the match criterion were obtained.

To obtain data on match criteria and discrimination capabilities, the search in each signature was carried out independently without reference to previous searches. To obtain data on selectivity and search effectiveness, a cumulative search was then made by selecting the numbers of compounds in the intersection of all previous searches with the current search; these data were tabulated on the search record shown in Appendix E.

Selected candidate compounds that were the intersection of the matches in all available signatures were entered on the candidate list shown in Appendix F.

Two series of tests were made. The first test was based upon the formulation of initial match criteria. The second tests were developed to explore the effects of modified match criteria on search precision.

a. Initial Match Criteria

The initial match criteria for the seven steps of the sequential search were:

1. State: solid/liquid/gas
2. Elements: combinations of significant elements as coded in column 32 of the ASTM punched cards for infrared spectra
3. Infrared spectroscopy: match on any two of the first three highest peaks of spectrum. The 3.4μ peak (C-H band) was omitted from the input data
4. Nuclear magnetic resonance: match on the highest peak; solvents were coded but where disregarded

5. Mass spectroscopy: match on any two of the four highest peaks; matches were made on the mass numbers of ranked peaks but not on the relative amplitudes

6. Gas chromatography: match on relative retention time of either one of two columns depending upon availability of data

7. Ultraviolet spectroscopy: match on the most intense band.

b. Modified Match Criteria

Following an evaluation of the precision of the search results obtained with the initial match criteria, several modifications of the criteria were made to assess their effect in improving precision.. These modifications included the following criteria used singly and in combinations:

1. State: no change

2. Elements: no change

3. Infrared spectroscopy: no change

4. Nuclear magnetic resonance: match on two out of the two highest peaks

5. Mass spectroscopy: a) same as initial criteria except the 1/4 and 4/1 input data/stored data peak matches were deleted; b) match on three out of four peaks

6. Gas chromatography: match on relative retention time of both columns A and B

7. Ultraviolet spectroscopy: no change.

4. Candidate Ratings

Ratings were used to evaluate the probability that an unknown compound was a compound on the candidate list in accordance with weights assigned to the signatures and with the closeness of the match.

The maximum ratings assigned to the five types of signatures and to the state and the element searches are:

MS	40	
IR	36	
NMR	24	
GC	18	(9 points per column)
UV	6	
Elements	4	
State	2	
	130	

To test the tolerance limits in each signature, the rating scale considered exact matches as contrasted to matches within the tolerance range. A skip was rated 0, and a default was rated 1.

The rating table for a candidate list is shown in Appendix G.

B. Search Procedures

1. Search Logic

A composite chemical signature of a compound consists of a representation of the significant parameters of its component signatures. The composite signature for a compound can be represented as

$$C_i = f\ (M_i,\ R_i,\ N_i,\ G_i,\ U_i,),\ i = (1,\ 2,\ \ldots,\ n)$$

where C = compound

M = mass spectrometry signature

R = infrared spectroscopy signature

N = nuclear magnetic resonance signature

G = gas chromatography signature

U = ultraviolet spectroscopy signature.

Each signature, in turn, can be represented by the parameters selected for the experiment:

$$M_i = f\ (M_{1_i} : a_{1_i},\ M_{2_i} : \frac{a_{2_i}}{a_{1_i}},\ M_{3_i} : \frac{a_{3_i}}{a_{1_i}},\ M_4 : \frac{a_{4_i}}{a_{1_i}}\ ),$$

where M_1 = maximum peak

a_1 = amplitude of maximum peak = unity

$M_j : \frac{a_j}{A_1}$ = rank order of 2nd, 3rd, and 4th peaks

$$R_i = f\ (b_{1_i},\ b_{2_i},\ b_{3_i}),$$

where $b_{1,2,3}$ = three highest absorption bands with transmittance equal to or greater than 10% of the most intense band.

$$N_i = f\ (h_{max_i}),$$

where $h_{max_i} = \emptyset_{max} = \frac{\Delta H}{H} \times 10^6$ = chemical shift.

$$G_i = f\ (g_{A_i} + g_{B_i}),$$

where g = relative retention time (dimensionless), and A and B are columns described in Table 6.

$$U_i = f\,(\mu_{max_i}),$$

where μ_{max} = the most intense peak.

The objective of a search aimed at identifying an unknown compound is to yield a minimal set of compounds that satisfies the Boolean expression

$$C_j = \left\{ \left\{C_{m_i}\right\} \cap \left\{C_{R_i}\right\} \cap \left\{C_{N_i}\right\} \cap \left\{C_{G_i}\right\} \cap \left\{C_{U_i}\right\} \right\},$$

where $\cap$ = logical AND, and the i and j subscripts designate the stored known and input unknown compounds respectively.

State and elements searches, S_i and E_i, preceded the composite signature search in the experiment.

2. Computer Search Design

A generalized search procedure capable of being implemented on a computer was developed to provide the Boolean search and to serve the objectives of the experiment as described previously. A data availability code was assigned to every unknown compound whose identification was sought and to every stored (known) compound. The code indicates the presence or absence of data upon which a match can be attempted. The absence of data for a parameter in an unknown compound, for example, precludes a search through the relevant parameters of the known compounds.

In a computer search, each step would start with a match on the data availability codes of an unknown compound and the candidate compounds. A search routine CSEAR would be

Table 11

SEARCH TABLE (CSTABL)

CSEAR 1	STATE
PARAM 1	Solid/Liquid/Gas
CSEAR 2	ELEMENTS
PARAM 2	As specified by unknown
CSEAR 3	MS
PARAM 3	M_1 (mass with highest amplitude) M_2 (mass with 2nd amplitude) M_3 (mass with 3rd amplitude) M_4 (mass with 4th amplitude)
CSEAR 4	IR
PARAM 4	μ_{max} μ_2 (second peak) μ_3 (third peak)
TOLER 4	±0.1μ
CSEAR 5	NMR
PARAM 5	PPM_{max}
TOLER 5	±C.1 PPM
(PARAM)	Solvent (not used at present)
CSEAR 6	GC
PARAM 6	Relative Retention Time, Column A Relative Retention Time, Column B
TOLER 6	(See Section IVD1d)
CSEAR 7	UV
PARAM 7	Strongest Peak
TOLER 7	±2mμ

called in from a search table CSTABL and the appropriate parameter(s) PARAM and tolerance(s) TOLER would be initialized in accordance with Table 11.

The set of compounds that matched in each stage would be placed in a candidate list together with ratings as per the rating table. Only the signatures of compounds on the candidate lists would be searched in subsequent stages. Surviving candidates at the end of all stages would be printed out in order of their accumulated ratings.

Upon completion of the search, the results would fall into one of four categories:

1. Zero selection - no compounds
2. Plural selection - more than one compound
3. Conditional selection - one compound isolated with less than "perfect" rating; identification cannot rule out other possibilities
4. Unique selection - positive identification of single compound having a "perfect" rating.

3. Termatrex Experiment

The search experiment described above was implemented by using an inverted file with optical coincidence (Termatrex system) to demonstrate the feasibility of identification by sequential signature matching. The logic and search procedures, which were designed for a computer, were carried out using drilled cards and appropriate tally forms.

C. Search Measures

1. Availability and Default Ratios

The feasibility of identifying an unknown compound by sequentially searching a data file consisting of a minimum number of significant data points in designated signatures depends on good discrimination in each signature search and on fine selectivity in screening the data base.

As used in this report, discrimination is the capability to separate sets of compounds on the basis of well-defined characteristics of a signature. Selectivity is the screening capability of a sequential search. Precision is the accuracy of identification.

Because data are not uniformly available for all compounds in all signatures, data availability must be considered as a factor in formulating measures for discrimination, selectivity, and precision. Accordingly, the following search actions are defined.

A match occurs when data stored for a given signature of a known compound are equal to or in the tolerance range of input data. A default occurs when input data for a given signature of an unknown are available and a stored compound does not have data for that signature. The nonavailability of stored data precludes obtaining a match but does not rule out the possibility that the compound with no stored data is a candidate. A skip occurs when input data for a given signature of an unknown compound are not available. In this case, a search of the stored data in the sig·ature file is not possible.

The availability ratio of a signature is the ratio of the number of compounds having data available to the total number of compounds in the data base. The default ratio of a signature is the ratio of the number of compounds not having data to the total number of compounds in the data base. These ratios can be expressed mathematically:

$\alpha = \frac{S}{C}$, $\beta = \frac{D}{C} = 1 - \alpha$, where

C is the number of compounds in the data base

S is the number of compounds with available data in a given signature

D is the number of compounds with no data in a given signature

α is the availability ratio

β is the default ratio.

Table 13 lists the availability and the default ratios of each signature in the 500-compound data base.

2. Discrimination

The discrimination factor of a signature is a function of the number of matches and the default ratio inasmuch as the number of possible successful matches is reduced by the non-availability of data. The defaults can be likened to a sea of uncertainty upon which successful matches float. The discrimination factor, δ, is expressed as:

$$\delta = M \times \beta$$

where M is the number of matches obtained in a signature search.

Table 12

SIGNATURE AVAILABILITY AND DEFAULT RATIOS

SIGNATURE	NUMBER OF COMPOUNDS (C)	NUMBER OF SIGNATURES AVAILABLE (S)	NUMBER OF DEFAULTS (D)	AVAILABILITY RATIO (α)	DEFAULT RATIO (β)
MS	500	478	22	0.956	0.044
IR	500	470	30	0.940	0.060
NMR	500	357	143	0.714	0.286
GC	500	308	192	0.616	0.384
UV	500	233	267	0.466	0.534
TOTALS	2500	1846	654	0.738	0.262

3. Selectivity

The selectivity of a sequential search measures the sieving capability of the search and deals with the intersections of the signature matches. Selectivity is defined as the ratio of the number of surviving candidates at the end of a stage to the total number of compounds in the data base. Inasmuch as compounds with defaults can be present in the candidate list, an adjustment for defaults is not necessary. Expressed mathematically,

$$\text{Selectivity}_k = \eta_k = \frac{M_k}{C} \; (k = 1, 2, \ldots n)$$

where

subscript k designates a search stage

n is the maximum number of stages

M_k is the number of matches at the end of the k-th stage.

The sequential search consisted of seven stages: state, elements, and the five signatures (MS, IR, NMR, GC, and UV). The final search results are independent of the order of search inasmuch as intersection is commutative. However, the search effectiveness, that is, the rapidity of screening and the total number of compounds that must be searched, is affected.

4. Precision

The most critical of the measures that describe the results of a sequential signature search is precision. Precision is the accuracy of identification. In the case of Input 1 and Input 3 compounds, perfect precision would result in the identification of the input compound and only that compound. In the case of Input 2 compounds, perfect precision would result in the rejection of all candidate compounds. The elimination of the "true" compound (or the presence of more than one candidate) in the former case and the "false" identification of one or more compounds in the latter case would decrease the precision.

Retention of a compound on the candidate list is an arbitrary decision that depends on the minimum number of signatures required for screening and on a threshold in the rating scale. By use of these two quantities, the precision can be increased or decreased. Similarly, enhancing discrimination by means of tightening match tolerances and criteria also increases precision.

Precision is measured by the number of candidates selected in identifying a compound.

D. Test Results with Initial Match Criteria

The results reported below are based upon the initial match criteria listed in Section IVA3a.

1. Discrimination

The range of matches, the average number of matches, the average number of candidates and the discrimination factor for each signature for the 100 Input 1 compounds are shown in Table 13. The match range column indicates the number of positive matches (less defaults) that occurred in each signature with "low" referring to the minimum number of matches obtained

Table 13

SIGNATURE DISCRIMINATION FOR 100 KNOWN COMPOUNDS

SIGNATURE	MATCH RANGE (LESS DEFAULTS)		AVERAGE NUMBER OF MATCHES	NUMBER OF DEFAULTS	AVERAGE NUMBER OF CANDIDATES	DISCRIMINATION FACTOR
	LOW	HIGH				
MS	1	153	49.83	22	71.83	2.19
IR	1	84	30.39	30	60.39	1.82
NMR	1	59	31.22	143	174.22	8.93
GC	1	39	11.74	192	203.74	4.51
UV	1	85	36.77	267	303.77	19.64

in the 100 searches and "high" referring to the maximum number.

There was at least one compound among the 100 compounds in each signature that was the sole selected candidate, exclusive of defaults. The largest range of matches occurred in the MS signature, the smallest range in GC. The distribution of signature matches is shown in Figure 2. Both the mean and the medians are listed together with the range of matches including defaults.

The effect of the defaults is clearly seen as the distributions of the signatures are shifted on the abscissa. The peak at 351-360 matches in the UV distribution is due to the 34 compounds that are transparent to UV.

a. MS Discrimination

The MS search ranked second in discrimination among the five signatures. An average number of 49.83 matches per search from an available data base of 478 compounds was found. With the 22 defaults, a total of 71.83 candidates were found, on the average, for each of the 100 known input compounds. The discrimination factor was 2.19.

The criterion of match was equality between at least two of four peaks. Because numerous permutations were found in the ranks of the four major peaks among different data sources for a single compound, any two matches from the 16 combinations of input and stored data peaks were accepted. The rating table (Appendix G) takes permuted matches into account, a match of input peak 1 against stored peak 1 as contrasted with a match

SIG.	MEAN	MEDIAN	RANGE
MS	71.83	48.00	23 - 175
IR	60.39	56.33	31 - 114
NMR	174.22	176.00	144 - 202
GC	203.74	200.40	193 - 231
UV	303.77	280.77	268 - 352

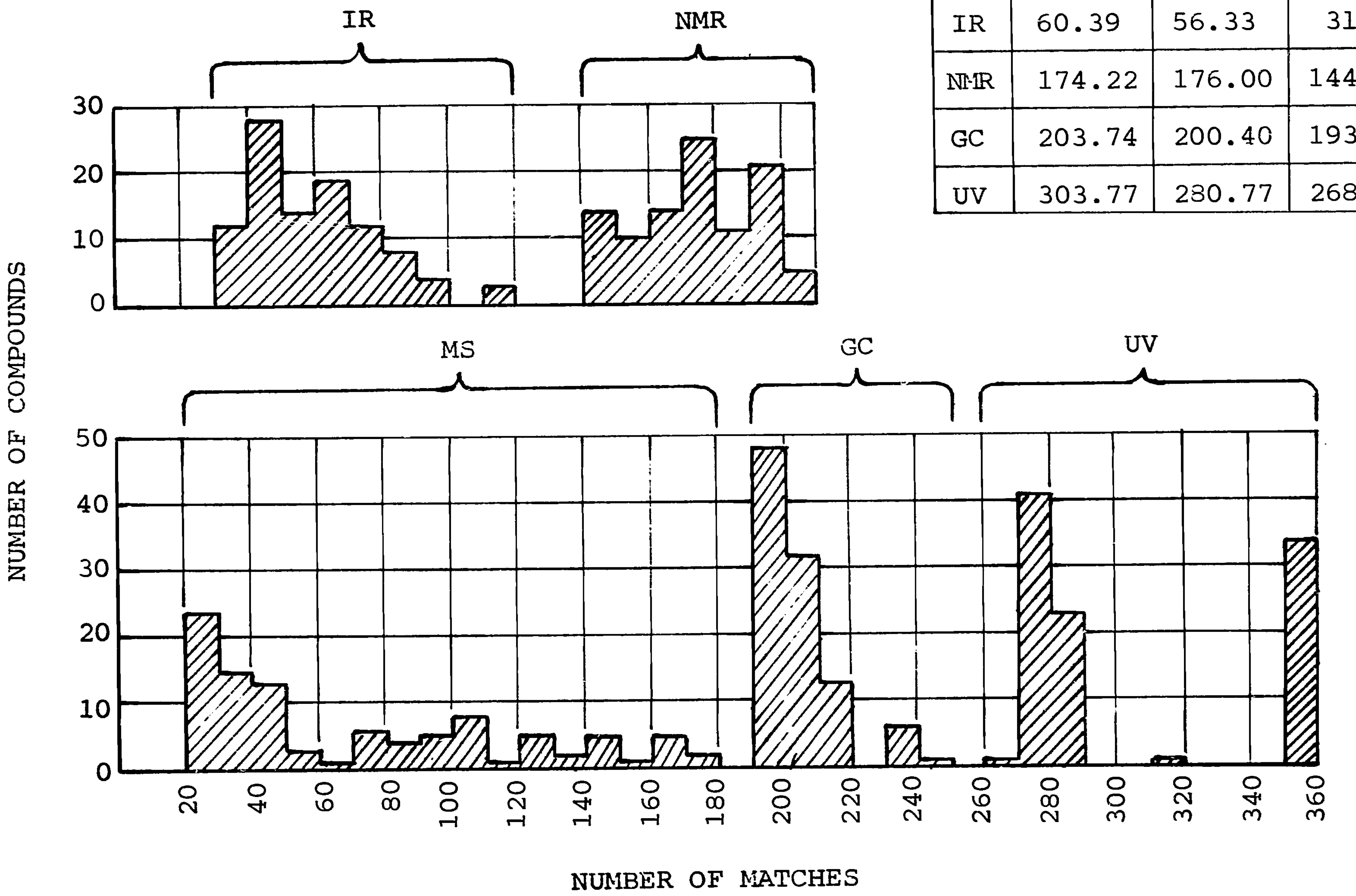


Figure 2

DISTRIBUTION OF SIGNATURE MATCHES

of input peak 1 against stored peak 4, for example.

A summary of the MS matches obtained on the permuted peaks of a sample of 25 known input compounds is presented in Table 14. The range indicates the low (minimum) and high (maximum) number of candidates obtained in the searches for the 25 compounds. The results of the 1 on 1, 3 on 3, and 4 on 4 matches were as expected, namely that the average number of matches was greatest in these permutations. Only in the 2 on 2 match was there a reversal of order with the 2 on 1 match. The 1 on 1 matches provided the largest average number of matches, followed closely by the 4 on 4 matches.

Since there is no tolerance on an MS peak, discrimination of the search can be increased (i.e., δ can be made smaller) by (1) increasing the number of peaks for which matches are required, e.g., three out of four; (2) increasing the number of peaks searched, e.g., from four to five or six, and raising the match criterion accordingly; (3) introducing relative amplitudes, and (4) eliminating some of the 16 combinations of peak searches now employed, e.g., input peak 1 must match on any one stored data peaks 1, 2, or 3, but not 4.

To evaluate the effect of changing the number of peaks required for MS matching, Table 15 was prepared. The table lists the average number of matches obtained with the varying criteria of 1 out of 4 to 4 out of 4 peaks, exclusive of defaults. The term "exclusive" in Table 15 refers to matches on one peak only, two peaks only, three peaks only, and four peaks only. The

Table 14

MS MATCHES ON PEAK PERMUTATIONS
(25 Input Compounds)

PEAKS		RANGE		AVERAGE NUMBER OF MATCHES
INPUT	STORED	LOW	HIGH	
1	1	2	78	34.76
	2	2	37	19.92
	3	0	68	16.24
	4	1	69	14.12
2	1	0	78	19.20
	2	2	37	15.52
	3	2	68	13.00
	4	0	69	12.08
3	1	0	78	12.20
	2	1	37	13.08
	3	2	68	20.00
	4	1	69	15.80
4	1	0	78	16.40
	2	0	37	16.92
	3	0	68	24.88
	4	1	69	29.16

term "inclusive" refers to every instance of one, two, three, or four matches; for example the number of matches on one peak includes the number of matches on two, three, and four peaks.

Table 15

NUMBER OF CANDIDATE COMPOUNDS OBTAINED
WITH DIFFERENT MS MATCH CRITERIA
(100 Known Input Compounds)

NUMBER OF PEAKS MATCHED	MATCH RANGE (INCLUSIVE)		AVERAGE NUMBER OF MATCHES	
	LOW	HIGH	EXCLUSIVE	INCLUSIVE
1	3	313	103.54	153.37
2[a]	1	153	40.23	49.83
3	2	38	7.85	9.60
4	1	6	1.75	1.75

[a]Criterion used for the experiment: matches on 2 or more peaks.

The average of 49.83 matches (exclusive of defaults) obtained with the criterion of matching on 2 out of 4 peaks would be reduced to 9.60 if 3 out of 4 peaks were the match criterion. A 4 out of 4 criterion would further reduce the average number of candidates per search to 1.75.

b. IR Discrimination

The IR search was the most discriminating of the five signatures. An average number of 30.39 matches per search was found from an available data base of 470 compounds. With the 30 IR defaults, a total of 60.39 candidates were found, on the average, for each of the 100 known input compounds. The discrimination factor was 1.82.

The peak at 3.4μ, characteristic of hydrocarbon compounds, had been eliminated from all input and stored data. The 5.8μ peak, however, was included in the data and was a valid search value. A match was registered if the stored data were equal to or within the tolerance range of the input data. Permutations of ranked peaks were searched as in the MS search, i.e., IR input peak 1 was searched against stored peaks 1, 2, and 3, etc.

A listing of the IR matches obtained on the permuted peaks of the same 25 compound sample measured in MS appears in Table 16. The average number of matches obtained on the 1 on 1 and 2 on 2 peaks were the highest in their respective categories. The number of 3 on 2 peak matches were slightly greater than the 3 on 3 matches.

Table 16

IR MATCHES ON PEAK PERMUTATIONS
(25 Input Compounds)

PEAKS		RANGE		AVERAGE NUMBER OF MATCHES
INPUT	STORED	LOW	HIGH	
1	1	10	71	41.72
	2	13	39	22.12
	3	11	24	18.44
2	1	1	71	24.36
	2	2	57	27.28
	3	2	67	26.20
3	1	2	64	26.40
	2	4	59	34.56
	3	9	67	33.92

The number of candidate compounds obtained with 1-, 2-, and 3-peak matches is shown in Table 17. The rarge and exclusive-inclusive definitions are the same as those accompanying Table 15.

Table 17

NUMBER OF DIFFERENT COMPOUNDS OBTAINED
WITH DIFFERENT IR MATCH CRITERIA
(100 Known Input Compounds)

NUMBER OF PEAKS MATCHED	MATCH RANGE (INCLUSIVE)		AVERAGE NUMBER OF MATCHES	
	LOW	HIGH	EXCLUSIVE	INCLUSIVE
1	38	282	161.28	191.67
2[a]	2	84	27.05	30.39
3	1	12	3.34	3.34

[a] Criterion used for the experiment: matches on 2 or more peaks.

The 2-peak match criterion resulted in an average number of 30.39 matches per search (not including defaults), whereas a single peak match would have resulted in an average of 191.67 matches. Greater discrimination would be achieved by a 3-peak match criterion with an average of 3.34 matches per search.

c. NMR Discrimination

The NMR signature, with a discrimination factor of 8.93, ranked fourth. An average number of 31.22 matches per search from an available data base of 357 compounds was found. Together with the 143 NMR defaults, an average of 174.22 matches per search was found. The criterion of search was a match on the strongest peak within the tolerance limit of ±0.1ppm. When two peaks were of the same magnitude, each peak was assigned a different compound number and the data for all signatures were entered for the two numbers.

Crotonaldehyde, for example, had nearly equal NMR peaks, after rounding, of 2.1 ppm and 2.0 ppm. The former value was assigned IITRI number 337; the latter was assigned 961. In the search for crotonaldehyde with an NMR peak of 2.0 ppm, compound 337 was selected, with a rating of 130 points. It is interesting that since compound 961 was within the tolerance range of 2.1 ppm, compound 961 was also selected, although its rating was 126, because of the tolerance match.

The discrimination of the NMR search can be increased by changing the criterion of match to two or more of the most intense peaks and by matching solvent data. Matching solvent data was not employed in the search, and it would not improve the discrimination to the degree that adding additional peaks would.

d. GC Discrimination

The criterion of the GC search was a match with the input data of one of two columns within the tolerance limit. Because of the wide range of relative retention times, the tolerance was made proportional to the time in the following manner:

0.01-0.99:	±0.01
1.0-9.9:	±0.1
10-99:	±1.

The GC signature provided the lowest average number of matches per search, 11.74, but the nonavailability of data on 192 compounds reduced the discrimination factor to 4.51; GC thus ranked third in discrimination. Table 18 summarizes the GC signature search.

The discrimination of the GC signature can be improved by using Column A alone or Column B alone instead of the inclusive OR criterion used in the experiment. Even greater discrimination would result from using the intersection of the two columns, as shown in Table 18.

Table 18

NUMBER OF CANDIDATE COMPOUNDS OBTAINED
WITH DIFFERENT GC MATCH CRITERIA
(100 Known Input Compounds)

MATCH CRITERION	INPUT DATA	NUMBER OF MATCHES RANGE		AVERAGE NUMBER OF MATCHES
		LOW	HIGH	
Column A	96	1	26	7.62
Column B	83	1	20	6.32
A and B	79	1	7	1.44
A or B[a]	100	1	39	11.75

[a] Criterion used for the experiment.

The GC signature provided the lowest average number of matches per search, 11.74, but the nonavailability of data o 192 compounds reduced the discrimination factor to 4.51; GC thus ranked third in discrimination. Table 18 summarizes th GC signature search.

The discrimination of the GC signature can be improved using Column A alone or Column B alone instead of the inclus OR criterion used in the experiment. Even greater discrimin tion would result from using the intersection of the two col as shown in Table 18.

Table 18

NUMBER OF CANDIDATE COMPOUNDS OBTAINED
WITH DIFFERENT GC MATCH CRITERIA
(100 Known Input Compounds)

MATCH CRITERION	INPUT DATA	NUMBER OF MATCHES RANGE		AVERAGE NUMBER OF MATCHES
		LOW	HIGH	
Column A	96	1	26	7.62
Column B	83	1	20	6.32
A and B	79	1	7	1.44
A or B[a]	100	1	39	11.75

[a] Criterion used for the experiment.

e. UV Discrimination

The UV signature was the least discriminating of the five signatures. Its average output of 36.77 candidates per search added to the 267 compounds for which UV data were not available resulted in an average of 303.77 candidates per search. The discrimination factor was 19.64.

2. Selectivity

a. State and Element Searches

The distribution of state and elements data for the 500-compound data base and the 100 test compounds is shown in Table 19. The liquid state category eliminated 15 of 100 compounds; the solid state eliminated 86. The gas category provided unique identification. The selectivity of the two stages is given in Table 20.

Table 19

STATE AND ELEMENTS DATA AVAILABILITY

SEARCH	PARAMETER	DATA BASE 500 COMPOUNDS	INPUT 1 100 COMPOUNDS
STATE	Solid	95	14
	Liquid	387	85
	Gas	18	1
ELEMENT*	Bromine	25	1
	Chlorine	43	7
	Fluorine	3	
	Iodine	9	
	Lead	1	
	Nitrogen	80	9
	Oxygen	300	63
	Sulfur	16	6
	No Hydrogen	7	1
	No designation	70	23

*Data base and Input 1 columns total more than 500 and 100 because more than one element was listed for many compounds.

Table 20

SELECTIVITY OF STATE AND ELEMENTS SEARCHES
(100 Known Input Compounds)

STAGE	AVERAGE NUMBER OF SURVIVING CANDIDATES	SELECTIVITY
State	342.43	0.685
Elements	195.11	0.390

The effectiveness of the preselection is indicated clearly by the 61 percent reduction in candidate compounds that was obtained at the end of the elements search.

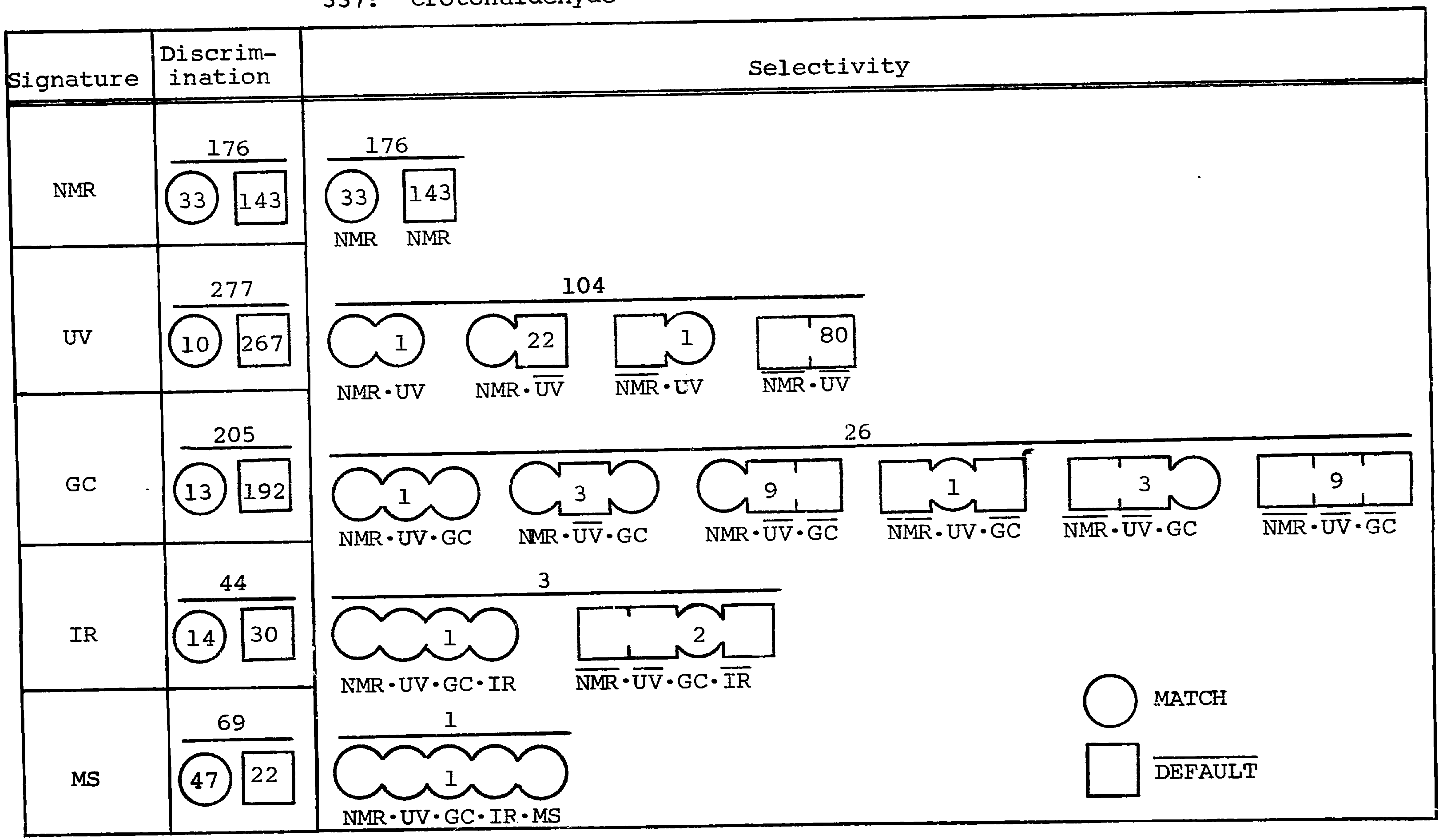
It was anticipated that the state and elements searches would serve only as a preselection tool to reduce the number of compounds to be searched in the signatures. Although this objective was achieved, it was found that the state and elements searches also aided in identification of a number of compounds.

b. Signature Searches

Two sequences were tested in preliminary experiments to test search effectiveness. The first sequence was ordered on the basis of the number of data points to be searched in accordance with the match criteria. The order chosen was NMR, UV, GC, IR, and MS. The selectivity for compound 337, crotonaldehyde, is shown in Figure 3. The state and elements searches were omitted from Figure 3. The second sequence was ordered on the basis of the default ratio, the signature with the fewest defaults being searched first. The search order chosen was MS, IR, NMR, GC, and UV. The selectivity for crotonaldehyde via this sequence is shown in Figure 4.

The presence of 143 defaults in the NMR data and 267 in the UV data accounted for 80 candidate compounds that were carried into the third stage of search, as indicated in Figure 3. The 22 and the 30 defaults in the MS and the IR data, respectively, gave rise to no double-default candidates, as

337: Crotonaldehyde



Signature
Discrimination
Selectivity
NMR
176
33
143
176
33
143
NMR
NMR
UV
277
10
267
104
1
22
1
80
$NMR \cdot UV$
$NMR \cdot \overline{UV}$
$\overline{NMR} \cdot UV$
$\overline{NMR} \cdot \overline{UV}$
GC
205
13
192
26
1
3
9
1
3
9
$NMR \cdot UV \cdot GC$
$NMR \cdot \overline{UV} \cdot GC$
$NMR \cdot \overline{UV} \cdot \overline{GC}$
$\overline{NMR} \cdot UV \cdot \overline{GC}$
$\overline{NMR} \cdot \overline{UV} \cdot GC$
$\overline{NMR} \cdot \overline{UV} \cdot \overline{GC}$
IR
44
14
30
3
1
2
$NMR \cdot UV \cdot GC \cdot IR$
$\overline{NMR} \cdot \overline{UV} \cdot GC \cdot \overline{IR}$
MS
69
47
22
1
1
$NMR \cdot UV \cdot GC \cdot IR \cdot MS$
MATCH
$\overline{DEFAULT}$

Figure 3

SELECTIVITY WITH SEQUENCE ORDERED BY MATCH CRITERIA

337: Crotonaldehyde

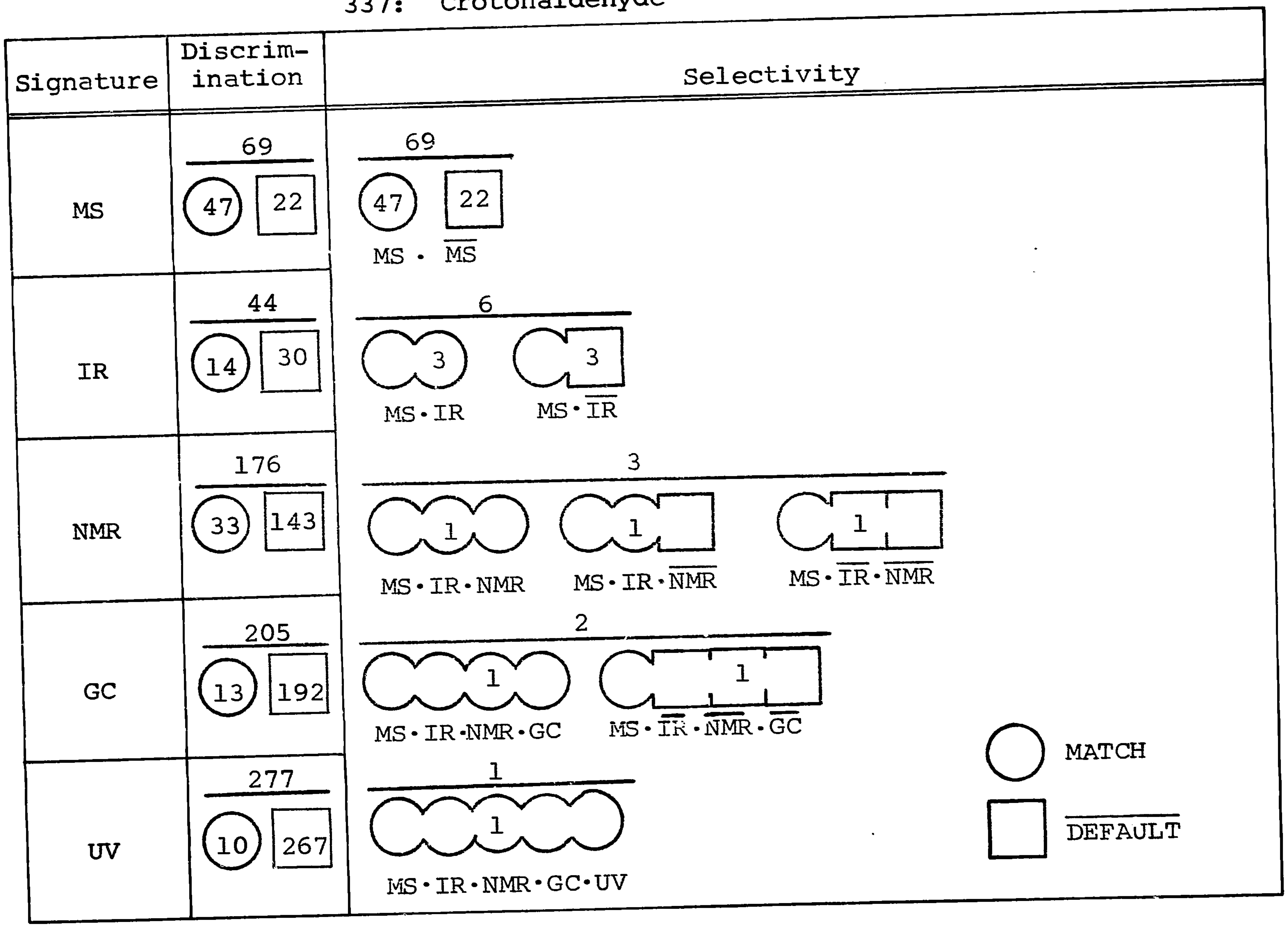

Signature
Discrim-
ination
Selectivity
MS
69
47
22
69
47
22
MS · $\overline{\text{MS}}$
IR
44
14
30
6
3
MS · IR
3
MS · $\overline{\text{IR}}$
NMR
176
33
143
3
1
MS · IR · NMR
1
MS · IR · $\overline{\text{NMR}}$
1
MS · $\overline{\text{IR}}$ · $\overline{\text{NMR}}$
GC
205
13
192
2
1
MS · IR · NMR · GC
1
MS · $\overline{\text{IR}}$ · $\overline{\text{NMR}}$ · $\overline{\text{GC}}$
UV
277
10
267
1
1
MS · IR · NMR · GC · UV
MATCH
$\overline{\text{DEFAULT}}$

Figure 4

SELECTIVITY WITH SEQUENCE ORDERED BY DEFAULT RATIO indicated in Figure 4. The discrimination of the searches, of course, is also an important factor in establishing the selectivity. The selectivity of the two sequences is summarized in Table 21.

Table 21

SELECTIVITY WITH DIFFERENT SEARCH ORDERS (Signatures Only)

STAGE	MATCH CRITERIA	DEFAULT RATIO
1	0.354	0.104
2	0.201	0.012
3	0.052	0.006
4	0.006	0.004
5	0.002	0.002

Inasmuch as the selectivity of search of Table 21 was characteristic of numerous compounds, the search sequence based on default ratios was employed in the experimental tests which were made using a manual Termatrex system. The search order was: 1) state, 2) elements, 3) MS, 4) IR, 5) NMR, 6) GC, and 7) UV.

Selectivity curves of 10 compounds are shown in Figures 5 and 6. The pattern of selection varies with each compound. A unique identification was not achieved until the final stage of search in four of the illustrated compounds, whereas with compound 522: benzonitrile, a single candidate was selected at the end of the fourth stage, after the MS and IR searches. For three of the compounds in Figure 6, unique identification was

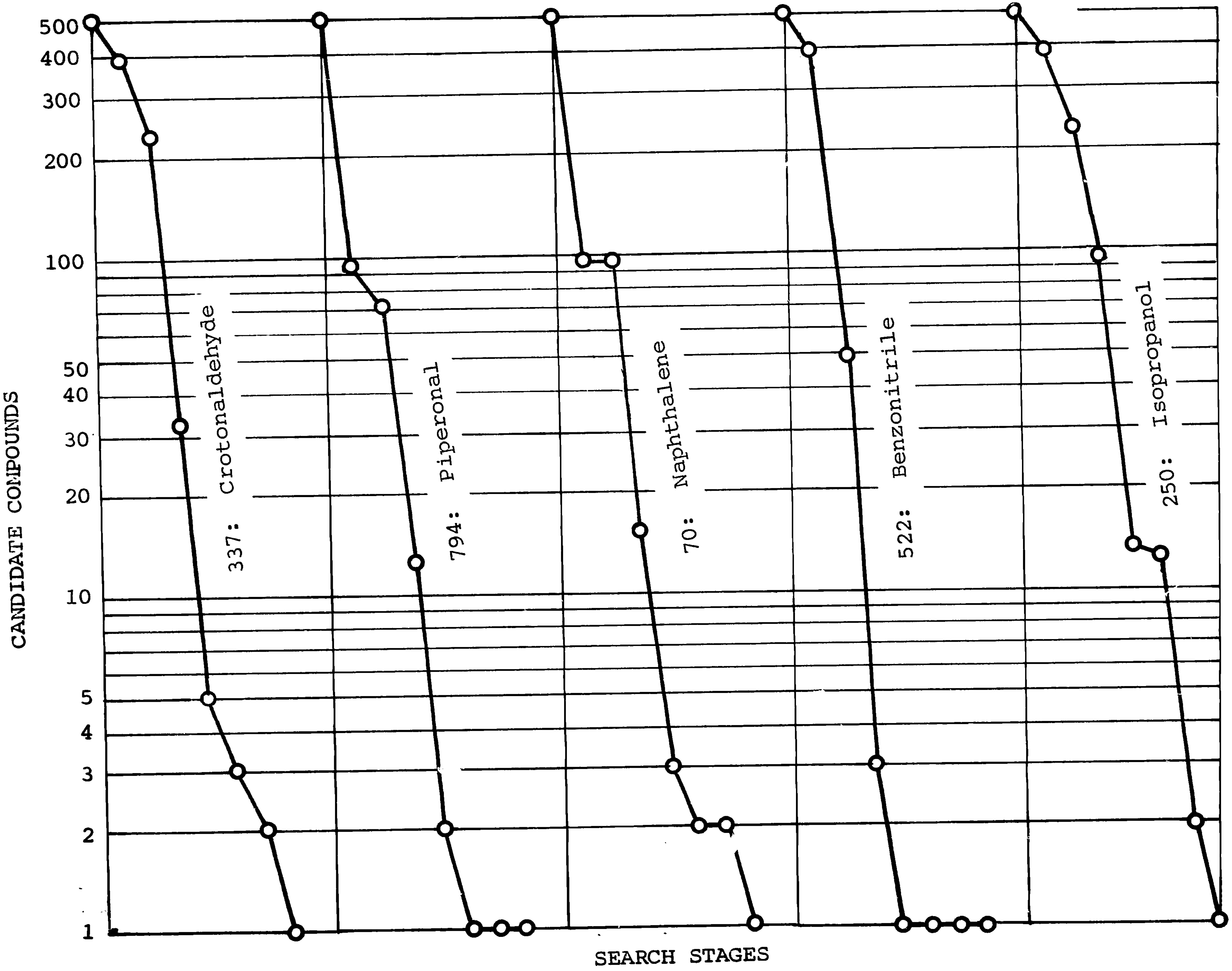


Figure 5

SELECTIVITY OF 7-STAGE SEARCHES (1)

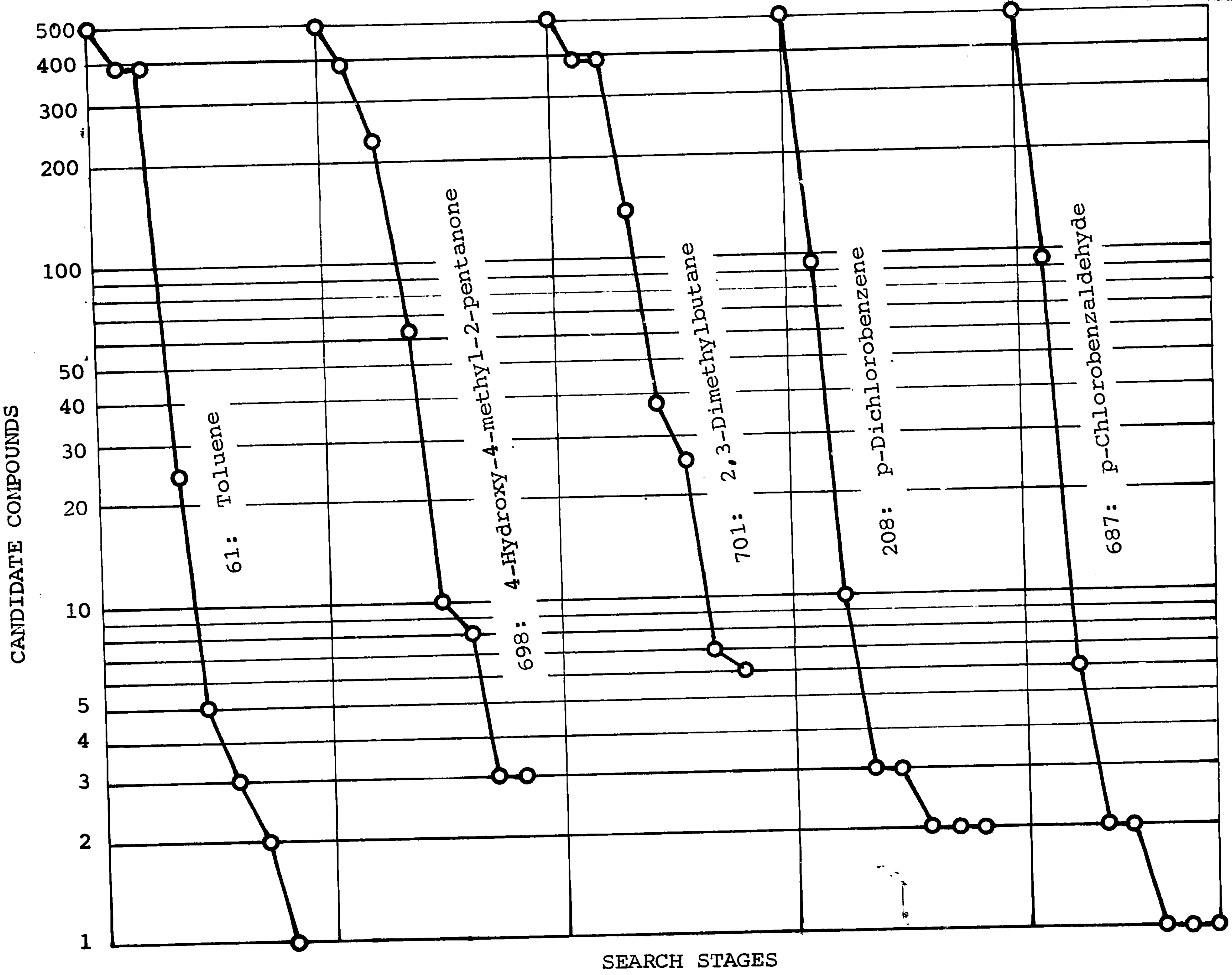


Figure 6

SELECTIVITY OF 7-STAGE SEARCHES (2)

not achieved. Compound 208: p-dichlorobenzene, for example, had 10 candidates at the end of the state and elements searches, 3 at the end of MS, 3 after IR, and 2 after NMR, GC, and UV.

A summary of selectivity for the 100 known input compounds in the 7-stage sequential search is listed in Table 22. At the end of the first signature search, MS, nearly 93 percent of the 500 compounds in the data base were eliminated. The IR search eliminated up to 98 percent of the 500 compounds. For the 100 test compounds. an average of 1.65 candidates survived following the final stage of search, UV.

The range column of Table 22 indicates the lowest and highest number of matches obtained in the respective stages for any of the 100 test compounds. These ranges differ from those of Table 13 where each signature was considered independent of the sequential search.

The two right hand columns of Table 22 list the stage at which unique identification was obtained. Six compounds were identified at the end of the first signature search, MS. Sixteen additional compounds for a cumulative total of 22 were identified following the second signature search, IR. At the conclusion of the 7-stage search, 68 of the 100 compounds were uniquely identified.

The selectivity of the searches varied with the functional groups as was expected. The selectivity of the five largest groups represented in the test compounds is listed in Table 23.

Table 22

SEARCH SELECTIVITY BY STAGES
(100 Known Input Compounds)

STAGE	AVERAGE NO. SURVIVING CANDIDATES	SELECTIVITY	CANDIDATE RANGE		NUMBER OF COMPOUNDS WITH ONE CANDIDATE	
			LOW	HIGH	NEW	CUM.
1. State	342.43	0.685	18	387	0	0
2. Elements	195.11	0.390	6	387	0	0
3. MS	37.72	0.075	1	143	6	6
4. IR	8.96	0.018	1	39	16	22
5. NMR	5.85	0.012	1	26	22	44
6. GC	1.85	0.0037	1	7	14	58
7. UV	1.65	0.0033	1	6	10	68

Table 23

SEARCH SELECTIVITY OF FUNCTIONAL GROUPS
(Known Input Compounds)

STAGE	AVERAGE NUMBER SURVIVING CANDIDATES				
	ALCOHOLS	ALDEHYDES	HYDRO-CARBONS	KETONES	POLY-FUNCTIONAL
1. State	340.47	350.50	345.56	357.80	241.00
2. Elements	210.47	207.37	345.56	227.00	113.60
3. MS	56.05	33.87	59.96	37.20	14.10
4. IR	13.73	5.62	14.00	10.80	2.70
5. NMR	9.68	4.12	8.91	8.37	1.80
6. GC	1.68	1.37	3.21	1.40	1.40
7. UV	1.63	1.12	2.65	1.40	1.20
No. of Test Compounds	19	8	23	10	10

The aldehydes had an average of 1.12 candidates per test compound at the end of the 7-stage search followed closely by the polyfunctionals with an average of 1.20 candidates. The hydrocarbons had poorer selectivity than all other groups except in the state and NMR searches.

3. Precision

a. Input 1 - 100 Known Compounds

As noted in Table 24, 68 out of the 100 test compounds were uniquely identified in the 7-stage search. Sixty-five extraneous candidates were selected for the remaining 32 compounds. The number of extraneous candidates ranged from one to five, with an average of 2.03 for the multicandidate compounds (1.65 average for the 100 test compounds).

Only two of the 65 extraneous compounds had data on all signatures available as shown in Table 24. Eleven had one signature missing, 45 had two signatures missing, and 7 had three signatures missing. The ratings for the 65 extraneous candidates ranged from 34 to 95 compared to maximum possible ratings of 117 to 130 as shown in Table 25.

The distribution of the number of candidates broken down by functional groups is shown in Table 26. The hydrocarbons had 13 multicandidate compounds out of 23 tested. The extraneous candidates ranged from one to five. The alcohols had 9 out of 19 multicandidate compounds, although only one or two extraneous candidates were found.

Table 24

DATA AVAILABILITY
65 EXTRANEOUS CANDIDATE COMPOUNDS

NUMBER SIGNATURES UNAVAILABLE	SIGNATURES UNAVAILABLE					NUMBER OF COMPOUNDS	SUB-TOTAL
	MS	IR	NMR	GC	UV		
0						2	2
1			X			4	
				X		4	
					X	3	11
2	X		X			2	
		X	X			1	
		X		X		8	
			X		X	4	
			X	X		16	
				X	X	14	45
3	X		X		X	1	
		X	X		X	1	
			X	X	X	5	7
Total							65

Table 25

DISTRIBUTION OF CANDIDATE RATINGS

TRUE COMPOUNDS (32)

Rating	No. Compounds	
130	18	(All data matched)
126	11	(No elements data)
121	1	(Only one GC column available)
117	2	(No elements data; only one GC column)

EXTRANEOUS COMPOUNDS (65)

34	1	60	1
35	1	62	5
39	3	64	2
40	1	65	3
42	2	66	1
43	3	67	3
44	2	70	2
45	1	72	1
46	4	74	1
47	1	75	1
48	1	*77	1
49	2	**78	2
53	1	80	1
54	5	84	1
56	1	86	1
57	1	87	1
58	5	93	1
59	1	95	1

*5 Signatures available

**5 Signatures available for one compound

Table 26

SEARCH PRECISION

NUMBER OF CANDIDATES	NUMBER OF COMPOUNDS	FUNCTIONAL GROUP												
		Alcohols	Aldehydes	Diols	Esters	Ethers	Halocarbons	Hydrocarbons	Ketones	Nitriles	Nitro	Polyfunctionals	Sulfides	Thiols
0	0													
1	68	10	7	1	2	2	4	10	8	3	6	9	5	1
2	15	6	1	2	2		1	2	1					
3	6	3						2				1		
4	8			1				6	1					
5	1							1						
6	2							2						
Subtotal	32	9	1	3	2		1	13	2			1		
Total	100	19	8	4	4	2	5	23	10	3	6	10	5	1

The 32 test compounds with multicandidates are listed in Table 27 by functional group. Table 28 shows the detailed test results for each of the multicandidate compounds. Notations below the test results for some compounds in Table 28 are discussed in the following section dealing with modified match criteria.

The letters X and M in Tables 28 designate "default" (no data available) and "match," respectively. A and B in the GC row designate the columns on which a match was made. The Y/Y designations in the MS and IR rows indicate the peaks that matched; the left digit designates the input data, the right digit designates the stored data. The symbols + and - designate matches on the tolerance limits. The "true" input compounds appear at the left of each set of candidates and are marked with an asterisk.

Table 27

COMPOUNDS WITH MULTIPLE CANDIDATES BY FUNCTIONAL GROUP

ALCOHOLS			CAND.
1.	244:	ethanol	2
2.	279:	nonanol	2
3.	761:	2-ethyl-1-hexanol	2
4.	246:	butanol	2
5.	760:	3-heptanol	2
6.	278:	octanol	2
7.	759:	2-ethyl-1-butanol	3
8.	247:	pentanol	3
9.	259:	cyclohexanol	3
ALDEHYDES			
1.	322:	propionaldehyde	2
DIOLS			
1.	284:	diethylene glycol	2
2.	783:	2,3-butanediol	2
3.	282:	1,2-propanediol	4
ESTERS			
1.	425:	diethyl malonate	2
2.	819:	butyl benzoate	2
HALOCARBONS			
1.	208:	p-dichlorobenzene	2

KETONES			CAND.
1.	305:	cyclopentanone	2
2.	296:	2-pentanone	4
HYDROCARBONS			
1.	2:	ethane	2
2.	53:	cis-decalin	2
3.	51:	cyclohexane	3
4.	50:	cyclopentane	3
5.	10:	n-decane	4
6.	12:	n-hexadecane	4
7.	7:	n-heptane	4
8.	9:	n-nonane	4
9.	6:	n-hexane	4
10.	56:	methyl cyclohexane	4
11.	52:	cycloheptane	5
12.	701:	2,3-dimethylbutane	6
13.	715:	3-methylpentane	6
POLYFUNCTIONALS			
1.	698:	4-hydroxy-4-methyl-2-pentanone	3

Table 28.1

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: ALCOHOLS (1)

1

Alcohol	244*: Ethanol		658: Acetal	
State	M	2	M	2
Elements	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	2/1 3/3 4/4	21
IR	1/1 2/2 3/3	36	1/3- 2/1-	9
NMR	M	24	M	24
GC	M : A, M : B	18	M : B+	6
UV	M	6	X	1
	Score	130	Score	67
			GC: No Match Col. A	

2

Alcohol	246*: Butanol		769: 3,3-Dimethyl-2-butanol	
State	M	2	M	2
Elements	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	2/4 3/3	14
IR	1/1 2/2 3/3	36	1/3 2/1-	12
NMR	M	24	X	1
GC	M : A, M : B	18	M : B	6
UV	M	6	M	6
	Score	130	Score	45
			GC: No Match Col. A MS: 2 of 4 Peaks	

3

Alcohol	278*: Octanol		182: Citronellol	
State	M	2	M	2
Elements	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	1/1 4/3	19
IR	1/1 2/2 3/3	36	1/2 2/1+	18
NMR	M	24	X	1
GC	M : A, M : B	18	X	1
UV	M	6	X	1
	Score	130	Score	46
			MS: 2 of 4 Peaks	

Table 28.1 (Continued):

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: ALCOHOLS (2)

4	Alcohol	279*: Nonanol		182: Citronellol	
	State	M	2	M	2
	Elements	M	4	M	4
	MS	1/1 2/2 3/3 4/4	40	1/1 3/3	24
	IR	1/1 2/2 3/3	36	1/2 3/1+	13
	NMR	M	24	X	1
	GC	M : A, M : B	18	X	1
	UV	M	6	X	1
		Score	130	Score	46
				MS: 2 of 4 Peaks	

5	Alcohol	760*: 3-Heptanol		776: Cyclopentyl-methanol	
	State	M	2	M	2
	Elements	M	4	M	4
	MS	1/1 2/2 3/3 4/4	40	2/3 3/1	13
	IR	1/1 2/2 3/3	36	X	1
	NMR	M	24	M+	20
	GC	M : A, M : B	18	X	1
	UV	M	6	M	6
		Score	130	Score	47
				NMR: 1 of 2 Peaks MS: 2 of 4 Peaks	

6	Alcohol	761: 2-Ethyl-1-hexanol		182: Citronellol	
	State	M	2	M	2
	Elements	M	4	M	4
	MS	1/1 2/2 3/3 4/4	40	3/1 4/3	7
	IR	1/1 2/2 3/3	36	1/2 2/1-	18
	NMR	M	24	X	1
	GC	M : A, M : B	18	X	1
	UV	M	6	X	1
		Score	130	Score	34
				MS: 2 of 4 Peaks	

Table 28.1 (Continued):

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: ALCOHOLS (3)

7

Alcohol	247*: Pentanol		182: Citronellol		256: 2,2-Dimethyl-1-butanol	
State	M	2	M	2	M	2
Elements	M	4	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	2/3 3/1	13	X	1
IR	1/1 2/2 3/3	36	1/1+ 2/2	27	2/1- 3/3	12
NMR	M	24	X	1	X	1
GC	M : A, M : B	18	X	1	M : B	9
UV	M	6	X	1	M	6
	Score:	130	Score	49	Score	35
			MS: 2 of 4 Peaks		GC: No Match Col. A	

8

Alcohol	259*: Cyclohexanol		690: 2-Butoxy-ethanol		182: Citronellol	
State	M	2	M	2	M	2
Elements	M	4	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	1/1 4/4	19	2/4 4/1	8
IR	1/1 2/2 3/3	36	1/2- 2/3-	15	1/1+ 2/2	27
NMR	M	24	X	1	X	1
GC	M : A, M : B	18	M : A+, M : B-	12	X	1
UV	M	6	X	1	X	1
	Score	130	Score	54	Score	44
			MS: 2 of 4 Peaks		MS: 2 of 4 Peaks MS: 4/1 Match	

9

Table 28.1 (Continued):

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: ALCOHOLS (4)

Alcohol	759*: 2-Ethyl-1-butanol		765: 2-Methyl-1-pentanol		770: 2-Heptanol	
State	M	2	M	2	M	2
Elements	M	4	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	1/1 3/3 4/4	28	1/2 3/4 4/3	21
IR	1/1 2/2 3/3	36	1/1 2/3- 3/2	28	1/1 2/2-	27
NMR	M	24	X	1	X	1
GC	M : A, M : B	18	M : A, M : B	18	M : B	9
UV	M	6	M	6	M	6
	Score	130	Score	87	Score	70
					GC: No Match Col. A	

Table 28.2

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: DIOLS (1)

10

Diol	284*: Diethylene glycol		282: 1,2-Propanediol	
State	M	2	M	2
Elements	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	1/1 3/3 4/4	28
IR	1/1 2/2 3/3	36	1/2+ 3/1	11
NMR	M	24	M+	20
GC	M : A, M : B	18	M : A+	6
UV	M	6	M	6
	Score	130	Score	77
			GC: No Match Col. B NMR: 1 of 2 Peaks	

11

Diol	783*: 2,3-Butanediol		765: 2-Methyl-1-pentanol	
State	M	2	M	2
Elements	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	2/3 3/1 4/4	17
IR	1/1 2/2 3/3	36	1/1 2/2	28
NMR	M	24	X	1
GC	M : A, M : B	18	M : A	9
UV	M	6	M	6
	Score	130	Score	67
			GC: No Match Col. B	

Table 28.2 (Continued):

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: DIOLS (2)

12

Diol	282*: 1,2-Propanediol		284: Diethylene glycol		765: 2-Methyl-1-pentanol		256: 2,2-Dimethyl-1-butanol	
State	M	2	M	2	M	2	M	2
Elements	M	4	M	4	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	1/1 3/3 4/4	28	2/1 4/4	13	X	1
IR	1/1 2/2 3/3	36	1/3 2/1-	12	1/1 2/2	30	1/1- 2/2+	24
NMR	M	24	M-	20	X	1	X	1
GC	M : A, M : B	18	M : A-	6	M : A+	6	M : A-	6
UV	M	6	M	6	M	6	M	6
	Score	130	Score	78	Score	62	Score	44
			GC: No Match Col. B NMR: 1 of 2 Peaks		GC: No Match Col. B MS: 2 of 4 Peaks		GC: No Match Col. B	

Table 28.3

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: MISCELLANEOUS GROUPS (1)

13

Polyfunctional	698*: 4-Hydroxyl-4-methyl-2-pentanone		797: Isobutyric acid		402: Ethylidene diacetate	
State	M	2	M	2	M	2
Elements	M	4	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	1/1 4/3	19	1/1 2/4	22
IR	1/1 2/2 3/3	36	2/1 3/2-	11	X	1
NMR	M	24	M	24	X	1
GC	M : B	9	X	1	M : B	9
UV	M	6	X	1	X	1
	Score	121	Score	62	Score	40
			MS: 2 of 4 Peaks NMR: 1 of 2 Peaks		MS: 2 of 4 Peaks	

Table 28.3 (Continued):

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: MISCELLANEOUS GROUPS (2)

14

Halocarbon	208*: p-Dichloro-benzene		755: 1-Bromo-4-chloro-benzene	
State	M	2	M	2
Elements	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	3/3 4/4	12
IR	1/1 2/2 3/3	36	2/2+ 3/3+	13
NMR	M	24	M+	20
GC	M : A, M : B	18	X	1
UV	M	6	X	1
	Score	130	Score	53
			MS: 2 of 4 Peaks	

15

Aldehyde	322*: Propionaldehyde		332: Acrolein	
State	M	2	M	2
Elements	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	3/4 4/1	7
IR	1/1 2/2 3/3	36	1/1+ 3/3	21
NMR	M	24	X	1
GC	M : A, M : B	18	M : A-	6
UV	M	6	X	1
	Score	130	Score	42
			GC: No Match Col. B MS: 4/1 Match MS: 2 of 4 Peaks	

Table 28.4

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: ESTERS

16

Ester	425*: Diethyl malonate		661: Isobutyric anhydride	
State	M	2	M	2
Elements	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	2/1 4/4	13
IR	1/1 2/2 3/3	36	1/1- 1/2+	24
NMR	M	24	M-	20
GC	M : A, M : B	18	X	1
UV	M	6	X	1
	Score	130	Score	65
			NMR: 1 of 2 Peaks MS: 2 of 4 Peaks	

17

Ester	819*: Butyl benzoate		818: Benzyl benzoate	
State	M	2	M	2
Elements	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	1/1 3/3	24
IR	1/1 2/2 3/3	36	1/1 2/2 3/3-	34
NMR	M	24	M	24
GC	M : A, M : B	18	X	1
UV	M	6	M+	4
	Score	130	Score	93
			NMR: 1 of 2 Peaks MS: 2 of 4 Peaks	

Table 28.5

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: HYDROCARBONS (1)

18

Hydrocaibon	2*: Ethane		3: Propane	
State	M	2	M	2
Elements	-	0	-	0
MS	1/1 2/2 3/3 4/4	40	1/2 2/3	21
IR	1/1 2/2 3/3	36	1/1 2/2+	27
NMR	M	24	X	1
GC	M : A, M : B	18	X	1
UV	M	6	M	6
	Score:	126	Score	58
			MS: 2 of 4 Peaks	

19

Hydrocarbon	53*: Cis-decalin		54: Trans-decalin	
State	M	2	M	2
Elements	-	0	-	0
MS	1/1 2/2 3/3 4/4	40	1/1 4/3	15
IR	1/1 2/2 3/3	36	1/1 2/2 3/3	36
NMR	M	24	M	24
GC	M : A, M : B	18	X	1
UV	M	6	M	6
	Score	126	Score	84
			MS: 2 of 4 Peaks	

Table 28.5 (Continued):
MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: HYDROCARBONS (2)

20

Hydrocarbon	50*: Cyclopentane		844: Cis-1,4-dimethyl-cyclohexane		845: Trans-1,3-dimethyl-cyclohexane	
State	M	2	M	2	M	2
Elements	-	0	-	0	-	0
MS	1/1 2/2 3/3 4/4	40	3/3 4/1	9	3/3 4/1	9
IR	1/1 2/2 3/3	36	X	1	X	1
NMR	M	24	M	24	M	24
GC	M : A	9	X	1	X	1
UV	M	6	M	6	M	6
	Score	117	Score	43	Score	43
			MS: 4/1 Match MS: 2 of 4 Peaks		MS: 4/1 Match MS: 2 of 4 Peaks	

Table 28.5 (Continued):

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: HYDROCARBONS (3)

21

Hydrocarbon	51*: Cyclohexane		844: Cis-1,4-dimethyl-cyclohexane		845: Trans-1,3-dimethyl cyclohexane	
State	M	2	M	2	M	2
Elements	-	0	-	0	-	0
MS	1/1 2/2 3/3 4/4	40	3/3 4/1	9	3/3 4/1	9
IR	1/1 2/2 3/3	36	X	1	X	1
NMR	M	24	M+	20	M+	20
GC	M : B	9	X	1	X	1
UV	M	6	M	6	M	6
	Score	117	Score	39	Score	39
			MS: 4/1 Match MS: 2 of 4 Peaks		MS: 4/1 Match MS: 2 of 4 Peaks	

22

Hydrocarbon	6*: n-Hexane		13: n-Heptadecane		36: Squalane		839: 2-Propanethiol	
State	M	2	M	2	M	2	M	2
Elements	-	0	-	0	-	0	-	0
MS	1/1 2/2 3/3 4/4	40	1/1 2/2 3/4	34	1/1 2/2 3/4	34	2/1 3/3	17
IR	1/1 2/2 3/3	36	1/1 2/2 3/3+	34	1/1 2/2	30	1/2+ 2/3	17
NMR	M	24	X	1	X	1	M	24
GC	M : A, M : B	18	X	1	X	1	X	1
UV	M	6	M	6	M	6	X	1
	Score	126	Score	78	Score	74	Score	62
							NMR: 1 of 2 Peaks MS: 2 of 4 Peaks	

Table 28.5 (Continued):

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: HYDROCARBONS (4)

23

Hydrocarbon	7*: n-Heptane		13: n-Heptadecane		839: 2-Propanethiol		36: Squalane	
State	M	2	M	2	M	2	M	2
Elements	-	0	-	0	-	0	-	0
MS	1/1 2/2 3/3 4/4	40	1/2 2/4 3/1	22	1/1 2/3	17	1/2 2/4 3/1	22
IR	1/1 2/2 3/3	36	1/1- 2/2 3/3	33	1/2 2/3	20	1/1 2/2	27
NMR	M	24	X	1	M	24	X	1
GC	M : A, M : B	18	X	1	X	1	X	1
UV	M	6	M	6	X	1	M	6
	Score	126	Score	65	Score	65	Score	59
					NMR: 1 of 2 Peaks MS: 2 of 4 Peaks			

24

Hydrocarbon	9*: n-Nonane		839: 2-Propanethiol		13: n-Heptadecane		36: Squalane	
State	M	2	M	2	M	2	M	2
Elements	-	0	-	0	-	0	-	0
MS	1/1 2/2 3/3 4/4	40	1/1 3/3	24	1/2 2/1 3/4	27	1/2 2/1 3/4	27
IR	1/1 2/2 3/3	36	1/2 2/3	20	1/1- 2/2 3/3	33	1/1 2/2	27
NMR	M	24	M	24	X	1	X	1
GC	M : A, M : B	18	X	1	X	1	X	1
UV	M	6	X	1	M	6	M	6
	Score	126	Score	72	Score	70	Score	64
			NMR: 1 of 2 Peaks MS: 2 of 4 Peaks					

Table 28.5 (Continued):

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: HYDROCARBONS (5)

25

Hydrocarbon	10*: n-Decane		839: 2-Propanethiol		13: n-Heptadecane		36: Squalane	
State	M	2	M	2	M	2	M	2
Elements	-	0	-	0	-	0	-	0
MS	1/1 2/2 3/3 4/4	40	1/1 3/3	24	1/2 2/1 3/4	17	1/2 2/1 3/4	17
IR	1/1 2/2 3/3	36	1/2+ 2/3+	15	1/1 2/2+	27	1/1 2/2+	27
NMR	M	24	M	24	X	1	X	1
GC	M : A, M : B	18	X	1	X	1	X	1
UV	M	6	X	1	M	6	M	6
	Score	126	Score	67	Score	54	Score	54
			NMR: 1 of 2 Peaks MS: 2 of 4 Peaks					

26

Hydrocarbon	12*: n-Hexadecane		13: n-Heptadecane		36: Squalane		839: 2-Propanethiol	
State	M	2	M	2	M	2	M	2
Elements	-	0	-	0	-	0	-	0
MS	1/1 2/2 3/3 4/4	40	1/1 2/2 3/3 4/4	40	1/1 2/2 3/3 4/4	40	2/1 4/3	12
IR	1/1 2/2 3/3	36	1/1 2/2 3/3	36	1/1 2/2	30	1/2+ 2/3	17
NMR	M	24	X	1	X	1	M	24
GC	M : A, M : B	18	X	1	X	1	X	1
UV	M	6	M	6	M	6	X	1
	Score	126	Score	86	Score	80	Score	57
							NMR: 1 of 2 Peaks MS: 2 of 4 Peaks	

Table 28.5 (Continued):

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: HYDROCARBONS (6)

27

Hydrocarbon	56*: Methylcyclohexane		836: 4-Methylvaleronitrile		834: 4-Methylpentanenitrile		846: Trans-1,4-dimethylcyclohexane	
State	M	2	M	2	M	2	M	2
Elements	-	0	-	0	-	0	-	0
MS	1/1 2/2 3/3 4/4	40	2/1 3/2	12	2/1 3/2	12	2/1 3/3	14
IR	1/1 2/2 3/3	36	1/1- 1/2+ 3/3-	28	1/1- 1/2+	24	X	1
NMR	M	24	M+	20	M+	20	M	24
GC	M : A, M : B	18	X	1	X	1	X	1
UV	M	6	X	1	X	1	M	6
	Score	126	Score	64	Score	60	Score	48
			MS: 2 of 4 Peaks		MS: 2 of 4 Peaks		NMR: 1 of 2 Peaks MS: 2 of 4 Peaks	

28

Hydrocarbon	52*: Cycloheptane		717: 2-Methyl-2-butene		726: Chlorocyclohexane		844: Cis-1,4-dimethylcyclohexane		845: Trans-1,3-dimethylcyclohexane	
State	M	2	M	2	M	2	M	2	M	2
Elements	-	0	-	0	-	0	-	0	-	0
MS	1/1 2/2 3/3 4/4	40	1/3 3/1	12	1/4 3/3	12	1/3 3/1	12	1/3 3/1	12
IR	1/1 2/2 3/3	36	1/1 3/2+	18	1/1 3/3-	18	X	1	X	1
NMR	M	24	M	24	M	24	M	24	M	24
GC	M : A, M : B	18	X	1	X	1	X	1	X	1
UV	M	6	X	1	X	1	M	6	M	6
	Score	126	Score	58	Score	58	Score	46	Score	46
			MS: 2 of 4 Peaks		MS: 1/4 Match MS: 2 of 4 Peaks		MS: 2 of 4 Peaks		MS: 2 of 4 Peaks	

Table 28.5 (Continued):

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: HYDROCARBONS (7)

29

Hydrocarbon	701*: 2,3-Dimethylbutane		836: 4-Methylvaleronitrile		36: Squalane		13: n-Heptadecane		705: 2,2-Dimethylbutane		28: 2,3,4-Trimethylpentane	
State	M	2	M	2	M	2	M	2	M	2	M	2
Elements	-	0	-	0	-	0	-	0	-	0	-	0
MS	1/1 2/2 3/3 4/4	40	1/3 3/2 4/4	18	1/2 3/4	18	1/2 3/4	18	1/1 2/2 3/3 4/4	40	1/1 3/4	22
IR	1/1 2/2 3/3	36	1/1 2/3-	24	1/1 2/2	30	1/1 2/2	30	X	1	2/2- 3/1- 3/3	10
NMR	M	24	M-	20	X	1	X	1	X	1	X	1
GC	M : A, M : B	18	X	1	X	1	X	1	M : B-	6	X	1
UV	M	6	X	1	M	6	M	6	M	6	M	6
	Score	126	Score	66	Score	58	Score	58	Score	56	Score	42
					MS: 2 of 4 Peaks		NMR: 1 of 2 Peaks MS: 2 of 4 Peaks				MS: 2 of 4 Peaks	

30

Hydrocarbon	715*: 3-Methylpentane		713: 2,3-Dimethylpentane		712: Isopentane		13: n-Heptadecane		36: Squalane		41: 1-Hexene	
State	M	2	M	2	M	2	M	2	M	2	M	2
Elements	-	0	-	0	-	0	-	0	-	0	-	0
MS	1/1 2/2 3/3 4/4	40	1/2 2/4 3/3	26	1/4 3/3	12	1/1 3/4	22	1/1 3/4	22	2/2 3/1	16
IR	1/1 2/2 3/3	36	1/1 2/2 3/3	36	1/1 2/2	30	1/1 2/2	30	1/1 2/2	30	1/3+ 3/2-	6
NMR	M	24	M	24	M	24	X	1	X	1	M+	20
GC	M : A, M : B	18	X	1	X	1	X	1	X	1	M : A	9
UV	M	6	M	6	M	6	M	6	M	6	X	1
	Score	126	Score	95	Score	75	Score	62	Score	62	Score	54
					MS: 1/4 Match MS: 2 of 4 Peaks		MS: 2 of 4 Peaks		MS: 2 of 4 Peaks		NMR: 1 of 2 Peaks MS: 2 of 4 Peaks	

Table 28.6

MATCHES OBTAINED WITH MULTICANDIDATE TEST COMPOUNDS: KETONES

31

Ketone	305*: Cyclo-pentanone		812: Propyl acrylate	
State	M	2	M	2
Elements	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	X	1
IR	1/1 2/2 3/3	36	1/1 3/3	24
NMR	M	24	X	1
GC	M : A, M : B	18	M : A-	6
UV	M	6	X	1
	Score	130	Score	39
			GC: No Match Col. B	

32

Ketone	296*: 2-Pentanone		373: Methacrylic acid		404: Allyl acetate		384: Isobutyl formate	
State	M	2	M	2	M	2	M	2
Elements	M	4	M	4	M	4	M	4
MS	1/1 2/2 3/3 4/4	40	3/2 4/1	7	1/1 4/2	18	1/1 4/4	19
IR	1/1 2/2 3/3	36	1/1+ 3/3-	19	1/2- 2/3	17	1/2- 3/1-	10
NMR	M	24	M-	20	X	1	X	1
GC	M : A, M : B	18	X	1	M : A-	6	M : A-	6
UV	M	6	X	1	X	1	X	1
	Score	130	Score	54	Score	49	Score	43
			NMR: 1 of 2 Peaks MS: 4/1 Match MS: 2 of 4 Peaks		GC: No Match Col. B MS: 2 of 4 Peaks		GC: No Match Col. B	

In general, it can be said that hydrocarbons and alcohols posed the greatest problem with respect to multiple candidates. The reasons for this appear to be correlated to compound type and lack of data. The summarized results in Tables 26, 27 and 29 point out the large number of alcohols and hydrocarbons that came through with multiple candidates. Tables 24 and 29 show that, in general, significant data were unavailable particularly in GC and NMR. Of all the signatures these two could be the most important for characterizing alcohols and hydrocarbons. This correlates with the fact that hydrocarbons and alcohols were the worst offenders among the extraneous candidates selected more than once.

The observation one can make is that a more complete data file, with particular emphasis on GC and NMR which are among the less abundant signatures, would greatly reduce (by perhaps 75 percent) the multiple candidate problem we experienced. The greatest influence of having these data would be greater discrimination in the alcohols and hydrocarbons.

b. Input 2 - Compounds Without Stored Data

Five compounds that do not have signature data stored in the data base were tested. These Input 2 compounds are:

68:	n-Prophylbenzene
270:	m-Cresol
8:	Octane
74:	Indene
92:	Thiophene.

Table 29

EXTRANEOUS CANDILATE COMPOUNDS
SELECTED MORE THAN ONCE

COMPOUND	NUMBER SELECTIONS	SIGNATURES UNAVAILABLE				
		MS	IR	NMR	GC	UV
13: n-Heptadecane	7			X	X	
36: Squalane	7			X	X	
182: Citronellol	5			X	X	X
839: 2-Propanethiol	5				X	X
844: cis-1,4-dimethyl-cyclohexane	3		X		X	
845: trans-1,3-dimethyl-cyclohexane	3		X		X	
765: 2-Methyl-1-pentanol	3			X		
836: 4-Methyl-valeronitrile	2				X	X
256: 2,2-Dimethyl-1-butanol	2	X		X		

Compounds 68, 92, and 74 had perfect precision; i.e., no candidate compounds survived.

The search for m-cresol resulted in the candidacy of one compound, 274 (2,4-dimethylphenol), which had a rating of 28 points. The GC signature search was omitted because no GC data were available for m-cresol and the 2,4-dimethylphenol passed through the NMR and the IR searches by default. Hence the candidate compound passed through only two signature searches by matching the negative tolerance on UV and the 2/2 and the 4/4 peaks in MS.

Compound 274 is an example of a compound that was very similar to the unknown and passed through because of chemical similarity and default. If GC data were available, the candidate compound surely would have been rejected, because homologues, such as m-cresol and 2,4-dimethylphenol, are readily resolved, particularly with the GC operating parameters used. Although these molecules are different, they are similar enough to produce very similar MS, IR, and NMR spectra, and, considering our match criteria, 2,4-dimethylphenol will probably pass through as a candidate based on these spectral data.

The search for n-octane, a hydrocarbon, resulted in selecting the six candidates shown in Table 30 with scores ranging from 72 to 30 out of a possible 130. The following key factors were significant in the failure of the screening process to reject the six candidates accompanying n-octane:

Table 30

CANDIDATE COMPOUNDS RETRIEVED IN SEARCH FOR OCTANE

COMPOUND				SEARCH RESULTS[a]					
NO.	NAME	STATE	ELEMENT	MS	IR	NMR	GC	UV	SCORE
19:	3-METHYLHEPTANE	M	SKIP	1/1 2/4 4/2	1/+1 2/+2	D	A	M	72
25:	2,5-DIMETHYLHEXANE	M	SKIP	1/1 2/3 4/2	1/1 2/+2 2/3	D	-B	M	71
839:	2-PROPANETHIOL	M	SKIP	1/1 2/3	1/+2 2/+3	M	D	D	68
763:	2-HEXANOL	M	SKIP	1/2 2/3	1/1 2/+3	D	A	M	63
13:	N-HEPTADECANE	M	SKIP	1/2 2/4 4/1	1/1 2/+2 3/3	D	D	M	56
29:	2.3.3-TRIMETHYLPENTANE	M	SKIP	1/1 4/4	D	D	D	M	30

[a] D = default; M = match.

(1) GC data on Carbowax 20M would have caused rejection of 2-propanethiol and 2-hexanol

(2) GC data on n-heptadecane would have caused rejection of this compound

(3) The chemical similarity among the four hydrocarbons (all saturated hydrocarbons, three of which have eight carbon atoms) would make it difficult to distinguish on the basis of our MS, IR, and UV match criteria. NMR may or may not have been selective

(4) The MS fragmentation behavior of hydrocarbons, 2-hexanol, and 2-propanethiol is very similar. In practice, even with entire mass spectra for the candidate compounds, it is often difficult to distinguish between them.

The interesting feature of the above examples is the similarity of the candidates to the input compound. This proves to be an advantage because if there are no data in the file for an unknown, the compounds selected give a strong indication of the nature of the compound and can be very helpful in identifying the compound type. Of course, the selection of "false" candidates depends not only on matches, but also on defaults; the more data available, the greater the selectivity, while the lack of data could permit more candidates to be carried through by default.

If a large number of default compounds are carried through, little help might be provided to the user, since there would be no reason to suspect that the default candidates are structurally related. No inferences should be made on the basis of default matches. Judgments should be made on the basis of true data matches.

Because the Input 2 compounds serve the same purpose as Input 1 compounds, no further tests were made on this group, and further evaluation of test results is included in the discussion of modified match criteria.

c. Input 3 - Compounds with Laboratory Measurements

The Input 3 compounds were used to evaluate measurement repeatability, data standards, and the tolerance limits that would be necessitated to accommodate variations in measured data.

Eleven compounds were chosen at random from the 100 "unknowns." All eleven compounds were analyzed by each of the five analytical techniques. Only one, diphenyl sulfide, lacked the relative retention data on the Carbowax 20M column.

Table 31 lists the eleven compounds along with the abstracted data generated in our laboratory. Included in the table are the abstracted data for the compounds contained in the file. The laboratory generated data were abstracted and entered into the data sheets for unknowns. These data were screened against the file according to the initial match criteria described in IVA3a.

The results of the screening process are tabulated in Table 32. Nine of the eleven came through as unique matches.

Table 31

COMPARISON BETWEEN STORED AND LABORATORY DATA

		MS (Mass)			IR (μ)			NMR (ppm)		GC (Rel. Ret. Time)			UV (mμ)	
Unknown	Compound	Param.	Stored	Lab	Param.	Stored	Lab	Stored	Lab	Param.	Stored	Lab	Stored	Lab
A	560:1-Butanethiol	Pk 1 Pk 2 Pk 3 Pk 4	56 41 27 90	56 41 90 47+	Pk 1 Pk 2 Pk 3	6.8 7.8 8.2	6.8 7.8 7.2*	0.9	0.9	Col A Col B	0.76 0.57	0.76 0.59x	NA	TP
B	327:Isobutyraldehyde	Pk 1 Pk 2 Pk 3 Pk 4	43 41 27 71	43 27 41 29+	Pk 1 Pk 2 Pk 3	5.8 6.8 3.7	5.8 6.8 3.7	0.9	1.0	Col A Col B	0.31 0.33	0.31 0.33	214	214
C	10:Decane	Pk 1 Pk 2 Pk 3 Pk 4	43 57 41 29	43 57 41 71+	Pk 1 Pk 2 Pk 3	6.8 7.2 3.9	6.8 7.2 13.7*	1.3	1.3	Col A Col B	2.7 0.70	2.7 0.68x	TP	TP
D	184:Menthone	Pk 1 Pk 2 Pk 3 Pk 4	41 112 69 55	112 69 41 55	Pk 1 Pk 2 Pk 3	5.8 6.8 7.3	5.8 6.8 7.3	NA	0.9	Col A Col B	6.0 6.0	6.1 6.1	NA	236
E	685:p-Anisaldehyde	Pk 1 Pk 2 Pk 3 Pk 4	135 136 77 92	135 136 77 76+	Pk 1 Pk 2 Pk 3	6.2 7.9 8.6	6.3 8.0 8.8*	3.9	3.8	Col A Col B	9.4 NA	9.4 45	274	274
F	323:n-Butyraldehyde	Pk 1 Pk 2 Pk 3 Pk 4	27 29 44 43	44 43 27 41+	Pk 1 Pk 2 Pk 3	5.8 3.7 6.8	8.7* 10.3* 5.8	1.0	0.9	Col A Col B	0.35 0.42	0.21x 0.41	290	290
G	458:Diphenyl Ether	Pk 1 Pk 2 Pk 3 Pk 4	170 51 77 141	170 51 77 141	Pk 1 Pk 2 Pk 3	8.1 6.3 6.7	6.7* 8.1 14.5*	7.0	7.0	Col A Col B	17 54	17 47	226	226
H	574:Diphenyl Sulfide	Pk 1 Pk 2 Pk 3 Pk 4	186 185 51 184	186 185 51 184	Pk 1 Pk 2 Pk 3	13.6 14.5 6.3	14.5 13.6 6.8*	7.2	7.2	Col A Col B	14 NA	14	250	250
I	249:n-Heptanol	Pk 1 Pk 2 Pk 3 Pk 4	41 56 70 43	56 70 41 43	Pk 1 Pk 2 Pk 3	3.0 9.5 6.9	3.1 9.5 6.9	1.3	1.3	Col A Col B	2.5 4.6	2.4 4.3x	TP	TP
J	540:Nitrobenzene	Pk 1 Pk 2 Pk 3 Pk 4	77 51 123 50	77 123 57+ 56+	Pk 1 Pk 2 Pk 3	6.6 7.4 14.3	7.4 6.6 14.2	7.6	7.6	Col A Col 3	4.9 17	4.5 17	260	258
K	528:o-Tolunitrile	Pk 1 Pk 2 Pk 3 Pk 4	117 116 90 89	117 116 90 89	Pk 1 Pk 2 Pk 3	13.1 4.6 6.7	13.1 6.7 4.5	2.6	2.5	Col A Col B	4 12	3.9 12	228	228

+ = No match on any one of four stored peaks.

* = Outside tolerance limits of any one of three stored peaks.

x = Outside tolerance limits.

NA = Not available.

TP = Transparent.

Table 32

IDENTIFICATION OF COMPOUNDS FROM LABORATORY DATA

Search	Compound A			Compound B			Compound C			Compound D		
	Discrim.		Select.	Discrim.		Select.	Discrim.		Select.	Discrim.		Select.
	M	M+D	Cand.	M	M+D	Cand.	M	M+D	Cand.	M	M+D	Cand.
State	387	387	387	387	387	387	387	387	387	387	387	387
Elements	16	16	14	300	300	227	-	-	387	300	300	227
MS	30	52	4	131	153	89	108	130	91	39	61	25
IR	77	107	2	28	58	11	48	78	34	89	119	4
NMR	42	185	2	50	193	11	49	192	20	41	184	3
GC	10	202	1	11	203	2	7	199	6	5	197	1
UV	85	352	1	6	273	1	85	352	5	5	272	1
Identification (True compound underlined)	Candidate		Score	Candidate		Score	Candidate		Score	Candidate		Score
	1-Butanethiol		104	iso-Butyraldehyde		103	n-Decane Squalane n-Heptadecane 2-Propanethiol 2,3,3-Trimethylpentane		107 71 67 67 29	Menthone		87

Search	Compound E			Compound F			Compound G			Compound H		
	Discrim.		Select.	Discrim.		Select.	Discrim.		Select.	Discrim.		Select.
	M	M+D	Cand.	M	M+D	Cand.	M	M+D	Cand.	M	M+D	Cand.
State	387	387	387	387	387	387	387	387	387	387	387	387
Elements	300	300	227	300	300	227	300	300	227	16	16	14
MS	6	28	11	138	160	96	17	39	18	1	23	1
IR	10	40	1	14	44	10	14	44	2	16	46	1
NMR	11	154	1	42	185	3	11	154	1	33	176	1
GC	4	196	1	10	202	0	3	195	1	3	195	1
UV	20	287	1	6	273	0	13	280	1	8	275	1
Identification (True compound underlined)	Candidate		Score	Cpndidate		Score	Candidate		Score	Candidate		Score
	p-Anisaldehyde		110				Diphenyl Ether		108	Diphenyl Sulfide		104

Search	Compound I			Compound J			Compound K		
	Discrim.		Select.	Discrim.		Select.	Discrim.		Select.
	M	M+D	Cand.	M	M+D	Cand.	M	M+D	Cand.
State	387	387	387	387	387	387	387	387	387
Elements	300	300	227	27	27	10	80	80	48
MS	103	125	68	13	35	2	5	27	4
IR	49	79	16	7	37	2	14	44	2
NMR	49	192	13	12	155	2	16	159	2
GC	7	199	1	7	199	2	12	204	1
UV	85	352	1	14	281	1	14	281	1
Identification (True compound underlined)	Candidate		Score	Candidate		Score	Candidate		Score
	n-Heptanol		101	Nitrobenzene		107	o-Tolunitrile		120

M = Matches
M+D = Matches+Defaults

One, n-butyraldehyde, was lost due to a "no-match" on infrared. A comparison was made of our infrared spectrum with the original data source (Sadtler IR #333). Sadtler's spectrum shows a strong band at 13μ which is totally absent in our spectrum. We strongly suspect an impurity in Sadtler's material making that whole spectrum suspect. Since we cannot, at this time, obtain a confirmation on Sadtler's spectrum we must carry it through as a mismatch.

Another unknown, n-Decane, encountered the same problem as the hydrocarbons encountered in the Input 3 category which were discussed earlier. N-Decane came through with 4 additional candidates. These four were the same compounds that came through with other hydrocarbons. These four candidates lacked gas chromatographic data. If gas chromatographic data had been available these four extraneous candidates would have been eliminated.

All but one of the proper candidates received a rating of over 100 out of 130 maximum. None received 130 points. Menthone, though uniquely selected, obtained a rating of 87. The reason for this was the lack of both NMR and UV data in the file. If these signatures had been present and there had been matches within the tolerance limits, 24 more points would have been obtained. In spite of the lack of these data, methone still came through as a unique candidate.

Two of the eleven searches resulted in a unique selection only after UV had been searched. Prior to the UV search two candidates remained. Although this suggests that UV does have its ability to be discriminating, a search based on the modified match criteria would probably have eliminated the need for UV. This supposition was not tested however.

d. Precision with Modified Match Criteria

The results obtained above using the initial match criteria described in Section IVA3a can be considered as a first pass based upon assumptions as to the efficacy of each signature in identifying an unknown compound. There were no explicit guidelines to determine the number of peaks to be searched in each signature, the number required to establish a match, and the tolerance limits that should be established.

Subsequent to an evaluation of the precision of identification obtained with the initial match criteria, additional tests were made using the modified match criteria described in Section IVA3b. These tests were made only on the 32 known input compounds for which multiple candidates were obtained. No effort was made to determine the effect of the modified criteria on the discrimination and selectivity measures derived from the tests using the initial criteria. The modified criteria would have the effect, in general, of improving both discrimination and selectivity. The availability and default ratios, however, would also be changed by the requirement for additional data in the NMR and GC signatures.

The results of imposing the modified match criteria on the 32 test compounds with multiple candidates are recorded beneath affected compounds in Table 28. The ratings have not been changed in accordance with the modified criteria, however. The number of candidates eliminated by the modified criteria, singly and in combination, are listed in Table 33.

Table 33

CANDIDATES ELIMINATED WITH MODIFIED MATCH CRITERIA

Match Criterion	Number of Candidates Eliminated
(1) GC: Columns A and B	7
(2) NMR: 2 of 2 Peaks	-
(3) MS: 3 of 4 Peaks	17
(4) MS: 1/4 or 4/1 match eliminated	-
(1) and (2)	2
(1) and (3)	5
(2) and (3)	11
(3) and (4)	7
(1) and (3) and (4)	1
(2) and (3) and (4)	1
Total	51

Fifty-one of the 65 extraneous compounds were eliminated leaving only 8 test compounds with multiple candidates. Of these 8 compounds, 2 have one extraneous candidate and 6 have 2 extraneous candidates. Seven of the test compounds were hydrocarbons and one was an alcohol. Each of five of the hydrocarbons had n-heptadecane and squalane as the extraneous candidates. Certainly a more complete data file on these compounds would have resulted in rejection also.

For the 100 test compounds, the modified match criteria resulted in unique identification of 92 of the compounds.

f. Additional Match Modifications

A number of additional match modifications can be considered for the improvement of search precision. These include the number of signatures to be searched, data tolerances, and number of peaks to be searched and matched. As the data base grows and encompasses thousands of compounds, some additional match criteria may be required to maintain satisfactory precision.

One of the objectives of the search experiment was to determine the minimum number of signatures necessary to achieve identification. Table 22 has indicated that the five signatures each aided in this task. To investigate the precision obtained with various combinations of signatures, ten compounds were selected and matches were made on combinations of 2, 3, 4, and 5 signatures. Only combinations were tested because the final results are independent of the order of search. The results are summarized in Table 34 and illustrated in Figures 7, 8, and 9.

In the 2-signature search, the average number of candidates for the 10 test compounds averaged over the 10 combinations of five signatures taken 2 at a time was 49.2. The MS·IR combination

Table 34

SELECTIVITY OF COMBINATIONS
OF SIGNATURES
(10 Known Input Compounds*)

NO. SIGN.	SIGNATURES	AVERAGE NUMBER OF SURVIVING CANDIDATES					
		Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	MEAN
2	MS·IR	73.2	15.3				49.2
	MS·NMR	73.2	30.3				
	MS·GC	73.2	22.1				
	MS·UV	73.2	45.1				
	IR·NMR	65.8	32.2				
	IR·GC	65.8	28.4				
	IR·UV	65.8	39.2				
	NMR·GC	175.3	55.6				
	NMR·UV	175.3	109.4				
	GC·UV	205.3	114.7				
3	MS·IR·GC	73.2	15.3	4.0			13.2
	MS·IR·UV	73.2	15.3	10.2			
	MS·IR·NMR	73.2	15.3	8.8			
	MS·NMR·GC	73.2	30.3	7.4			
	MS·NMR·UV	73.2	30.3	21.1			
	MS·GC·UV	73.2	22.1	11.0			
	IR·NMR·GC	65.8	32.2	12.3			
	IR·NMR·UV	65.8	32.2	20.9			
	IR·GC·UV	65.8	28.4	11.2			
	NMR·GC·UV	175.3	55.6	25.0			
4	MS·IR·NMR·UV	73.2	15.3	8.8	5.4		3.8
	MS·NMR·GC·UV	73.2	30.3	7.4	3.6		
	IR·NMR·GC·MS	65.8	32.2	12.3	2.9		
	IR·GC·UV·MS	65.8	28.4	11.2	2.1		
	NMR·GC·UV·IR	175.3	55.6	25.0	5.0		
5	IR·NMR·GC·MS·UV	65.8	32.2	12.3	2.9	1.5	1.5

*All compounds had only one surviving candidate in 7-stage search.

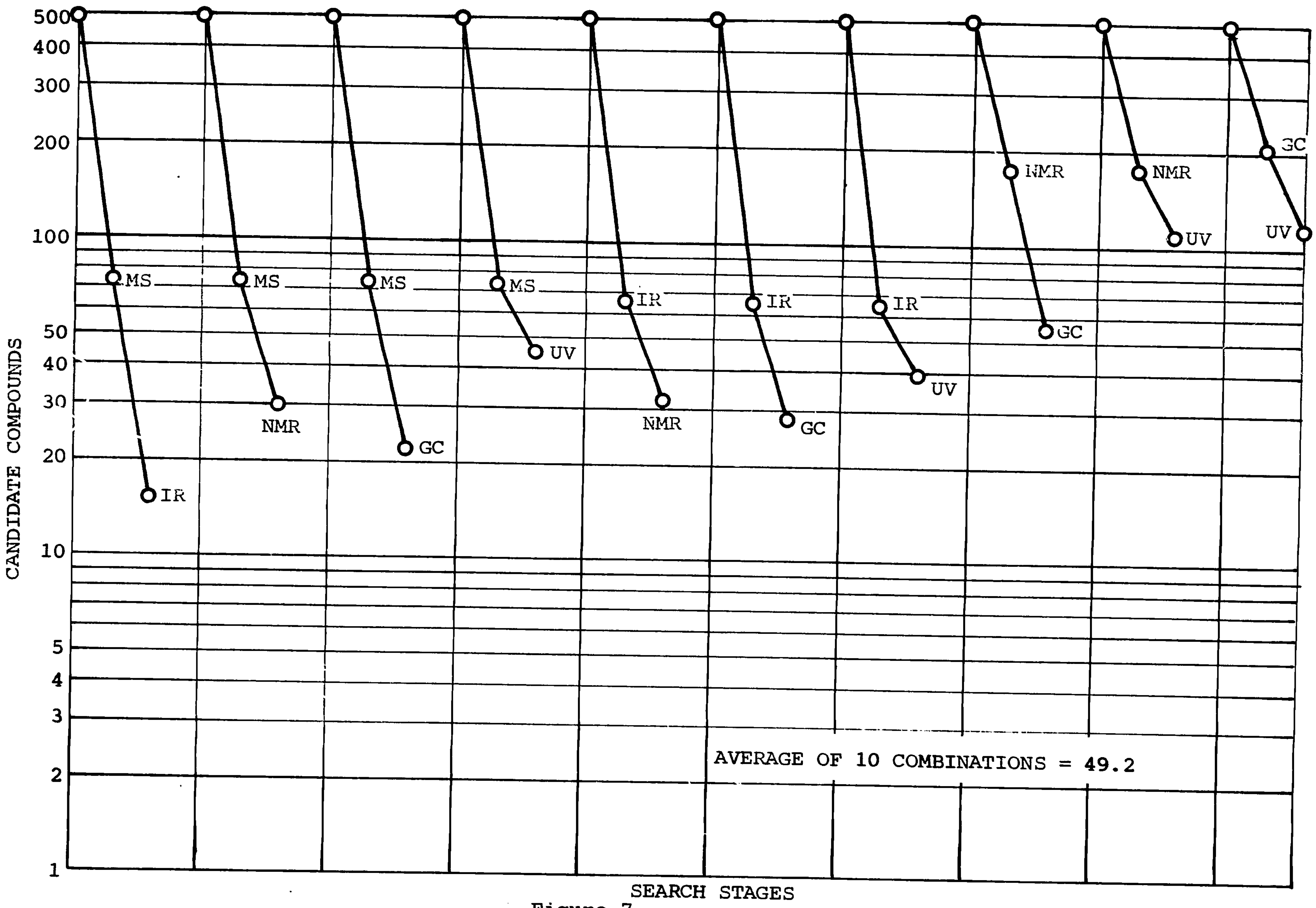


Figure 7

SELECTIVITY OF 2-SIGNATURE SEARCHES
AVERAGES OF 10 KNOWN COMPOUNDS

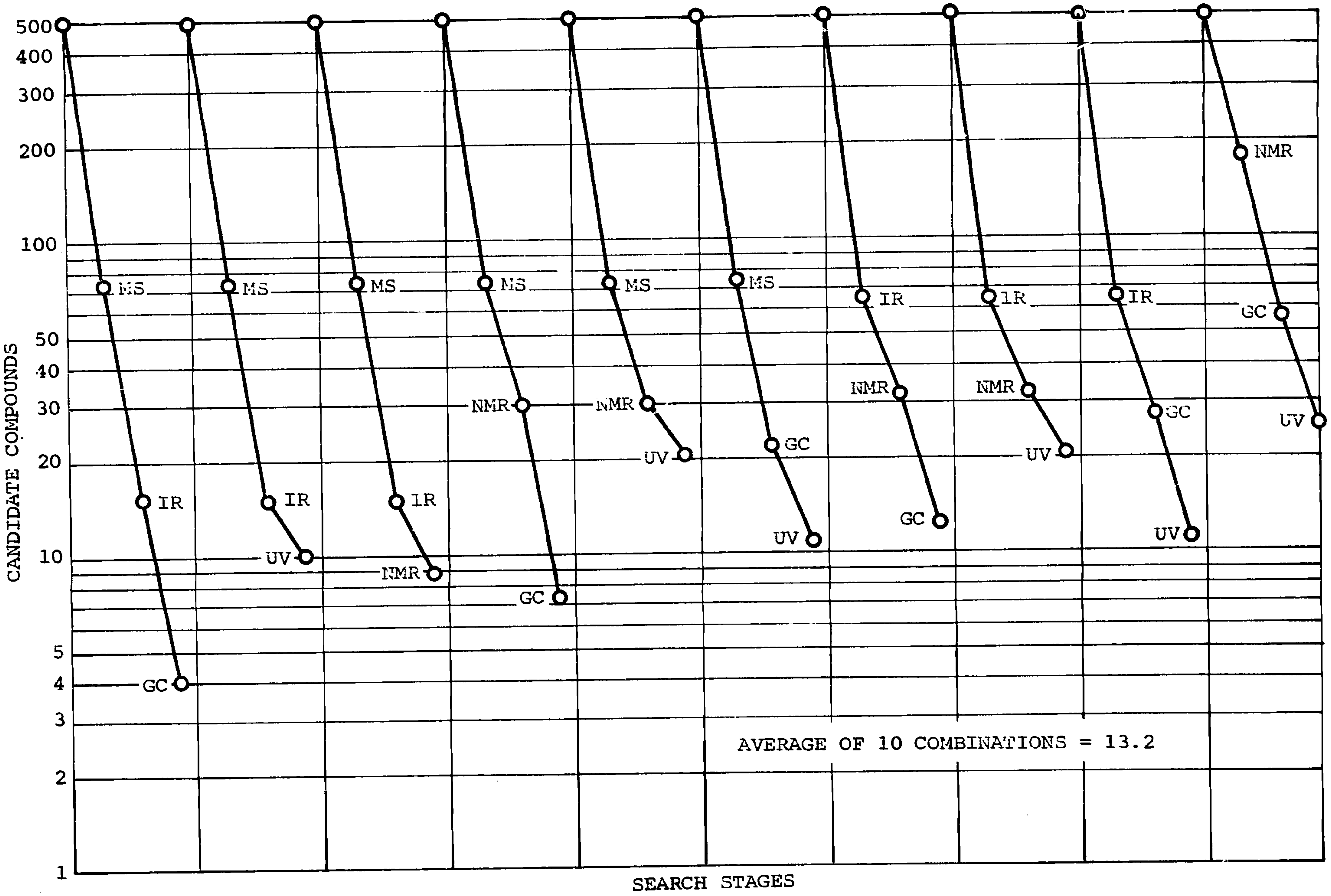


Figure 8

SELECTIVITY OF 3-SIGNATURE SEARCHES
AVERAGES OF 10 KNOWN COMPOUNDS

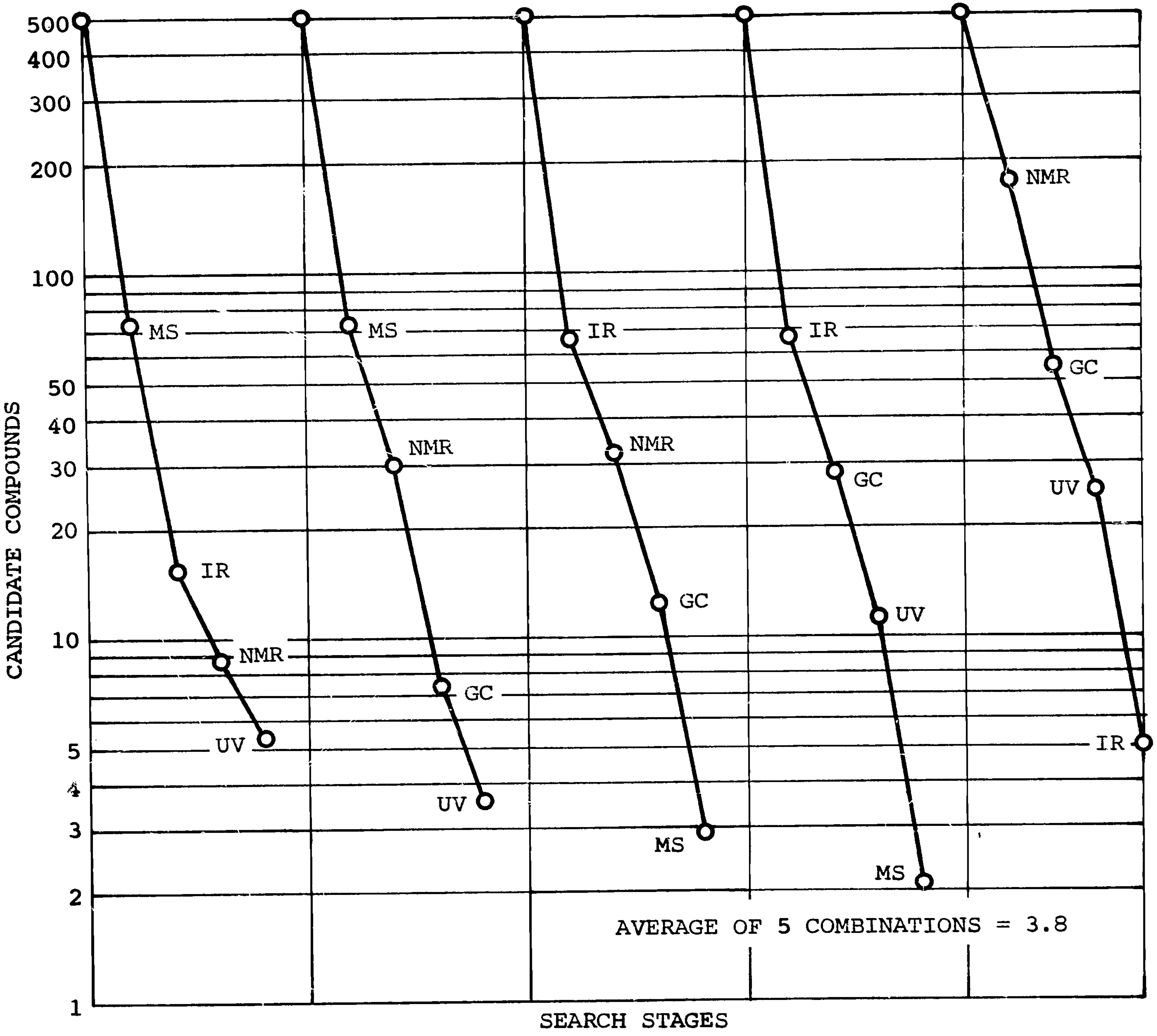


Figure 9

SELECTIVITY OF 4-SIGNATURE SEARCHES
AVERAGES OF 10 KNOWN COMPOUNDS

with an average of 15.3 candidates per compound was the most selective. The GC·UV combination with an average of 114.7 was the least selective attributed mainly to the large number of defaults in both techniques.

The MS·IR·GC combination with an average of 4.0 candidates per compound was the most selective of the 10 3-signature searches, followed by MS·NMR·GC with 7.4 candidates and MS·IR·NMR with 8.8. The average of the 10 combinations was 13.2 candidates per :ompound.

The IR·GC·UV·MS combination was most selective among the 4-signature combinations with an average of 2.1 candidates per compound. The IR·NMR·GC·MS combination was second with an average of 2.9 candidates. For the 5 combinations, 3.8 was the average number of candidates.

The search on all five signatures resulted in an average of 1.5 candidates per compound thus showing once again that each signature contributed to the precision. Inasmuch as all of the 10 compounds searched in the combinatorial tests resulted in unique identifications in the 7-stage search, it is concluded that the state and element searches also contributed to the identification.

It can also be concluded that each of the 7 stages made a contribution to identification in the experiment, not all equally, however. It is possible that the modified match criteria, which eliminated many extraneous candidates, would alter the results of Table 34 sufficiently to permit elimination of a signature. The selectivity shown in Tables 22 and 32 suggests that four signatures would be the minimum number to be searched using the number and ranking of peaks as employed in the modified match criteria.

V. SUMMARY AND CONCLUSIONS

The objective of the research study reported herein was to determine the capability of a composite signature to identify an unknown compound. The composite signature is made up of a minimum number of data points obtained from five analytical techniques. Identification was accomplished by matching the composite signature of the unknown against a file of composite signatures. In a sequential search, candidates were selected by obtaining the logical intersection of the individual signature matches.

To determine the feasibility of the above concept a data base of 500 representative organic compounds was established and a search experiment was conducted. The 500 compounds represented 23 functional chemical groups and were selected from 838 compounds for which data were collected. The data consisted of five signatures obtained from the following analytical techniques:

1. Mass spectrometry
2. Infrared spectrometry
3. Nuclear magnetic resonance
4. Gas chromatography
5. Ultraviolet spectrophotometry.

The 500 test compounds were selected on the basis of maximum data availability. One hundred compounds had data on five signatures, 174 on 4 signatures, 195 on 3 signatures, and 31 on 2 signatures. Signature availability and distribution among chemical groups are summarized in Tables 1 to 5 and a list of the 500 compounds appears in Appendix A.

The signature data were collected primarily from published sources. Standard measurements and classifications were used when available. Gas chromatography signatures for more than 100 compounds were obtained from laboratory measurements.

The search experiment was designed to determine the number of data points to be associated with each signature, the tolerances that would be required to accommodate instrument variability, and the match criteria that would facilitate positive identification.

The match criteria used in the first part of the experiment were as follows:

MS: Two out of four of the most intense peaks

IR: Two out of three of the most intense bands

NMR: The most intense peak

GC: Relative retention time of either of two columns:
Carbowax 20M or Silicone SE-30

UV: Strongest absorption band.

A search on state: solid, liquid, or gas, and on elements was conducted prior to the signature searches to facilitate the elimination of compounds from the candidate list.

Three categories of input compounds were tested. Input 1 consisted of 100 known compounds having all five signatures, the data of which were included in the file. Input 2 consisted of five compounds for which data were available but not included in the data file. Input 3 consisted of 11 compounds for which data on each of the five signatures were generated in IITRI laboratories and data in the file were obtained from published sources.

Several search measures were defined to aid in the evaluation of feasibility. Discrimination is the capability to separate sets of compounds on the basis of well-defined characteristics of a signature. Selectivity is the screening capability of a sequential search. Precision is the accuracy of identification. Because data were not uniformly available for all compounds in all signatures, data availability was a factor in the measurements.

The availability ratio of a signature is the ratio of the number of compounds having data to the total number of compounds in the data base. The default ratio is one minus the availability ratio. Signature availability and default ratios for 100 test compounds are listed in Table 12. A match by default occurs when input data for a given signature of an unknown are available and a stored compound does not have data for the signature. The discrimination factor is the product of the number of matches obtained in a signature search multiplied by the default ratio.

Discrimination factors for the signatures obtained by searching for the 100 Input 1 compounds are listed in Table 13. Infrared was the most discriminating signature with an average of 30.39 matches per search out of 470 available signatures. Mass spectrometry was second with an average of 49.83 matches per search out of 478 available signatures. Gas chromatography was third with an average of 11.74 matches per search out of 308 available signatures. Nuclear magnetic resonance was fourth and ultraviolet photospectrometry was the least discriminating.

Searches ordered on the basis of data availability provided more rapid screening than searches based on the number of data points to be matched. Accordingly, the order of search in all the experiments was:

1. State
2. Elements
3. MS
4. IR
5. NMR
6. GC
7. UV

The search experiment was designed for processing on a computer but was conducted using an inverted file system in which data were drilled on cards (Termatrex system). Matches were established by optical coincidence.

Ratings were used to evaluate the probability that an unknown compound was on the candidate list in accordance with weights assigned to the signatures and with the closeness of the match. The maximum ratings assigned to the signatures and to the state and elements searches were:

MS	40
IR	36
NMR	24
GC	18 (9 points per column)
UV	6
Elements	4
State	2
Total	130

A candidate list rating table showing the scoring procedure appears in Appendix F.

The most critical measure in establishing feasibility is precision. Of the 100 known Input 1 compounds, none were rejected in the search process. Using the match criteria listed above, 68 compounds were uniquely identified by the seven-stage search. The average number of surviving candidates for the 100 searches was 1.65 (see Table 22). Sixty-five extraneous candidates appeared on the candidate lists of the 32 compounds with multiple candidates, ranging from 1 to 5 extra candidates. The aldehydes (8 compounds) had the lowest average of surviving candidates, 1.12, and the hydrocarbons had the highest average, 2.65. The alcohols with an average of 1.63 candidates ranked second highest.

Only 2 of the 65 extraneous compounds had data on all signatures (see Table 24). Eleven had one signature missing, 45 had 2 signatures missing, and 7 had 3 signatures missing. Ratings for the candidates ranged from 34 to 95 compared to maximum possible ratings of 117 to 130. The unavailability of GC and NMR signatures contributed in large measure to the low precision in the alcohols and hydrocarbons. Two compounds appeared as candidates on seven compound lists.

The search for the five Input 2 candidates resulted in three cases of perfect precision, that is, no candidate compounds survived. One search resulted in the candidacy (score of 28) of a closely related compound but data on three signatures were unavailable. The search for a hydrocarbon obtained six candidates.

Nine of the 11 Input 3 compounds were uniquely identified. A hydrocarbon had four extraneous candidates. Only one compound was erroneously rejected because of failure to match on infrared. It is suspected that the compound's file data was in error because of a significant discrepancy between the published data and IITRI laboratory data.

To improve precision, the following modified match criteria were imposed on the 32 Input 1 compounds that had multiple candidates:

1. Both olumns in GC were required

2. Three out of four MS peaks were required, and the matches on permuted input peak 1 against stored peak 4 and input peak 4 against stored peak 1 were eliminated

3. Two out of two NMR peaks were required.

The modified match criteria eliminated 51 of the 65 extraneous candidates leaving only 8 test compounds with multiple candidates. Of these 8 compounds, 2 had one extraneous and 6 had 2 extraneous candidates. Seven of the test compounds were hydrocarbons and one was an alcohol.

In summary:

- 92 of 100 known compounds were uniquely identified, and none of the 100 were erroneously rejected
- 2 of 5 "unknown" compounds were not rejected when they should have been, but the surviving candidates were closely related
- 10 of 11 compounds whose signatures were measured in IITRI laboratories were identified, 9 of them uniquely; one compound was erroneously rejected because of suspected error in the published data.

The high precision attained and the absence of erroneous rejections, especially when signature data were completely available, indicate that identification of organic compounds by the matching of composite chemical signatures is feasible. The screening of a large data file in a sequential search is effective and is readily amenable to computer operations.

The ratings associated with the closeness of matches and the weights assigned to each signature provided a measure for ranking the candidates. Multiple candidates identifications can act as indicators of chemical groups and can, therefore, provide clues for unique identification.

Refinement of the match criteria, signature parameters, and tolerance limits can be utilized to obtain high precision in an enlarged data file.

The importance of obtaining reliable data from standardized analytical techniques was recognized early in the program and the test results corroborate this view. The diversity of indexing between techniques and the lack of coordination among them poses many problems in developing a composite signature. The establishment of a data file along the lines described in this report can serve as an impetus for relating the techniques and coordinating the data sources.

VI. RECOMMENDATIONS

Our investigations have given us an awareness of the shortcomings of existing analytical data information files and the contribution of our proposed computerized file system can make to overcoming these. The principal weaknesses are: (1) that existing files are not tied together in any way, and (2) that many of the compounds for which one type of data have been collected, e.g., NMR spectra, are not found in data collections of other types. This is the reason for the large number of defaults in our present file.

The input for such an operational system should come from existing compilations and an attempt will be made to coordinate the efforts of those organizations producing such compilations so that data gaps (defaults) will be minimized in the future. Thus, we offer the following recommendations pertaining to the development of our system and its utilization by the scientific community for organic compound identification and as an information file.

1. A computerized system for storing, searching, and retrieving compound identification data should be established in a subsequent phase of research.

2. The present file of 500 organic compounds should be expanded to a minimum of 3,000 compounds in the second phase. This and the first recommendation constitutes a pilot scale study to establish the mechanics for operating a larger file. A computerized system of 3,000 compounds would constitute a workable and useful system. Actual application of the file to solving users identification problems and information needs can begin after the pilot study.

3. The mode of operation of the system should be centralized with respect to data storing, searching and retrieving. However, data collecting points might be located near major data sources such as Sadtler,

ASTM and NBS. The centralized mode for operating the file will offer better control over and coordination of the data file and its operations.

4. Where possible, data for all signatures should be obtained for all compounds. Our past research pointed out the need for complete data files. For the final operating system, this will require generation of some data mainly in gas chromatography, nuclear magnetic resonance and to a lesser extent in mass spectrometry. Generating these data should be undertaken by existing data-generating organizations. For the pilot study some gas chromatographic data may have to be generated.

5. The investigation of the employment of automatic data reduction and digitization is recommended. Equipment of this nature will speed data processing and improve the accuracy of the data.

6. The information file for organic chemical signatures will obtain data from ongoing data operations such as NBS, ASTM and various privately-operated data services. It is not our objective to provide complete spectra or yet another competing data compilation. Our operation will supply the user with a compound identification and references directing him to the original or primary data source if he wishes to obtain the complete spectra. In this way we hope to increase the usefulness of presently

existing data services.

7. Our information file will be, in essence, a subsystem of American Chemical Society's discipline-based chemical information system. It can be so classified because of the tie-in of each compound in our file to Chemical Abstract Services' information service through the registry number. CAS has provided, and promised to provide in the future, registry numbers for all compounds used in our file making our ACS subsystem entirely feasible.

8. Although we have included ultraviolet spectrophotometric data in our signatures it is recommended that, in future developmental work, the UV signature be excluded in favor of a more complete file of signatures for each compound. We feel that while some discrimination was shown by the UV signature, its elimination can be compensated for by more complete data for the other four signature techniques.

9. A survey of a representative sampling of potential users should be conducted to determine the level of interest in and financial support of our information file and the growth of usage we could expect for the operating system.

APPENDIX A

500 COMPOUNDS IN DATA BASE

KEY

D	= Data available
T	= Transparent
A	= GC Column A data available
B	= GC Column B data available
AB	= GC data for both columns available
(Blank)	= No data available

*Preceding compound name designates Input 1 test compound used in search experiment

IITRI No.		Compound (CA name)	Availability of Spectra				
			UV	MS	GC	IR	NMR
1		Methane (same)	T	D		D	D
2	*	Ethane (same)	T	D	AB	D	D
3		n-Propane (same)	T	D		D	
4		n-Butane (same)	T	D	AB	D	
6	*	n-Hexane (same)	T	D	AB	D	D
7	*	n-Heptane (same)	T	D	AB	D	D
9	*	n-Nonane (same)	T	D	AB	D	D
10	*	n-Decane (same)	T	D	AB	D	D
12	*	n-Hexadecane	T	D	AB	D	D
13		n-Heptadecane	T	D		D	
14		n-Octadecane	T	D	B	D	
16		n-Eicosane (same)	T	D		D	
17		Isobutane (2-Methylpropane)	T	D		D	
18		2-Methylheptane	T	D	AB		
19		3-Methylheptane	T	D	AB	D	
25		2,5-Dimethylhexane	T	D		D	

IITRI No.		Compound (CA name)	Availability of Spectra UV	MS	GC	IR	NMR
28		2,3,4-Trimethylpentane	T	D		D	
29		2,3,3-Trimethylpentane	T	D			
34		3,4-Dimethylhexane	T	D	AB	D	
36		Squalane (2,6,10,15,19,23-Hexamethyltetracosane)	T	D		D	
37		Isobutene (2-Methylpropene)		D		D	D
40		Isoprene (same)		D	AB	D	D
41		1-Hexene		D	AB	D	D
42		trans-2-Hexene		D	A	D	D
47		Butadiene (1,3-Butadiene)		D		D	
48		Cyclopropane (same)		D		D	D
50	*	Cyclopentane	T	D	A	D	D
51	*	Cyclohexane (same)	T	D	B	D	D
52	*	Cycloheptane	T	D	AB	D	D
53	*	cis-Decalin (cis-Decahydronaphthalene)	T	D	AB	D	D
54		trans-Decalin (trans-Decahydronaphthalene)	T	D		D	D
56	*	Methylcyclohexane	T	D	AB	D	D

IITRI No.		Compound (CA name)	Availability of Spectra				
			UV	MS	GC	IR	NMR
57		Cyclohexene		D	AB	D	D
59		Dicyclopentadiene (3a,4,7,7a-Tetrahydro-4,7-methanoidene)		D	AB	D	
60	*	Benzene (same)	D	D	AB	D	D
61	*	Toluene (same)	D	D	AB	D	D
62	*	o-Xylene (same)	D	D	AB	D	D
63	*	m-Xylene (same)	D	D	AB	D	D
64	*	p-Xylene (same)	D	D	AB	D	D
65	*	Ethylbenzene (same)	D	D	AB	D	D
66		Vinylbenzene (Styrene)	D	D	AB	D	
67	*	Isopropylbenzene (Cumene)	D	D	AB	D	D
69		p-Isopropyltoluene (p-Cymene)	D	D		D	D
70	*	Naphthalene (same)	D	D	A	D	D
71		Anthracene (same)	D	D		D	
73		Azulene	D	D			
77		Indane	D	D		D	D
78		Tetralin (1,2,3,4-Tetrahydronaphthalene)		D	A	D	D

IITRI No.	Compound (CA name)	Availability of Spectra				
		UV	MS	GC	IR	NMR
79	cis-1,2-Dimethylcyclohexane	T	D		D	D
80	trans-1,2-Dimethylcyclohexane	T	D		D	D
84	Ethylene oxide (same)		D	AB	D	D
85	Propylene oxide (same)		D	AB	D	D
88	Furan (same)		D	AB	D	D
89	Tetrahydrofuran (same)		D	AB	D	D
90	Pyrrole (Azole)	D	D		D	D
91	Pyrrolidine		D		D	D
95	Pyridine (same)	D	D		D	D
96	Piperidine (same)		D	B	D	D
98	3-Methylpyridine	D	D			D
99	4-Methylpyridine (4-Picoline)	D	D			D
101	3-Methylpiperidine		D		D	D
104	Isoquinoline (same)	D	D		D	D
105	Acridine	D	D		D	D
106	Indole	D	D		D	

IITRI No.	Compound (CA name)	Availability of Spectra				
		UV	MS	GC	IR	NMR
107	Skatole (3-Methylindole)		D		D	D
110	Pyrazole	D			D	D
112	Dioxane (p-Dioxane)		D	AB	D	D
113	Imidazole		D		D	D
125	Tryptophan (same)	D		NP	D	D
127	Phenylalanine (same)			NP	D	D
134	α-Alanine (same)			NP	D	
135	Glycine			NP	D	D
146	Caffeine (same)		D	NP		D
147	Morphine (same)		D	NP		
151	Novocain	D		NP		D
154	Cholesterol (same)		D		D	D
164	Progesterone (same)		D		D	
173	α-Pinene (2-Pinene)		D	A	D	D
176	Alloöcimene	D	D		D	
177	β-Pinene (2(10)-Pinene)				D	D

IITRI No.	Compound (CA name)	Availability of Spectra				
		UV	MS	GC	IR	NMR
178	Camphene (same)		D		D	D
179	α-Terpineol (p-Menth-1-en-8-ol)		D	AB	D	
180	Geraniol (3,7-Dimethyl-trans-2,6-octadien-1-ol)		D	A	D	D
181	Citral (3,7-Dimethyl-2,6-octadienal)	D	D	AB	D	
182	Citronellol (3,7-Dimethyl-6-octen-1-ol)		D		D	
183	Linaloöl (3,7-Dimethyl-1,6-octadien-3-ol)		D	AB	D	
184	Menthone (p-Menthan-3-one)		D	AB	D	
185	Menthol (same)		D	AB	D	D
194	Methylene dichloride (Dichloromethane)		D	AB	D	D
195	Chloroform (same)		D		D	D
196	Carbon tetrachloride (same)		D		D	T
197	Chloroethane (same)		D	AB	D	D
198	Trichloroethylene (same)		D	B	D	D
200	1,1,2-Trichloroethane (same)		D	AB	D	D
201	cis-1,2-Dichloroethylene		D	AB	D	D
202	Tetrachloroethylene (same)		D		D	T

IITRI No.		Compound (CA name)	Availability of Spectra				
			UV	MS	GC	IR	NMR
203		Hexachlorobenzene (same)		D		D	T
204		n-Propyl chloride		D	AB	D	
205	*	Chlorobenzene (same)	D	D	AB	D	D
206	*	o-Dichlorobenzene (same)	D	D	AB	D	D
207		m-Dichlorobenzene	D	D	AB	D	
208	*	p-Dichlorobenzene (same)	D	D	AB	D	D
211		Bromoform (Tribromomethane)		D	AB	D	D
213	*	Bromobenzene (same)	D	D	AB	D	D
214		Trichlorofluoromethane (same)		D		D	T
215		Dichlorodifluoromethane (same)		D		D	T
216		Hexafluorobenzene		D		D	T
217		Methyl iodide (Iodomethane)		D	A	D	D
222		Allyl chloride (3-Chloropropene)		D	AB	D	
223	*	Benzylchloride (α-Chlorotoluene)	D	D	A	D	D
225		Benzotrichloride (α,α,α-Trichlorotoluene)		D		D	D
226		Benzal chloride (α,α-Dichlorotoluene)	D	D		D	D

IITRI No.		Compound (CA name)	Availability of Spectra				
			UV	MS	GC	IR	NMR
231		1,5-Dibromopentane (same)		D	A	D	
235		o-Chlorotoluene (same)		D	AB	D	D
236		p-Chlorotoluene (same)	D	D		D	D
239		3-Bromopropyne (same)		D	AB	D	D
240		Bromochloromethane		D	A	D	D
241		1-Iodonaphthalene (same)		D	A	D	
242		2-Iodopropane (same)		D	AB	D	D
243	*	Methanol (same)	T	D	AB	D	D
244	*	Ethanol (Ethyl alcohol)	T	D	AB	D	D
245	*	Propanol (Propyl alcohol)	T	D	AB	D	D
246	*	Butanol (Butyl alcohol)	T	D	AB	D	D
247	*	Pentanol (Pentyl alcohol)	T	D	AB	D	D
249	*	Heptanol (Heptyl alcohol)	T	D	AB	D	D
250	*	Isopropanol (Isopropyl alcohol)	T	D	AB	D	D
251	*	Isobutanol (Isobutyl alcohol)	T	D	AB	D	D
253		2-Methyl-1-butanol (same)	T	D	AB	D	

IITRI No.		Compound (CA name)	Availability of Spectra				
			UV	MS	GC	IR	NMR
254		2-Methyl-2-butanol (tert-Pentyl alcohol)	T	D	AB	D	
255		3-Methyl-2-pentanol	T	D	AB	D	
256		2,2-Dimethyl-1-butanol	T		AB	D	
257		2,3-Dimethyl-2-butanol	T		AB	D	
258	*	Cyclopentanol (same)	T	D	AB	D	D
259	*	Cyclohexanol (same)	T	D	AB	D	D
260		cis-2-Methylcyclohexanol	T	D	AB	D	
262		2-Propyn-1-ol (same)		D	AB	D	D
266		2-Methyl-3-butyn-2-ol (same)		D	AB	D	D
267		3-Methyl-1-pentyn-3-ol (same)		D	AB	D	
268	*	Phenol (same)	D	D	AB	D	D
269		o-Cresol (same)	D	D		D	D
271		p-Cresol (same)	D	D		D	
272		α-Naphthol (1-Naphthol)	D	D		D	D
273		β-Naphthol (2-Naphthol)	D	D		D	
274		2,4-Dimethylphenol (2,4-Xylenol)	D	D			

IITRI No.		Compound (CA name)	Availability of Spectra UV	 MS	 GC	 IR	 NMR
276		2,5-Dimethylphenol (2,5-Xylenol)	D	D		D	
277		3,5-Dimethylphenol (3,5-Xylenol)	D	D		D	D
278	*	Octanol (Octyl alcohol)	T	D	AB	D	D
279	*	Nonanol (Nonyl alcohol)	T	D	AB	D	D
281		Ethylene glycol (same)	T	D	AB	D	
282	*	1,2-Propanediol (same)	T	D	AB	D	D
283		1,3-Propanediol	T	D	AB	D	
284	*	Diethylene glycol	T	D	AB	D	D
285		1,2-Butanediol	T	D	AB	D	
286		1,4-Butanediol (same)	T	D	AB	D	
288		Catechol (Pyrocatechol)	D	D		D	D
289		Resorcinol (same)	D	D		D	D
291		2,3-Dimethyl-2,3-butanediol (same)		D	AB	D	
293		2,4-Pentanediol	T	D	AB	D	
294	*	Acetone (same)	D	D	AB	D	D
295	*	Methyl ethyl ketone (2-Butanone)	D	D	AB	D	D

IITRI		Compound	Availability of Spectra				
No.		(CA name)	UV	MS	GC	IR	NMR
296	*	2-Pentanone (same)	D	D	AB	D	D
297		3-Pentanone (same)		D	AB	D	D
298		3-Methyl-2-butanone		D	AB		
299		2-Hexanone		D	AB	D	
300		3-Hexanone	D	D	AB	D	
301		3-Methyl-2-pentanone		D	AB	D	
302		2-Heptanone (same)		D	AB	D	D
304		4-Heptanone (same)	D	D	AB	D	
305	*	Cyclopentanone (same)	D	D	AB	D	D
306	*	Cyclohexanone (same)	D	D	AB	D	D
307		3-Buten-2-one (same)		D	AB	D	
308		5-Hexen-2-one (same)		D	AB	D	
309		3-Methyl-3-buten-2-one		D	AB	D	
310		2,3-Butanedione	D	D	AB	D	
312	*	2,4-Pentanedione (same)	D	D	AB	D	D
313	*	Acetophenone (same)	D	D	A	D	D

IITRI No.		Compound (CA name)	Availability of Spectra				
			UV	MS	GC	IR	NMR
314	*	Propiophenone (same)	D	D	AB	D	D
315		Benzophenone (same)	D			D	D
317		Benzil (same)	D			D	D
318		p-Benzoquinone (same)	D	D		D	D
320		Formaldehyde (same)	D	D	AB	D	
321	*	Acetaldehyde (same)	D	D	AB	D	D
322	*	Propionaldehyde (same)	D	D	AB	D	D
323	*	Butyraldehyde (same)	D	D	AB	D	D
324		Valeraldehyde (same)		D	AB	D	
325		Hexanal		D	AB	D	
326		Heptanal (same)		D	AB	D	
327		Isobutyraldehyde (same)		D	AB	D	D
328		Isovaleraldehyde (2-Methylbutyraldehyde)		D	AB	D	D
331		2-Ethylhexanal (same)		D	AB	D	
332		Acrolein (same)		D	AB	D	
333		Methacrolein (Methacrylaldehyde)		D	AB	D	

IITRI No.		Compound (CA name)	Availability of Spectra				
			UV	MS	GC	IR	NMR
336		2,4-Hexadienal (Sorbaldehyde)	D	D			
337	*	Crotonaldehyde (same)	D	D	AB	D	D
339	*	Benzaldehyde (same)	D	D	A	D	D
340	*	Furfural (2-Furaldehyde)	D	D	AB	D	D
344		Paraldehyde (same)		D	AB	D	D
347		Formic acid (same)	D	D		D	D
349		Propionic acid (same)		D		D	D
350		Butyric acid (same)		D		D	D
351		Valeric acid (same)		D		D	D
352		Hexanoic acid (same)		D		D	D
353		Heptanoic acid (same)		D		D	D
357		Benzoic acid (same)	D	D		D	D
358		Phenylacetic acid (same)	D	D		D	D
359		o-Toluic acid (same)	D	D		D	D
360		m-Toluic acid (same)	D	D		D	D
361		p-Toluic acid (same)	D	D		D	D

IITRI No.	Compound (CA name)	Availability of Spectra				
		UV	MS	GC	IR	NMR
362	β-Naphthoic acid (2-Naphthoic acid)	D	D		D	
365	Succinic acid (same)		D		D	D
367	Adipic acid (same)		D		D	D
368	o-Phthalic acid (Phthalic acid)	D	D		D	D
369	m-Phthalic acid (Isophthalic acid)		D		D	D
370	Terephthalic acid (same)	D	D		D	D
373	Methacrylic acid (same)		D		D	D
374	2,4,6-Mesitylenecarboxylic acid		D			D
375	Stearic acid (same)	D	D		D	
379	Methyl formate (Formic acid, methyl ester)		D	AB	D	D
381	Propyl formate (Formic acid, propyl ester)		D	AB	D	
382	Butyl formate		D	AB	D	
384	Isobutyl formate		D	AB	D	
385	Methyl acetate (Acetic acid, methyl ester)		D	AB	D	D
387	Propyl acetate (Acetic acid, propyl ester)		D	AB	D	D
388	Butyl acetate (Acetic acid, butyl ester)		D	AB	D	D

IITRI No.	Compound (CA name)	Availability of Spectra UV	MS	GC	IR	NMR
389	Isopropyl acetate (Acetic acid, isopropyl ester)		D	AB	D	D
390	Isobutyl acetate (Acetic acid, isobutyl ester)		D	AB	D	
391	Methyl propionate		D	AB	D	D
392	Ethyl propionate (Propionic acid, ethyl ester)		D	AB	D	
393	Propyl propionate (Propionic acid, propyl ester)		D	AB	D	D
395	Isopropyl propionate		D	AB	D	
396	Isobutyl propionate		D	AB	D	D
398	sec-Butyl acetate (Acetic acid, sec-butyl ester)		D	AB	D	
402	Ethylidene diacetate		D	AB		
403	Vinyl acetate		D	AB	D	D
404	Allyl acetate		D	AB	D	
405	Vinyl butyrate (Butyric acid, vinyl ester)		D	AB	D	
407	Methyl methacrylate (Methacrylic acid, methyl ester)		D	AB	D	D
410	Allyl propionate		D	AB	D	
411	Isopropenyl acetate		D	AB	D	D
412	2-Ethyl-1-hexyl acetate (Acetic acid, 2-ethylhexyl ester)		D	AB	D	

IITRI No.		Compound (CA name)	Availability of Spectra UV	 MS	 GC	 IR	 NMR
415		Isopentyl propionate		D	AB	D	D
418	*	Methyl benzoate (Benzoic acid, methyl ester)	D	D	A	D	D
419		Ethyl benzoate (Benzoic acid, ethyl ester)	D	D		D	D
421		Dimethyl-o-phthalate (Phthalic acid, dimethyl ester)	D	D	A	D	D
423		Dimethyl-p-phthalate	D	D	A	D	
424		Diethyl-p-phthalate	D	D		D	
425	*	Diethyl malonate (Malonic acid, diethyl ester)	D	D	AB	D	D
426		Diethyl succinate (Succinic acid, diethyl ester)		D	AB	D	D
427		Diethyl oxalate (Oxalic acid, diethyl ester)		D	AB	D	
432		Ethyl methacrylate (Methacrylic acid, ethyl ester)		D		D	D
434		Ethyl acrylate (Acrylic acid, ethyl ester)		D	AB	D	D
435		Methyl cinnamate (Cinnamic acid, methyl ester)	D	D		D	D
438		Methyl-o-toluate (o-Toluic acid, methyl ester)		D			D
439		Methyl-p-toluate (p-Toluic acid, methyl ester)		D		D	D
440		Methyl-m-toluate (m-Toluic acid, methyl ester)		D		D	D
441		Dimethyl ether (Methyl ether)		D	AB	D	D

IITRI No.	Compound (CA name)	Availability of Spectra UV	MS	GC	IR	NMR
443	Diethyl ether (Ethyl ether)		D	AB	D	D
444	Methylpropyl ether		D	AB		
445	Dipropyl ether (Propyl ether)		D	AB	D	D
446	Butylethyl ether		D	AB	D	
449	Isopropyl ether (same)		D	AB	D	D
450	Allylethyl ether		D	AB	D	
452	Ethylvinyl ether (same)		D	AB	D	
453	Butylvinyl ether (same)			AB	D	
454	2-Ethyl-1-hexyl vinyl ether (2-Ethylhexyl vinyl ether)		D	AB	D	
456	Diallyl ether		D	AB	D	D
457	Dibenzyl ether (Benzyl ether)	D	D		D	D
458	* Diphenyl ether (Phenyl ether)	D	D	AB	D	D
459	* Anisole (same)	D	D	AB	D	D
464	Butylamine (same)		D		D	D
466	Hexylamine (same)		D		D	D
467	Dimethylamine (same)		D		D	

IITRI	Compound	Availability of Spectra				
No.	(CA name)	UV	MS	GC	IR	NMR
468	Diethylamine (same)		D		D	D
471	Triethylamine (same)		D	AB	D	D
472	Trimethylamine (same)		D		D	D
475	Allylamine (same)		D		D	D
476	Aniline (same)	D	D		D	D
477	N-Methylaniline (same)	D	D		D	
478	N,N-Dimethylaniline (same)	D	D		D	D
480	o-Toluidine (same)	D	D		D	D
484	m-Phenylenediamine (same)	D			D	D
485	p-Phenylenediamine (same)	D	D		D	D
486	Diphenylamine (same)	D			D	D
490	o-Tolidine (3,3'-Dimethylbenzidine)	D			D	D
491	Formamide (same)		D		D	D
492	Acetamide (same)		D		D	D
493	Propionamide (same)		D			D
494	Butyramide		D		D	

IITRI No.		Compound (CA name)	Availability of Spectra UV	 MS	 GC	 IR	 NMR
498		Benzamide	D	D		D	D
501		Dimethylformamide (N,N-Dimethylformamide)		D		D	D
515		Methacrylamide (same)				D	D
516		N-Methylformanilide	D	D		D	
522	*	Benzonitrile (same)	D	D	A	D	D
524		Acetonitrile (same)		D	AB	D	D
525		Propionitrile (same)		D	AB	D	D
526		Dicyanopropane		D		D	D
527	*	p-Tolunitrile	D	D	AB	D	D
528	*	o-Tolunitrile	D	D	AB	D	D
529		Valeronitrile (same)		D	AB	D	
530		3-Butenenitrile (same)		D	A	D	D
531		Acrylonitrile (same)		D		D	D
532		Malonitrile (same)		D	A	D	
534	*	Nitromethane (same)	D	D	AB	D	D
535	*	Nitroethane (same)	D	D	AB	D	D

IITRI No.		Compound (CA name)	Availability of Spectra				
			UV	MS	GC	IR	NMR
536	*	Nitropropane (1-Nitropropane)	D	D	AB	D	D
537	*	2-Nitropropane (same)	D	D	AB	D	D
540	*	Nitrobenzene (same)	D	D	AB	D	D
542		o-Nitrotoluene		D	A	D	D
544	*	p-Nitrotoluene (same)	D	D	AB	D	D
545		m-Nitrotoluene (same)	D			D	
547		1-Nitronaphthalene (same)	D			D	D
557		Propylmercaptan (1-Propanethiol)		D		D	
560		Butylmercaptan (1-Butanethiol)		D	AB	D	D
562	*	Thiophenol (Benzenethiol)	D	D	AB	D	D
567		Dimethyl sulfide (Methyl sulfide)	D	D		D	D
568	*	Diethyl sulfide	D	D	AB	D	D
569	*	Ethyl disulfide (same)	D	D	AB	D	D
572		tert-Butyl sulfide (same)	D	D	AB	D	
573	*	Carbon disulfide (same)	D	D	AB	D	T
574	*	Diphenyl sulfide (Phenyl sulfide)	D	D	A	D	D

IITRI No.	Compound (CA name)	Availability of Spectra				
		UV	MS	GC	IR	NMR
575	Phenyl disulfide		D		D	D
578	Dimethyl sulfoxide (Methyl sulfoxide)		D	A	D	D
587	Dimethyl sulfone		D		D	
594	2-Chloroethanol		D	AB	D	D
595	3-Hydroxy-2-butanone	D	D		D	
600	2-Methoxyethylvinyl ether		D	AB	D	
610	Phthalic anhydride (same)	D	D		D	
611	Succinic anhydride (same)		D		D	D
613	Acetyl chloride				D	D
618	Trichloroacetic acid		D		D	D
626	o-Nitrophenol (same)	D	D	A	D	
628	o-Methoxybenzaldehyde (o-Anisaldehyde)		D	A	D	D
633	Chloral (same)		D		D	
655	Tetraethyl lead		D		D	
658	Acetal (Acetaldehyde, diethyl acetal)		D	AB	D	D
659	Ketal (Acetone, dimethyl acetal)		D	AB		

IITRI No.		Compound (CA name)	Availability of Spectra UV	MS	GC	IR	NMR
661		Isobutyric anhydride (same)		D		D	D
662		Benzoic anhydride (same)	D	D		D	D
663		Maleic anhydride (same)	D	D		D	D
664	*	Coumarin (same)	D	D	B	D	D
665		Chloroacetic acid (same)		D		D	D
666		Dichloroacetic acid (same)		D		D	D
667		m-Nitrobenzoic acid (same)	D	D		D	D
668		Diethoxymethane		D	AB	D	D
670	*	o-Methoxyphenol (same)	D	D	AB	D	D
671	*	m-Methoxyphenol (same)	D	D	A	D	D
672	*	p-Methoxyphenol (same)	D	D	A	D	D
673		3-Bromopropionitrile		D		D	D
674		3-Bromopropionic acid		D		D	D
675		2-Bromobutyric acid (same)		D		D	D
676		2-Bromo-4-chlorophenol	D	D		D	D
677		p-Bromophenol (same)	D	D		D	D

IITRI		Compound	Availability of Spectra				
No.		(CA name)	UV	MS	GC	IR	NMR
678		p-Chlorobenzonitrile	D	D			
679		m-Chlorophenol (same)	D	D		D	D
680	*	1-Chlorophenol	D	D	A	D	D
681	*	Salicylaldehyde (same)	D	D	A	D	D
682		3-Ethoxy-4-hydroxybenzaldehyde (same)	D	D		D	D
683		2-Furan acrolein		D		D	D
684	*	2,4-Dichlorobenzaldehyde	D	D	A	D	D
685	*	p-Anisaldehyde (same)	D	D	A	D	D
687	*	p-Chlorobenzaldehyde (same)	D	D	AB	D	D
688		3-Methoxy-1-butanol			AB	D	D
689		2-Ethoxyethanol (same)		D	AB	D	
690		2-Butoxyethanol (same)		D	AB	D	
691		2-Methoxy-1-propanol		D	AB		
692		1-Methoxy-2-propanol (same)		D	AB	D	
693		2-Ethoxyethyl acetate (2-Ethoxyethanol, acetate)		D	AB	D	
695		Dipropyl acetal (Acetaldehyde, dipropyl acetal)		D	AB	D	

IITRI No.		Compound (CA name)	Availability of Spectra UV	MS	GC	IR	NMR
698	*	4-Hydroxy-4-methyl-2-pentanone (same)	D	D	B	D	D
699		2-Methoxyethanol (same)		D	AB	D	D
700		bis(2-Ethoxyethyl)ether		D	AB	D	D
701	*	2,3-Dimethylbutane (same)	T	D	AB	D	D
702		o-Diethylbenzene		D	AB	D	D
703		p-Diethylbenzene (same)		D	AB	D	D
704	*	m-Diethylbenzene (same)	D	D	AB	D	D
705		2,2-Dimethylbutane (same)	T	D	AB		
706		1-Methylcyclohexene (Methylcyclohexene)		D		D	D
707		4-Vinyl-1-cyclohexene (4-Vinylcyclohexene)		D	AB		D
711		3-Phenylpropene	D	D		D	D
712		Isopentane (2-Methylbutane)	T	D		D	D
713		2,4-Dimethylpentane (same)	T	D		D	D
714		2-Methylpentane (same)	T	D		D	D
715	*	3-Methylpentane (same)	T	D	AB	D	D
716		2,3-Dimethyl-1-butene		D		D	D

IITRI	Compound	Availability of Spectra				
No.	(CA name)	UV	MS	GC	IR	NMR
717	2-Methyl-2-butene		D		D	D
718	* Biphenyl (same)	D	D	B	D	D
720	3,5-Dimethylpyrazole		D		D	D
721	2,6-Dimethylpyridine	D	D		D	D
722	4-Vinylpyridine		D		D	D
723	2-Methylfuran		D	AB	D	D
726	Chlorocyclohexane		D		D	D
727	1-Chlorohexane		D		D	D
728	1,2-Dichloropropane (same)		D		D	D
729	1,3-Dichloropropane (same)		D	AB	D	D
732	1,2,4-Trichlorobenzene (same)		D	A	D	D
733	Bromocyclohexane (same)		D		D	D
734	2-Bromobutane (same)		D	AB	D	D
735	1-Bromo-3-methylbutane (same)		D		D	D
736	1,2-Dibromobutane		D		D	D
737	Bromoethane (same)		D	AB	D	D

IITRI No.	Compound (CA name)	Availability of Spectra UV	 MS	 GC	 IR	 NMR
738	1-Bromopentane (same)		D	AB	D	D
739	1-Bromopropane (same)		D	AB	D	D
740	2-Bromoprcpane (same)		D	AB	D	D
741	2-Bromo-2-methylpropane (t-Butyl bromide)		D		D	D
742	1,2-Dibromopropane (same)		D		D	D
743	1,3-Dibromopropane (same)		D		D	D
744	2-Bromo-1-propene		D		D	D
745	2,3-Dibromo-1-propene		D		D	D
746	Iodocyclopentane		D		D	D
747	1-Iodobutane		D		D	D
748	2-Iodobutane		D		D	D
749	Iodoethane (same)		D	AB	D	D
750	1-Iodopropane (same)		D	AB	D	D
751	Iodobenzene		D		D	D
754	1-Bromo-2-chlorobenzene		D		D	D
755	1-Bromo-4-chlorobenzene		D		D	D

IITRI No.		Compound (CA name)	Availability of Spectra UV	MS	GC	IR	NMR
757	*	2-Pentanol (same)	T	D	AB	D	D
758	*	2,2-Dimethyl-1-propanol (same)	T	D	AB	D	D
759	*	2-Ethyl-1-butanol (same)	T	D	AB	D	D
760	*	3-Heptanol	T	D	AB	D	D
761	*	2-Ethyl-1-hexanol (same)	T	D	AB	D	D
763		2-Hexanol	T	D	AB	D	
764		3-Hexanol	T	D	AB	D	
765		2-Methyl-1-pentanol (same)	T	D	AB	D	
767		4-Methyl-2-pentanol	T	D	AB	D	
768		2-Methyl-3-pentanol	T	D	AB	D	
769		3,3-Dimethyl-2-butanol	T	D	AB	D	
770		2-Heptanol	T	D	AB	D	
771		4-Heptanol	T	D	AB	D	
772		2,2-Dimethyl-1-pentanol	T	D	AB	D	
773	*	2-Octanol (same)	T	D	AB	D	D
774		2-Methyl-2-propen-1-ol		D	AB	D	

IITRI		Compound	Availability of Spectra				
No.		(CA name)	UV	MS	GC	IR	NMR
775		2,6-Dimethylphenol		D		D	D
776		Cyclopentylmethanol	T	D			D
777		Benzyl alcohol (same)		D	A	D	D
778		t-Butanol	T		AB	D	D
779		2,5-Hexanediol	T	D		D	D
780		2-Methyl-2,4-pentanediol (same)	T	D	AB	D	
781	*	1,3-Butanediol (same)	T	D	AB	D	D
783	*	2,3-Butanediol	T	D	AB	D	D
785		4-Methyl-2-pentanone (same)		D	AB	D	D
786		2-Nonanone		D	AB	D	D
787	*	Camphor (same)	D	D	AB	D	D
788		4-Methylcyclohexanone	D	D		D	D
789	*	Butyrophenone	D	D	A	D	D
791	*	Cinnamaldehyde (same)	D	D	AB	D	D
792		o-Tolualdehyde		D	A		D
793		2-Ethylbutanal		D	AB	D	D

IITRI No.	Compound (CA name)	UV	MS	GC	IR	NMR
		Availability of Spectra				
794 *	Piperonal (same)	D	D	AB	D	D
795	Cyclohexanecarboxylic acid (same)		D		D	D
796	2-Methylbutyric acid (same)		D		D	D
797	Isobutyric acid (same)		D		D	D
798	Senecioic acid		D		D	D
799	Pentyl formate		D	A	D	
800	Hexyl formate		D	AB	D	
801	Allyl formate		D	AB	D	
803	Heptyl acetate (Acetic acid, heptyl ester)		D	AB	D	
804	Cyclohexyl acetate		D	AB	D	
805	Methyl butyrate		D	AB	D	
806	Propyl butyrate (Butyric acid, propyl ester)		D	AB	D	D
807	Isopropyl butyrate		D	AB	D	
808	Butyl butyrate (Butyric acid, butyl ester)		D	AB	D	
809	Pentyl butyrate		D	AB	D	
810	Isopentyl butyrate (Butyric acid, isopentyl ester)		D	AB	D	

IITRI No.		Compound (CA name)	Availability of Spectra				
			UV	MS	GC	IR	NMR
811		Butyl isobutyrate (Isobutyric acid, butyl ester)		D	AB	D	
812		Propyl acrylate			AB	D	
814		Isopentyl acetate		D	AB	D	D
815		Hexyl acetate (Acetic acid, hexyl ester)		D	AB	D	D
816		Pentyl propionate		D	AB	D	D
817		Ethyl butyrate (Butyric acid, ethyl ester)		D	AB	D	D
818		Benzyl benzoate (Benzoic acid, benzyl ester)	D	D		D	D
819	*	Butyl benzoate (Benzoic acid, butyl ester)	D	D	AB	D	D
820		Pentyl ether		D	AB	D	D
821		Butylmethyl ether		D	AB	D	
822		Butyl ether (same)		D	AB	D	
823		Hexyl ether (same)		D	AB	D	
824		Isobutylvinyl ether (same)		D	AB	D	
825		Ethoxybenzene	D	D			
826		p-Ethoxytoluene	D	D		D	D
827		N,N-Dimethylcyclohexylamine (same)		D		D	D

IITRI No.	Compound (CA name)	Availability of Spectra				
		UV	MS	GC	IR	NMR
828	tert-Butylamine (same)		D		D	D
829	sec-Butylamine		D		D	D
830	Isobutyronitrile (same)		D	AB	D	D
831	Butyronitrile (same)		D	AB	D	D
832	Fumaronitrile		D		D	D
833	Hexanenitrile		D	AB	D	D
834	4-Methylpentanenitrile		D		D	D
835	Phenylacetonitrile (same)		D	A	D	D
836	Oleic acid (same)		D		D	D
837	m-Tolunitrile		D	A	D	D
838	Benzylmercaptan (α-Toluenethiol)		D	AB	D	D
839	2-Propanethiol		D		D	D
840	1,4-Butanedithiol		D		D	D
841	* Butyl sulfide	D	D	AB	D	D
842	2-Methyl-1-butene		D		D	D
843	cis-1,3-Dimethylcyclohexane	T	D			D

IITRI No.	Compound (CA name)	Availability of Spectra				
		UV	MS	GC	IR	NMR
844	cis-1,4-Dimethylcyclohexane	T	D			D
845	trans-1,3-Dimethylcyclohexane	T	D			D
846	trans-1,4-Dimethylcyclohexane	T	D			D
850	trans-2-Methylcyclohexanol	T	D	AB		

APPENDIX B

AVAILABILITY OF SIGNATURE DATA

TABLE 1

MAJOR SPECTRA SOURCES

MS SPECTRA

MAJOR SOURCE	SPECTRA AVAILABLE	FORM OF SPECTRA	COST	SPECTRA SOURCE
ASTM	3,200 spectra	Indexed by strongest peaks	$13 for DS-27	Industrial, Commercial, Literature
		3,200 IBM cards-strongest peak and indices ($96)		
	9,000 (European collection to be added in June)			Miscellaneous Collections

MS DATA COMPILATIONS

American Society for Testing & Materials. Mass Spectral Data Index, DS27. 248 pp., 1964. $13.

Mass Spectral Data
DS27-1a. Mass Spectral Data, 3,200 cards. $96.
DS27-1b. Mass Spectral Name and Formula, 3,500 cards. $115.

Cornu, A., and Massot, R. Compilation of Mass Spectral Data, Index of Spectres de Masse. Heyden and Son Limited, Presses Universitaires de France, 1966.

Garvin, D., and Rosenstock, H. M. National Bureau of Standards. Two National Bureau of Standards Data Centers: Chemical Kinetics and Mass Spectrometry. J. Chem. Doc. 7, No. 1, 31-4, 1967.

Rosenstock, H. M. (NBS Mass Spectrometry Data Center, Washington, D. C.). Information from 1,000 basic documents published since 1955. Inf. Retr. Newsl. 2, No. 2, p. 5, Dec. 1966.

TABLE 2

MAJOR SPECTRA SOURCES

IR SPECTRA

MAJOR SOURCE	SPECTRA AVAILABLE	FORM OF SPECTRA	COST	SPECTRA SOURCE
ASTM	100,000 spectra IR - 2μ to 16μ; growth potential 15,000-20,000/year Indices DS-24 through 10th Supplement available.	Coded spectra No complete spectra	Complete indexed collection, $2,250 - index to collection additional. Available on tapes for IBM 7090 up through the 9th supplement; being prepared for IBM 360 (model 20)	API, MCA, TRC, DMS, NBS, Sadtler, IRDC-Japan. Literature - Domestic and Foreign.
Sadtler Research Laboratories	32,000 spectra	Complete spectra - hard copy or microfilm. All Sadtler spectra identified and indexed by ASTM.	$4,000. Spec-finder additional.	Sadtler spectra.
Coblentz Society	5,000 spectra	Complete spectra only. All Coblentz spectra identified and indexed by ASTM.	10¢/sheet. Sold by Sadtler Research Laboratories.	Industrial and University Research Laboratories.
TRC	Approximately 4,000 spectra	Complete spectra. All TRC spectra identified and indexed by ASTM. TRC index to be available Summer, 1968.	30¢/sheet.	Industrial and University Research Laboratories.

IR DATA COMPILATIONS

American Society for Testing & Materials. Numerical list of abstracted infrared spectra indexed on Wyandotte-ASTM punched cards, December 1961.
First supplement - Jan. 1963
Second supplement - Jan. 1964
Third supplement - Apr. 1965
Fourth supplement - Apr. 1967.

American Society for Testing & Materials. Serial number list of compound names and references, published IR spectra 1963.
Fourth supplement to serial number lists - DS29-S4, April, 1967.

American Society for Testing & Materials
1916 Race Street
Philadelphia, Pennsylvania 19103
Indexes - 100,000 spectra of organic compounds.

Anderson Physical Laboratory
609 South Sixth Street
Champaign, Illinois
Mostly organic IR - 0.7μ to 3.2μ
2 spectra per 8½" x 11" loose leaf sheet
1,200 Anderson Near Infrared Spectra
500 polynuclear compounds.

Coblentz Society. Pure compounds and commercial products.
IR only, 2μ to 30μ
5,000 spectra available on 8-1/2" x 11" sheets

Coblentz Society spectra - obtained from Sadtler Research Laboratories, 3316 Spring Garden Street, Philadelphia 2, Pa.
$100 per 1,000 spectra.

Dobriner, K., et al. Infrared Absorption Spectra of Steroids. Interscience Publishers, Inc., New York. 1953. An Atlas - Vol. I. 308 spectra on Perkin-Elmer Model 21 double beam spectrometers, NACL or Calcium Fluoride prism.

Roberts, G., Gallagher, B. S., and Jones, R. N. Infrared Absorption Spectra of Steroids - An Atlas - Vol. II. 1958. 453 spectra.

Documentation of Molecular Spectroscopy (DMS). Published jointly by Butterworth's Scientific Publications and Verlag Chemie GmbH. Weinheim/Bergstrosser.
Spectra available from:
Spex Industries, Inc.
205-02 Jamaica Avenue
Hollis 23, New York

Fisk University, Nashville, Tennessee (Infrared Spectroscopy Research Laboratory).
Maintains a library of 100 volumes on infrared spectroscopy and gas chromatography.

Hershenson, H. M. IR Absorption Spectra. Index for 1958-1962. Academic Press. 1964.

Jonker Business Machines, Inc., Gaithersburg, Maryland.
ASTM Infrared Optical Coincidence Index.
Data retrieval and correlation for organics in the following IR collections:

API Project #44	1,702 compounds
NRC-NBS	2,495 compounds
Coblentz	1,859 compounds
MCA	170 compounds
IRDC-Japan	1,197 compounds
All hydrocarbons from all other ASTM indexed collections	1,828 compounds

Korobeinicheva, I. K., Petrov, A. K., and Koptyng, V. A. Spectral Atlas for Aromatic and Heterocyclic Compounds No. 1, Infrared and Ultraviolet Absorption Spectra of Polyfluoro Aromatic and Polyfluoro Heterocyclic Compounds. Nauk: Hovosibirsk. 171 pages, 1967.

LeCentre d'Information de l'Infra-rouge (CIR). Organic compounds and Inorganic compounds.
Sources:
Sadtler
DMS
API Research Project #44
Coblentz collections
Institute Francois de Petrole, RNU Renault, Rhone - Poulenc, St. Gobain.
CIR published 200 spectra then discontinued publication.

Ministry of Aviation. An index of published infrared spectra. Her Majesty's Stationery Office, London, 1960.

National Research Council - National Bureau of Standards. Spectra Data Project.
Mostly organic compounds - 2,510 spectra on 8-1/2" x 11" sheets
Project terminated and spectra distributed to universities and not for profit research organizations as of December 1967.

Sadtler Research Laboratories, Inc.
3316 Spring Garden Street
Philadelphia, Pennsylvania 19104
IR - 32,000 complete spectra organic compounds

Szymanski, H. A. (Ed.). Infrared Band Handbook. Plenum Press. 1965. Supplements 1-2, Plenum Press, 1965.

Szymanski, H. A. (Ed.). Infrared Band Handbook and Supplements and Interpreted Infrared Spectra. Plenum Press. New York. 1963. $7.50.

White, R. G. Handbook of Industrial IR Analysis. Plenum Press. 1964.

Wright-Patterson Air Force Base, Ohio. Aliphatic and Aromatic Hydrocarbons, bromohydrocarbons. IR 15μ-35μ. Published as WADC TR 57-359, 57-413, 58-198 and supplement 1 to 58-198.

TABLE 3

MAJOR SPECTRA SOURCES

NMR SPECTRA

MAJOR SOURCE	SPECTRA AVAILABLE	FORM OF SPECTRA	COST	SPECTRA SOURCE
Sadtler Research Laboratories	4,000 spectra	Complete spectra	$800 for collection	Sadtler spectra
TRC	1,260 spectra	Complete spectra	30¢/sheet	Industrial and University Research Labs.

NMR DATA COMPILATIONS

Bhacca, N. S., Johnson, L. F., and Shoolery, J. N. (Instrument Division of Varian Associates). NMR Spectra Catalog, Vol. 1 and 2, National Press, 1962.

Howell, M. G., Kende, A. S., and Webb, J. S. (Eds.). Formula Index to NMR Literature Data, Vol. 2. Plenum Press. New York. 518 pp. 1966. CA 62, 1956a. $22.50.

NMR Data - Chamberlain, privately circulated by ESSO Research and Engineering. Tables of chemical shift values and coupling constants for representative structures.

Sadtler Research Laboratories, Inc.
3316 Spring Garden Street
Philadelphia, Pennsylvania 19104
4,000 NMR spectra available in printed form or microfilm at a cost of $680.

Varian Spectra Catalogs, Vol. 1 and 2. Varian Associates, Palo Alto, California. A total of 700 spectra of representative compounds, indexed by name, by a functional group code, and by chemical shifts.

TABLE 4

MAJOR SPECTRA SOURCES

GC DATA

MAJOR SOURCE	SPECTRA AVAILABLE	FORM OF SPECTRA	COST	SPECTRA SOURCE
ASTM	4,000	Indexed compilation	---	Commercial Industrial Literature

GC DATA COMPILATIONS

American Society for Testing & Materials. Compilation of Gas Chromatographic Data. STP 343. 626 pp. 1963. $20.00. DS25A published 1967 (see reference: Schupp, O. E.).

McReynolds, W. O. Gas Chromatography Retention Data. Preston Technical Abstracts Co., Chicago, Illincis. 1966. $25.

Preston, S. T., Jr., and Gill, M. (Preston Tech. Absts. Co., Chicago, Ill.). Bibliography and index to the literature on gas chromatography. 1966.

Schupp, O. E., III, and Lewis, J. S. Compilation of Gas Chromatographic Data. ASTM Data Series Publ. No. DS25A, 2nd Ed., ASTM, 732 pp. 4,000 compounds. 1967.

Signeur, A. V. Guide to Gas Chromatography Literature. Plenum Press, New York. 359 pp. 1964. $12.50.

TABLE 5

MAJOR SPECTRA SOURCES

UV DATA

MAJOR SOURCE	SPECTRA AVAILABLE	FORM OF SPECTRA	COST	SPECTRA SOURCE
ASTM	25,749 spectra		$754 complete collection Indices additional	Commercial Industrial Literature
Sadtler Research Laboratories	22,000 spectra	Complete spectra 8½" x 11" or microfilm	$2,278	Sadtler Research Laboratories
TRC	1,246 spectra	Complete spectra	30¢/sheet	Industrial and University Labs.

UV DATA COMPILATIONS

American Society for Testing & Materials
1916 Race Street
Philadelphia, Pennsylvania 19103
Indexed through the 7th supplement. 25,749 organic compounds.
Cost $754 for complete index.

Atlas of Organic Compounds. Photoelectric Spectrometry Group and the Institut for Spektrochemie and Angewandte Spektroskopie. Plenum Press, New York. 5 volumes. 1966.

American Society for Testing & Materials. Numerical List of abstracted Ultraviolet and Visible Spectra Indexed on Wyandotte-ASTM punched cards. March 1963. Second supplement to The Numerical List of Abstracted Ultraviolet and Visible Spectra Indexed on Wyandotte-ASTM Punched Cards. September 1966.

Hiragma, K. Handbook of UV and visible spectra of organic compounds. Plenum Press, New York. 616 pp. Data on 8,443 compounds from literature. 1967.

Hershenson, H. M. Ultraviolet and Visible Absorption Spectra. Index 1960-1963. Academic Press, New York. 1966.

Korobeinicheva, I. K., Petrov, A. K., and Koptyng, V. A. Spectral Atlas for Aromatic and Heterocyclic Compounds No. 1, Infrared and Ultraviolet Absorption Spectra of Polyfluoro Aromatic and Polyfluoro Heterocyclic Compounds. Nauk: Novosibirsk. 171 pp. 1967.

Lang, L. (Ed.). Absorption Spectra in the UV and Visible Region. Academic Press, New York. 1966.

Lang, L. (Ed.). Absorption Spectra in the UV and Visible Region, Hungarian Academy of Sciences. 1961-1963.

Sadtler Research Laboratories, Inc.
3316 Spring Garden Street
Philadelphia, Pennsylvania 19104
Printed form, or microfilm available of complete spectra.
22,000 UV spectra available at cost of $2,278.

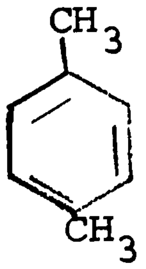
APPENDIX C

DATA MASTER SHEET

p-xylene	106.160	106,423
COMPOUND NAME	MOL. WT.	CAS REG. NO.

TRIVIAL AND TRADE NAMES: Xylol

STRUCTURE:



FORMULA: C_8H_{10}

M.P.: 13.3°C

B.P.: 138.4°C

STATE: Liquid

ANALYTICAL DATA

GC

CONDITION A	CONDITION B
RV_g= 1.63	RV_g= 1.56
I_x= 889	I_x= 1180

Data Source - McReynolds, "GC Retention Data," 1966, pp 51 and 145

MS

MASS:CHARGE:	91	106	105	39	51	77	27	78	50	92
REL. INTENSITY:		624	297	164	162	137	118	82	77	77

DATA SOURCE-API Project 44, #422

NMR ppm: ____ ____ ____ ____ ____ ____ ____ ____ ____ ____

0	1	2	3	4	5	6	7		
-	-	30	-	-	-	-	05		

STRONGEST: 2.3

SOLVENT: $CDCl_3$

DATA SOURCE-Varian, Vol. 1, 1962, #203
(arranged in order of increasing shift value from TMS)

IR MICRONS:

12.6, 6.6, 6.9

Phase - liquid
Cell - 0.01 mm

DATA SOURCE-Sadtler #2276

UV MILLIMICRONS:

$_{max}$ = 2.2 mμ

Solvent - Cyclohexane

DATA SOURCE-Sadtler #2276N

APPENDIX D

CHEMICAL SIGNATURE SEARCH EXPERIMENT

PROJECT C6104

CHEMICAL SIGNATURE SEARCH EXPERIMENT

DATA SHEETS FOR "UNKNOWN" COMPOUNDS

Compound name________________ Number________________

STATE Solid Liquid Gas

ELEMENTS NOH O N Br Cl ___ ___ ___ ___ ___

NMR
SKIP
TRANSPARENT

____________	____________	____________
-0.1 ppm	Max. Peak	+0.1 ppm

UV
SKIP
TRANSPARENT

____________	____________	____________
-2 mμ	Peak	+2 mμ

GC
SKIP
NOT POSSIBLE

Column A:	____________	____________	____________
	* -	Time	* +
Column B:	____________	____________	____________
	-	Time	+

IR
SKIP

Peak 1:	____________	____________	____________
	-0.1 μ	Peak 1	+0.1 μ
Peak 2:	____________	____________	____________
	-0.1 μ	Peak 2	+0.1 μ
Peak 3:	____________	____________	____________
	-0.1 μ	Peak 3	+0.1 μ

MS
SKIP

Peak 1:	____________	Peak 3	____________
Peak 2:	____________	Peak 4	____________

Coder____________________ Date________________

*Magnitude of tolerance based on a sliding scale.

APPENDIX E

CHEMICAL SIGNATURE SEARCH RECORD

PROJECT C6104

Record 1 (#R1)

CHEMICAL SIGNATURES SEARCH RECORD

Unknown ____________________

Experiment ____________________

Date ____________________

Operator ____________________

Routine	Parameter Tally														Search Tally				
															Skip	M Match	D Default	M+D	Cum. Match
State																			
Elements																			
NMR																			
UV																			
GC	NP	Col. A	Col. B	A·B															
IR	1-1	1-2	1-3	2-1	2-2	2-3	3-1	3-2	3-3	1M	2M	3M	Match = 2M + 3M						
MS	NP	1-1	1-2	1-3	1-4	2-1	2-2	2-3	2-4	3-1	3-2	3-3	3-4						
	4-1	4-2	4-3	4-4	1M	2M	3M	4M	Match = 2M + 3M + 4M										

REMARKS:

APPENDIX F

CANDIDATE LIST

Record 14 (#R14)

CANDIDATE LIST

Unknown ____________________ Date ____________________

Experiment ____________________ Analyst ____________________

Compound		Data Match													Rank
		State	Element	NMR		UV		GC			IR		MS		
No.	Name	M	No. M	M	D	M	D	1 Col.	2 Col.	D	Peaks M	D	Peaks M	D	

APPENDIX G

CANDIDATE LIST RATING TABLE

CANDIDATE LIST

RATING TABLE

SEARCH	PARAMETER	RATING
STATE	-	2
ELEMENTS	-	4
NMR	± TOLERANCE	20
	EXACT PEAK	24
GC	COLUMN A:	
	± TOLERANCE	6
	EXACT	9
	COLUMN B:	
	± TOLERANCE	6
	EXACT	9
	MAXIMUM	18
IR	1 to 1	18
	1 to ±1	15
	1 to 2	12
	1 to ±2	9
	1 to 3	6
	1 to ±3	3
	2 to 1	8
	2 to ±1	6
	2 to 2	12
	2 to ±2	9
	2 to 3	8
	2 to ±3	6
	3 to 1	2
	3 to ±1	1
	3 to 2	4
	3 to ±2	3
	3 to 3	6
	3 to ±3	4
	MAXIMUM	36

SEARCH	PARAMETER	RATING
MS	1-1	16
	1-2	12
	1-3	8
	1-4	4
	2-1	9
	2-2	12
	2-3	9
	2-4	6
	3-1	4
	3-2	6
	3-3	8
	3-4	6
	4-1	1
	4-2	2
	4-3	3
	4-4	4
	MAXIMUM	40
UV	± TOLERANCE	4
	EXACT PEAK	6
Any	SKIP	0
Any	DEFAULT	1
	MAXIMUM	130

DISTRIBUTION LIST

This report is being distributed as follows:

Copy No.	Recipient
1-100	National Science Foundation 1800 G Street Washington, D. C. Attention: Mr. T. Quigley
101	IIT Research Institute Section Files
102	IIT Research Institute Editors, M. J. Klein, Main Files
103	IIT Research Institute G. E. Burkholder
104	IIT Research Institute M. E. Williams
105	IIT Research Institute P. Llewellen
106	IIT Research Institute E. S. Schwartz
107	IIT Research Institute H. J. O'Neill
108	IIT Research Institute B. E. Edwards
109	IIT Research Institute W. M. Linfield
110	IIT Research Institute R. G. Scholz