

This article was downloaded by:

On: 17 January 2011

Access details: Access Details: Free Access

Publisher Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Critical Reviews in Analytical Chemistry

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title~content=t713400837>

Computer-Aided Identification of Organic Molecules by their Molecular Spectra

Lev A. Gribov; Mikhail E. Elyashberg; J. T. Clerc

To cite this Article Gribov, Lev A. , Elyashberg, Mikhail E. and Clerc, J. T.(1979) 'Computer-Aided Identification of Organic Molecules by their Molecular Spectra', Critical Reviews in Analytical Chemistry, 8: 2, 111 — 220

To link to this Article: DOI: 10.1080/10408347908542711

URL: <http://dx.doi.org/10.1080/10408347908542711>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

COMPUTER-AIDED IDENTIFICATION OF ORGANIC MOLECULES BY THEIR MOLECULAR SPECTRA

Authors: Lev A. Gribov
Laboratory of Molecular Spectroscopy
and Quantum Chemistry
Vernadsky Institute of Geochemistry and
Analytical Chemistry
U.S.S.R. Academy of Sciences
Moscow, U.S.S.R.

Mikhail E. Elyashberg
All-Union Research Institute of Organic
Synthesis
Moscow, U.S.S.R.

Referee: J. T. Clerc
Department of Organic Chemistry
Swiss Federal Institute of Technology
Zurich, Switzerland

TABLE OF CONTENTS

- I. Introduction
- II. Computer Storage and Search Systems
 - A. General Characteristics of Search Systems
 - B. Information Retrieval System of the Siberian Division of the U.S.S.R. Academy of Sciences
 - C. Combined Use of IR, UV, PMR, and Mass Spectra in Search Systems
 - D. Account of Shape and Intensities of Bands
 - E. Identification of Compounds in Mixtures
 - F. Comparison of Spectra by Correlation Coefficient Method
 - G. Deconvolution of a Spectrum of a Mixture into Spectra of Its Components
- III. Systems Based on Pattern Recognition Methods
 - A. Elements of the Pattern Recognition Theory
 - 1. Essence of the Pattern Recognition Method
 - 2. Probability Approach
 - 3. Geometric Approach
 - 4. Characteristics of Recognition Systems
 - B. Alternative Recognition and Linear Classifiers

- C. Recognition by Means of Simultaneous Use of IR and Mass Spectra
- D. Some Recognition Algorithms and Establishment of the Most Important Spectral Features

IV. Artificial Intelligence Approach

- A. General Features of Artificial Intelligence Approach
- B. Elements of Symbolic Logic
- C. Basic Principles of the Graph Theory
- D. Problem of Structural-Group Analysis
- E. The Structure Recognition System (STREC)
 - 1. Block Diagram of STREC System
 - 2. Mathematical Synthesis of Molecular Structures
 - 3. Algorithm for Analysis of Structural Formulas
 - 4. Library of Standard Fragments
 - 5. Results
- F. Use of Computational Methods in the Identification of Molecules
- G. CONGEN Program
- H. Various Strategies for Identification of Molecular Structure

V. Summary

References

I. INTRODUCTION

In recent years the different branches of analytical chemistry, especially those dealing with the identification of individual molecules, as well as qualitative and quantitative analysis of molecular mixtures, have become increasingly important. Questions in these areas often arise in solving a number of global problems. For example, the analysis of compounds produced in chemical reactions under ordinary conditions is common, and more work is being done under conditions of plasma, photochemistry under different interaction energies, laser synthesis, and so on. The analysis of molecules and molecular associates is also encountered with regard to environmental pollution: organic impurities in water, the state of the atmosphere, the state of biological objects, and many others. It should be remembered that molecules in natural media enter into various reactions and change with time, thus forming innumerable derivatives. Moreover, a problem for a number of industries is automatic protection of the final product from contamination by different kinds of by-products formed in the course of production. At some stage in the research, all these problems call for exact and express analyses aimed at investigating both individual molecular systems and their associates as well as mixtures. The need for mass-scale analysis inevitably leads to the creation of automatic identification systems. In some respects, this reminds us of the situation which existed a few decades back when the rapid advances in metallurgy and the creation of diverse alloys called for the development of highly reliable and express methods of analyzing metals and alloys at the final stage and during production. This problem

was essentially solved after an effective quantitative analysis method had been developed on the basis of emission spectral analysis and a series of automatic analyzers (of which the best is the quantometer) capable of performing quantitative analysis of metals and alloys (often simultaneously for several dozen compounds) in a few seconds and sometimes in a fraction of a second.

Similar problems will also have to be solved in the near future with regard to the analysis of molecular systems. Automation of these processes has only recently begun. For example, about 10 years ago only occasional papers dealing with the automated analysis of molecules were published. Today, however, there are not only a large number (already more than a hundred) of fundamental reports on this topic, but also real prototypes of several systems for solving certain problems in identifying molecules belonging to a particular class, as well as systems capable of solving a number of extremely complicated problems.

It would be no exaggeration to say that we are witnessing the birth of a new science, which may be tentatively called analytical molecular physics. Considerable progress has already been made in this field, and definite research trends have taken shape.

New problems and future development prospects are gradually coming to the fore. The aim of this review is to outline the present state-of-the-art in this field and to draw certain conclusions about its development prospects in the near future.

First, it may be noted that the principal difficulties are encountered in designing rapid automatic methods of the analytical chemistry of molecules. The task is undoubtedly more complex than the identification of individual atoms, because molecules as systems are incomparably more complicated than atomic systems, and they are, in addition, characterized by numerous different features. These compounds cannot, therefore, be frequently identified with the help of just one of these methods on its own.

All the known methods, developed specifically for the identification of molecular systems can roughly be divided into two groups: chemical methods and physical techniques. The latter is also known as the instrumental method. The methods of optical spectroscopy, resonance techniques, mass spectroscopy, and the recently developed photoelectron spectroscopy all come under physical methods. These techniques are increasingly being used in analyses, and they do, in fact, underlie almost all of the sophisticated automatic systems designed for analytical purposes. This is not accidental.

The analytical use of purely chemical methods stems from the fact that molecules of a particular class can enter into one chemical reaction or another. If a molecule is sufficiently large, then its ability to enter into a reaction of a specific kind depends on the presence of a proper reaction center in this molecule. The presence or the absence of this reaction center is precisely determined in the course of the chemical reaction. The environment of this center, particularly the remote surroundings, may not have any bearing on the pattern of the chemical reaction, and therefore, the fact that a reaction does take place is of no help in inferring information about the structure of this molecule. This is precisely the reason why a chemical reaction usually can only help in determining the class of compounds. Complete identification of a molecule as a whole is, therefore, often a complicated scientific problem, and calls for a large number of diverse and intricate chemical reactions to be carried out. Such a problem does not frequently yield to automatization.

Although, theoretically, we may conceive of the construction of automatic chemical identification systems for certain molecules or a specific class of molecules, it is probably beyond our imagination to perceive the complexity of a universal system that could concurrently identify, say, 30 to 40 different classes of molecules and their struc-

tures as well. This does, perhaps, explain why instrumental methods of analysis are finding wider and wider application in analytical practice. These instrumental methods have certain important merits: first, they are universal (for instance, IR spectra can be recorded even in a trace concentration and in any aggregate state of the specimen), they are rapid (just a few seconds or sometimes even a fraction of a second is sufficient to record the IR spectrum by means of modern equipment), and it is easy to connect spectral instruments to computers, thus making it possible to transmit the information to computing centers where the information is rapidly analyzed with respect to many different features.

The main superiority of instrumental methods over purely chemical techniques is that the specimen is exposed to such a weak influence of radiation that it hardly changes in the course of analysis, whereas in a chemical analysis, the specimen is almost completely destroyed and cannot be further used. In the instrumental methods, the same specimen can be first used in IR spectroscopic analysis, then for recording the UV, visible, or NMR spectra, and finally for mass spectroscopic examination.

It would be possible to give a long list of the other merits of instrumental methods; however, they can be found in many sources in the literature. Apparently, as they are improved, these instrumental methods will become so effective that no other approach may ever be able to compete with them. Nevertheless, one should not be misled into believing that instrumental methods will prove to be a push-button procedure for solving analytical molecular problems. It is not as simple as that.

At a later stage we shall dwell on different kinds of approaches to this problem. At present, it is only possible to say that all these methods call for the recording not only of various spectra, but also mathematical treatment of these spectra which will finally give the necessary information.

In the mathematical treatment of spectra, use is often made of the techniques of the pattern recognition theory, symbolic logic, graph theory, and some other approaches. To solve the problem of spectrum identification, a special algorithm had to be designed, a program written for the synthesis of molecular structures, and theoretical spectra constructed which may have different degrees of complexity and approximation, etc. These questions will be dealt with in due course.

The complexity of the problem encountered in a systematic molecular analysis lies in the fact that in a large number of cases it is impossible to create a completely automatic system, and it may be undesirable. The most rational system comprises man and his intellect as integral components which participate in the logical treatment of the data obtained directly from the instruments and computer.

At present, we can distinguish the following trends in the development of special methods for automatic molecular analysis by spectra. These methods are based on the atlas approach realized in information retrieval systems, pattern recognition techniques, and methods involving the construction of a so-called artificial intelligence, which is supposed to be capable of executing several complicated logical operations usually carried out by man.

In our review, we shall outline the ideology which underlies each method and the principles of the approach which determine the realization of the method, the merits and demerits of each method, and the future development prospects of each. We shall not touch upon questions pertaining to the equipment used in automatic systems. This is rather a rapidly developing and changing area, which is increasingly being supplanted by newer and newer equipment. Therefore, this topic requires a separate survey. All that can be said here is that today's equipment of automated systems includes a variety of facilities which provide for instrument-computer interconnection, diverse

display devices, and appliances which make possible man's interaction with the computer.

Instrument-computer interconnection devices include a vast number of different analyzers, analog-digital converters, equipment for primary spectra processing, memory and storage devices, and noise suppressors which detect weak signals against a noisy background and thus give an opportunity to record spectra in unfavorable conditions. For example, special instruments exist today which can record IR spectra of organic molecules dissolved in water in trace amounts. A few years ago, the recording of spectra of aqueous solutions was an extraordinarily complex, and at times insurmountable, problem in IR spectroscopy.

Display devices include a variety of digital printers, data plotters, black and white and color displays, etc. Man-machine systems include terminal typewriters, light pens, devices for feeding graphic information into the computer, and many others.

All these devices form the functional components of highly developed automatic systems designed to identify complex compounds. They provide rapid data input and processing, man's thinking, and other processes up to final information output, which in many cases is delivered in a few seconds, and in the most typical cases, a few minutes after the start of the experiment.

In regard to the delivery time of the final result, it varies within rather wide limits, depending upon the devices available and the computer potentialities. The range of these devices, of course, greatly depends on the degree of instrumentation of the research center. In this respect, there is no universal approach. Indeed, we can assert that there is no scientific problem; there are only problems of a technical and organizational nature. Therefore, as already mentioned, these problems will be touched on only briefly.

This review will consist of three sections and cover the following topics. In Section I, information retrieval systems will be described. Section II, systems in which identification of compounds is effected by means of pattern recognition will be considered. Section III is devoted to systems based on the ideas underlying the design of the so-called artificial intelligence. Our professions naturally compelled us to focus attention mainly on those works dealing with optical spectra, although other trends have also been given treatment. A large number of works concerned with the automatization of mass spectroscopic analysis have not been included in this review.

We cannot claim to have given complete coverage to all the extensive literature available today in the field, because this is a rather difficult task. Such attempts are being made in different countries, and considerable progress has already been made in this respect.

We can say with confidence that, while in the early stages of computer application in chemistry, the percentage of correct answers given by the spectroscopist in the identification of molecules by spectra was slightly higher than the percentage obtained in a computer-assisted analytical system. Today the situation is reversed, and frequently the machine is apparently "cleverer" than even an experienced specialist.

This review is more a critical survey than a collection of abstracts. We have, therefore, endeavored to dwell in detail on those works which, in our opinion, clearly outline one trend or another. Consequently, some works are dealt with in detail, while others are only briefly touched upon so that the reader may have a clear idea of the works of others.

In the conclusions, an attempt has been made to put forward our opinion regarding the future prospects and problems in this area of science.

II. COMPUTER STORAGE AND SEARCH SYSTEMS

A. General Characteristics of Search Systems

As already pointed out in the introduction, there are several trends in the development of methods to automate molecular spectral analysis based on the use of spectral instruments and computers.

The first trend is that of developing the so-called information retrieval systems (IRS). By IRS we mean the set of all linguistic and technical means designed for storage, search, and output of the information needed. IRS provide for the automatic search for information. They are capable of executing two types of functions: giving out specific information on a particular question and searching for the information needed by comparing the features coded in the question and those stored in the computer.

The first class of problems include, for example, output of answers in the form of spectra, absorption band frequencies, intensities, etc. directly by the name or structural formula of the compound fed into the computer.

The second class of problems is only solved when there is a need to find out whether the IRS storage contains a spectrum identical or similar to the spectrum fed into the computer.

The need for such search systems naturally arises in molecular spectroscopy because the data contained in the different kinds of atlases already run to more than one hundred thousand spectra. Evidently, such a vast amount of information can only be handled conveniently when it is stored in a computer and appropriate algorithms are available for the search for compounds. Manual search for information from such a data bank stored in the form of ordinary atlases or punched card catalogs is almost impossible.

At first glance it seems that such IRS are capable of solving a large number of analytical problems by a formal comparison of input spectra with the spectra contained in the atlas. Thus, when the input and the atlas spectra coincide, they establish the identity of compounds and solve the analytical problem. However, in practice it is not so simple as that. For example, consider the IR spectrum shown in Figure 1. It consists of a set of bands characterized by various frequencies, intensities, and half widths. In order to feed the spectral curve into the computer without any loss of information contained in the spectrum, we have to represent the information in the form of a large number of points, each of which is defined by a number, and then store this sequence of these numbers in the computer.

Suppose that these points are located every 5 cm^{-1} . The spectrum in this case will consist of a set of points as shown in Figure 1. It is evident that this set of points represents the spectrum in sufficient detail, but the number of points in the interval, say, from 400 to 3600 cm^{-1} , is fairly large (more than 600).

If we were to code 100,000 spectra by this method, the computer memory would then contain the information in the form of more than 60 million numbers. The point here is not that this number is rather large for the memory of a modern computer, but it is very difficult to compare a spectrum represented in such a way with analogous spectra stored in the computer. The search in real time may, therefore, take quite a while and it may in many respects be devoid of sense.

Hence, we may adopt a different procedure, that is, represent the spectrum of a compound in the form of an analytical expression, say, the sum of Gaussian or any other similar curves which correspond to the whole spectral curve. In this case, the number of digits needed to describe the spectrum is considerably reduced. For example, the spectrum shown in Figure 1 can now be characterized by only 66 parameters

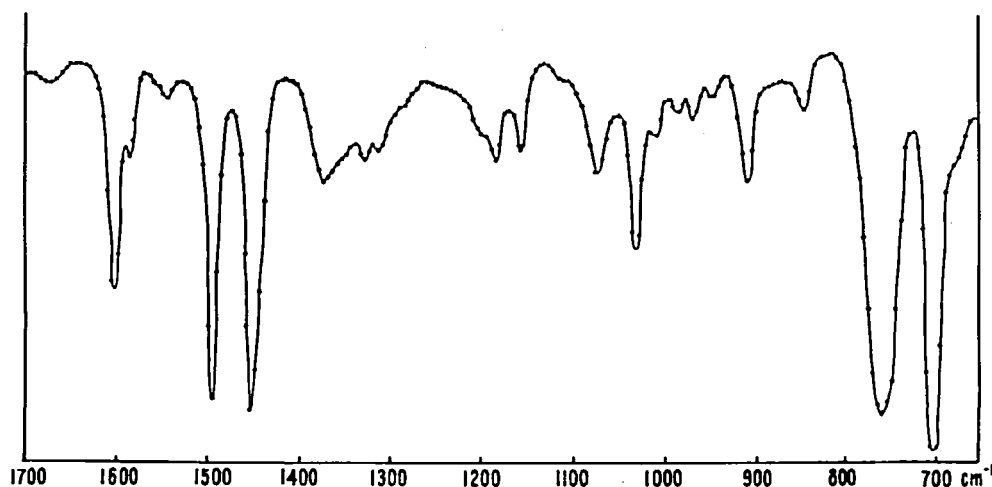


FIGURE 1. A part of the IR spectrum of polystyrene.

of 22 Gaussian bands almost without any loss of information. Thus, the spectral data are compressed. Nevertheless, this approach is not very advantageous in that considerable computer time has to be spent in deriving an analytical expression for each spectrum stored in the computer. These circumstances naturally suggest that in any IRS, as a rule, we have to code not the whole spectrum, but just certain individual features of it, frequently, the position of peaks on the frequency scale. In more modern systems, the position of peaks, their intensities and half widths are also coded in a specific scale. This additional compression of information invariably leads to certain information losses and, consequently, to a situation in which complete identity may escape detection in the course of comparison of the atlas spectrum and experimental spectrum, which should also be coded by the same number of features. In this case, we are faced with the intricate problem of designing special algorithms which could collate the roughly encoded atlas spectra with the experimental spectra and thus give out similar spectra for further investigation by man. In those cases where the number of similar spectra is not too large, for instance, about 50, the problem may be regarded as resolvable. If the number of spectra similar to the experimental one amounts to several hundred, then man is in no position to collate the spectra. Moreover, man's participation, at least in the last stage of collation of the experimental and atlas spectra, implies the availability of a library of atlases, the creation of which is in itself by no means a simple task, and is beyond the reach of many laboratories conducting analytical research.

These considerations naturally suggest that, although they have several obvious merits, the IRS are, nonetheless, not such a powerful tool as it would first seem.

It should also be mentioned that, despite the large number of spectra found in different kinds of atlases, there are, nevertheless, far less of them than the total number of compounds known today (about four million). Moreover, in laboratories engaged in creating new substances, we are faced with another type of problem, i.e., identification of a compound the spectrum of which cannot be found in any atlas. In such cases, therefore, IRS are altogether ineffective.

B. Information Retrieval System at the Siberian Division of the U.S.S.R. Academy of Sciences

A typical example of IRS intended for extensive use in molecular spectroscopy,

which we shall dwell upon in detail, is the system designed by Professor V. A. Koptug and colleagues at the Institute of Organic Chemistry and the Computing Center of the Siberian Division of the U.S.S.R. Academy of Sciences.¹⁻⁸ This system is intended for use with the BESM-6 and Minsk-32 computers and is capable of conducting the following operations: input, verification and entry of spectrum files on magnetic tape coded by special methods, entry of amendments and corrections into the files in storage, search for the spectra of the compounds according to given features and print-out of this information, search for compounds whose spectra are close to a given spectrum, statistical treatment of archives for the purpose of obtaining different kinds of correlation dependences, and print-out of these correlation tables which could subsequently be used in designing, say, a system of artificial intelligence for constructing diagnostic tables in the dictionary of features. The name of the compound, its serial number in the atlas, empirical formula, molecular weight, melting point, boiling point (if available in the atlas), and the spectrum description are entered into the system catalog.

The programs which form the software of this system includes a control, a subroutine for comparison of the inquiry with spectral entries in the computer storage, a program for evaluating the similarity between input and atlas spectra, a program for extraction of search results, calculation of correlation tables, duplication and print-out of information files, and a program for editing inquiries. The search program operates with entries of a fixed format. The file in which the search is effected is stored on a separate magnetic tape. Each zone of the tape contains entries which must not be split between zones. Information regarding each compound consists of a fixed set of signs. The system operates with three types of signs: Boolean signs, single-digit integers, and symbolic signs. The file structure description carries all the data necessary for the search in a given information file. The structure description of any file is formed by a special subroutine of file structure description. Information is retrieved by means of a search program which scans the magnetic tape file and thus searches for an entry consistent with the inquiry. The search inquiry contains an obligatory and desirable spectral signs.

Since the data have to be entered in a discrete form in the memory of a digital computer, the question naturally arises as regards the breakdown or simplification of spectral curves; in particular, the number of discrete spectral features which should desirably be used in the search for compounds by their spectra.

In a general form, this question can be formulated as follows,¹ Let us suppose that the number of frequency intervals is n and the number of gradations of the absorption band intensities is m . Out of all the spectral intervals, it is desirable to choose only those intervals s in which the absorption bands have sufficient intensity. This eliminates the information noise. Use of only these absorption bands provides for variety in the search. This variety is expressed by the following formula:

$$N = (m - 1)^s C_n^s$$

where C_n^s is the number of modes which allows us to select s intervals from n intervals, $(m - 1)^s$ is the number which shows how many different spectra may be coded using s intervals, considering that each interval can include one band with an intensity having m gradations, and N is the number of different spectra which may be coded using this procedure.

Taking the logarithm, we obtain

$$\begin{aligned} H(s, m) &= \log N = s \log (m - 1) + \log C_n^s \\ \log C_n^s &= \log \frac{n!}{s! (n - s)!} = \log n! - \log s! - \log (n - s)! \end{aligned}$$

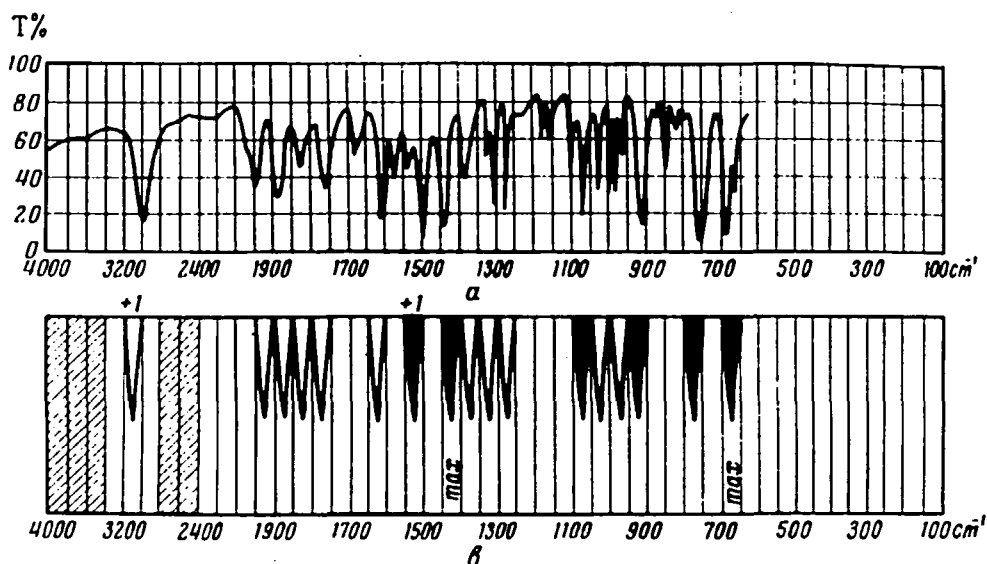


FIGURE 2. IR spectrum of diphenylacetylene (a) and the search inquiry filled for its search (b). (From Drobyshchev, Yu., P., Nigmatullin, R. S., Lobanov, I. K., Korobeinicheva, I. K., Bochkarev, V. S., and Koptug, V. A., *Vestn. Akad. Nauk SSSR*, 8, 75, (1970). With permission.)

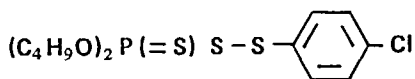
It is clear that even when $m = 2$, i.e., in using data on the presence or absence of an absorption band in the spectral region without any regard for the quantitative data on its relative intensity, 10 to 20 spectral intervals having intensive absorption bands provide a range of $N = 10^{10} - 10^{13}$, which is sufficient to describe the spectra in the existing collections.

Proceeding from these considerations, the authors of this IRS chose the following procedure for coding the IR spectra. The whole frequency range of spectra was divided into 48 intervals of steps 50 cm^{-1} in the range from 100 to 2000 cm^{-1} , and of steps 200 cm^{-1} in the range from 2000 to 4000 cm^{-1} . Sixteen strong absorption bands are coded and the intervals in which they occur are noted. The coded bands are likewise characterized by the relative intensity (five gradations) and half widths (six gradations). Also, information on the melting point, molecular weight, empirical formula, and molecular structure are coded as a supplement to the spectral information. In the structural description, note is also made of the presence or absence of special structural fragments if such preliminary information is available. In searching for a compound using IR spectra, the operator fills in a standard search questionnaire bearing 48 positions corresponding to 48 spectral features (see Figure 2). Note is made of the spectral intervals at which the spectrum of the unknown compound should have the obligatory features (shown in Figure 2 by black triangles) and the spectral intervals at which the spectrum of the unknown compound has absorption bands, but it is not excluded that they will be masked due to absorption of the solvent or may not be present due to some other reason in the reference spectrum. These desirable spectral features are shown by light triangles. The spectral intervals at which the spectrum of the unknown compound has no absorption bands (negative features) are shown by a broken line. If the absorption bands, which are subsequently used as obligatory features, exist near the boundary of the spectral interval, and if the recording conditions change (for example, when the solvent changes), they may shift to the neighboring interval. Then a plus or a minus sign is placed over the triangle denoting this band (to indicate that this band may fall into the preceding or succeeding spectral interval). The inquiry is then fed into the

computer. The search program is so designed that the spectra chosen by the computer on the basis of spectral features (positive and negative features) are ordered in the increasing number of coinciding desirable features. In the answer, the first 126 spectra that have been chosen from the atlas of spectra are numbered and ranked in order of probability. The computer-selected spectra are then divided into groups with respect to the number of coinciding desirable features.

The number of spectra chosen by the computer essentially depends on the number of obligatory spectral features. Thus, if only three spectral features are given, the number of selected spectra lies in the range from 350 to 2300. On increasing the obligatory features to five, the number of selected spectra decreases by four- to sixfold. A further increase in the obligatory features results in a further reduction in the selection spectra, but at the same time there is an increased probability that the reference spectra of the initial compounds in the file will not satisfy any of the obligatory features. Experience shows that five to seven obligatory features are optimal for this system.

A change in the number of desirable spectral features does not very much affect the number of selected spectra, but has a strong influence on the position or ranking of the unknown spectrum given out by the computer. Ambiguity in the answer can be considerably reduced if additional data (molecular weight, probable presence or absence of certain structural fragments) are included in the inquiry besides the spectral features. For example, in searching for the compound



by means of only spectral features (five obligatory and seven desirable), the computer would give out 185 compounds. For the same inquiry which also carries information on the presence of a benzol nucleus in the molecule, the number of selected compounds drops to 55. The number falls to 17 if the inquiry includes an additional condition that the molecule contains a chlorine atom. On including the tentative molecular weight of the compound to an accuracy of about ± 20 atomic units, the answer contained only two substances of similar structure. In general, inclusion of information about one structural element in the inquiry, in addition to the obligatory spectral features, reduces the number of selected spectra by 2 to 3 times, information about two structural elements by 8 to 15 times, and three structural elements by 20 to 50 times. If the inquiry carries information about three structural elements plus the tentative molecular weight, then the answer contains only one to five compounds.

Since the data on molecular weight perceptibly decrease the number of selected spectra, a special algorithm has been incorporated into the system. This algorithm makes it possible to design a satisfactory filter on the basis of molecular weight, and thus to discard any molecule that does not satisfy this filter.

Besides the information on IR spectra, data about electron spectra are also introduced into the system. In this case, the spectrum is coded by approximation with the help of a broken line. The broken line is so constructed that all the characteristic points of the spectral curve (maxima, minima, inflections) are the nodal points (Figure 3). In this coding procedure, each spectrum in the machine catalog is presented by a table of numbers characterizing the coordinates of the selected nodal points. Since in this coding procedure the spectra have no reference point, comparison of an unknown compound with the machine catalog data means that the encoded spectra have to be reestablished in the form of broken lines. A special algorithm has been designed for this purpose. The search program has been constructed on a block principle. The set of

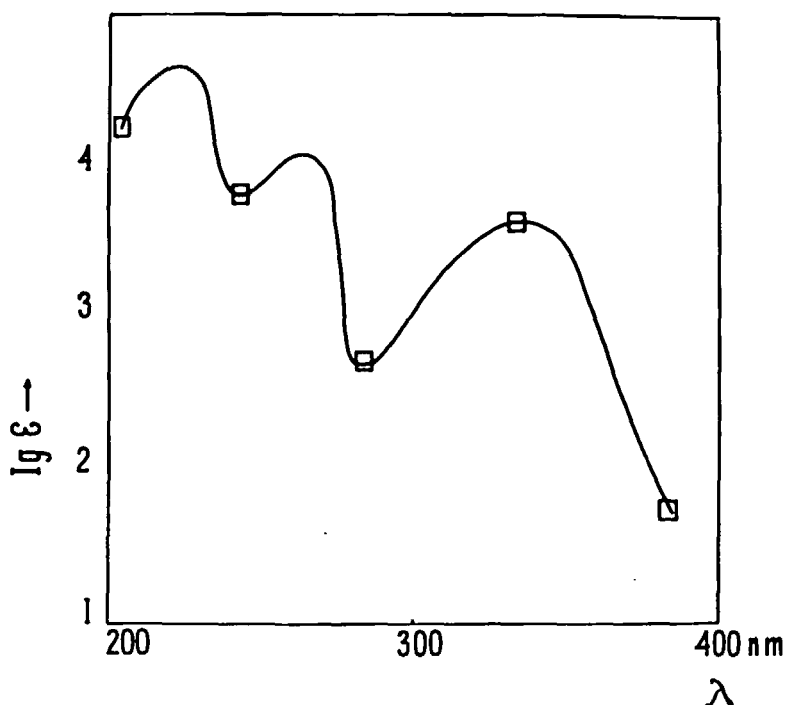


FIGURE 3. Coding of UV spectrum with the help of coordinate rectangles.

routines makes it possible to carry out the search, using the following criteria: the serial number of the compounds in the main catalog file, the position of the maximum and the logarithm of molar absorptivity of the most intensive absorption bands shown by the coordinate rectangle in which this maximum should occur (see Figure 3); the positions of the maxima and minima in the spectral curves given by a set of coordinate rectangles, which this curve should be drawn through; the code of the chromophore groups; presence or absence of certain elements in the empirical formula; the range of molecular weight; and melting points and boiling points in a given temperature range.

The following facts immediately attract our attention: since coordinate rectangles are used in coding the spectra, the spectra of not only the compound being identified but also those of similar compounds, may in principle, fall in these coordinate rectangles. This is especially true of the UV spectrum which is, as a rule, characterized by fairly wide bands and poorly expressed fine structure.

Figure 4 shows a number of spectra which all satisfy a given set of coordinate rectangles, but belong to different compounds. This ambiguity in the results is characteristic of any atlas approach and probably cannot be avoided altogether. Even in those cases where the entire curve is subjected to comparison, the difference in spectrum-recording conditions may lead to a situation where the atlas and input curves may not be identical. This is one of the demerits of this search system.

Of course, with the introduction of additional data, as already pointed out, the answer becomes less ambiguous. Use of molecular weight and an empirical formula, in particular, alleviate the situation. Without these additional data, the IRS may produce an unduly large number of answers. It is highly likely that this set of answers may not at all contain a single spectrum similar to the input one.

Side by side with the data on IR and UV spectra, information on NMR is also fed into the search system. Spectra have been chosen which contain signals falling within

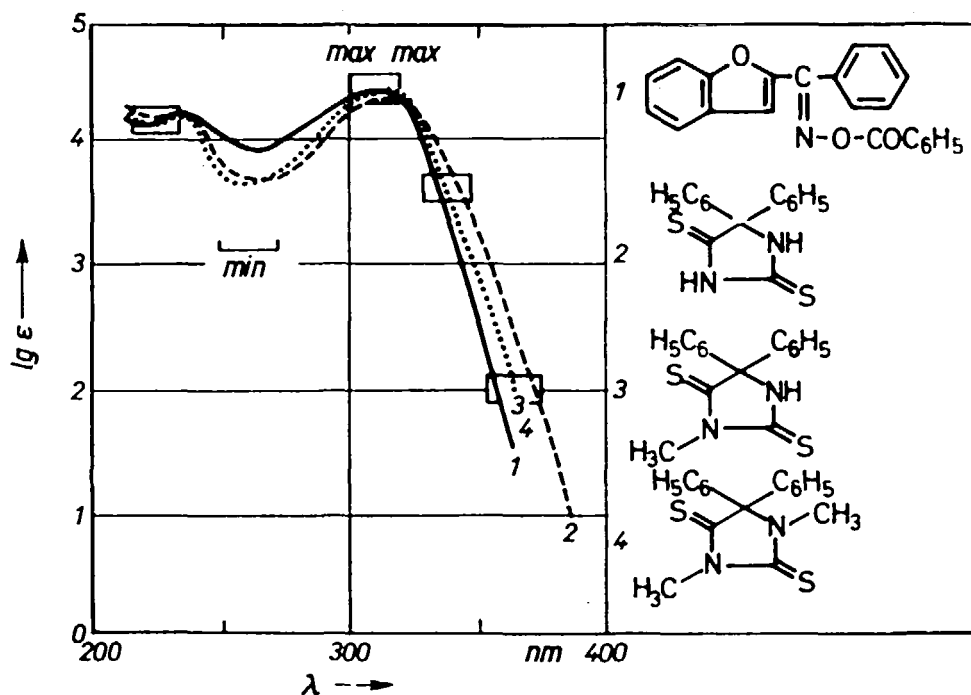


FIGURE 4. Set of spectra which have passed through the coordinate rectangles of the inquiry. (From Barkhash, V. A., Sokolov, S. P., Sekerina, L. F., Drobyshev, Yu. P., and Koptug, V. A., *Izv. Sib. Otd. Akad. Nauk SSSR, Ser. Khim.*, 14(6), 111 (1974). With permission.)

at least one of the following intervals: 0.95 to 1.05, 2.03 to 2.13, 4.05 to 4.15, 7.43 to 7.53, and 8.01 to 10.11 ppm. The corresponding routines provide for a search in two stages. In the first stage, the machine analog search is conducted by a complete set of signals of the specimen under investigation, where the positions of the signals are given with possible experimental errors. The next stage in the analysis of the chosen spectra, using the criteria of signal intensity and the presence of structural fragments common to all compounds, increases the degree of certainty of the answer.

This IRS also contains a catalog of mass spectra. On the whole, the computer memory contains about 63,000 IR spectra, 5000 UV spectra, more than 15,000 NMR spectra, and 17,000 mass spectra. For the sake of ease in handling, dialogue with the computer is also incorporated into the system, so that the inquiry can be fed directly into the computer with the help of a special typewriter. About 4 to 5 min are sufficient to get the answer when the BESM-6 computer has to scan through 20,000 spectra. If the operator so desires, the search can also be conducted only within a preset class of spectra if it is known that the compound to be identified belongs to a certain class.

This system has been in use over the last few years and is quite efficient in structure recognition.

The system is built up by connecting the following three types of terminal complexes to the processor via a multiplex channel:

1. Complexes for investigation by different types of molecular spectroscopy
2. Such complexes for investigation by the chemical or physical methods that are capable of communicating with the computer by means of a dialogue in solving computational or search problems

3. Special-purpose complexes for the formation of machine catalogs; for example, catalogs of search systems including spectral, structural, and other data.

These terminal complexes considerably widen the field of computer utilization, especially in conducting experiments and in the subsequent processing of experimental results. From the viewpoint of hardware, these complexes have almost the same structure and constitute the standard peripheral equipment of the Minsk-32 computer designed for man-machine communication. In those cases where it is necessary, they are supplemented by special interfaces for connecting nonstandard devices to the computer (for example, various types of spectrometers). The interfaces are based on integral circuits of standard CAMAC.

The range of terminal complexes can easily be varied, depending on the nature of the problem to be solved, by connecting the various auxiliaries of the multiplex channel of the Minsk-32 computer (punched card or punched tape devices, alpha-numeric printers, data plotters, etc.). Experimental data can be recorded and processed with on-line or off-line modes.

The interfaces designed on the CAMAC principles make it possible to use CAMAC devices for surveying the experimental data, and consequently, to

1. Considerably widen the scope of experimental work due to the utilization of special equipment (even a minicomputer) in the terminal complexes
2. Rapidly rearrange the terminal complexes to suit the problem being solved
3. Retain the same software even when the system is being expanded

Moreover, the CAMAC devices provide for easy passage to multi-machine systems.

C. Combined Use of IR, UV, PMR, and Mass Spectra in Search Systems

Information retrieval systems based on other principles are described in.⁹⁻¹² Machine atlases of IR, UV, NMR, and low-resolution mass spectra are used side by side in this modification. All the spectra in this system are coded by a binary code, i.e., unity stands for the presence of a spectral feature in a specific interval, and zero for the absence of the same. For each reference compound, the codes of all the types of spectra are grouped in one line, the signature of the corresponding compound. Spectral intervals are chosen within fairly wide limits (for IR spectra from 50 to 300 cm^{-1} and UV spectra through 20 nm), and the intensities of the strongest bands are marked in each interval of IR and UV spectra in a three- or four-grade scale. For NMR spectra, the presence or absence of signals and their multiplicities are marked in 16 intervals on the chemical shift scale. The position and intensity of peaks are coded for the mass spectra. The signature of an unknown compound is constructed in a similar manner.

In a simple comparison of the signature of an unknown compound, X , with the signature of the i th reference spectrum, C_i , it is possible to calculate the number of positions, S_i , in which the features for X and C_i coincide. Then the reference compounds, which have the maximum S_i , will be taken as the compounds most resembling substance X . This simple comparison procedure has been improved upon for the following reasons. Comparison of two features encoded by binary codes may lead to four different results, namely:

1. The unknown compound and the reference compound exhibit the same feature.
2. A feature is absent in both the signatures compared.
3. The feature of the unknown compound is absent in the reference compound.

4. The feature of the reference compound is absent in the spectrum of the unknown compound.

Moreover, on the basis of the empirical estimate of spectral and information significance, each feature is assigned a weight factor. Thus, the result of each elementary comparison of features was interpreted with the help of the following estimate matrix:

		Code of unknown compound	
		0	1
Code of reference compound	0	R_2	R_4
	1	R_3	R_1

where the weights, R_i , are so determined that $R_1 \geq R_2 \geq R_3 \geq R_4$. Depending upon which one of the four alternatives is realized, the corresponding weight factors, R_i , are summed up. Thus, we obtain a set of reference compounds with the highest S , and consequently, they exhibit a high degree of similarity to the spectrum of the unknown specimen. This procedure proves highly effective, especially when the data bank does not contain the spectrum of the substance under identification, and we are required to find a substance of close structure, or in those cases where the spectrum has been recorded with a high amount of distortions or the specimen is contaminated with impurities.

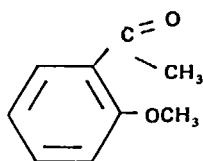
Comparison of the signature of a specimen with all the spectra in the data bank calls for a fairly long machine time. It can be reduced if a strategy is used in which the comparison is stopped when the signatures do not agree. In this system, prior to starting the comparison, the maximum possible contribution to the sum, S , of each feature is determined for each element of the signature by means of previously computed values of R_1 and R_2 . This procedure makes it possible to establish a priority sequence for the features, depending upon their significance during comparison of signature X , and to determine some threshold magnitude for S .

Search begins with a comparison of the most important features, and the current value of S is compared with the threshold value. If S drops below a certain lower limit, comparison is stopped. As a result, the obvious differences are detected quite early, and thus the machine time is saved.

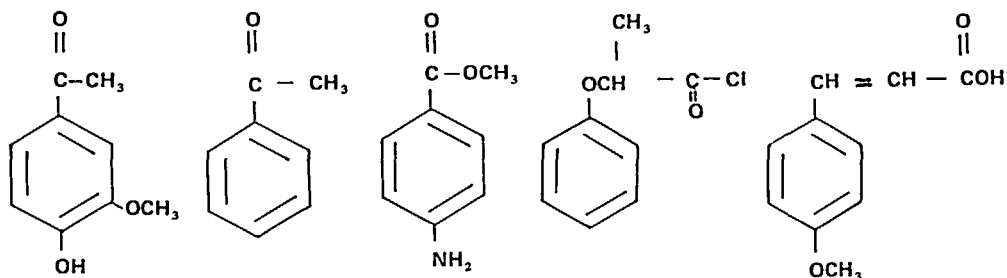
In accordance with the similarity criterion introduced, the computer gives out 20 compounds ranked in the order of decreasing similarity.

Erni and Clerc¹⁰ pointed out that the data bank has to be formed with due regard for the demands of constant users. This system automatically counts the number of times each spectrum is chosen as the reference spectrum, and thus eliminates from the bank those spectra that are of little use to the user. At the same time, the system takes into account the discriminating efficiency of the features. As a result of regular correction, the data bank is constantly updated to keep in line with the user's changing interests.

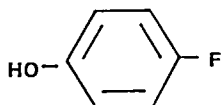
An experimental run of the system was conducted using a data bank containing spectra of about 1000 compounds. The operation modes were tested by conducting the search by one spectrum type or by simultaneous use of IR, UV, NMR, and mass spectra. The tests gave satisfactory results which can be judged by the following examples quoted by the authors. Unknown compound:



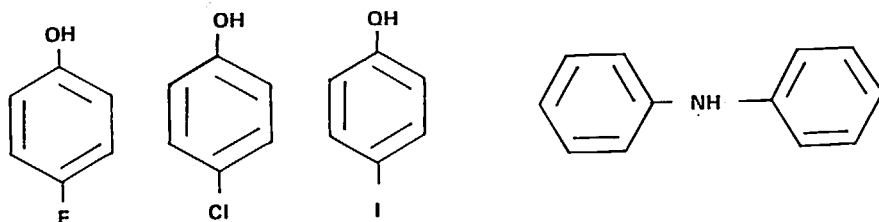
Compounds given out by the system:



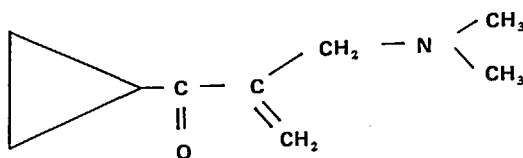
Unknown compound:



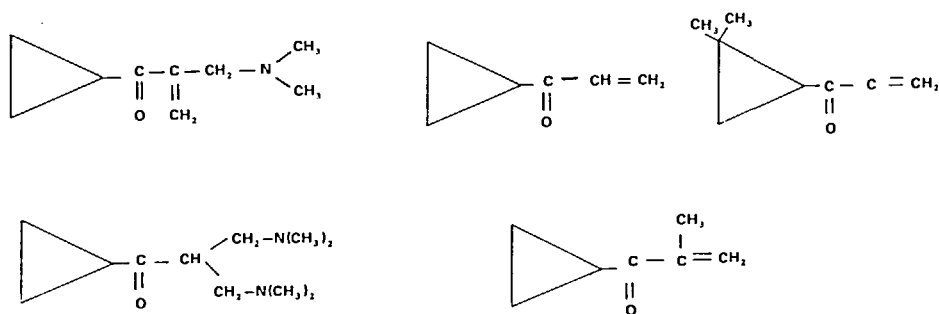
Compounds given out by the system:



Unknown compound:



Compounds given out by the system:



D. Account of Shape and Intensity of Bands

Since in the atlas comparison, the identification result depends upon the degree of thoroughness to which the spectra have been coded and upon the quality of the recording of spectra (of the compound to be identified and of the reference compounds), it is worthwhile developing systems in which these circumstances are given due consideration and the corrections are introduced automatically by the program itself. Such an attempt has been made.¹³

Penski, Padowski, and Bouck selected 60 IR spectra of organophosphor compounds. They applied a special method for coding the spectral features of these compounds. The bands were divided into eight classes with respect to the intensities and half widths, and each class was coded by a special code, for example, strong and narrow, strong and very broad, weak and narrow, weak and very broad, etc. Each band class was assigned a weight which is taken into account during comparison of the corresponding peaks in the spectra of the unknown compound and the reference compound. In comparing the peaks belonging to different classes, the comparison result is assigned a weight factor according to a matrix which gives these factors for all conceivable cases. The specimen preparation method (gas, liquid, KBr pellet, solution in CCl_4 , etc.) is coded by a special code, and the spectral region (in μm) not suitable for analytical work is shown in each case. The cell material, spectrometer type (grating, NaCl prism, LiF prism, etc.) are also coded in a similar manner. The program only collates those peaks that fall within the operating range in the spectra of the unknown and the reference compounds. A measure for the coincidence of two spectra is regarded as the quantity computed by a formula which takes account of the special data coding procedure, i.e., it is assumed beforehand that coincidence of maxima of two narrow-peaks is not equivalent to the coincidence of the maximum of a broad weak band with the maximum of a narrow strong peak, etc. The coincidence measure, M_i , is expressed in percent.

The system was tested as follows. The data bank comprising 60 spectra was so formed that it can include spectra of the same substance recorded under various conditions (different types of spectrometers, cells, and other factors). Each spectrum is considered to be a separate unknown entity in conducting the search in the general file. The degree of coincidence is estimated by the magnitude of M_i . Here it should be expected that if the spectra of the unknown compound and the reference compound had been recorded under identical conditions, M_i should be 100. Some typical results listed in Tables 1 and 2 show that, in comparing the spectra of the same compound, M_i always assumes a value greater than 80 to 90. This discrepancy from 100 is not due to the shortcomings of the system, but is caused by the differences in the spectrogram recording methods. Moreover, it was found that the identification was quite successful even with a small number of peaks (say, up to four).

The authors investigated how recognition is affected by the shifts of the whole spectrum which might occur due to the differences in the calibration of spectrometers or in coding. Shifts within 0.03 to 0.07 μm were found to improve coincidence by 10%. It is interesting to note that the attempts to recognize a substance absent in the machine atlas resulted in obtaining a compound of close structure with $M_i = 60$.

The experimental results have shown that the procedure proposed by the authors to overcome the difficulties caused by the differences in the recording conditions of spectra of the specimen and the reference is promising, and the method should be given due consideration in designing new atlas systems. This approach may, however, need a long machine time, and it may therefore be impracticable for large machine catalogs containing tens of thousands of spectra.

In addition to the papers reviewed above, there are several reports dealing with the

TABLE 1

Matching Results of "Unknown Spectrum" (Diethylchlorophosphate, Liquid Film)

Compound	Compared peaks	Match result (M _I)	Cell	Spectrometer
Diethylchlorophosphate (liquid film)	14	100.0	CsI	2.5—40 μ m grating
Diethylchlorophosphate (liquid film)	12	82.0	NaCl	NaCl prism
Triethylphosphate (liquid film)	12	70.3	KBr	NaCl prism
Triethylphosphate (liquid film)	14	64.2	CsI	2.5—40 μ m grating

Reprinted with permission from Penski, E. C., Padowski, D. A., and Bouck, J. B., *Anal. Chem.*, 46, 955 (1974). Copyright by the American Chemical Society.

TABLE 2

Search using the Spectrum of Tri-*n*-butylphosphate (Liquid Film, CsI Cell, and Grating Spectrometer [2.5—40 μ m] as the "Unknown")

Compound	Spectrum type	Match result (M _I)	Compared peaks
Tri- <i>n</i> -butylphosphate	Liquid film, CsI cell, (2.5—40 μ m) grating	100.0	14
Tri- <i>n</i> -butylphosphate	Liquid film, KBr cell, (2.5—25 μ m) grating	88.6	14
Tri- <i>n</i> -butylphosphate	Liquid film, KBr cell, prisms (2—25 μ m)	82.0	14
Di- <i>n</i> -butyl- <i>n</i> -butyl- phosphonate	Liquid film, KBr cell, (2.5—25 μ m) grating	77.7	14

Reprinted with permission from Penski, E. C., Padowski, D. A., and Bouck, J. B., *Anal. Chem.*, 46, 955 (1974). Copyright by the American Chemical Society.

various aspects of development of atlas systems.¹⁴⁻²⁰ We cannot dwell on them in detail due to lack of space. Therefore, we shall pass on to information retrieval systems designed to solve the important and complicated problem of identification of compounds in mixtures.

E. Identification of Compounds in Mixtures

Considerable difficulties are encountered in isolating a substance in its pure state. Even if one succeeds in doing so, the substance is often obtained in such minute quantities that it is impossible to record all the spectra needed. Consequently, the approach based on an atlas comparison of spectra of mixtures with reference spectra of pure substances seems particularly attractive. The first results of such investigations are reported in the papers.²¹⁻²³ Sebesta and Johnson²¹ were the first to point out one such procedure for identification of components in a mixture without preliminary separation. The MIRET system developed by them is based on the following assumptions:

(1) the spectrum of a mixture can be represented in the form of a linear sum of spectra of the components, (2) the intensity of the absorption peaks follows Beer's Law, and (3) the changes in the spectrum of a component in the mixture should not exceed certain specific limits. Hence, it follows that the spectrum of a mixture can be decomposed into subsets of spectra, each giving a solution to the problem; moreover, it is highly likely that the solution will fall into the set of spectra of substances to which a small number of absorption bands corresponds. The main difficulty here is how to reduce the number of solutions (number of subsets). The following criteria are applied in the MIRET system to screen out the superfluous solutions: (1) a good solution should give a better fit to the pattern of the spectral curve of the mixture than others do, and (2) the solution should contain the maximum possible number of bands coinciding with the bands of the specimen. The use of these two criteria is illustrated in the example shown in Figure 5, where (a) and (b) are the spectra of individual components, (c) is the spectrum of their mixture, and (d) represents the best approximation to the spectrum of the mixture under the assumption that it is the spectrum of an individual substance. Evidently, the spectra (a), (b), and (d) may be regarded as the solutions of our problem, but (d) will be taken to be the most probable one.

These general principles were realized in the form of an appropriate program which was tested by analyzing artificial mixtures, using IR spectra. The data bank comprised the machine atlases, ASTM, and the mixtures were made of only those substances whose spectra are contained in the atlases. The spectra of mixtures were coded, using enumeration of strong peaks, other peaks, and regions in which there is no absorption, and the spectral regions discarded in the investigation. In order to eliminate the influence of instrumental distortions, the positions of the strong peaks were widened to $\pm 0.1 \mu\text{m}$.

A provision is made in the MIRET system for the utilization of additional chemical information. This is quite understandable because a chemist can often predict which functional groups and chemical elements should not be present in the reaction products and which groups or elements will most likely or necessarily be present in the final products. In this system, it is possible to set information regarding 312 chemical properties, the code words being "present", "absent", and "no information".

The search strategy of solution consists of the following. First, using a negative logic, the maximum possible number of standard spectra are discarded which do not satisfy the chemical data or have peaks in the range where there is no absorption in the mixture spectrum. Then potential components are eliminated by discarding the compounds, which have simple, weak absorption bands in the IR spectrum. For each solution, the significance factor is calculated by comparing the solution with the spectrum of the mixture in which the maxima of the strongest bands have experimental wavelengths (without spreading by $\pm 0.1 \mu\text{m}$). The final result is stored in the form of three sets of the best solutions for 50 spectra in each set, the sets being formed with the help of three different methods of accounting for the chemical information. The list of 50 best solutions, ranked in decreasing order of the significance factor, is printed out.

The MIRET system was tested in the identification of individual substances and components of artificial mixtures. An analysis of a mixture containing anthracene and benzoic acid is treated in detail by the authors.

Of the 50 alternatives given out by the computer, four compounds (including anthracene) were characterized by a significance factor of 100. Benzoic acid was found in the next group consisting of nine compounds with a significance factor of 90.

Although it is somewhat early to speak of the practical utility of this system, none-

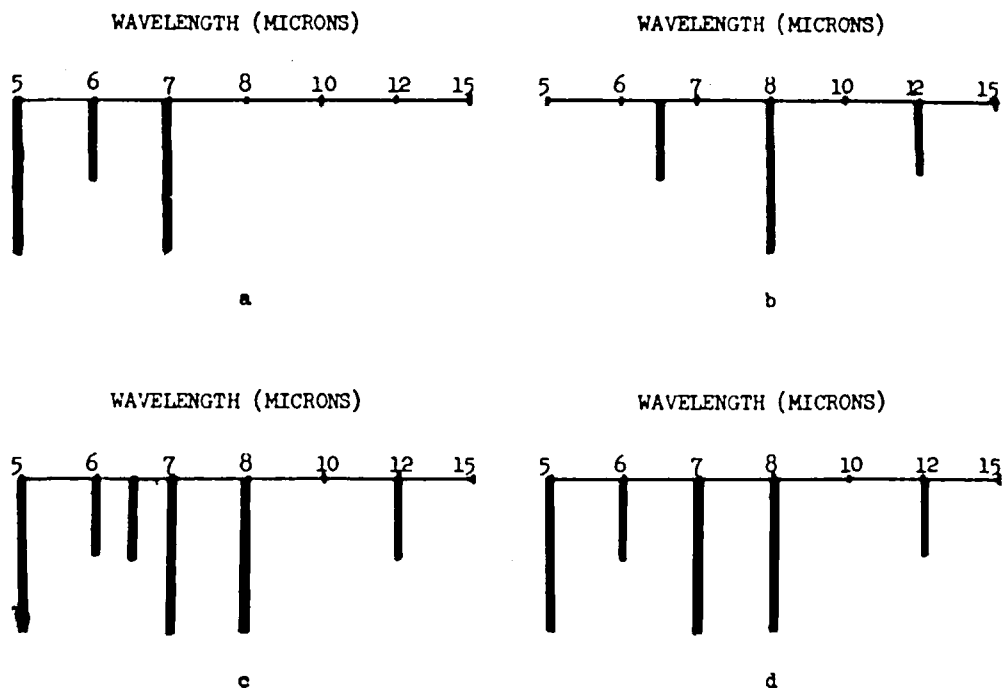


FIGURE 5. A simple explanation of MIRET multicomponent solution criterion. (Reprinted with permission from Sebasta, R. W. and Johnson, G. G., *Anal. Chem.*, 44, 260 (1972). Copyright by the American Chemical Society.)

theless, the restricted number of components (not more than about 20) should serve as a basis for planning further work in this direction.

An IRS designed for identification of organic compounds, both individual substances or mixtures, has been described by Fox in a recent paper.²¹

This system is based on machine atlases of IR spectra in which the position and intensity (in a three-grade scale) of peaks are coded. The degree of coincidence of the spectra of an unknown specimen and of the reference compound is estimated empirically, using a numerical factor in which the higher the value of the factor, the better the coincidence. The peaks are compared for the wavelengths and peak intensities. If the spectrum of a reference compound does not contain any one of the bands in the spectrum of the specimen, the reference spectrum is usually not considered as a competitive alternative. However, depending on the search strategy used, it may also be retained as a probable substance.

If the system is to operate effectively, it is important that there should be a proper balance between the intensities in the spectra of the reference compounds, i.e., the most intensive bands should not go beyond the scale. Special attention is paid to the coding of the strongest and medium intensities in the spectrum of the specimen, although as experience has shown, the program is not very sensitive to coding accuracy.

Several search strategies can be used with this system, the strategy chosen depending on the nature of the problem and the characteristics of the specimen. We shall now take up this question in brief.

The identification strategy for an individual compound lies in comparing the spectra of the specimen and the reference, and then discarding those reference spectra which have strong or medium peaks in the range where there is no absorption of the specimen. A similar procedure is used for the identification of mixtures, the only difference

being that the reduction in the band intensities of an individual substance in the solution is given due consideration.

If the spectrum of the specimen contains bands which are rarely found in the spectra stored in the data bank, this circumstance is also given due consideration in the program.

Frequently in identifying a mixture, it happens that several compounds having a close structure with a high degree of similarity are selected even after the first scanning. Thus, after identifying one of the components in the mixture, the remaining peaks in the specimen spectrum are artificially intensified by giving higher ratings (5, 6, 7 instead of the usual 1, 2, 3). This strategy helps in "pulling out" the spectrum of correct components of the mixture.

Fox called the last two strategies the liberal strategy and the simplest spectrum strategy, respectively. The liberal strategy is applied in those cases where the unknown spectrum is obviously not contained in the machine library, or when it is required to establish a correlation between certain peaks and functional groups and thus to detect allied compounds. Liberal strategy lifts those limitations imposed by the main strategy and does not admit elimination of references. The simplest spectrum strategy is always used in combination with the liberal strategy. In this strategy, each reference spectrum is assigned a number proportional to the number of peaks in the spectrum. Higher degrees of coincidence are attributed to those references which contain certain desirable peaks but have the least number of bands (simplest spectra). This strategy is efficient in searching for spectral structural correlations.

The system was tested using the atlases containing the IR spectra of commercial specimens. An experimental run has shown that these strategies, when used in different combinations, give good results. Individual substances almost always yielded to identification if the corresponding reference is found in the data bank. A stage-wise analysis was found necessary in analyzing mixtures. For example, in identifying pyridine-dimethylphthalate mixture, the first search was conducted with a strategy intended for mixtures, and thus dimethylphthalate was detected. Then a strategy was applied in which the intensities of the peaks not belonging to dimethylphthalate were intensified, thereby leading to the identification of pyridine. Another example is also given which illustrates how a commercial mixture of two oils was analyzed using only one strategy.

On the basis of experience, Fox suggested that identification of an unknown specimen should be conducted in four consecutive searches, using the following strategies: (1) strategy for the search of an individual substance, (2) the same strategy with due regard for the unusual peaks, (3) liberal strategy in combination with the strategy of simplest spectra (to detect allied compounds), and finally, (4) the strategy for analysis of mixtures. The final result of the search should be checked and analyzed by man.

The system developed by Fox is still under experimentation. At times, as Fox himself has noted, the results of analysis are obscure, and as such, the system needs further improvement. Nonetheless, as compared to the MIRET system, one step forward has already been made in this system: a flexible search strategy has been developed based on man-computer interaction. Judging by the examples cited, it is apparent that the system features high selectivity. Further advancement is probably to be expected in this direction.

F. Comparison of Spectra by Correlation Coefficient Method

As is evident from the works examined above, success in atlas comparison largely depends upon the similarity criterion used in the system. In turn, the choice of the criterion depends on the method employed for representing the spectra of the specimen and the reference compound. As has been demonstrated,¹³ a detailed description of

spectra carrying information on the band intensities and half widths and due consideration for the weight of the result of each comparison result certainly give the only spectrum and sometimes several similar spectra from the catalog. It is, therefore, very tempting to use the whole spectral curve if there is an appropriate method available for estimating the similarity between two complete spectra.

Recently Tanabe and Saëki²⁴ have reported some important results in this field. They have studied the feasibility of comparing the spectral curves, using the correlation coefficients. For this purpose, they recorded the IR spectra of 110 liquids, using a Perkin Elmer spectrophotometer Model 180 equipped with a digital recorder. A liquid cell 0.015 mm thick with KBr window was used. From these spectra, 6000 pairs were chosen, and the correlation coefficients between two spectra were calculated by the formula:

$$r = \frac{\sum x_i y_i - \frac{1}{n} (\sum x_i) (\sum y_i)}{(\left[\sum x_i^2 - \frac{1}{n} (\sum x_i)^2 \right] \left[\sum y_i^2 - \frac{1}{n} (\sum y_i)^2 \right])^{1/2}}$$

where x_i , y_i are the ordinates of two spectra (X and Y) at the i th point of the abscissa, and n is the number of reference points on the spectral curves.

The use of correlation coefficients reveals a number of important circumstances. A comparison of the histograms (Figures 6 and 7) shows that in the range 1750 to 650 cm^{-1} , ten pairs of spectra assume a value of 0.95 for the correlation coefficient, whereas in the range 1200 to 650 cm^{-1} , only one pair has a correlation coefficient exceeding 0.95. Hence, we can say that the correlation coefficient technique is most effective in the fingerprint region (this result was to be expected). This small but highly informative range constitutes only one sixth of the whole IR spectrum (200 to 3650 cm^{-1}), and this range was employed in comparing the two spectra.

The number of reference points needed for the calculation of r can be further reduced if we can determine the maximum intervals of quantization of a spectral curve in which the correlation coefficients do not suffer reduction. The dependences of r on the width of quantization interval for hexane-heptane, cumene-secondary butylbenzene, hexane-secondary butylbenzene are shown in Figure 8. It is obvious, that r begins to change when $\Delta\nu > 10 \text{ cm}^{-1}$. Hence, the range 1200 to 650 cm^{-1} can be described without any loss of information, using the ordinates at only 56 points.

Tanabe and Saëki also studied the effect of relative shifts in the frequency scale of two spectra of the same substance. From Figure 9, it is obvious that for r to be greater than 0.95, the frequency shift should not exceed 3 cm^{-1} . For greater shifts, corrections have to be introduced, and this calls for careful calibration of the spectrophotometer.

It is known that the pattern of the spectral curve changes, depending on the thickness of the absorbing layer. The spectra of the same substance recorded in cells of thicknesses differing by five to ten times may not exhibit any common features in a formal comparison. Therefore, in the literature it is recommended that the spectrum of the specimen to be identified by atlas comparison should be recorded in conditions where the absorbance of the strongest bands is no more than 1.0 or 1.5. Tanabe and Saëki have found that if the spectra are recorded in the absorbance scale (rather than in the transmission scale) for a variation in the cell thickness from 0.025 to 1 mm, the correlation coefficient between the corresponding spectral curves remains quite high (above 0.99).

Figures 10 and 11 show the dependence between the correlation coefficient and the purity of the specimen. The plots show that r does not decrease uniformly as the concentration of the impurity in the specimen increases. In the examples cited, r attained

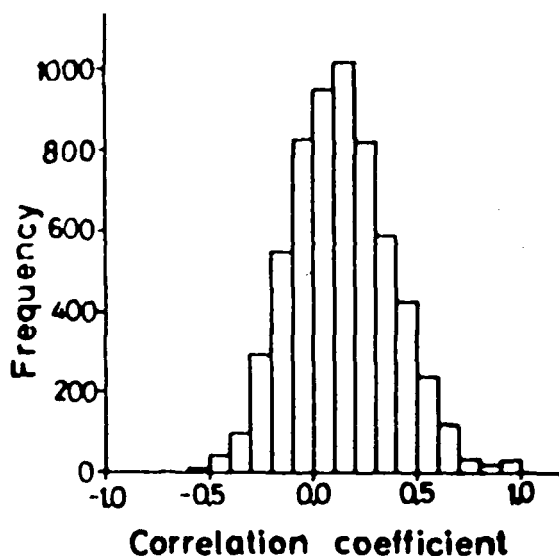


FIGURE 6. Histogram of correlation coefficients for the wave number range from 1750 to 650 cm^{-1} . (Reprinted with permission from Tanabe, K. and Saeki, S., *Anal. Chem.*, 47, 118 (1975). Copyright by the American Chemical Society.)

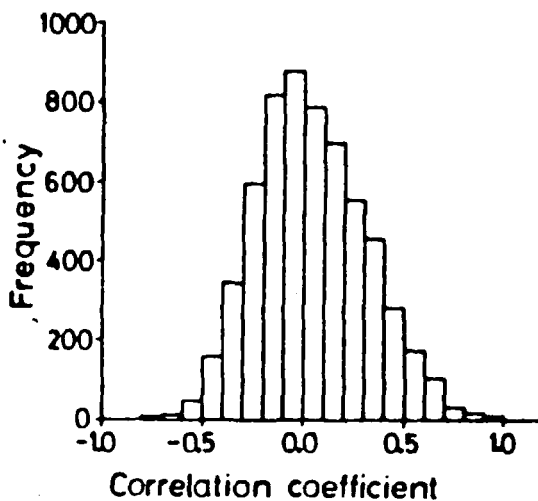


FIGURE 7. Histogram of correlation coefficients for the wave number range from 1200 to 650 cm^{-1} . (Reprinted with permission from Tanabe, K., and Saeki, S., *Anal. Chem.*, 47, 118 (1975). Copyright by the American Chemical Society.)

a value greater than 0.95 at a purity of 53, 80, 90, and 96%, respectively. It is therefore rather difficult to discern the permissible limits for the impurities. The authors nevertheless believe that for a purity of more than 95%, the correlation coefficient should be greater than 0.95 in all cases.

Moreover, they found that the resolving power of a modern grating spectrophoto-

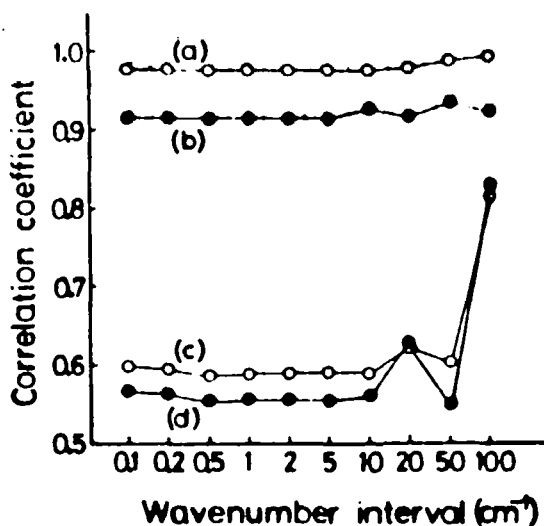


FIGURE 8. Correlation coefficients vs. wave number range for (a) hexane-heptane, (b) cumene-sec-butyl benzene, (c) hexane-sec-butyl benzene, and (d) heptane-cumene. (Reprinted with permission from Tanabe, K. and Saeli, S., *Anal. Chem.*, 47, 118 (1975). Copyright by the American Chemical Society.)

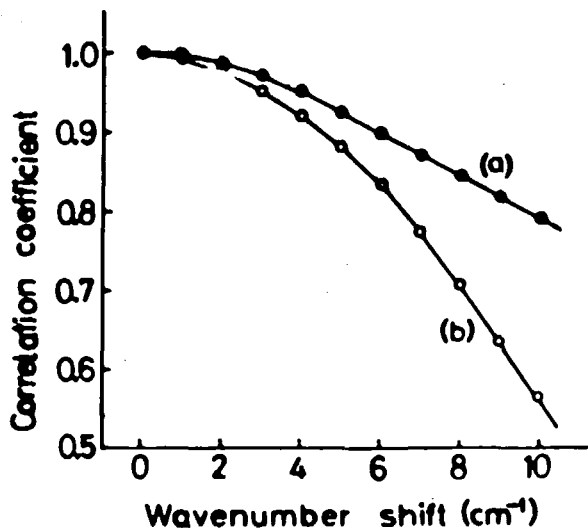


FIGURE 9. Correlation coefficient vs. wave number shift for (a) anisole, and (b) cyclohexyl bromide. (Reprinted with permission from Tanabe, K. and Saeli, S., *Anal. Chem.*, 47, 118 (1975). Copyright by the American Chemical Society.)

meter (1 to 2 cm^{-1} at 1000 cm^{-1}) is quite sufficient to retain a satisfactory similarity between the spectra of the solids and the liquids. The correlation coefficient, however, is rather sensitive to the method of preparation of the specimen, especially if account is taken of the differences in the spectra of a gas, liquid, and solid. In such cases, r proves to be far below 1.0. The correlation coefficient for spectra of liquids and solu-

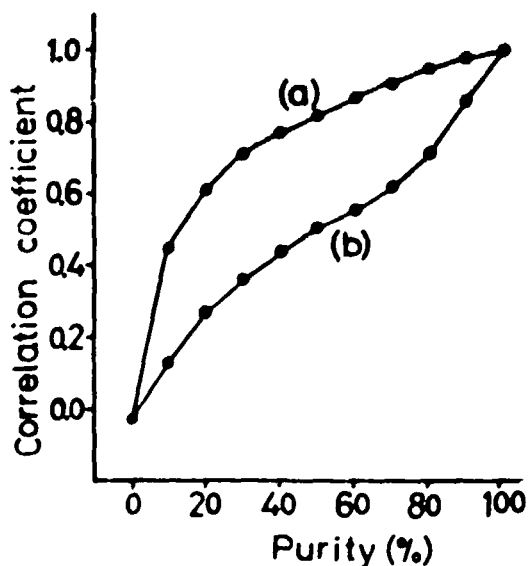


FIGURE 10. Correlation coefficients between (a) pure and impure *o*-xylene, and (b) pure and impure *m*-xylene. (Reprinted with permission from Tanabe, K. and Saeki, S., *Anal. Chem.*, 47, 118 (1975). Copyright by the American Chemical Society.)

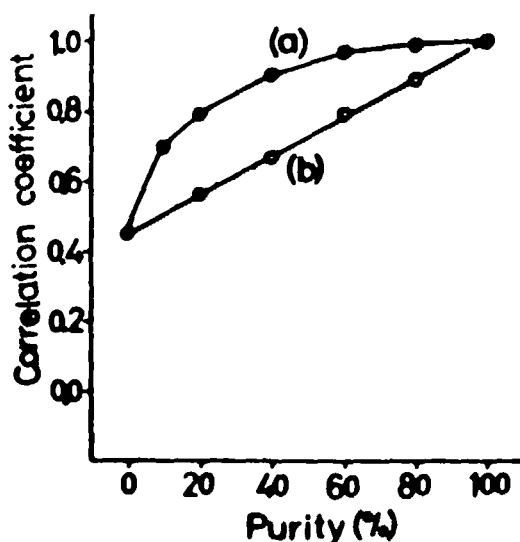


FIGURE 11. Correlation coefficients between (a) pure and impure propyl chloride, and (b) pure and impure propyl bromide. (Reprinted with permission from Tanabe, K. and Saeki, S., *Anal. Chem.*, 47, 118 (1975). Copyright by the American Chemical Society.)

tions is slightly less than unity. It is therefore imperative to record spectra either under certain standard conditions or the atlases should contain the reference compounds for all the possible recording methods. Since the latter procedure is not advantageous be-

cause in this case the computer memory and machine time have to be unduly large, the authors rightly tend to believe in recording the spectra of liquids without dilution and the solids in KBr pellets.

Systems in which the correlation coefficient is employed as the similarity criterion seem to be promising. In the comparison range 650 to 1200 cm^{-1} and quantization interval 10 cm^{-1} , they do not require an unduly large number of ordinates for spectrum coding. In the opinion of Tanabe and Saeki, not more than a few minutes are required to scan an atlas containing about 100,000 spectra. The main merit of these systems is that the identification result is not in the least affected by the various subjective factors that arise in considering the results of comparison. Moreover, for feeding the spectrum into the computer in on-line mode, the operator is called upon only to correctly prepare the specimen, without making any assessment of the spectral band parameters.

G. Deconvolution of a Spectrum of a Mixture into Spectra of its Components

To date, considerable experience has been gained in designing IRS for identification of individual compounds, and routes have been opened for the development of these systems for identification of components in a mixture. Since at present it is difficult to foresee the potentialities of mixture analysis (number of components, complexity of spectra, and other factors), a search has to be made for various ways of establishing the structure of the unknown components of a mixture.

The emergence of devices possessing high technical capabilities and digital outputs has stimulated the development of techniques for recording difference spectra. In a situation where the number and the concentration of $(n-1)$ components of an n component mixture are known, there is a definite possibility of finding the spectrum of the n th component by subtracting the resultant curve of the known components from the spectral curve of the mixture. The spectrum thus separated may be subjected to further identification by some IRS.

Vasil'ev et al.²⁵⁻²⁷ have worked out several methods of qualitative analysis for multicomponent mixtures by their IR spectra. They have designed such an algorithm and program suitable for quantitative analysis of the mixture, when not all of its components are known, regarding the spectrum of the unknown components as the background. After determining the concentration of the components, the background is separated in the form of a spectral curve which is used in identifying the unknown components. It should be mentioned that the computational procedures employed in this method are rather tedious and cumbersome and are based on linear and quadratic programming techniques as well as algebraic correction for the background. Moreover, this method presupposes that the main components in the mixture are known beforehand.

In a case where the number of components in a mixture and their concentrations are not known (such a situation is more frequent in practice), it appears that the components of a mixture cannot be easily identified. Hirshfeld, however, suggested an ingenious approach to this problem.³⁰ He has developed a method in which separation of a mixture is replaced by separation of the IR spectrum into its components. The method consists in the following.

Obviously, the IR spectrum $M(\nu)$ of the initial mixture can be represented in the form of the sum of the spectra of its components, $f_n(\nu)$:

$$M(\nu) = \sum_{n=1}^N f_n(\nu)$$

where N is the number of components (for the time being unknown). After redistri-

buting the concentrations of the components by partial fractionization with the help of evaporation, extraction, filtration, etc., the spectral curve of the mixture may be represented as

$$M_i(\nu) = \sum_{n=1}^N a_n f_n(\nu)$$

Now express the ratio of the spectrum as follows:

$$R(\nu) = \frac{M_i(\nu)}{M(\nu)} = \frac{\sum a_n f_n(\nu)}{\sum f_n(\nu)}$$

The function, $R(\nu)$, will have a constant magnitude, a_n , in those frequency ranges where $f_n(\nu)$ dominates, and consequently, the graph of this function can be represented in the form of flat sections of height, a_n , alternating with the curve. Using this plot, we can find the value of a_n and the number of unknown components, N . Of course, each component should have at least one flat section. This requirement is the main constraint on the number of possible components which can be identified by this method because there is an increasing danger of superposition of the bands with the increasing number of components in the mixture. This naturally reduces the reliability of the results. A weaker constraint is contained in the requirement that $a_n \neq a_r$. In the contrary case, both the substances (N and X) will give rise to a resultant spectrum in the course of separation. Moreover, we should bear in mind that always $a_n \neq 1$.

This method, as the author points out, always works out if $N = 2$; it works out in the majority of cases if $N = 3$, and is rarely effective if $N \geq 4$.

In general, after determining the number of components, N , partial fractionization is repeated $(N-2)$ times, and thus the set of $(N-1)$ spectra are obtained:

$$M_i(\nu) = \sum a_{i,n} f_n(\nu)$$

with the help of which the values of $R_i(\nu)$ are computed:

$$R_i(\nu) = \frac{\sum a_{i,n} f_n(\nu)}{\sum f_n(\nu)}$$

Then a system of linear equations is derived for $f_n(\nu)$ ($M_i(\nu)$ and $a_{i,n}$ are found experimentally). The spectrum of each component is determined from the solution of this system. Thereafter, the spectra can be identified, using any one of the IRS methods.

This procedure, according to the author, needs a spectrometer of resolving power of not less than 0.01 cm^{-1} , high signal-to-noise ratio, and computer output. The IR Fourier spectrometer Digilab FTS may be used for this purpose. High precision spectrometers are evidently needed not only in fundamental research, but also in identification of mixture composition. Such a change in the outlook on the purpose of instruments is very essential.

Hirshfeld illustrated his method with an example of separation of the spectra of a mixture containing toluene, cyclohexane, and hexane. The results obtained (see Figures 12 through 15) confirm the validity and the efficiency of the method. It may be expected that with the increasing availability of high precision spectrometers this method will find ever increasing application, and in the case of simple mixtures containing two or three components, it may even come out as a serious competitor to chromatography due to high rapidity, ease, convenience, and reliability of identification.

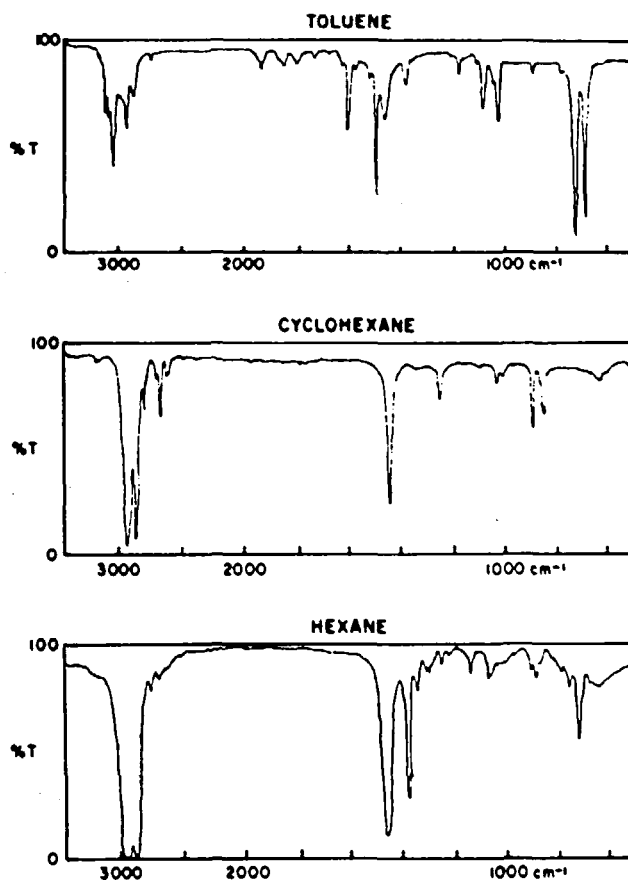


FIGURE 12. Reference spectra of pure components of ternary mixture. (Reprinted with permission from Hirschfeld, T., *Anal. Chem.*, 48, 721 (1976). Copyright by the American Chemical Society.)

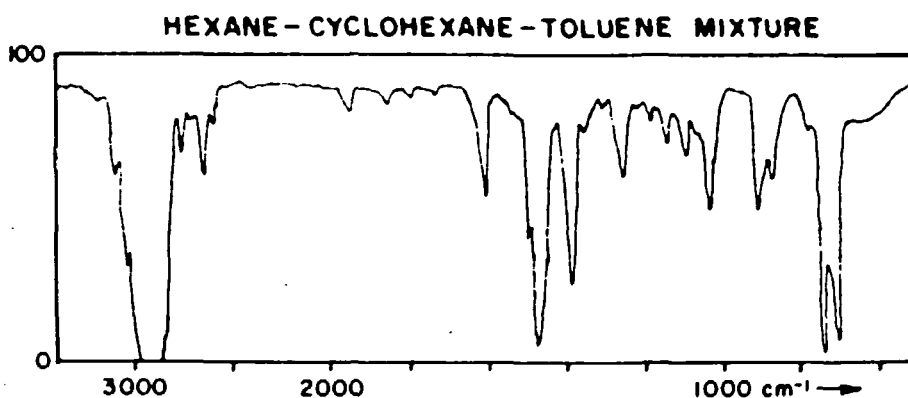


FIGURE 13. Spectrum of toluene-cyclohexane-hexane mixture. (Reprinted with permission from Hirschfeld, T., *Anal. Chem.*, 48, 721 (1976). Copyright by the American Chemical Society.)

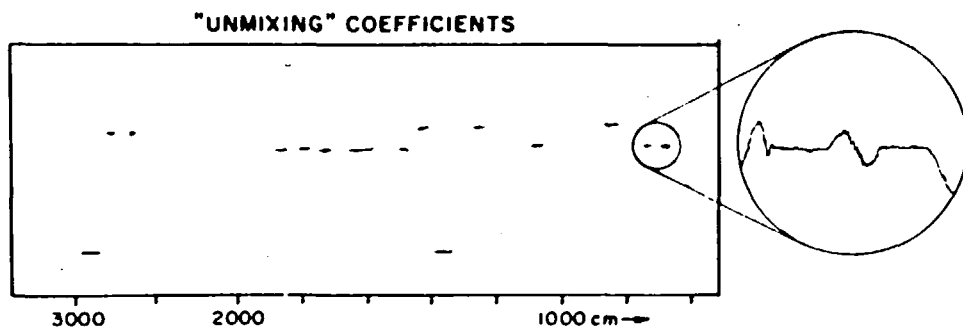


FIGURE 14. Ratio of absorbance spectra of two successive mixtures. (Reprinted with permission from Hirschfeld, T., *Anal. Chem.*, 48, 721 (1976). Copyright by the American Chemical Society.)

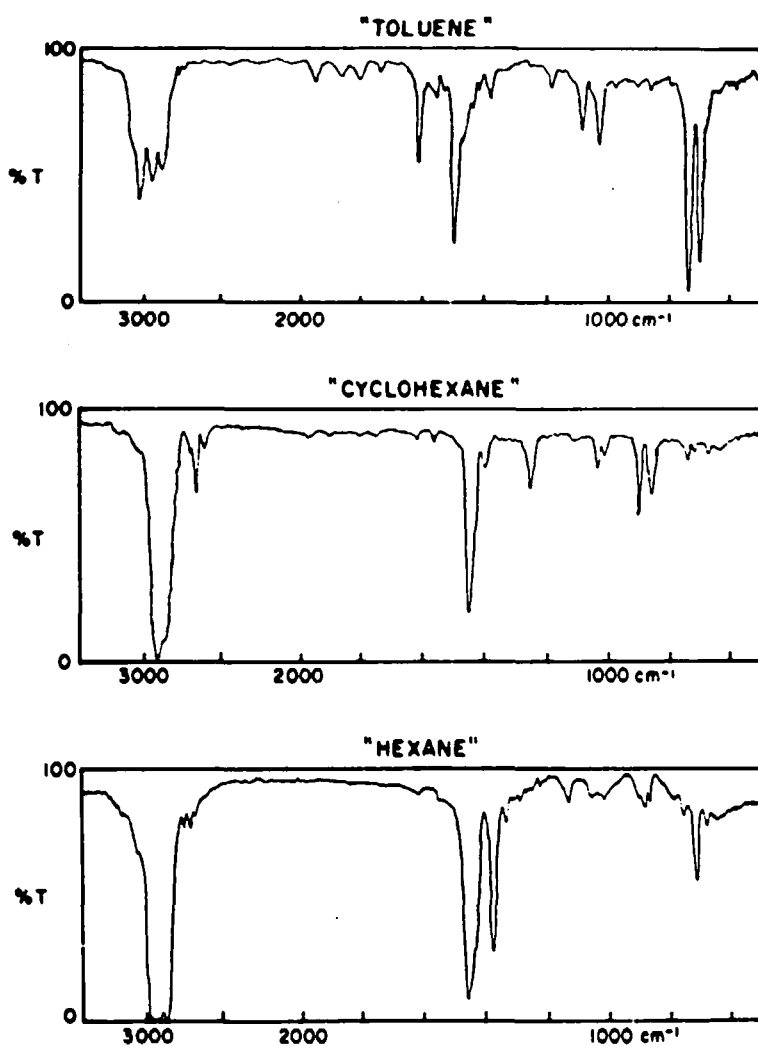


FIGURE 15. Computer-resolved spectra of mixture components. (Reprinted with permission from Hirschfeld, T., *Anal. Chem.*, 48, 721 (1976). Copyright by the American Chemical Society.)

III. SYSTEMS BASED ON PATTERN RECOGNITION METHODS

A. Elements of the Pattern Recognition Theory

1. Essence of the Pattern Recognition Method

To date, quite a number of papers have been published dealing with pattern recognition methods on various levels, ranging from simple theoretical consideration to a description of algorithms.³¹⁻³⁵

In mathematical language, a pattern is an ensemble of phenomena with common properties. Thus, the IR spectra of a particular class of compounds, say, of the type $\text{HC} \equiv \text{CR}$, may be grouped into one class if they have an absorption band corresponding to vibrations of $\text{C} \equiv \text{C}$ bond in the interval from 2140 to 2100 cm^{-1} . This criterion may be used in identifying these compounds.

Learning always precedes recognition. It consists of the following. The computer is fed with a sufficiently large number (a few hundred) of spectra of chemical compounds forming the training set. The elements of this set are usually so chosen that about half of them contain a given atomic group which determines their state of belonging to class A. Class B, in which the second half of the remaining elements are grouped, does not contain a given structural element. Then an analysis is made of the distribution of the points of classes A and B in a multidimensional space of the spectral features. Here a distinction is made between learning with a teacher and self-learning. In the latter case, algorithms are chosen that automatically classify the objects presented for learning. On the basis of certain statistical algorithms, the computer reveals the typical features of each class of compounds, and thus sorts out the spectra into several subsets. If these features are already known, then they can be stored beforehand in the computer.

We shall now illustrate the construction of the system based on a pattern-recognition algorithm with the following example.

Suppose that a compound has a number of features that can be characterized quantitatively; for example, the position of intense lines on the spectral scale, melting point, solubility, etc. Each of these features can be plotted on some coordinate axis, and the set of all these features form a point in a multidimensional space. We shall illustrate this by a simple example of identification of carbonyl compounds and monoalkylacetylenes.³⁶ We shall characterize each compound by the percentage of the IR absorption band in the range 1700 to 1760 cm^{-1} and 3300 cm^{-1} . On plotting the percentage of absorption at 3300 cm^{-1} along the ordinate, we obtain the plot shown in Figure 16. Points corresponding to different compound classes fall into different regions: those corresponding to carbonyl compounds fall in region 1 and those of substituted acetylenes in region 2. These two regions may be separated by a straight line. On receiving the data on the percentage of absorption at 1700 and 3300 cm^{-1} , the computer checks on which of the regions a particular point falls, and then answers to which of the classes the given compound belongs.

In a real system, the number of features characterizing a compound may be quite large, sometimes more than 250. If the points corresponding to compounds belonging to one class in such a multidimensional space are grouped in a particular domain without overlapping with the domain defining other classes of compounds, then by determining the location of the points of the compound to be identified, we can solve the analytical problem. The boundaries of the domains in which the points belonging to different compound classes fall are recognized by learning.

Unfortunately, the domains characterizing different classes often overlap in a great many cases. In spectral analysis and in other methods, not all the specimens exhibit

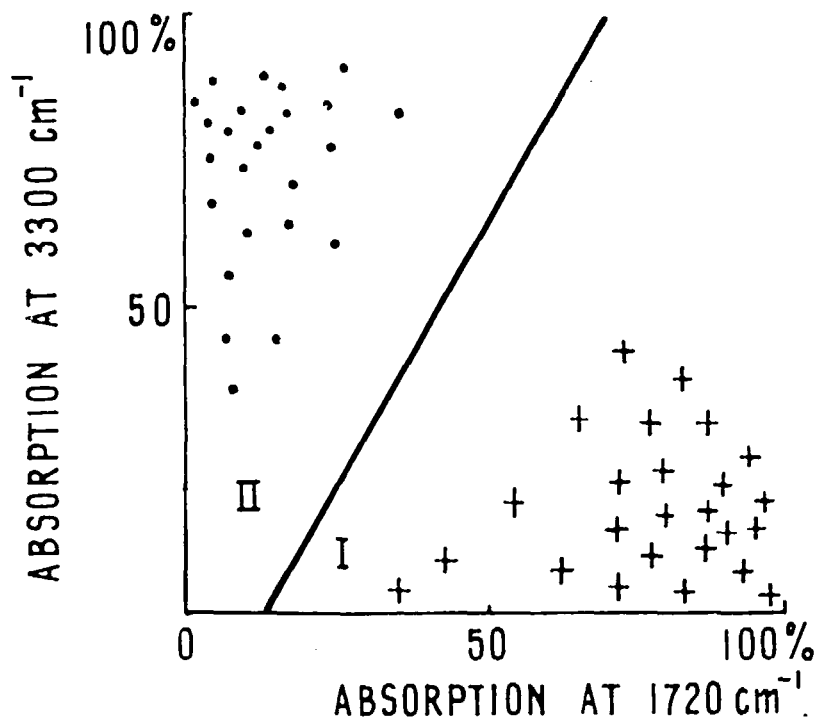


FIGURE 16. Separation of acetylene derivatives and carbonyl compounds in the space of features.

distinct boundaries, and consequently, the recognition programs cannot always guarantee errorless classification.

We shall now take up the techniques of the pattern recognition theory as applied in analytical practice.

2. Probability Approach

If a specific probability measure, P_v , defined in a description space is put in correspondence with each pattern, then the image, $S(X)$, may be regarded as some (or all) of the points or bands in the spectrum that have to be used in determining the probability that this image belongs to one of the patterns. The conditional probability that an object belongs to one of the classes A or B (probability of the pattern, V_i , provided the feature X_i is exhibited) is calculated from the statistical data by the formula:

$$P(V_j / X_i) = P(V_j) [P(X_i / V_j) / P(X_i)]$$

Application of algorithms in the probability approach is restricted due to the need at the learning stage for extensive experimental data if the information on the distribution pattern is to be reliable. If the features are mutually correlated, as is often the case in practice, the volume of computations needed in decision-making becomes excessively large. This is an important point. In fact, in spectroscopic practice, we are rarely faced with a situation in which a few hundred spectra or more are used for learning. Even most of the large search systems, which can be utilized in learning, have at their disposal data on a comparatively narrow range of compounds of a definite class. This is particularly true of newly synthesized compounds.

3. Geometric Approach

The geometrical approach to pattern recognition is based on the assumption that in the feature space (in a multidimensional space where each coordinate represents a known property of a compound expressed in the form of numbers) one domain pertaining to one compound class is distinctly separated sufficiently from a domain characterizing another class. In the learning sequence, the set of n -dimensional vectors, $[XA_p]$, represent molecules belonging to one pattern, and the vectors $[XB_p]$, molecules belonging to another pattern. Moreover, there should be test sampling for each of the patterns represented by the respective vectors. An image under test, $V_k (V_1, V_2 \dots V_n)$, which is represented by a point in a multidimensional space, is attributed to the patterns, $[XA]$ or $[XB]$, after its similitude to some references A or B has been established:

$$V_k \in \{XA\} \text{ if } R(V, A) < R(V, B)$$

In simple cases, the distance in the Euclidean space is used as a measure of similarity:

$$R(V, A) = \left\{ \sum_{n=1}^N [W_n (V_n - A_n)]^2 \right\}^{1/2}$$

where W_n is a weight factor and A_n is a reference coordinate vector.

In the potential function method, each point corresponding to a particular object is assigned some function similar to the electrical potential. This potential is maximum at a certain point and decreases in all directions from this point. An example of such a potential function is

$$\psi_R = \frac{1}{1 + \alpha R^2}$$

where α is the decrease rate factor, and R is the distance between the source point and the point at which the potential is calculated.

The quantity Ψ in this case may serve as the measure for similarity between a given point and the source point. For the references A and B, we may take certain volumes in a multidimensional space which are specified either by man (learning with a teacher) or found with the help of a separate algorithm and program (self-learning).

The geometrical approach is a common technique in pattern recognition. At the learning stage it offers a possibility of varying the choice of similarity measure, minimizing criterion, dimensionality of the initial space, and optimal references. Besides the similarity measures which we considered earlier, namely, the distance in the Euclidean space and elementary nonlinear potential functions, there are many other measures, too. For example, if there is a reference defined by the vector, $V (\{X_1^0, X_2^0, \dots X_n^0\})$, and the test object, $S (\{X_1, X_2, \dots X_n\})$, then the question whether the test S belongs to the class represented by reference V is decided upon after calculating the angle between the vectors:

$$R(S, V) = \arccos \frac{SV}{|S| |V|} = \arccos \frac{\sum_{i=1}^n x_i x_i^0}{|S| |V|}$$

For binary features it is more convenient to use the simple similarity measures, i.e., distance in the Hemming space or the scalar products of vectors:

$$R = |x_1 - x_1^0| + |x_2 - x_2^0| + \dots + |x_n - x_n^0|$$

$$R = x_1 x_1^0 + x_2 x_2^0 + \dots + x_n x_n^0$$

Frequently, recourse is taken to the correlation method in which

$$R = (x_1 x_1^0 + x_2 x_2^0 + \dots + x_n x_n^0) - \frac{1}{n} (x_1 + x_2 + \dots + x_n) (x_1^0 + x_2^0 + \dots + x_n^0)$$

An essential property of recognition systems is their ability to generalize the data derived from a comparatively small number of representatives. In this respect, the pattern recognition theory is an excellent example of the fundamental idea underlying scientific progress — data compression without loss of the information contained in them.

A comparison of the dispersion of the various features describing the pattern under test may reveal the extent to which each feature belongs to a given pattern. The choice of references representing a pattern has a great bearing on the recognition results. As a reference, we may take a point in a multidimensional space, the coordinates of which are the mean value of the sampling coordinates. The references should be such that in terms of “proximity measure” they should be as far away as possible from the extraneous patterns and as close to their own patterns as possible. In the recognition of complex patterns, a pattern is sometimes represented by several references.

4. Characteristics of Recognition Systems

The identification systems based on pattern recognition principles may be characterized by several parameters, of which the most important are the recognition power, R , predictive ability, P (probability of correct prediction), reliability, and convergence rate.³⁷ The recognition power is the percentage of members of the training set properly classified in the course of learning.

If, for example, the training set has been completely separated, then obviously $R = 100\%$. The probability of making correct predictions is a quantitative measure of the ability of the method to correctly classify new patterns which had not taken part in the learning process, whereas reliability characterizes the ability to correctly classify elements of the training set when there are distortions (for instance, spectra recorded with the help of various types of spectrometers or samples prepared under different conditions). The convergence rate of the classification process determines the rate at which the preset goal is attained in the course of learning. This parameter gives a means to judge how far a particular method is economical in practice.

These characteristics of a system cannot be theoretically calculated beforehand. They not only depend on the classification methods used, but also to a large extent on the method of formation of the training set and preliminary mathematical processing of the initial data. It is desirable that the training set be as large as possible, but at the same time, it should be remembered that the set size has a significant influence on the convergence rate. The order in which the elements are arranged in the training set affects the quality of the learning process. The training set is therefore randomized before starting the work with an algorithm, i.e., the elements of the training set are arranged in a random manner. Of equal importance is the proper choice of the dimensionality of the feature space (the number of descriptors) and the method chosen for the preliminary data processing. Such methods as scaling, weighting of features, Karunen-Loeve expansion, and other techniques (it is not possible to dwell on them in detail due to lack of space) are treated in detail by Kowalski,³⁸ but so far there is no theoretical criterion which can be used for choosing the method of preliminary processing. As a rule, it is decided upon empirically.

The basic ideas described above are employed both in learning with a teacher and in self-learning as well. The brief analysis made above has demonstrated that as a rule we can distinguish a small number of different patterns. Thus, using the algorithms of the pattern recognition theory, we can determine to which class a given compound belongs, but as a rule, we cannot identify a separate compound. If the whole analysis ends in determining the fact that a given compound belongs to a particular class, then methods based on the pattern recognition theory may be used very effectively. The problem today, however, is quite a different one, namely, the complete identification of compounds. For such purposes, pattern recognition techniques often prove unsuitable because the algorithms are so cumbersome that it is impossible to use them in practice. Here, in our opinion, lie the main limitations due to which these methods do not seem very promising.

We shall now take up certain methods for the purpose of solving some spectroscopy problems.

B. Alternative Recognition and Linear Classifiers

Quite interesting is the classification of IR spectra,³⁹ using an alternative recognition program with a learning algorithm. The main idea consists in the following: all the possible combinations of several spectral features (frequencies) present in the spectra contained in the training set are examined, and then the most informative combinations of features are chosen. These combinations are then used in working out a decision-making rule, according to which a given spectrum is assigned to one of the two classes. The effectiveness of the algorithm was verified by determining the number and places of the branches in the paraffin chains and the number and places of the substituents in aromatic compounds. An interesting result is that the spectral intervals reported in the literature were found to be unsuitable for use in classifying benzene derivatives with respect to the 12 possible types of substitution.

Quite a large number of papers^{37,38,40-48} have been devoted to pattern recognition based on division of the spectra into two sets by means of a linear classifier. We shall now discuss the essence of this method.

Suppose that certain objects to be sorted into two classes are given in a d -dimensional feature space. Place a vector $X(X_1, X_2, \dots, X_d)$ in correspondence with each object. Then we attempt to find a plane in the feature space that will divide the objects into two categories. We shall impose a condition that this hyperplane should pass through the origin. For this, add a new $(d + 1)$ th dimensionality that would give a new vector, $Y(Y_1, Y_2, \dots, Y_d, Y_{d+1})$. Take the Y_{d+1} component of each vector (pattern) to be unity.

Now we have to find a criterion that can be used to determine on which side of the plane a given object lies. For this purpose, take a vector, W , through the origin normal to the separating plane. Call this vector the weight vector. Since W is perpendicular to the plane, the scalar product of the vector, W , and the vector of the pattern, Y , will determine the position of this pattern relative to the hyperplane. Thus, we have

$$S = W \cdot Y = |W| |Y| \cos \theta$$

where θ is the angle between the two vectors. Since $|W|$ and $|Y|$ are always positive

$$S > 0 \quad \text{if} \quad -90^\circ < \theta < 90^\circ$$

$$S < 0 \quad \text{if} \quad 90^\circ < \theta < 270^\circ$$

Hence, it is obvious that the position of the object relative to the hyperplane is deter-

mined by the sign of S . Thus, this position can readily be found with the help of a computer:

$$S = W \cdot Y = w_1 y_1 + w_2 y_2 + \dots + w_d y_d + w_{d+1} y_{d+1}$$

In order to derive the decision-making rule for this classification, it is necessary to use a training set, the elements of which are assigned beforehand to one of the two classes. The first element of the training set is taken in an arbitrary manner, and then the sign of S is calculated. If this sign is correct, i.e., the compound falls in its proper class for the vector, W , taken, then we proceed to choose the next element. This procedure is continued until we find an object with the wrong sign of S (evidently, such an object does not fall within its class according to the position relative to the decision surface). In order that the classification may not be false, correction is introduced into the position of the partition plane by means of a feedback as follows: the hyperplane is shifted along a perpendicular drawn from the wrongly classified point until this point is located at the same distance (S) from the plane, but now is on the correct side.

The new position of the hyperplane is now characterized by a new weight vector, W' , so that

$$W \cdot Y_i = S, \quad W' \cdot Y_i = -S \quad (1)$$

We shall find the vector W' in a form

$$W' = W + c Y_i \quad (2)$$

where c is a certain factor.

From Equations 1 and 2, we find that

$$S' = W' \cdot Y_i = (W + c Y_i) \cdot Y_i$$

hence

$$c = \frac{S' - S}{Y_i \cdot Y_i}$$

Since by assumption that $S' = -S$, we have

$$c = \frac{-2S}{Y_i \cdot Y_i}$$

the expression for the new weight vector takes the form:

$$W' = W - \left(\frac{2S}{Y_i \cdot Y_i} \right) Y_i$$

For the new vector, W' , we have to verify whether the classification of the preceding elements, Y_1, Y_2, \dots, Y_{i-1} has not been altered. An appropriate correction is introduced in case the previous classification has been disarranged. This process is continued until all the elements of the training set are classified in a proper way.

After the problem has been solved and a vector, W^* , that properly classifies all the elements of the training set has been found, it may be taken that the computer has

learned to recognize these objects. In order to assign a new object not contained in the training set, it is sufficient to calculate $S = W^* \cdot Y_*$; the sign of S will determine to which class a new element belongs.

Although the possibility of linear classification is not proved by the theory, practical experience evidently shows that this classification can often be effected within reasonable machine time, using a finite number of feedbacks. In those cases where a large number of iterations fail to divide the objects into classes, it is not clear whether a desirable result can be obtained by continuing the calculations. Obviously, the algorithm described can also be used in classifying the initial set into a large number of classes.

The linear classification algorithm was tested by solving the problem of identification of the structural elements of organic molecules by different kinds of spectra. A series of papers published since 1969 deal with this topic. In this review we shall take up just a few of them. The main aim of these works was to elucidate the way in which recognition is affected by various factors like the size of the training set, dimensionality of the feature space, the choice of the initial values of coordinate vectors, etc.

Kowalski et al.⁴⁰ have examined the feasibility of applying linear classifier in interpreting IR spectra and in separating chemical compounds into classes according to spectral features. For this purpose, the first 4500 spectra that correspond to compounds of a composition of no more than $C_{10}O_4N_3H_n$ were selected from the Sadtler atlas. The vectors of the patterns were selected in a space of 130 descriptors (including the $(d + 1)$ th). This means the quantization interval was $0.1 \mu m$. The peak intensities were estimated in a four-grade scale: the strongest peak in the spectrum was assigned an amplitude equal to 3, the strongest peak in the interval $1 \mu m$ was given 2, and other descriptors were assigned either 1 (band exists) or 0 (no band). In order to feed the data into the computer, they were preliminarily compressed with due regard for the large number of zeros among vector components. Thus there was a 20-fold saving in machine time.

For the purpose of learning, 500 spectra were chosen at random; 3500 spectra not used in learning were employed in assessing the predicting ability. In the course of learning, the training set was divided into 19 classes by the presence of one of the 19 functional groups in the molecular structure ($COOH$, $CH_2 - OH$, NH_2 , etc.). Classification was effected with 100% result. The predicting ability P , however, varied from 75 to 95% for different groups, a percentage far higher than the results of man-assessed intuitive classification (the effectiveness of man-made classification was preliminarily assessed by the authors).⁴⁰

The authors have also studied the effect of the training set size on the probability of correct prediction in identification of substances containing a carbonyl group. From Table 3, it is evident that a gradual increase in the number of spectra in the training set from 10 to 500 gives only a slow increase in the prediction ability from 62 to 71%. It was therefore inferred that a training set composed of 300 spectra is quite sufficient for finding a satisfactory weight vector. Consequently, the authors chose a training set and a predictive set, each containing 300 spectra, for their further experiments. With the help of these sets, they found that if the initial weight vector is so taken that all its components are equal either to $+1$ or -1 , the predictive ability is approximately the same in both cases.

The authors have pointed out that there is a region near the partitioning hyperplane in which the points fall that are difficult to classify. These regions remain outside the range of application of the method. But the reliability with which an element is assigned to a class can be estimated by means of the numerical value of S . Indeed, the value S shows how close a given point is to the plane.

TABLE 3

Training for Carboxylic Acids with Training Sets of Different Sizes

Training set size	Number of feedbacks	Prediction percentage
10	5	62
20	10	63
30	10	61
40	15	56
50	22	61
100	36	67
150	103	71
200	167	71
250	199	72
300	341	73
350	398	71
400	597	71
450	1816	68
500	—	71

Reprinted with permission from Kowalski, B. R., Jurs, P. C., and Isenhour, T. L., *Anal. Chem.*, 41, 1945 (1969). Copyright by the American Chemical Society.

Table 4 lists the correct and incorrect classification of compounds containing carboxyl groups, depending on the amplitude of the scalar product. It is clear that the degree of confidence in the classification grows rapidly as S increases and attains 100% for large amplitudes.

C. Recognition by Means of Simultaneous Use of IR and Mass Spectra

It is known that the use of various physical methods for solving structural problems considerably enhances the effectiveness of the experiments due to the large amount of information obtained. It is therefore quite natural to inquire what is to be expected in this respect if a machine is taught to recognize patterns. Such a study was attempted by Jurs et al.⁴¹ Using certain cases, they investigated the effectiveness of recognition by means of the combined use of IR and low resolution mass spectra and information about the boiling and melting points. For this purpose, they selected about 300 compounds for which all these data were available. The IR spectrum was represented by a 130-component vector constructed per Kowalski et al.⁴⁰ The mass spectrum was defined by 132 positions with amplitudes in the range from 10 to 100. About 200 compounds were included in the training set, and 100 substances were used in prediction. Learning was conducted in a space of combined (IR and mass) patterns described by vectors composed of 262 components. In the course of learning, an investigation was concurrently made to find how a decrease in the number of descriptors affects the recognition results.

Compounds containing C=C double bond, ethyl, and vinyl groups were subjected to identification by this method. In identifying double bonds, it was found (see Table 5) that each of the two methods (IR and mass spectra) taken separately gives an extremely modest result. In joint utilization, however, the learning process and the recognition power depend on the order of magnitude of the intensities in the IR and mass

TABLE 4

Confidence Intervals for Carboxylic Acid Prediction

Calculated scalar	Number of correctly assigned spectra	Number of wrongly assigned spectra	Total	Confidence percentage
-16	1	0	1	100
-14	8	0	8	100
-12	11	1	12	92
-10	22	2	24	92
-8	21	1	28	75
-6	39	9	48	82
-4	41	13	54	62
-2	41	17	58	57
0	37	29	66	
+2	48	24	72	67
+4	47	27	74	63
+6	43	11	54	80
+8	27	7	34	79
+10	26	5	31	84
+12	13	4	17	76
+14	11	2	13	85
+16	5	0	5	100
+18	2	0	2	100
+20	0	0	0	
+22	1	0	1	100

Reprinted with permission from Kowalski, B. R., Jurs, P. C., and Isenhour, T. L., *Anal. Chem.*, 41, 1945 (1969). Copyright by the American Chemical Society.

TABLE 5

Detection of Double Bonds

Number of descriptors	IR P	Mass P	IR + Mass I(M) > I(IR)		IR + Mass I(M) < I(IR)		IR + Mass I(M) ~ I(IR)	
			P	M/IR	P	M/IR	P	M/IR
262			86	136/126	78	136/126	89	136/126
162			84	99/6	80	52/110	88	76/86
125	79	87	87	92/33	83	24/101	89	61/64
100	83	88	85	88/12	80	11/89	90	46/54
70	82	85	85	69/1	81	1/69	89	30/40
50	77	84	85	50/0	78	0/50	89	23/27
30	62	81	79	30/0	69	0/30	92	13/17
20	47	81	82	20/0	52	0/20	88	9/11
10	50	69	61	10/0	52	0/10	75	6/4
5	69	56	61	5/0	55	0/5	62	4/1

Note: P stands for prediction percentage, M/IR is the ratio of numbers of descriptors in the mass and IR spectra, and I(M) and I(IR) are the intensities of the mass and IR spectra, respectively.

Modified from Jurs, P. C., Kowalski, B. R., and Isenhour, T. L., *Anal. Chem.*, 41, 1949 (1969). Copyright by the American Chemical Society.

spectra. Moreover, recognition is mainly effected by that spectrum in which the intensities are represented by numbers of higher orders of magnitude. On normalizing each type of spectra so that the amplitudes of the major peaks are equal, the degree of utilization of both spectra in recognition is the same. This can be judged by the number of IR and mass spectral parameters retained.

In the course of filtering the superfluous descriptors, certain vectors were found which, even for a comparatively low dimension (about 20 to 50), are capable of guaranteeing a sufficiently satisfactory prediction ($\sim 90\%$). The recognition was hardly affected on inclusion of the boiling and melting points as additional descriptors, but they were retained to the end of the learning process as the number of components was gradually decreased.

With respect to the identification of the ethyl group, joint utilization of IR and mass spectra did not lead to any perceptible increase in the prediction percentage as compared with the results obtained using mass spectra alone. The only advantage of combined utilization is that the convergence rate becomes slightly faster.

Experiments have shown that the main role in the identification of the vinyl group is played by the IR spectrum, which gives satisfactory convergence and recognition when 20 descriptors are used. The authors note that the inclusion of mass spectra will in no way lead to loss of information, the fact being that a mass spectrum has no discriminating power in such a case.

This has demonstrated the convenience of linear classification in elucidating the significance of intensity during the recognition process. Identification of a double bond is not at all more difficult if zero-unity vectors (presence or absence of a given feature in a given interval) are used in learning.

The results^{40,41} show that an approach based purely on a formal mathematical technique may give results which are difficult to interpret from a commonsense viewpoint, and extreme caution has to be exercised in applying such a method. At the same time, certain conclusions, which were inferred in the formal approach, are not objectionable to the spectroscopist. For instance, the account that the intensity has hardly any effect on the identification of double bonds is not an unexpected result. Even a weak band in the range from 1640 to 1670 cm^{-1} serves as an indication of the presence of the $\text{C}=\text{C}$ bond.

D. Some Recognition Algorithms and Establishment of the Most Important Spectral Features

For the sake of convenience in applying the pattern recognition method, it seems tempting to work out such algorithms for constructing a weight vector that, due to rapid convergence and high recognition, might lead to the vector \mathbf{W} which depends on a small number of descriptors. A whole series of techniques were tested⁴²⁻⁴⁴ with this end in view. Since it is impossible to foresee the result in every case, each method was tested empirically. Here we shall outline these approaches.⁴²⁻⁴⁴

In developing the basic method of binary pattern classification, a new approach was proposed⁴² which was called the method of learning with a normalized weight vector. It consists in the following. As already shown elsewhere, the position of a point relative to the decision hyperplane is determined by the scalar product of the vectors, \mathbf{X} and \mathbf{W} . Let us introduce a threshold number Z , and impose a condition that the scalar product, $\mathbf{W} \cdot \mathbf{X}$ be greater than Z until the vector falls into its own class, i.e.,

$$Z < |\mathbf{W} \cdot \mathbf{X}|, \quad Z < |\mathbf{W}| |\mathbf{X}| \cos \theta \quad (3)$$

However, $|\mathbf{X}| \cos \theta$ is just the distance between \mathbf{X} and the surface along the normal. Call $|\mathbf{X}| \cos \theta$ the half thickness of the decision surface. From Equation 3, we get

$$t = \frac{Z}{|W|}$$

If a specific thickness, $2t$, is assigned to the decision surface, then the learning process can be continued until all the points of the training set are found on a particular side of an infinitesimally thin separation surface, and no point exists inside the separation hyperplane of given thickness. In order to determine t , it is necessary to preset Z and the initial weight vector. However, t is related to Z through W , and W undergoes variations due to feedback. Consequently, t also suffers variations. Hence, although t is specified prior to the beginning of the learning process, the user cannot influence its final value. If, however, the vector W is normalized to a particular value after each feedback, the magnitude of t will remain constant until the end of the learning process. Such an approach gives a weight vector with better predictive ability.

The superfluous components have to be eliminated from the weight vector in order to reduce its dimensionality. In other words, the only spectral features retained are those that guarantee fast and correct identification.

The corresponding algorithm, proposed by Preuss and Jurs,⁴² consists in the following. The distance of the point, X_i , from the decision surface along the normal can be defined by d_i . Depending upon whether the point has been correctly or incorrectly classified, d_i may be either positive or negative. The half thickness of the decision surface can then be determined as the minimum value of d_i in the training set, i.e.,

$$d_i = \pm \frac{|W \cdot X_i|}{|W|}$$

$$t = \min(d_i)$$

It is obvious that the half thickness, t , so determined is the maximum thickness typical of a given decision surface, provided all the elements of the training set have been correctly classified.

In order to verify the significance of a descriptor, it is temporarily discarded from the data set, and then all d_i and t are calculated. It is assumed that the elimination of a significant descriptor will lead to a considerable decrease in t , while elimination of a superfluous descriptor will give rise to an imperceptible change in t . All the descriptors are tested by this procedure, and the results compared. The algorithm presupposes that the maximum value of t will correspond to a descriptor that can be discarded more easily than others. Then that descriptor is discarded from the data set. The most important spectral features are retained by repeated application of this procedure.

One more version of the classification algorithm is known as the fractional correction method (FCM).⁴³ The learning weight vector, W , in this case is decomposed into two components:

$$W_t = W_p + W_n$$

where p applies to the positive class and n to the negative class. Consequently,

$$S = W \cdot X_i$$

and the classification is effected as usual by the magnitude and the sign of S , the only difference being that the weight vector is not changed by the feedback in the course of classifying the elements of the training set. Instead, the following procedure is adopted. Two sums

$$S_p = \sum_{i=1}^{n_p} S_i$$

$$S_n = \sum_{i=1}^{n_n} S_i$$

are formed, where the summation in S_p is carried out for those patterns which belong to the positive class but are classified in the wrong way (their number is equal to n_p), and those elements in S_n are summed up which are wrongly classified (their number is equal to n_n). The feedback equation in this case takes the form:

$$W'_p = W_p + C_i f_i^p X_i$$

$$W'_n = W_n + C_i f_i^n X_i$$

where

$$f_i^p = \frac{1.5 (S_i - Z)}{S_p}$$

$$f_i^n = \frac{1.5 (S_i - Z)}{S_n}$$

$$C_i = \frac{-2 (S_i - Z)}{X_i \cdot X_i}$$

and the factor 1.5 is a given constant. Feedback is effected for W_p and W_n , respectively, n_p and n_n times. The procedure is continued until all the elements of the training set are classified correctly.

Other techniques are also available, for example, MAX,⁴³ but we shall not deal with them here.

The predicting power of these above-mentioned methods has been compared.⁴³ Learning was conducted in a space of 131 descriptors on a training set consisting of 126 arbitrarily chosen spectra. The predicting power was determined by means of 86 spectra that were not included in the training set. In the process of classification, a program for identification of seven classes of chemical compounds had to be learned. The experimental results are listed in Table 6. It can be seen from this table that the first three methods give very similar values for the predicting power, whereas the fractional correction technique gives slightly better results for some classes (from 93 to 100%). Of the four methods, the MAX technique is the least effective.

Interesting results were obtained in studying the way in which the predictive ability is affected on decreasing the number of descriptors, i.e., only the most important spectral features are retained.

Teaching a machine to select the most important features in the IR spectra of carbonyl acids, esters, and primary amines was investigated by Preuss and Jurs.⁴² Table 7 lists the predictive abilities that are attainable by decreasing the number of descriptors characterizing the spectra of these compounds, while Figures 17 through 19 show the components of the weight vectors for the groups COOH, CCOOC, and NH₂ after the superfluous components have been eliminated. Downward-directed components cor-

TABLE 6

Predictive Abilities of Four Binary Pattern Classifiers (BPC) Using 131 Peaks

Functional group	Basic BPC	BPC with nor- malization	Fractional correction method	MAX
Alcohols	98.8	97.7	97.7	98.8
Benzenes	90.7	90.7	93.0	84.7
Carbonyls	100.0 ^a	100.0	100.0	100.0 ^{b,c}
Carbonyl acids	96.5	97.0	100.0	97.7
Esters	97.7	98.8	98.8	97.7
Ethers	98.8	97.7	98.8	100.0
Ketones	97.7	96.5 ^b	96.5	95.4

^a Prediction of least populous class low.^b Prediction invalid.^c Did not classify all training set correctly.

Modified from Liddell, R. W. and Jurs, P. C., *Anal. Chem.*, 46, 2126 (1974).
Copyright by the American Chemical Society.

TABLE 7

Training and Feature Selection with Carboxylic Acids, Esters, and Primary Amines

-COOH		$\begin{array}{c} \text{O} \\ \parallel \\ \text{C}-\text{C} \\ \diagup \quad \diagdown \\ \text{O} \quad \text{O}-\text{C} \end{array}$		C-NH ₂	
Number of descriptors	Prediction percentage	Number of descriptors	Prediction percentage	Number of descriptors	Prediction percentage
128	95.9	128	96.6	128	95.2
34	95.6	36	95.5	32	95.5
22	93.5	24	94.4	24	96.0
18	94.5	14	93.5	18	95.1
16	91.7	12	93.2	12	92.3
14	90.7	10	91.5	10	92.6
12	91.2	8	—	8	—
10	92.6	—	—	—	—
8	88.5	—	—	—	—
6	—	—	—	—	—

Modified from Preuss, D.R. and Jurs, P. C., *Anal. Chem.*, 46, 520 (1974). Copyright by
the American Chemical Society.

respond to positive correlation when the corresponding functional group is present, while the upward-directed components correspond to negative correlation. From the table it can be seen that a weight vector containing about 18 to 20 descriptors is sufficient to guarantee a predictive ability of 95%. Certain interesting conclusions can be drawn from Figures 17 through 19.

The weight vector of the carboxyl group contains 34 components. A negative correlation is observed in the range from 2.4 to 2.8 μm which the authors attribute to the presence of a free hydroxyl group having no relation with the group COOH in this

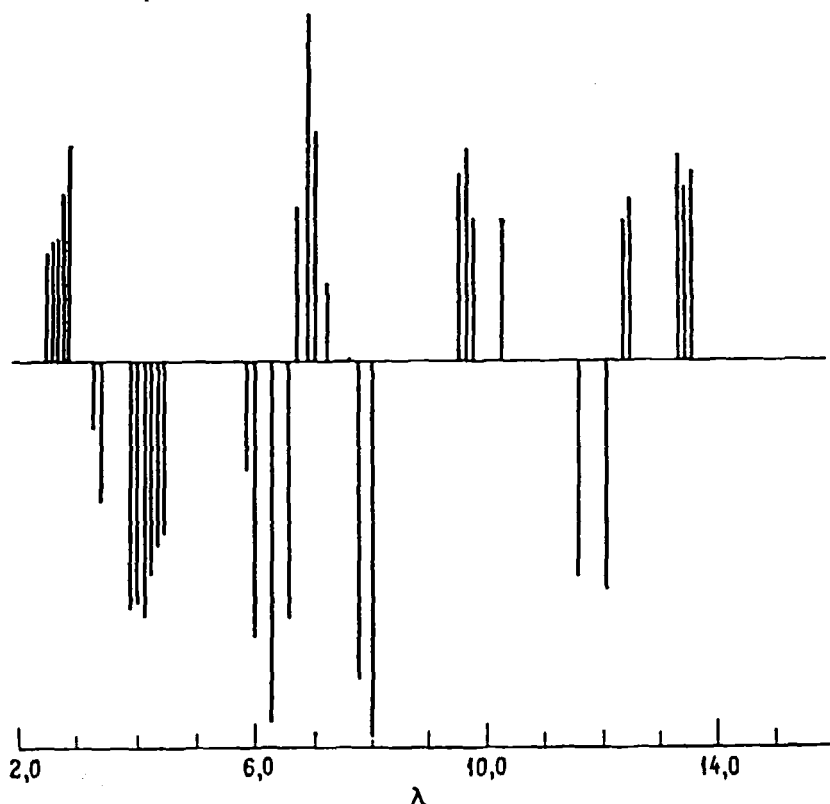


FIGURE 17. Weight vector map of carboxylic acids. (Reprinted with permission from Preuss, D. R. and Jurs, P. C., *Anal. Chem.*, 46, 520 (1974). Copyright by the American Chemical Society.)

frequency range; in contrast, the correlation is positive in the range of absorption of associated acids (3.2 to 4.3 and 5.8 to 5.9 μm). It is positive in the vicinity of 8 μm as well, a fact which is understandable. However, the negative correlations observed for other spectral ranges and the positive correlation in the vicinity of 12 μm cannot be given any physical interpretation and are probably due to the imperfections of the method itself or to imperfections due to a biased training set. An analogous conclusion can be inferred from weight vectors constructed for esters and primary amines or from the data reported⁴³ for carboxyl, carbonyl, and alcohol groups (Figures 20 and 21). The latter figure shows the positive (upward) and negative (downward) components of weight vectors. For example, in the opinion of the authors,⁴³ the negative components observed for alcohols are a distinctive feature of the weight vector of this group, and this makes it possible to identify them easily from other classes.

The latter results, indeed, are very interesting and promising. Despite the high level of data noise, the information obtained by these methods on the vibrations of certain groups in the IR range may prove to be very useful in determining the characteristic frequencies of the functional groups. Here it has to be assumed that the accidental and true correlations derived from the weight vector diagrams can readily be differentiated by comparing the vector components with the calculated frequencies and shapes of the vibrations of the corresponding structural fragments.

A knowledge of the frequencies calculated and the shapes can confidently be applied in judging whether a given component of the weight vector is associated in any way with the vibrations of the functional group under investigation.

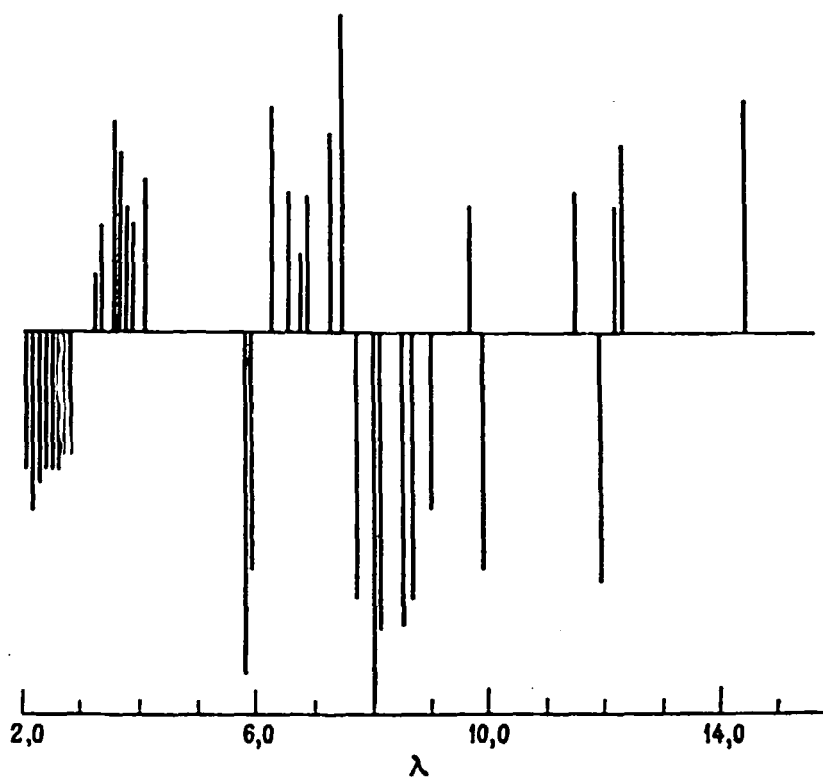


FIGURE 18. Weight vector map of esters. (Reprinted with permission from Preuss, D. R. and Jurs, P. C., *Anal. Chem.*, 46, 520 (1974). Copyright by the American Chemical Society.)

IV. ARTIFICIAL INTELLIGENCE APPROACH

A. General Features of Artificial Intelligence Approach

The third group of methods, which are employed for solving analytical problems, is based on the approach known as the linguistic method, or sometimes, as the artificial intelligence method. The so-called dictionary of attributes is used in this approach. This dictionary may contain, for instance, the following correspondences:

Structural element	Attributes in IR spectra
$-\text{CH}_3$	1460, 1380 cm^{-1}
$\text{R}_1 - \overset{\text{O}}{\parallel} \text{C} - \text{O} - \text{R}_2$	1740 cm^{-1}
$\text{R}_1 - \overset{\text{O}}{\parallel} \text{C} - \text{R}_2$	1715 cm^{-1}

With the help of this attribute dictionary, a relationship is established between the presence or absence of some structural element in a molecule and the corresponding absorption bands in the spectrum or some other measurable characteristics.

Of course, it is essential that this relationship between two attributes should be unique and should not admit any other interpretation. If such an attribute dictionary

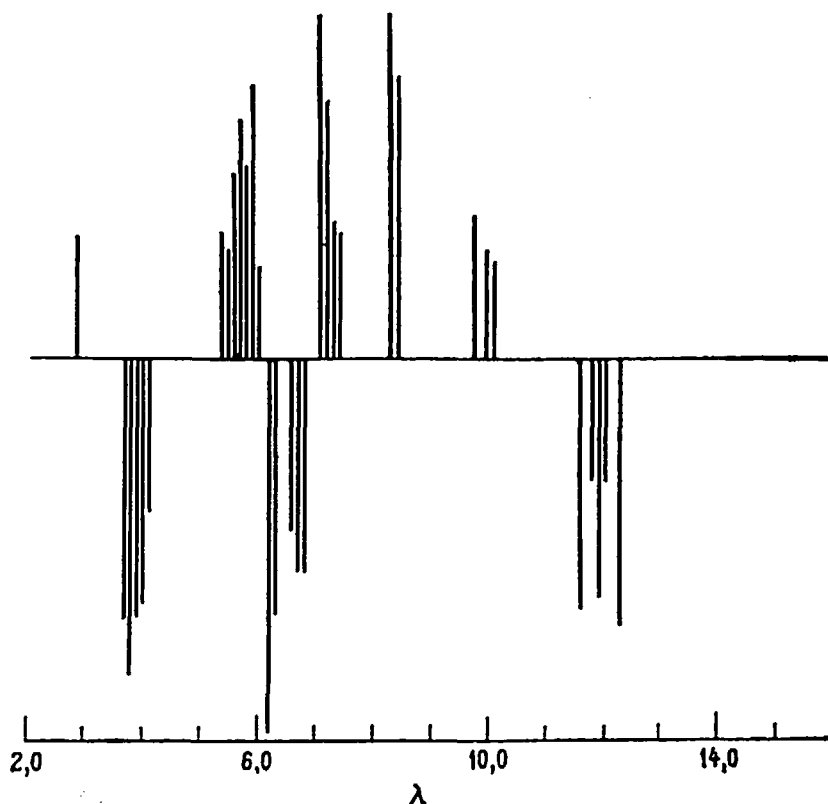


FIGURE 19. Weight vector map of primary amines.(Reprinted with permission from Preuss, D. R. and Jurs, P. C., *Anal. Chem.*, 46, 520 (1974). Copyright by the American Chemical Society.)

is available (a simple example being the tables of characteristic frequencies in IR, Raman, and UV spectra or the characteristic chemical shifts in NMR spectra), then there is a formal possibility of utilizing these attributes in the search and identification of individual compounds and the structural elements by means of certain mathematical techniques underlying the so-called proposition algebra or the Boolean algebra.

The attribute dictionary requires quite a small memory size for its storage, and what is most important is that it can easily be duplicated with the help of various types of computers for the dissemination of an already constructed system to different users. Of course, the quality of recognition depends essentially on the reliability and quality of the attribute dictionary. An automatic method based on a data bank, similar to the technique described by Nigmatullin and Smirnov,⁶ may conveniently be used in constructing these attribute dictionaries. Experience shows, however, that such dictionaries should not be constructed as all-purpose ones, but should be goal-oriented with due regard for the specific interests of the user.

Nonetheless, it should be mentioned that the application and development of such linguistic approaches give satisfactory recognition results only when the investigator has the empirical formula of the sample at his disposal. In the contrary case, the number of alternatives given by the computer will be extremely large. Atlas approaches do not necessarily need the use of empirical formulas, and as such are effective in those cases where the artificial intelligence method fails. Nevertheless, as already pointed out, it is not always possible to obtain an answer by the atlas method. The atlases are

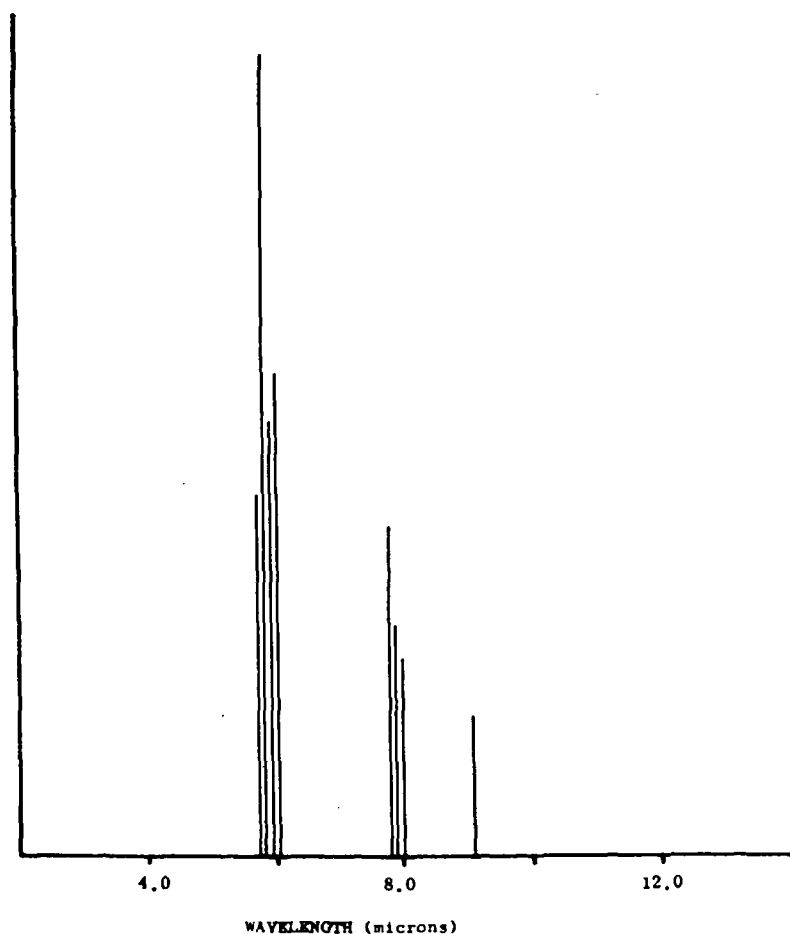


FIGURE 20. Weight vector map of carboxyls.(Reprinted with permission from Liddell, R. W. and Jurs, P. C., *Anal. Chem.*, 46, 2126 (1974). Copyright by the American Chemical Society.)

revised from time to time and tens of thousands of organic compounds are included; however, relatively few inorganic and organo-metallic compounds are among them.

At present, quite a large amount of empirical data have been collected regarding the characteristic features of the functional groups in the spectra of polyatomic molecules. This information from various journals has been generalized in several monographs,⁴⁹⁻⁵⁶ where it is presented in the form of correlation tables and diagrams. Correlation tables are the main source of information on the mutual relationship between spectra and the molecular structure needed in qualitative structural group analysis. Certain serious difficulties are encountered in applying these methods. We shall point out some of them.

Elucidation of the molecular structure by means of spectra is based on a logical comparison of the spectrum of the compound under investigation with the dependencies already known between the structural elements and their spectral features. These dependencies are, however, frequently extremely complicated, of a multistage nature, and include a large number of logical relations whose word formulation may often prove to be rather vague and ambiguous. To retain these relations in memory and to deduce the intermediate conclusions at all stages of the analysis is rather difficult and

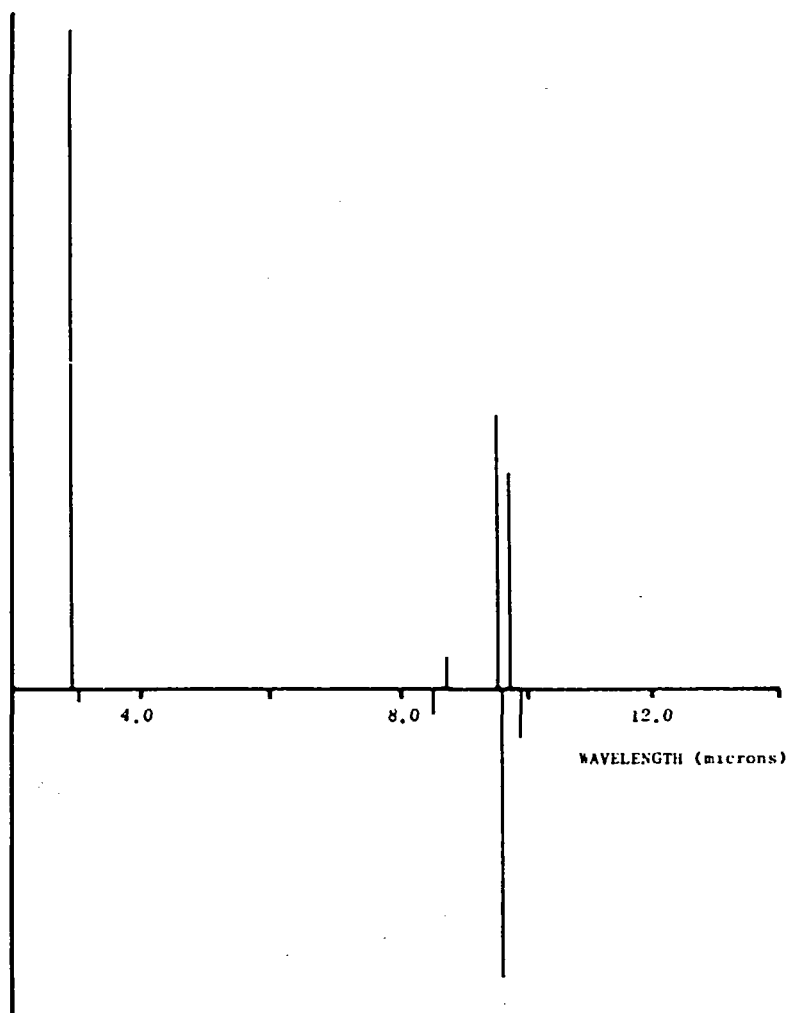


FIGURE 21. Weight vector map of alcohols. (Reprinted with permission from Liddell, R. W. and Jurs, P. C., *Anal. Chem.*, 46, 2126 (1974). Copyright by the American Chemical Society.)

is often beyond man's capacity. In consequence, there is a substantial dependence of the results of spectral data analysis on the experience and intuition of the specialist. This in its turn leads to the absence of an unbiased measure of the probability of a given set of structural elements occurring in the sample. This type of investigation can be facilitated using special mathematical techniques in which human argumentation is replaced by exacting computer-aided calculations.⁵⁷⁻⁶¹

This requirement is satisfied by the principles of symbolic logic, the science of discrete, structural simulation of finite discrete systems. Application of symbolic logic is highly fruitful in studying complicated objects of a discrete nature. Since a molecule is an ensemble of structural elements and its vibrational spectra are discrete objects, there is ground for believing that symbolic logic is the best mathematical tool for describing the mutual relationship between the spectrum and the structure of a molecule. Here this description may be looked upon as the "black box". Since the specific techniques of Boolean algebra are not familiar to chemical analysts, we shall begin this section with a brief description of the basic concepts and definitions of propositional

calculus⁶² followed by examples showing its usage. Basic principles of the graph theory will also be given.

B. Elements of Symbolic Logic

A proposition is a complete statement of which it may definitely be said that its content is either true or false. For example, the statement, "frequency ω is contained in the spectrum" is a proposition, but "which vibration of the functional group does this frequency belong to?" cannot be said to be a proposition. "True" and "false" are the two logical values of any proposition and are denoted by the letters T and F, respectively. None of the propositions can be both true and false. Complex propositions or Boolean functions are constructed from simple propositions by means of logical operations. These operations are often called the logical conjunctions or connections.

We shall consider the basic operations over propositions. Each simple proposition shall be assigned a letter, and the meaning of a propositional variable shall be attributed to it.

The simplest logical operation over propositions is negation. Negation of proposition a shall be denoted by \bar{a} , and it stands for a proposition whose truth value is opposite to the truth of the proposition a . If a is false, then \bar{a} is true; and \bar{a} is false if a is true. Negation of a implies the same as the propositions "not a ", "it is untrue that a ", and " a is false". For example, if the proposition "the frequency ω is contained in the spectrum" is denoted by ω , then $\bar{\omega}$ shall stand for the proposition "the frequency ω is not contained in the spectrum."

The action of each operation is represented in the form of truth table which defines this operation.

The truth table of a negation operation takes the following form:

a	\bar{a}
T	F
F	T

Conjunction (logical multiplication) of propositions a and b is denoted by $a \wedge b$.^{*} It means that proposition $a \wedge b$ is true if and only if each proposition, a and b , is true. Conjunction $a \wedge b$ also implies that "both the propositions a and b are true." Accordingly, it corresponds to the logical conjunction "and". For example, the Boolean function, $\omega_1 \wedge \omega_2$ may serve as the symbolic representation for the proposition "the frequency ω_1 is contained in the spectrum and the frequency ω_2 is contained in the spectrum."

Conjunction has the following truth table:

a	b	$a \wedge b$
T	T	T
T	F	F
F	T	F
F	F	F

Disjunction (logical addition) of the propositions a and b is denoted by $a \vee b$.^{**} It

^{*} Conjunction is usually denoted by the sign " \cdot " or " \wedge ".

^{**} Disjunction is also denoted by the sign " $+$ ".

stands for a proposition which is true if and only if at least one of propositions a or b is true. This operation also implies that "at least one of the propositions a or b is true," and in meaning corresponds to the inclusive connective "or". For example, the Boolean function, $\omega_1 \vee \omega_2$, is the symbolic representation for the proposition "the spectrum contains either the frequency ω_1 or the frequency ω_2 or both the frequencies."

The truth table of disjunction is of the form:

a	b	$a \vee b$
T	T	T
T	F	T
F	T	T
F	F	F

Exclusive disjunction of propositions a and b is denoted by $a \vee \cdot b$. In meaning it corresponds to the exclusive connective "or". For example, the statement $\omega_1 \cdot \vee \omega_2$ implies "the spectrum contains either the frequency ω_1 or the frequency ω_2 , but does not contain both frequencies together."

This operation has the following truth table:

a	b	$a \vee \cdot b$
T	T	F
T	F	T
F	T	T
F	F	F

Implication $a \rightarrow b$ denotes a statement that is false if and only if the statement a is true and the statement b is false. Implication $a \rightarrow b$ has the meaning of the statements "if a , then b ," "from a it follows that b ," " a implies b ," " b provided that a ". It corresponds to the logical connective "if ..., then...". The proposition $a \rightarrow b$ is also read as " a implies b ". The first term of the implication (a) is called the antecedent, and the second, the consequent. An example of implication is the statement "if a molecule contains the functional group A, then its spectrum contains the frequency ω ."

The truth table of implication is as follows:

a	b	$a \rightarrow b$
T	T	T
T	F	F
F	T	T
F	F	T

Equivalence of the propositions a and b is denoted by $a = b$; it stands for a statement which is true if and only if the statements a and b have the same truth values. Equivalence $a = b$ also implies the statements " a if and only if b ", "in order that a , it is necessary and sufficient that b ." For example, $A = \omega$ corresponds to the statement "the frequency ω is contained in the spectrum if and only if the molecule contains the group A".

This operation has the following truth table:

a	b	$a = b$
T	T	T
T	F	F
F	T	F
F	F	T

Tautology has a special role in the propositional calculus. Tautology is a complex proposition which is always true, irrespective of the truth or falsehood of its constituent propositions; for example, "the frequency ω is contained in the spectrum, or the spectrum does not contain the frequency ω ." Identically true formulas are also called tautology, and are denoted by the symbol I . The negation of I , i.e., I is called an always false statement and is denoted by the symbol 0 ("the spectrum contains the frequency ω , and the spectrum does not contain the frequency ω ").

On the other hand, certain complex statements, which are not tautologies, may be true in particular conditions of a problem. They are called the genuinely true statements and are denoted by the same symbol as for tautology.

If the truth tables of two different formulas coincide, these two formulas represent the same Boolean function. Such formulas are called equivalent ones. They are usually denoted by the sign $=$ or \equiv . Using the truth table, it is easy to verify that all other logical operations can be expressed in terms of the three fundamental operations of negation, conjunction, and disjunction:

$$(a \vee b) = (a \wedge \bar{b}) \vee (\bar{a} \wedge b)$$

$$(a \rightarrow b) = (\bar{a} \vee b)$$

$$(a = b) = (a \rightarrow b) \wedge (b \rightarrow a) = (a \wedge b) \vee (\bar{a} \wedge \bar{b})$$

Thus, for any formula, an equivalent formula can be found which contains only the signs of negation, conjunction, and disjunction.

Due to lack of space, without going into the details of Boolean algebra, we shall restrict ourselves to a mere listing of the most important equivalences with which a Boolean function can be reduced to a simple form:

$$\begin{aligned} a \vee a &= a \\ a \wedge a &= a \\ a \vee I &= I \\ a \wedge I &= a \\ 0 \vee a &= a \\ 0 \wedge a &= 0 \\ a \wedge \bar{a} &= 0 \\ a \vee \bar{a} &= I \\ \overline{(a \wedge b)} &= \bar{a} \vee \bar{b} \\ \overline{(a \vee b)} &= \bar{a} \wedge \bar{b} \\ a \wedge (b \vee c) &= (a \wedge b) \vee (a \wedge c) \\ a \wedge (a \vee b) &= a \\ (a \vee (a \wedge b)) &= a \\ [a \vee (b \wedge c)] &= (a \vee b) \wedge (a \vee c) \\ [a \vee (\bar{a} \wedge b)] &= a \vee b \end{aligned}$$

It should, however, be pointed out that in the practical application of the propositional calculus we come across considerable difficulties in converting the natural language into the language of the symbolic logic. This transformation, the first step in exact formulation of any logical problem, is not always obvious; this is largely due to the ambiguous usage of language. Therefore, in constructing the Boolean functions from statements expressed in natural language, a perspicuous understanding of the sense underlying these statements in a given context is imperative.

Now we shall take up the description of some computational methods in Boolean algebra which help us considerably in solving logical problems. The basic idea underlying these methods is the following. Some definite binary number is put into correspondence to each proposition contained in the Boolean function. All operations, which are to be carried out over the propositions, are performed over the corresponding binary numbers which are called the designation numbers. The final numeric results are then retransformed into a proposition form.

The starting point in constructing designation numbers is the truth table. Consider, for example, the truth table of a function of three variables: $a \vee (\bar{b} \wedge c)$

<i>a</i>	F	T	F	T	F	T	F	T
<i>b</i>	F	F	T	T	F	F	T	T
<i>c</i>	F	F	F	F	T	T	T	T
$a \vee (\bar{b} \wedge c)$	F	T	F	T	T	T	F	T

In this table substitute 1 for T and 0 for F: thus, we get

<i>a</i>	0	1	0	1	0	1	0	1
<i>b</i>	0	0	1	1	0	0	1	1
<i>c</i>	0	0	0	0	1	1	1	1
$a \vee (\bar{b} \wedge c)$	0	1	0	1	1	1	0	1

The rows in this numerical table are precisely the designation numbers of the propositions in the corresponding rows of the truth table. The set of designation numbers of a given number of logical variables form the logical basis. For example, the basis for three logical variables, *a*, *b*, and *c* takes the form:

	0	1	2	3	4	5	6	7
# <i>a</i> =	0	1	0	1	0	1	0	1
# <i>b</i> =	0	0	1	1	0	0	1	1
# <i>c</i> =	0	0	0	0	1	1	1	1

Here the symbol #*a* stands for the designation number of the proposition *a*, #*b*, the designation number of the proposition *b*, etc., while the number at the top of each column is the serial number of the column in the basis set. For a system consisting of *n* simple propositions, there are 2^n binary digits in each designation number. Consequently, the basis contains 2^n columns. With due regard for the correspondence between the basis and the truth table, the columns in the basis may be rearranged, thereby obtaining other bases which still remain valid. The columns in the basis consist of a total of 2^n possible combinations of digits, each of which is either unity or zero. This is in exact correspondence with all the possible truth values of *n* simple propositions. By virtue of permutations of the columns, we find that there can be in all $2^n!$ different bases. Among all the possible bases, only one is chosen which is called the standard logical basis. In the case of three variables, the table given above is the standard basis. The merit of this type of basis is that for any number of simple propositions it can be written as follows: zeros and unities alternate in the designation number of the first

simple proposition, pairs of zeros and unities alternate in the designation number of the second proposition, four zeros and unities alternate in the designation number of the third proposition; eight zeros and eight unities alternate in the designation number of the fourth simple proposition, etc. In the k th row of the standard basis ($l = 0, 1, \dots, n - 1$), the alternating groups of digits contain 2^l zeros or unities each. The columns in the basis are numbered from the left to the right beginning from zero to $2^n - 1$.

A specific feature of the standard basis is that each column read from the bottom upward forms a binary number equal to the serial number of the column in the decimal system. Unlike the truth table, the order of occurrence of the columns in the standard basis is strictly fixed. Usually the standard basis is expressed as follows: $B[a, b, c, \dots]$, the order of the elements in the square brackets coinciding with the order of the rows in the standard basis.

We shall now show how the designation numbers of complex propositions or Boolean functions are calculated.

In order to find the designation number $\#(a \vee b)$, we have to carry out the logical addition of the corresponding places of $\#a$ and $\#b$ without carrying over to the higher places, using the following rule: $0 + 0 = 0$; $1 + 1 = 1$; $0 + 1 = 1$; $1 + 0 = 1$. This addition is denoted by $\#(a \vee b) = \#a + \#b$, where the plus sign on the right stands for addition of designation numbers. For example,

$$\begin{array}{rcccc} \#a & = & 0 & 1 & 0 & 1 \\ \#b & = & 0 & 0 & 1 & 1 \\ \hline \#(a \vee b) & = & 0 & 1 & 1 & 1 \end{array}$$

The designation number $\#(a \wedge b)$, by analogy, is determined with the help of logical multiplication, using the rule: $0 \cdot 0 = 0$; $0 \cdot 1 = 0$; $1 \cdot 0 = 0$; $1 \cdot 1 = 1$.

Then, we have $\#(a \wedge b) = \#a \cdot \#b$

$$\begin{array}{rcccc} \#a & = & 0 & 1 & 0 & 1 \\ \#b & = & 0 & 0 & 1 & 1 \\ \hline \#(a \wedge b) & = & 0 & 0 & 0 & 1 \end{array}$$

In order to find $\#\bar{a}$, all the zeros in the digits of the number $\#a$ have to be replaced by unities and vice versa. For example, $\#a = 0101$, $\#\bar{a} = 1010$.

With the help of these fundamental operations over designation numbers, the designation number of any given Boolean function can be determined by means of these operations carried out in a consecutive sequence. For example, let us calculate $\#(a \vee (\bar{b} \wedge c))$ in the basis $B[a, b, c]$:

$$\begin{array}{rcccccccc} \#\bar{b} & = & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ \#c & = & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ \hline \#(\bar{b} \wedge c) & = & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ \#a & = & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ \hline \#(a \vee (\bar{b} \wedge c)) & = & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \end{array}$$

Note that $\#I = 1111\dots$, i.e., I has unities in all digits, while $\#0$ contains zeros in all digits. For the two functions, X and Y , $\#X = \#Y$, if and only if $X \equiv Y$, and $X \rightarrow Y$ if and only if $\#Y$ has unities at least in those digits which carry unities in $\#X$.

Several systematic methods are available for finding the various forms of symbolic representation of a given designation number.

For example, the number 0111 0100 can be interpreted in the following equivalent forms:

1. A complete disjunctive normal form (CDNF):
 $(a \wedge \bar{b} \wedge \bar{c}) \vee (\bar{a} \wedge b \wedge \bar{c}) \vee (a \wedge b \wedge 4\bar{c}) \vee (a \wedge \bar{b} \wedge c)$
2. Complete conjunctive normal form (CCNF):
 $(a \vee b \vee c) \wedge (a \vee b \vee \bar{c}) \wedge (a \vee \bar{b} \vee \bar{c}) \wedge (\bar{a} \vee \bar{b} \vee \bar{c})$
3. Simplest form as the sum of products:
 $(a \wedge \bar{b}) \vee (b \wedge \bar{c})$
4. Simplest form as the product of sums:
 $(4a \vee b) \wedge (\bar{b} \vee \bar{c})$
5. Mixed form:
 $((a \vee b) \wedge \bar{c}) \vee (a \wedge \bar{b} \wedge c)$

The CDNF is the logical sum of a certain number of elementary products. An elementary product is the conjunction of all the simple propositions from a given basis; any simple proposition may be contained in this conjunction either in the affirmative or in the negative form. For the basis $B[a, b, c]$, we shall write all the elementary products and their corresponding designation numbers:

$\#(\bar{a} \wedge \bar{b} \wedge \bar{c})$	=	1	0	0	0	0	0	0	0
$\#(a \wedge \bar{b} \wedge \bar{c})$	=	0	1	0	0	0	0	0	0
$\#(\bar{a} \wedge b \wedge \bar{c})$	=	0	0	1	0	0	0	0	0
$\#(a \wedge b \wedge \bar{c})$	=	0	0	0	1	0	0	0	0
$\#(\bar{a} \wedge \bar{b} \wedge c)$	=	0	0	0	0	1	0	0	0
$\#(a \wedge \bar{b} \wedge c)$	=	0	0	0	0	0	1	0	0
$\#(\bar{a} \wedge b \wedge c)$	=	0	0	0	0	0	0	1	0
$\#(a \wedge b \wedge c)$	=	0	0	0	0	0	0	0	1

It is seen that the designation number of an elementary product contains only one unity, while the total number of elementary products for n simple propositions is equal to 2^n .

The CDNF is the sum of only those elementary products, each of which represents unity occurring in the designation number of a given function. For example, the number 0100 0011 has unities in the first, sixth, and seventh digits, and consequently, it can be obtained by logically summing up those elementary products whose designation numbers contain unity in the first, sixth, and seventh digits. Therefore, CDNF of this number will be of the form:

$$(a \wedge \bar{b} \wedge \bar{c}) \vee (\bar{a} \wedge b \wedge c) \vee (a \wedge b \wedge c)$$

The CCNF is determined on the basis that the designation number of the logical sum of all the elementary propositions (with their "signs") in a given basis has just one zero. For the case of three variables, we have

$\#(\bar{a} \vee \bar{b} \vee \bar{c})$	=	1	1	1	1	1	1	1	0
$\#(a \vee \bar{b} \vee \bar{c})$	=	1	1	1	1	1	1	0	1
$\#(\bar{a} \vee b \vee \bar{c})$	=	1	1	1	1	1	0	1	1
$\#(a \vee b \vee \bar{c})$	=	1	1	1	1	0	1	1	1
$\#(\bar{a} \vee \bar{b} \vee c)$	=	1	1	1	0	1	1	1	1
$\#(a \vee \bar{b} \vee c)$	=	1	1	0	1	1	1	1	1
$\#(\bar{a} \vee b \vee c)$	=	1	0	1	1	1	1	1	1
$\#(a \vee b \vee c)$	=	0	1	1	1	1	1	1	1

These sums are called the elementary sums, and their number for n propositions is

equal to 2^n . Evidently, the CCNF corresponding to a given designation number is the product of those elementary sums to which the zeros in the designation number correspond. The CCNF and CDFN are the most important Boolean forms, as they are easily constructed. A more complicated problem is that of finding the simplest forms containing as few letters as possible. This problem is dealt with in many papers, but its solution has so far not been found.

Quite a number of logical problems encountered in practice can be solved with the help of Boolean equations or a system of Boolean equations. These equations are the symbolic representation of the information contained in the problem. The solution of these equations lies in such a transformation of this information which gives a means of answering certain questions relating to the terms appearing in the initial premises. By way of example, consider the equation $X \wedge (a \vee b) = (a \wedge b \wedge c)$, where X is the unknown Boolean function of the variables a , b , and c . On substituting the solution $X = X(a, b, c)$ back in the initial equation, we find that the initial equation is transformed into an identity (equivalence). It is easy to verify that the solutions of this equation are the following Boolean functions: $a \wedge b \wedge c$, $(a \wedge b \wedge c) \vee (\bar{a} \wedge \bar{b} \wedge \bar{c})$, $c \wedge ((a \wedge b) \vee (\bar{a} \wedge \bar{b}))$, $(\bar{a} \wedge \bar{b}) \vee (a \wedge b \wedge c)$.

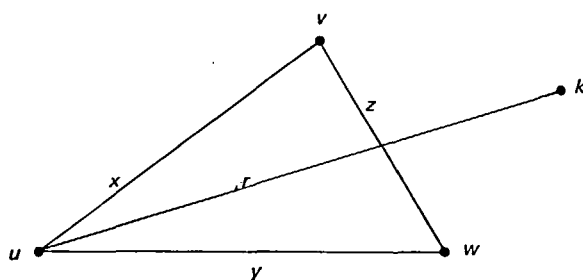
The solutions of Boolean equations or their systems are found with the help of matrix methods of Boolean algebra.⁶²

C. Basic Principles of the Graph Theory

The graph theory^{63,64} is another mathematical tool often used in artificial intelligence systems. The basic concepts of this theory are the vertex and the edge of a graph. Vertices are represented on a plane by points and the edges by a line joining points.

A graph (G) consists of finite nonvoid set (V) containing p vertices, and a given set (X) whose elements are q pairs of different vertices from V . Each pair of vertices $x = \{u, v\}$ is called the edge of the graph (G) , and x is said to join u and v . Symbolically this is expressed as follows: $x = u, v$ where the vertices u and v are called adjacent vertices. The vertex u and the edge x are incident. If two edges, x and y , are incident at the same vertex, they are called adjacent edges.

These definitions are illustrated, for example, by the graph:

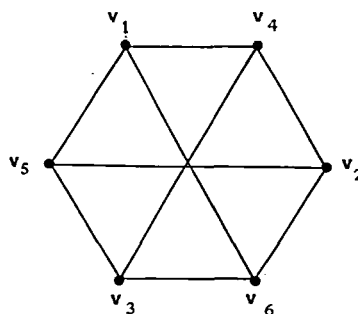
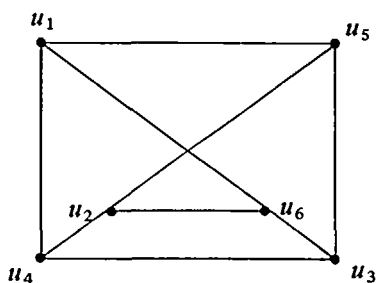


The edges z and r intersect, but their point of intersection is not used as a vertex.

A multigraph is a graph in which a pair of vertices may be joined by more than one edge. These edges are then said to be multiple edges.

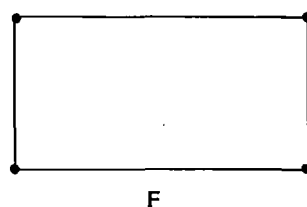
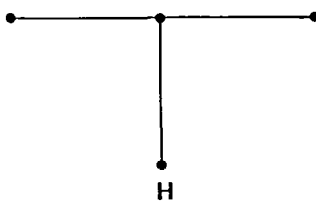
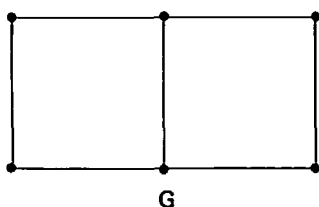
A graph is said to be labeled or numbered if its vertices are distinguishable by some label, for instance, u_1, u_2, \dots, u_k .

Two graphs, G_1 and G_2 , are said to be isomorphous if a mutual one-to-one correspondence exists between the sets of their vertices, and this correspondence retains their contiguity. For example, the graphs



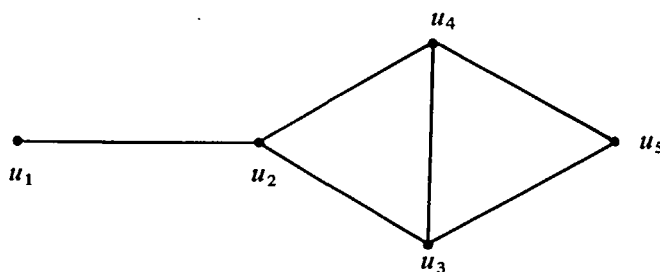
are isomorphic.

A subgraph of graph G is a graph in which all the vertices and edges belong to G . For example, H and F are subgraphs of G :



A sequence in a graph is an alternating series of vertices and edges $u_1, x_1, u_2, x_2, \dots, u_n$ which begins and ends in a vertex, and each edge being incident to the vertex that is its immediate predecessor and successor. A sequence is said to be closed if $u_1 = u_n$, and open if $u_1 \neq u_n$. A sequence is called a path if all of its edges are different. It is called a simple path, if in addition all of its vertices are different. A closed path is also called a cycle. A cycle is simple if all of its vertices are different.

In the graph

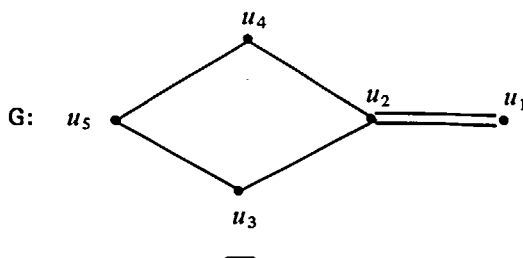


u_1, u_2, u_4, u_2 is a sequence but not a path, u_1, u_2, u_4, u_3 is a simple path, and u_2, u_4, u_5, u_3, u_2 is a simple cycle.

A graph is said to be connected if any one pair of its vertices is joined by a simple path. A graph is acyclic if there are no cycles in it. A connected acyclic graph is called a tree.

The incidence matrix $A = \|a_{ij}\|$ of labeled graph G containing p vertices is called a square matrix of p th order in which the elements $a_{ij} = r$ if the vertices u_i and u_j are joined together by edges of multiplicity equal to r , and $a_{ij} = 0$ in the contrary case. For example,

$$A = \begin{pmatrix} 0 & 2 & 0 & 0 & 0 \\ 2 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$



A mutually one-to-one correspondence exists between a labeled graph and its incidence matrix.

If the vertices of graph G are renumbered in some way, then another matrix will correspond to it which can be derived from A by rearrangement of rows and columns. Obviously, for a graph with p vertices, $p!$ such enumerations are possible, i.e., the number of possible incidence matrices is $p!$

In principle, the methods of the graph theory and symbolic logic can be applied to information of diverse nature, for example, spectral chemical and other data.

D. Problem of Structural-Group Analysis

The problem of structural-group spectral analysis may be formulated in the language of symbolic logic as follows:⁵⁷⁻⁶¹ suppose we have, for example, a vibrational (IR or Raman) spectrum of some sample, and the relationships between the frequencies in the spectrum and the structural elements of the molecule (for instance, correlation tables) are known. We shall assume that among the frequencies occurring in the spectrum, there are certain characteristic frequencies.* It is required to find all the possible combinations of the functional groups that may be present in the given sample consistent with the spectrum.

Consider the following sets: $\tilde{\omega} = \{\omega_j\}$, $j = 1, 2, \dots, m$ —the set of spectral features (frequencies) observed in the spectrum of the sample; $A = \{A_i\}$, $i = 1, 2, \dots, N$ is the set of structural elements (functional groups) having at least one characteristic frequency occurring in the set $\tilde{\omega}$; set A may also contain the functional groups for which their presence or absence in the sample under investigation (analysis for a given set of groups) has to be established; $\bar{\omega} = \{\omega_j\}$, $j = m+1, \dots, M$ is the set of frequencies which correspond to some $A_i \in A$, but are not present in the spectrum; $\omega = \tilde{\omega} \cup \bar{\omega} = \{\omega_j\}$, $j = 1, 2, \dots, M$ is the set of all frequencies which are to be considered in solving a particular problem; $A \cup \omega = \{a_v\}$, $v = 1, 2, \dots, M+N$ is the universal set which limits the functional groups and frequencies applicable to the problem under consideration.

We shall stipulate that the elements of these sets as logical elements shall be assigned the meaning of elementary propositions:

1. A_i : molecule contains group A_i
2. \bar{A}_i : molecule does not contain group A_i
3. ω_j : spectrum contains frequency ω_j
4. $\bar{\omega}_j$: spectrum does not contain frequency ω_j

Now we have to write all the logical connections between the elements of the sets A and ω , and also express the additional relationships arising in considering information of a nonspectral nature in the form of logical functions. In a general case where a set

* Obviously, in the contrary case, the structural-group analysis is devoid of sense.

of n characteristic frequencies is attributed to the group, their relation with group A_i is written as follows:*

$$A_i \rightarrow \omega_1^{(i)} \wedge \omega_2^{(i)} \wedge \dots \wedge \omega_n^{(i)}, \quad \omega_\alpha^{(i)} = \omega_j \in \omega \quad (4)$$

$$(\bar{\omega}_1^{(i)} \vee \bar{\omega}_2^{(i)} \vee \dots \vee \bar{\omega}_n^{(i)}) \rightarrow \bar{A}_i \quad (5)$$

Expression 5 means that if a molecule contains group A_i , then the set of frequencies $\omega_1^{(i)}, \omega_2^{(i)}, \dots, \omega_n^{(i)}$ shall occur in the spectrum. Indeed, the implication is that Expression 4 is the symbolic representation of the experimentally established fact that the vibrations are characteristic in vibrations. According to Expression 4, the presence of the conjunction of frequencies $\omega_1^{(i)} \wedge \omega_2^{(i)} \dots \wedge \omega_n^{(i)}$ is a necessary condition for the molecule to contain structural group A_i . Expression 5, equivalent to Expression 4, shows that the sufficient condition for the absence of the structural element in the sample under investigation is that the spectrum should not contain at least one of the characteristic frequencies of this group.

If the frequency ω_j is contained in the sets of the characteristic frequencies of the groups $A_1^{(i)}, A_2^{(i)}, \dots, A_\gamma^{(i)}, A_\beta^{(i)}$ element of A , then

$$\omega_j \rightarrow (A_1^{(i)} \vee A_2^{(i)} \vee \dots \vee A_\beta^{(i)} \vee \dots \vee A_\gamma^{(i)}), \quad A_\beta^{(i)} = A_i \in A \quad (6)$$

This expression means that if the frequency ω_j is detected in the spectrum, the molecule contains at least one of the functional groups, $A_1^{(i)}, A_2^{(i)}, \dots, A_\gamma^{(i)}$. Expressions 4 and 6 represent the fundamental types of logical relationships between functional groups and spectral features in a structural-group analysis.

Note that, by nature, an implication shows that one proposition follows from another, admitting that the same conclusion can also be drawn from other premises. Obviously, in principle, the conjunction of the frequencies, $\omega_1^{(i)}, \omega_2^{(i)}, \dots, \omega_n^{(i)}$, may be the consequence of the presence of not only the group A_i , but also due to the presence of some other groups or their combinations in the sample. On the other hand, it may happen that the disjunction, $A_1^{(i)} \vee A_2^{(i)} \vee \dots \vee A_\gamma^{(i)}$ is implied not only by the frequency, ω_j , but also by some other spectral features. Hence, it follows that fundamental formulas, 4 and 6, indeed represent quite exactly the nature of the relationship between the functional groups and the corresponding characteristic frequencies in the vibrational spectra.

Let $T(A, \omega)$ denote the Boolean function which describes all the mutual relationships between the frequencies and the structural elements in a given problem. The vibrational spectrum as a combination of frequencies may be represented by the Boolean function, $R(\omega)$.

The logical analysis of this situation lies in the joint processing of the functions, $T(A, \omega)$ and $R(\omega)$, so that as a result, we establish the type of the function, $f(A)$, that describes the whole set of structural elements which may be present in the specimen under test. Symbolically, this is expressed as

$$T(A, \omega) \rightarrow \{R(\omega) \rightarrow f(A)\} \quad (7)$$

Relationship 7, which is the most general formulation of the qualitative spectral problem in the language of Boolean algebra, is the logical equation for the function, $f(A)$.

* Here $\omega^{(i)}$ are the frequencies which occur in the characteristic intervals $(\Delta\omega)^{(i)}$ corresponding to the given group A_i .

Solution of the logical problem of structural-group analysis lies in calculating the function, $f(A)$.

It should be emphasized that Equation 7 has a high degree of generality, and as such may be used in describing the methodology of qualitative analysis based on other physical principles.

We shall illustrate the basic idea underlying the algorithm of the solution of Equation 7 with reference to the following simple example ($M = N = 3$).

Let there be a product of some chemical reaction in which, by the conditions of synthesis, the C=C bond is supposed to be present. For the sake of definiteness, we shall assume that the IR spectrum of this compound is available and that it contains bands at 1460 and 1380 cm^{-1} (if the Raman spectrum is used, all the arguments are applied in an analogous manner). A structural-group analysis of this sample has to be carried out.

It is known that absorption near 1650 cm^{-1} is the spectral indication of the presence of the C=C double bond. Absence of this bond in this spectral range, however, does not mean that the C=C double bond is absent in the sample. In fact, if this group exists at the center of the symmetry of the molecule, its stretching vibrations are inactive in the IR spectrum.^{50,51} Hence, introducing the notation

$$\omega_1 \equiv (1460 \text{ cm}^{-1})$$

$$\omega_2 \equiv (1380 \text{ cm}^{-1})$$

$$\omega_3 \equiv (1650 \text{ cm}^{-1})$$

$$A_1 \equiv (\text{CH}_2)$$

$$A_2 \equiv (\text{CH}_3)$$

$$A_3 \equiv (\text{C}=\text{C})$$

we obtain

$$\begin{aligned} A &= \{A_1, A_2, A_3\} \\ \omega &= \{\omega_1, \omega_2\} \\ \bar{\omega} &= \{\omega_3\} \\ \tilde{\omega} &= \{\omega_1, \omega_2, \omega_3\} \end{aligned}$$

The logical relationship between the elements of sets A and ω , for the case under consideration, may be represented in the form of the following Boolean function:*

$$\varphi_1 = (\omega_3 \rightarrow A_3)$$

$$\varphi_2 = (\omega_1 \rightarrow A_1 \vee A_2)$$

$$\varphi_3 = (\omega_2 \rightarrow A_2)$$

$$\varphi_4 = (A_1 \rightarrow \omega_1)$$

$$\varphi_5 = (A_2 \rightarrow \omega_1 \wedge \omega_2)$$

and the functions $T(A, \omega)$ and $R(\omega)$ are written as:

$$T(A, \omega) = \varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \varphi_4 \wedge \varphi_5$$

* For group A_3 , we could also have written $\varphi_6 = (A_3 \rightarrow \omega_3 \vee \bar{\omega}_3)$, but since $(\omega_3 \vee \bar{\omega}_3) = 1$, there is no sense in including φ_6 in $T(A, \omega)$.

$$R(\omega) = \omega_1 \wedge \omega_2 \wedge \bar{\omega}_3$$

From the elements of sets A and ω construct a standard logical basis, $B[A_1, A_2, A_3, \omega_1, \omega_2, \omega_3]$, whose columns are the sets of all the possible conjunctions of $A_i \in A$ and $\omega_j \in \omega$ ($i = 1, 2, 3; j = 1, 2, 3$) (see Table 8).

Evidently, not all possible combinations of the elements of the logical basis can always be realized in practice. For instance, the conjunction of $A_1 \wedge A_2 \wedge A_3 \wedge \bar{\omega}_1 \wedge \bar{\omega}_2 \wedge \bar{\omega}_3$ is devoid of physical sense because it describes a situation where the spectrum of a compound containing CH_2 and CH_3 groups does not exhibit the bands near 1380 and 1460 cm^{-1} . The superfluous columns are eliminated from the logical basis in accordance with the function $T(A, \omega)$. For this purpose, it is necessary to calculate this function and delete from Table 8 those columns which correspond to the zeros in the designation number of this function. Since $\#T(A, \omega) = \# \varphi_1 \cdot \# \varphi_2 \cdot \dots \cdot \# \varphi_s$, we shall determine the designation numbers of each of the functions $\varphi_1, \varphi_2, \dots, \varphi_s$, and then multiply them together. Prior to calculating the designation numbers, we have to interpret the meaning of the implication, with due regard for the fact that $(a \rightarrow b) \equiv (a \vee b)$. Carrying on these operations, we obtain:

$$\#T(A, \omega) = 1000 \ 1000 \ 0100 \ 0100 \ 0000 \ 0000 \ 0011$$

$$0011 \ 0000 \ 1000 \ 0000 \ 0100 \ 0000 \ 0000 \ 0000 \ 0011$$

After deleting from the basis, $B[A_1, A_2, A_3, \omega_1, \omega_2, \omega_3]$ the columns corresponding to the digits of $\#T(A, \omega)$, which carry zeros, we obtain the constrained logical basis, $B^0[A_1, A_2, A_3, \omega_1, \omega_2, \omega_3]$ which is written as follows:

A_1	0011	0101	0101
A_2	0000	1111	0011
A_3	0101	0011	1111
ω_1	0011	1111	0111
ω_2	0000	1111	0011
ω_3	0000	0000	1111

It is seen that in this case the constrained basis consists of 12 columns containing only the combinations of functional groups and frequencies that can, in principle, be realized. Regarding the rows of the constrained basis as the new designation numbers of its elements, we shall calculate $\#R(\omega)$ in the basis, $B^0[A_1, A_2, A_3, \omega_1, \omega_2, \omega_3]$,

$$\#R(\omega) = 0000 \ 1111 \ 0000$$

Consider those columns of the constrained basis which correspond to unities in $\#R(\omega)$. The sets of $A_i \in A$ contained in these columns form a total ensemble of the conjunctions forming the function $f(A)$ in the basis, $B[A_1, A_2, A_3]$

$$f(A) = (\bar{A}_1 \wedge A_2 \wedge \bar{A}_3) \vee (A_1 \wedge A_2 \wedge \bar{A}_3) \vee (\bar{A}_1 \wedge A_2 \wedge A_3) \vee (A_1 \wedge A_2 \wedge A_3) \quad (8)$$

In general, the function $f(A)$ will contain k ($1 \leq k \leq 2^n$) terms, each of which describes one of the equiprobable sets of structural elements that may occur in the sample under test. The probability of the occurrence of each set is equal to $1/k$. It does not depend on the subjective opinion of the investigator, but is wholly determined by the set of initial data.

From Equation 8 it follows that the probability of the presence of group A_2 is 1, while $p(A_1) = p(A_3) = 1/2$. The probability of the presence of each group is calculated

TABLE 8

Standard Logical Basis

A_1	0101 0101 0101 0101 0101 0101 0101 0101 0101 0101 0101 0101 0101 0101 0101 0101
A_2	0011 0011 0011 0011 0011 0011 0011 0011 0011 0011 0011 0011 0011 0011 0011 0011
A_3	0000 1111 0000 1111 0000 1111 0000 1111 0000 1111 0000 1111 0000 1111 0000 1111
ω_1	0000 0000 1111 1111 0000 0000 1111 1111 0000 0000 1111 1111 0000 0000 1111 1111
ω_2	0000 0000 0000 0000 1111 1111 1111 1111 0000 0000 0000 0000 1111 1111 1111 1111
ω_3	0000 0000 0000 0000 0000 0000 0000 0000 1111 1111 1111 1111 1111 1111 1111 1111

by the formula, $p(A_i) = q_i/k$, where q_i is the number of elementary products containing group A_i in affirmative form.

The indeterminacy in the analysis results may be characterized by entropy. If the sample under test is known to be an individual compound, the sets contained in $f(A)$ will be mutually exclusive, and the entropy will therefore be equal to $\log_2 k$. In a complete structural-group analysis of the sample composition, the numerical value of the information is equal to the initial entropy.

Thus, the establishment of the shape of $f(A)$ gives a numerical measure to the additional information essential for the unique determination of the sets of structural groups present in the molecules of a sample. This gives a means of designing further experiments in an optimal way.

Evidently, the logic of structural-group analysis lies in finding the logical consequences of a set of experimental and theoretical data represented in the form of Boolean functions regarded as the set of initial premises. In this sense, the formulation of the initial data in solving one particular structural problem or another is equivalent, generally speaking, to the construction of some axiomatic "micro-theory". In such an approach, analysis of the initial data is possible from the point of view of consistency, logical independence, and completeness.

E. The Structure Recognition System (STREC)

As far as we know, Elyashberg and Gribov⁵⁷⁻⁶¹ were the first to apply symbolic logic techniques in solving spectral-analytical problems. Subsequent developments in this direction resulted in the creation of a system of algorithms and programs for automatic molecular spectral analysis of a very general nature. Since our system is one of the highly developed and typical systems based on artificial intelligence principles, we shall describe it in detail.⁶⁵⁻⁷³ Here we can point out one important aspect of it. The structural-group analysis of molecules, based on the totality of their molecular spectra, is one version of the so-called inverse spectral problems formulated in a manner characteristic of incompletely defined problems. These can therefore be solved by the introduction of a number of constraints and imposition of additional conditions on the solutions, using these during the search for a solution. Consequently, none of the approaches can, in principle, afford to ignore this important aspect, which should be given due regard to some extent in identifying a molecule by any method available. This is taken into account in our system in the form of various supplementary constraints.

A specific feature of our method is that the feedback is effected by constructing a theoretical vibrational spectrum and then comparing it with the experimental spectrum for the purpose of evaluating the possible structure with regard to the degree of reliability. This algorithm has already been tested in practice and has proved its high reliability and rapid action. The limitation is that it works out only in those cases where a pure substance is identified. This, indeed, is a rather stringent restriction. Nevertheless, all the identification systems in existence today also suffer from the same drawback.

all the identification systems in existence today also suffer from the same drawback.

1. Block Diagram of STREC System

The algorithm and the system of structure recognition programs designated as STREC consists of the components presented in Figure 22. A brief description of the component applications is given below.

The block for structural-group analysis (Block 1) automatically constructs and solves logical equations reflecting relationships between a spectrum and the structure^{60,61,68}. The block operation is effected by use of an empirical formula and a library of standard fragments (LSF) containing spectrum-structure correlations for IR spectroscopy and a vibrational spectrum (presented in the form of a logical sum of IR and Raman spectra). The operation of Block 1 terminates with the output of all fragment sets which fit the spectrum and are not at variance with the empirical formula. It should be noted that this block may use, instead of IR, an NMR spectrum together with corresponding correlations.

Within the STREC system, the task of Block 1 consists in reducing the dimension of the problem through the choice of such fragments as would "absorb" a maximum number of the skeleton atoms.

The combinatorial block (Block 2) (Figure 22) then provides for a further reduction in the problem dimension by constructing from a given set of fragments combinations which may include several identical fragments on condition that the empirical formula requirements are met. This allows recognition problems for molecules with a large number (more than 20) of skeleton atoms to be solved, if several bulky identical fragments (benzene rings, etc.) are present.

Block 3 carries out mathematical synthesis of the structural formulas of all isomers on the basis of fragment sets chosen by means of the MASS system.^{65,66} In this case, any chemical information other than empirical formulas may be taken into account. This allows restrictions to be imposed on the structures synthesized.

Block 4 (Figure 22) exposes the structure of fragments contained in the incidence matrices in the form of "macro-atoms" (structural discrete units (SDU)). At this stage, a complete atomic structure of the molecule is obtained, with the nonequivalence of peripheral atoms of some fragments being taken into consideration (e.g., for $-\text{COO}-$, the following structures are formed: R_1-COOR_2 and R_2-COOR_1).

Block 5 checks the structures being synthesized for the presence or absence of each fragment from the library of standard fragments. When a fragment is found, its characteristic frequency intervals will be compared with the experimental spectrum. If the spectrum confirms the possibility of the presence of all fragments detected in the structure, the structure is tape recorded. Otherwise (should the presence of any fragment not be confirmed by the spectrum), further analysis of the given structure is terminated, and the program passes over to checking a subsequent incidence matrix. Block 5 makes it possible to solve problems of small dimension (up to seven or eight skeleton atoms) without the use of Blocks 1, 2, and 4. The screening of all isomers synthesized by the MASS system, which conform with a given empirical formula, is done by Block 5.

If in addition to a vibrational spectrum, NMR, UV, and mass spectra are known for the unidentified specimen, the structures selected by Block 5 are checked for the presence or absence of fragments included in the additional library of standard fragments (ALSF). This library is composed of NMR, UV, and mass fragments with appropriate characteristic spectral features (see the section entitled Library of Standard Fragments). Structures confirmed by NMR, UV, and mass spectra will be printed.

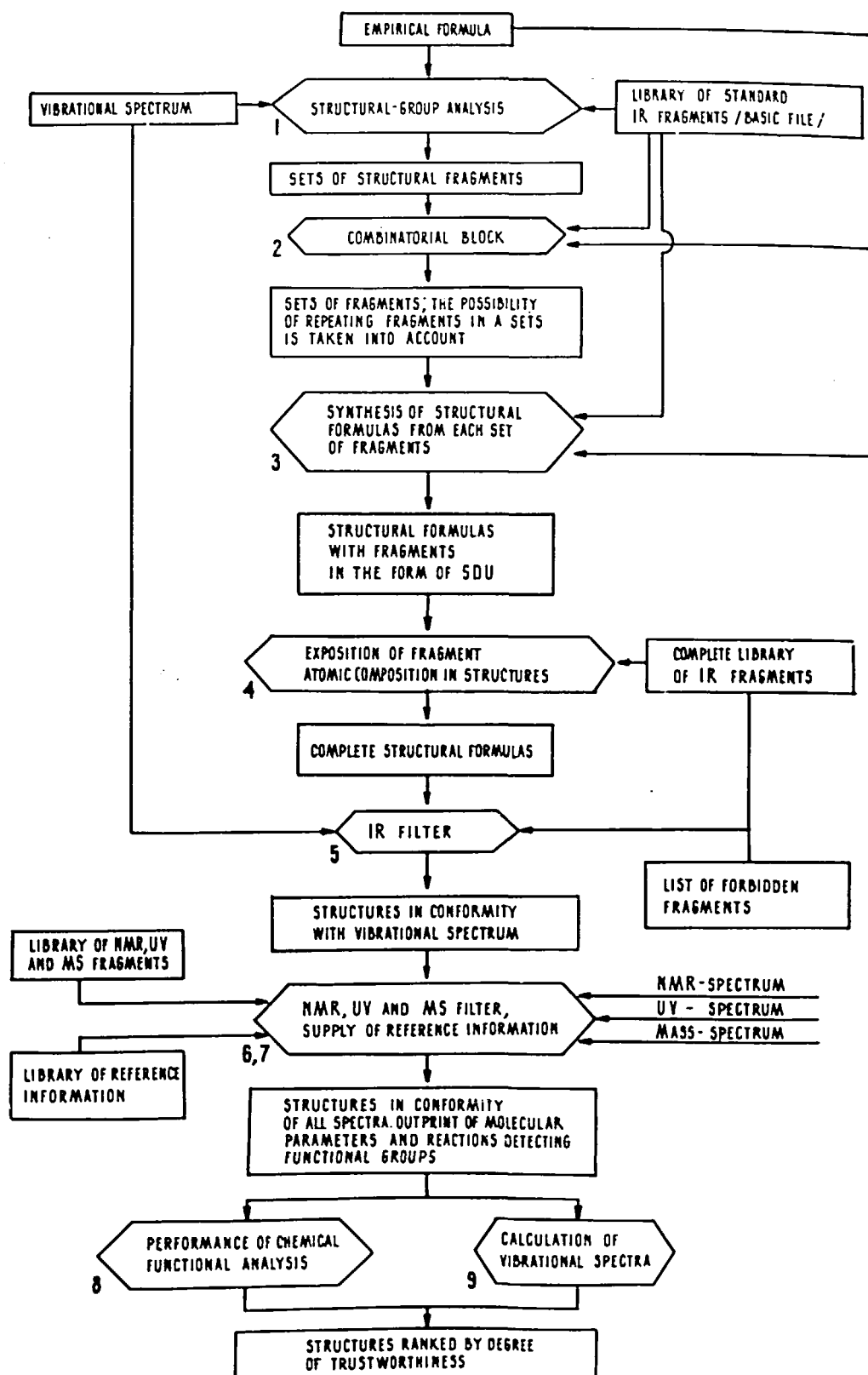


FIGURE 22. Block diagram of STREC system.

Structural formulas which are not confirmed by the mass spectrum are also printed along with the causes of contradiction.

If Block 6 does not supply a single solution, then at the operator's demand the informational library of standard fragments (ILSF) is accessed. This library comprises information on qualitative reactions for common functional groups as well as geometric, force, and electro-optical parameters for major fragments in organic chemistry.

The program searches the ILSF for fragments in the structural formulas obtained and prints out data relative to each structure.

In cases where qualitative reactions should lead to proper discrimination of structures (e.g., by bromination of O-O or C=C bonds), chemical analyses may be made by the procedures printed by Block 7.

The last block (Block 9) of the STREC system is intended to calculate the vibrational spectra of structures which cannot be distinguished by other methods. The spectra are calculated with the help of parameters supplied by Block 7 and compared by means of a formal criterion with the experimental spectra, whereupon the structures are ranked according to the degree of reliability. Obviously, Block 9 cannot be operated unless all data necessary for calculation are available, and calculations can be made only for compounds that do not have too many atoms. All of the algorithms of the system are in FORTRAN(IV).

In the following sections, the major algorithms of the STREC system are described. These are the algorithm for structural-group analysis, the algorithm for mathematical synthesis of all structural formulas, and the algorithm which analyzes structural formulas for the presence or absence of given fragments.

The mathematical formulation of an algorithm which decodes the group composition of polyatomic molecules from their spectra is based, as mentioned elsewhere, on the assumption that there is a certain relationship between individual structural units and the molecular spectra. Here the structural unit not only means groups containing a few atoms, but even whole molecules. This approach can therefore be extended to spectral catalogs, and as such, assumes a very general character. A collection of structural discrete units makes up one set. Another set is formed by the parameters inherent to the first set, e.g., band frequencies, chemical shifts, etc. Between these two sets a relationship is established whereby at least one object from the second set should correspond to each object in the first set and vice versa. With due regard for this aspect, the problem of structural-group analysis can be formulated strictly in terms of symbolic logic. As the general ideas relating to the solution of this problem have already been discussed, we shall not dwell on them here.

Experience in solving the structural-group problems has shown that in most cases the answer proved to be ambiguous, and the solution to be unstable. Such results are characteristic of inverse spectral problems. In this case, to obtain stable and unique solutions, a number of prior independent restrictions must be introduced which are specified by the type of problem. It is also necessary that each set of SDU detected can be used to synthesize mathematically a molecular structure corresponding both to the expected empirical formula and to valence theory.

The appropriate algorithm, placed in the third block of the system, is described in the following section.

2. Mathematical Synthesis of Molecular Structures

The algorithm for mathematical synthesis of molecular structures on the basis of atoms and fragments forming SDU is constructed on the following principles.^{65,66} To generate and analyze the structures, graph theory is used.^{63,64}

The structural formula of a compound is considered as a connected finite multigraph

represented by its incidence matrix. Atoms or fragments which can be introduced in the form of a "macro-atom" with a given valence, i.e., SDU, correspond to the vertices of the multigraph, whereas the bonds correspond to the edges. The multiplicity of the edges and the bonds is the same. The program generates all incidence matrices which satisfy both the empirical formula and the predetermined distribution of valences for the structural units. The matrices obtained are then checked for connectivity. From all the connected incidence matrices fitting a given isomer, the program selects one which is called a canonical.

Each structure has a corresponding set of incidence matrices.⁶³ Any matrix A_r of this set can be altered to any other matrix of the same set by renumbering the vertices of the graph. The A_r matrix is symmetrical, and its diagonal elements are zeros. For each vertex, there is a corresponding row and column of the matrix. If a square B_r submatrix which includes only incidence elements corresponding to SDU is selected, and if a column for hydrogen atoms is added on the right, then the values in this column will indicate the number of hydrogen atoms related to each SDU. Below, an incidence matrix is defined as a B_r matrix. Let us introduce the following number corresponding to the B_r matrix:

$$K_r = \sum_i f_i \sum_{ij} b_{ij} f_j \quad (i, j = 1, 2, \dots, n)$$

where i is the number of the row, j is the number of the column, b_{ij} is the value of a B_r matrix element (bond multiplicity), and f_i , f_j are the weighting factors for a row and column. The essential condition for the weighting factor is

$$f_i > f_{i+1} \cdot V$$

where V is the maximum multiplicity of the bonds observed in the samples. In the MASS system, $f_1 = 5 \times 10^9$; this value is sufficient for solving problems containing not more than 20 SDU.

A larger K_r represents a "larger" matrix. Generation of all possible structures involves constructing a B_m matrix regarded as the "maximal" one for a given empirical formula, and then successively subtracting from B_m the smallest possible numbers until K_r becomes zero.

To obtain a B_m matrix, the B_r rows are scanned successively downwards and from left to right, beginning with the element $(i, i+1)$. Each element is filled with the largest possible number (consistent with valence rules) until the sum of all the entries in the row equals the valence of the SDU corresponding to the row, or until the row ends (in the latter case the remaining valence units make up an element of an additional "hydrogen" column).

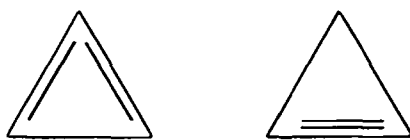
The "subtraction" is effected by working backward through the B_r until the first nonzero b_{ij} value is found; the b_{ij} value is then reduced by unity, and a "maximal" matrix is constructed beginning from $(b_{i, i+1})$.

The algorithm developed to check the graph for connectivity is called "graph vertex convergence." The concept of the algorithm is realized by successively uniting the adjacent vertices of the graph in one vertex, the edges remaining as they are except for those connecting the above vertices. If this operation results in only one vertex, the graph is connected; otherwise it is not.

Each graph has $n!$ incidence matrices (n is the number of vertices).⁶³ To single out one incidence matrix of set A , the set must first be ranked, and then one matrix must be chosen according to the principle of ranking.⁷⁴ This matrix is called the canonical matrix. In accordance with the algorithm of generation, canonicity is attributed to the "maximal" matrix along those possible for the graph. Thus, an algorithm for checking

canonicity must analyze whether the B_r rows could be renumbered in such a way as to produce a B_p matrix with $K_p > K_r$. If this proves to be impossible, the matrix obtained may be classified as canonical. There is no need for the inequality to be checked by calculating K_r and K_p ; the numbers obtained and the f_i value needed are too large. It is sufficient to calculate the sequences of $K_{r,i}$ and $K_{p,i}$ values for the B_r and B_p rows and to compare them; the matrices may also be compared element by element.

The system allows for human control of generation by means of subroutines which, depending on the problem, make it possible to generate structures pertinent to a certain class of organic compounds as well as to a combination of classes. These subroutines also allow isomers containing a certain number of predetermined functional groups or fragments to be included or excluded. For example, any output of isomers which are unlikely to be found in their natural form can be forbidden. Data on forbidden isomer structures are input beforehand as a computer catalog of the corresponding fragments, e.g.,



It should be noted that rotational, optical, and stereo-isomers are not distinguished during the generation, and molecular geometry is not taken into account. Therefore, the same incidence matrix will be valid both for *cis*-1 and *trans*-isomers.

An important feature of the MASS system in applications to identification problems should be emphasized. When the empirical formula contains chemical elements not incorporated in the STREC system or when the unknown compound includes *a priori* some fragments absent from the library of standard fragments (LSF), the MASS system may be used as a self-contained system. Under such conditions, the system is entered via fragments whose presence is confirmed by spectral and chemical considerations. At the same time, restrictions are imposed which explicitly forbid some combinations of fragments or of atoms in the empirical formula. Structures synthesized mathematically under these conditions are printed out and reviewed visually in an attempt to estimate their reliability. This variant of the system has been checked experimentally with positive results.⁷⁰

3. Algorithm for Analysis of Structural Formulas

As was mentioned in Subsection 1, the system is characterized by a feedback which compares the experimental and the theoretical spectra.

Theoretical spectra are constructed at two levels: (1) approximate "characteristic" spectra which are produced by additive summarizing of the spectra of discrete units forming a "suspected" structure (obviously, the LSF must contain all the indispensable discrete units) and (2) exact spectra which are constructed, only for IR spectroscopy, on the basis of molecular vibration theory by means of programs described earlier.⁷⁵

The STREC algorithm allows the construction of additive "characteristic" spectra of probable structures. This procedure is carried out by analysis of structural formulas in order to detect the presence or absence of fragments from the IR-LSF and/or the additional LSF. The structural formulas are also examined for certain atomic groups in order to use the informational LSF, which, in particular, contains data essential for forming vibrational equations and calculating intensities in IR spectra.

For this purpose, a general algorithm which detects predetermined fragments in structural formulas was developed. The concepts are as follows.⁶⁷

Let us assume a certain set of structural units, \mathcal{A} (chemical graph), represented by its incidence matrix A ; the latter includes only those incidence elements corresponding to the skeleton atoms or, in the general case, to SDU with a given valence. Any incidence matrix A_r obtained from A by r th renumbering of SDU is provided with the number K_r , which is used to check matrices for canonicity. The present problem consists in establishing the presence or absence of predetermined fragments (subgraphs) in all the \mathcal{A} graph.

It is assumed that the initial set which defines the incidence matrix A of the \mathcal{A} graph comprises N structural discrete units, and that $N = \sum_{k=1}^f n_k$, where f is the number of SDU types ($T_1, T_2, \dots, T_k, \dots, T_f$) and n_k is the number of SDU of the K type. For example, the compound $C_5N_2O_3H_{10}$ is characterized by $C \in T_1, N \in T_2, O \in T_3, f = 3, n_1 = 5, n_2 = 2, n_3 = 3$. The requirement is to establish the presence or absence in the \mathcal{A} graph of a subgraph \mathcal{B} , the vertices of which contain P structural discrete units ($P \leq N$); $P = \sum_{k=1}^f P_k$ (P_k = number of SDU of the K type in subgraph \mathcal{B}). For subgraph \mathcal{B} , an incidence matrix B , called a subgraph matrix, is constructed. Obviously, the order of matrix B is equal to P , and the number K_{rB} can be calculated for this matrix by the method described in Section IV.E.2.

From matrix A , rows and columns corresponding to the rows and columns of subgraph matrix B are selected by using the following criterion: a row of incidence matrix A is considered to correspond to a given row of matrix B , if for each element b_{rT_j} in the subgraph matrix B row, an element a_{rT_j} can be found in a row of matrix A , such that $b_{rT_j} = a_{rT_j}$. Here and below, i and j are the values of K ($K = 1 \div f$) for which $P_k > 0$.

A submatrix D is then constructed by combining the rows chosen in incidence matrix A and the columns which coincide by number with these rows. Suppose that submatrix D comprises d_i vertices of the T_i type, d_j vertices of the T_j type, etc. Obviously, if at least one T_k does not obey the inequality $d_k \geq P_k$, then the subgraph \mathcal{B} must be absent from graph \mathcal{A} ($d_k < P_k$ will mean that the number of K -type SDU in graph \mathcal{A} under analysis is insufficient for constructing graph \mathcal{B}).

If the above condition does not indicate the absence of subgraph \mathcal{B} , the following procedure must be applied. Since the inequality of $d_k \geq P_k$ ($P_k > 0$) is valid for every $T_k \in D$, all combinations of d_i vertices P_i at a time ($C_{d_i}^{P_i} = C_i$), of d_j vertices P_j at a time ($C_{d_j}^{P_j} = C_j$), etc. are listed, and an arbitrary choice is made of one combination from C_i , one from C_j , etc. The chosen combinations are united in one set S_u of vertices. Obviously, as $\sum_k P_k = P$, the number of vertices in set S_u equals the number of SDU in subgraph \mathcal{B} . In accordance with the number of possible methods for forming set S_u , the symbol u will run through the values of 1 to M , where $M = C_i \cdot C_j \cdot \dots = \prod_k C_k$ ($k = i, j$; the k values taken are such that $P_k > 0$).

Then a recursive procedure R is applied to the family of sets S_u ($u = 1 \dots M$): $S_u = RS_{u-1}$. The procedure R consists of successive ranking of all sets S_u differing at least in one element. At each step, the rows and columns of matrix D are chosen according to the meanings of the elements of set S_u (the numbers designating the vertices). These rows and columns are used to form matrix D_u . The examination of all of the S_u sets may be interpreted by operation of a counter, the j th place of which has C_j positions. Conditionally, the subsets $C_{d_i}^{P_i}, C_{d_j}^{P_j}, \dots$ are arranged in some order (e.g., in increasing order of the numbers in the incidence matrix rows; these numbers correspond to a given type of SDU). Thus, successive use of the R procedure initially generates all combinations of the $C_{d_i}^{P_i}$ subset elements, P_i at a time, and $S_u = RS_{u-1}$ is formed. Then, after C_i steps, a subsequent combination of the $C_{d_i}^{P_i}$ is generated, with another S_u being formed. This is followed by repeated generation of all $C_{d_k}^{P_k}$ and so on until all the counter places are filled up, i.e., until sets S_u differing in at least one element

have been examined. It is clear that the number of possible procedures R is equal to M.

Matrix D_u obtained at the u th step is checked for connectivity by the method described above. If matrix D_u proves to be connected, all the matrices $D_{u,l}$ ($l = 1, \dots, P_i!P_j! \dots$) are then generated, each matrix resulting from the permutation of rows pertinent to one of the SDU types in matrix D_u . For each matrix $D_{u,l}$, $K_{r,u,l}$ values are found. If for a given value l , $K_{r,u,l} = K_{r,B}$, then graph \mathcal{G} under analysis does not contain subgraph \mathcal{B} . Otherwise, if $K_{r,u,l} \neq K_{r,B}$ for all the l values, a further step of recursive procedure R is required. The total number of row permutations for all the D_u formed, which is needed to confirm the absence of a \mathcal{B} subgraph in graph \mathcal{G} , is equal to $M \cdot L$, where $M = \prod_k C_k$ and $L = \prod_k P_k!$. From this,

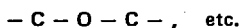
$$M \cdot L = \prod_k (C_k \cdot P_k!) = \prod_k \frac{d_k!}{(d_k - P_k)!} = \prod_k A_{d_k}^{P_k}$$

where $A_{d_k}^{P_k}$ is the number of arrangements of d_k vertices, P_k at a time.

The algorithm described above is very general and mathematically rigorous.⁴¹ Though it contains a substantial number of computational operations, the computer time required is negligible compared to the time needed to check the incidence matrices for canonicity during the generation of structures with predetermined topological properties. In the STREC system, the overwhelming majority of fragments contained in the correlation tables⁴⁹⁻⁵³ meet this requirement. Moreover, to describe dynamic interactions in molecules, it is sufficient in practically all cases to consider structural units containing no more than five atoms. Therefore, the use of the algorithm described does not offer significant difficulties.

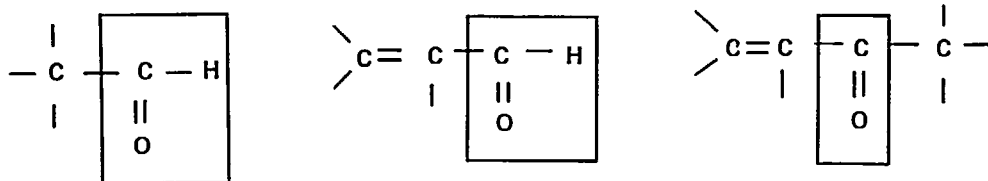
4. Library of Standard Fragments

The libraries of standard fragments for each spectral type have been composed from literature data. The IR LSF is composed of structural units taken from correlation tables.⁴⁹⁻⁵¹ The fragments are distributed between two files. The basic file receives fragments which possess sufficiently stable and highly informative characteristic spectral features, such as $-\text{C}-\text{CO}-\text{C}-$, $-\text{CH}_2\text{OH}$, $-\text{C}\equiv\text{CH}$, $\text{H}_2\text{C}=\text{CH}-\text{C}$, etc. These are used to automatically form logical equations of structural-group analysis and as a filter for the examination of hypothetical structures in order to detect contradictory features. The fragments of the auxiliary file participate only in the structure filtering. Two kinds of fragments are incorporated in this file: (1) those having only a small number of relatively uninformative features e.g.,



and (2) major functional groups without the indication of near environment, but with broad frequency intervals ($\text{C}=\text{O}$, 1640 to 2300 cm^{-1} ; OH , 3100 to 3650 cm^{-1} , etc.).

To accurately identify a class of chemical compounds which has its own fragment within the LSF (each fragment belongs to one class), a nucleus which includes the integral part of the given structural unit is selected from the fragment structure. The neighboring atoms with a given multiplicity of their valence bonds are also separated, e.g.,



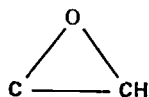
Empirical formulas of the nucleus and the environment are indicated separately. Environmental requirements are uniquely defined by specifying a list of forbidden fragments (LFF) and a list of obligatory fragments (LOF), which accompany the given nucleus.

The LOF enumerates the possible permissible variants of the nucleus environment which do not cause a change in any of the characteristic frequency intervals belonging to the nucleus. The LFF imposes necessary bans, e.g., if the fragment is a disubstituted benzene ring, addition of hydrogen atoms to the ring is forbidden.

The structure of fragments is represented by their incidence matrices. Thus, the fragment C_2CHOH (secondary alcohol) is characterized by the nucleus incidence matrix (a) and matrix (b) from the LOF pertaining to the fragment:

	Atoms	Connections					# of H
Free connections	01		1				1
	03	1					1
			2				
a							
	Atoms	Connections					# of H
Free connections	01		1	1	1		1
	01	1					
	01	1					
	03	1					1
b							

Here carbon atoms are designated by 01 and oxygen atoms by 03; the last column indicates the number of hydrogen atoms added to the corresponding SDU. SDUs can be either skeleton atoms or fragments, provided that their valence is given. The fifth row of the nucleus incidence matrix contains elements, the values of which are equal to the number of valences not utilized by the SDU in the nucleus (in this example, the nucleus carbon has two free valences). If a fragment contains nonequivalent atoms, this is fixed by inserting several matrices corresponding to the nucleus. Thus, in the fragment



the carbon atoms are not equivalent. Therefore, an epoxy group is described by two matrices:

01		1	1			1
01	1		1			
03	1	1				
	1	2				

01		1	1			
01	1		1			1
03	1	1				
	2	1				

Because of the LOF and LFF built into the system, certain groups, such as $-CH_3$, $C(CH_3)_2$, and $C(CH_3)_3$, can be confidently distinguished as independent. The system

is capable of handling the notion "benzene ring," without the latter being related to any specific type of substitution.

The vibrational spectrum is introduced into the computer as a sequence of experimental frequencies.

In forming the library of standard NMR fragments, data on the chemical shift interval for protons^{52,56,76} were used. Along with the chemical shift intervals, the expected multiplicities *M* were indicated on the assumption that they correspond to separations characteristic of first-order spectra. In cases when confident prediction of the multiplicity was impossible, constraints by *M* were not imposed. The experimental NMR spectra were expressed in terms of chemical shift values and corresponding multiplicities. When reliable determination of value *M* was impossible, only boundaries within which the complex signal occurred were specified. With the algorithm, the presence of a fragment in the structure was regarded as confirmed if the chemical shift in the spectrum fell within the characteristic interval of the fragment, provided that there was no contradiction from the *M* values.

For UV spectra, the positions of the band maxima were used. Wavelength intervals characteristic of the fragments were taken from the literature.^{55,56}

The library of mass fragments was formed on the basis of correlation tables listing characteristic mass numbers.⁵⁴ Each *m/e* value was provided with a corresponding set of probable fragments. Low-resolution mass spectrum was introduced as a sequence of the most intensive peaks (not more than 30).

The list of forbidden fragments (LFF) forms a file in the library and is common to the whole STREC system; LFF is analogous to BADLIST.⁷⁹⁻⁸⁰ As indicated above, the LFF includes unstable atom groups and those not found naturally. At the operator's option, all the available chemical information may be used in solving any specific problem. This can be achieved by entering additional atomic groups in the common LFF and indicating a list of obligatory fragments (LOF). Additional constraints may be introduced during data preparation and during the operation of the STREC program, particularly after output of a logical equations solution.

The library of reference information is based on the principle of the LSF. It contains the functional groups commonly encountered, as well as chemical reactions for their detection⁷⁷ and the geometrical, force, and electro-optical parameters characteristic of standard fragments of organic molecules.⁷⁵

The STREC system library consists of tape or disk files, each of which is called successively to the computer memory. The library can therefore contain practically any number of fragments together with related data. During tests of the STREC system, the sizes of the LSF were IR = 80, NMR = 150, mass spectra = 270, and UV = 20 fragments.

5. Results

To estimate the efficiency of the STREC spectroscopic recognition of molecular structures, more than 150 spectral problems taken from manuals and spectral atlases were solved by operating the system under various conditions. Molecules containing 4 to 16 skeleton atoms served as samples. During the first tests, the ability of the system to solve problems when only a vibrational spectrum was available was checked. Table 9 presents examples showing the number of solutions obtained in identifying various compounds from their IR spectra only. Table 10 provides data on the structure of isomers which could not be distinguished by the system from their characteristic IR frequencies. In most cases, analysis of IR spectra provides a small number of isomers, and these almost always contain the required structure. Incorrect answers were given when the spectrum contained features characteristic of fragments absent from the sam-

TABLE 9

Results of Structure Recognition by IR Spectroscopy

Empirical formula	Compound	Number of solutions
C ₂ H ₇ NO	2-Aminoethanol	1
C ₆ H ₁₂	2-Methylpentene-4	1
C ₇ H ₈	Toluene	1
C ₄ H ₁₀ O	Butanol-1	1
C ₄ H ₁₀ O	<i>t</i> -Butanol	1
C ₃ H ₆ O	Acetone	1
C ₄ H ₈ O	Methyl ethyl ketone	1
C ₇ H ₅ NO	Phenyl isocyanate	1
C ₈ H ₈ O ₂	<i>p</i> -Methoxybenzaldehyde	1
C ₄ H ₅ N	Allyl cyanide	1
C ₃ H ₆ O	Allyl alcohol	1
C ₉ H ₁₀ O	Methyl benzyl ketone	1
C ₈ H ₈ O	Acetophenone	1
C ₃ H ₆ O ₂	Propionic acid	1
C ₉ H ₁₂	Isopropylbenzene	1
C ₄ H ₉ N	Pyrrolidine	1
C ₆ H ₆ O	Phenol	1
C ₃ H ₇ N	<i>N</i> -Methyl-ethyleneimine	1
C ₇ H ₈ O	<i>m</i> -Cresol	1
C ₇ H ₉ N	Benzylamine	1
C ₆ H ₁₀ O	Cyclohexanone	1
C ₄ H ₁₁ N	<i>t</i> -Butylamine	1
C ₃ H ₆ O ₃	1-Methoxyacetic acid	1
C ₈ H ₁₁ N	<i>N,N</i> -Dimethylaniline	1
C ₈ H ₁₀	Ethylbenzene	2
C ₈ H ₁₀	<i>p</i> -Xylene	2
C ₅ H ₁₀ O ₂	Trimethylacetic acid	2
C ₁₁ H ₁₆ O	<i>p</i> - <i>t</i> -Butylanisole	2
C ₆ H ₆ O ₂	Pyrocatechol	2
C ₅ H ₁₀ N	3-Dimethylaminopropionitrile	3
C ₁₂ H ₁₀	1-Phenylhexadiyne-2,4	3
C ₄ H ₉ NO	Methoxypropionitrile	3
C ₄ H ₄ N ₂	Succinonitrile	3
C ₁₃ H ₁₁ N	Benzilene anilene	4
C ₄ H ₈ N ₂	2-(Methylamino) propionitrile	4
C ₁₃ H ₁₀	Fluorene	1
C ₃ H ₆ O ₂	Glycidol	1
C ₈ H ₄ N ₂	Isophthalonitrile	1
C ₈ H ₄ N ₂	Terephthalonitrile	1
C ₂ H ₅ NO ₂	Nitroethane	1
C ₃ H ₇ NO ₂	Nitropropane	1
C ₄ H ₁₀ O	Butanol-2	1
C ₇ H ₅ N	Benzonitrile	1
C ₇ H ₇ N	<i>o</i> -Methylaniline	1
C ₃ H ₆ N ₂	Dimethylcyanamide	1
C ₁₂ H ₁₀ O	Diphenyl ether	1
C ₂ H ₅ N	Ethylenimine	1
C ₃ H ₇ N	Trimethylamine	1
C ₆ H ₆	Hexadiyne-1,5	1
C ₄ H ₁₁ N	Diethylamine	2
C ₄ H ₈ O ₂	Ethyl acetate	2
C ₅ H ₁₀ O ₂	Valerianic acid	2
C ₅ H ₈ O ₂	Allyl acetate	2
C ₆ H ₁₄	2-Methylpentane	2

TABLE 9 (continued)

Results of Structure Recognition by IR Spectroscopy

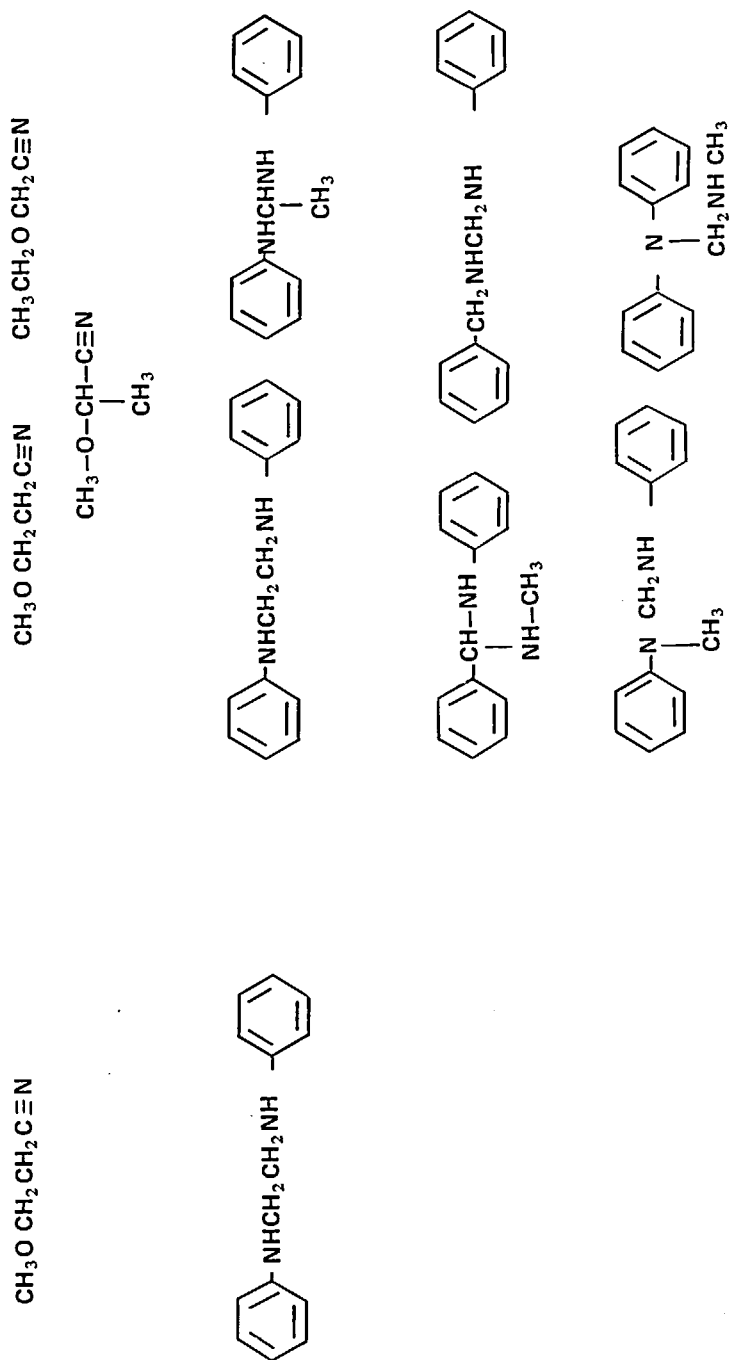
Empirical formula	Compound	Number of solutions
$C_3H_{10}O_2$	Isovaleric acid	2
C_3H_8NO	3-Aminopropanol-1	2
$C_3H_{10}O_2$	Acetopropanol	2
$C_4H_{11}N$	<i>n</i> -Butylamine	2
C_7H_8O	<i>p</i> -Cresol	2
C_7H_8O	<i>o</i> -Cresol	2
$C_3H_4O_2$	Epoxypropionic aldehyde	2
C_6H_{10}	Hexadiene-1,5	2
C_4H_8O	Tetrahydrofuran	3
$C_8H_4N_2$	Phthalonitrile	3
C_8H_7N	<i>m</i> -Tolunitrile	3
C_6H_{12}	Methylcyclopentane	4
$C_4H_8O_2$	1,4-Dioxane	6
$C_{13}H_{11}NO$	Benzilidene- <i>o</i> -aminophenol	6
$C_{14}H_{16}N_2$	<i>N,N</i> -Diphenylethylenediamine	6
$C_9H_{12}O_2$	Phenylglycidolic ether	10

TABLE 10

Examples of Isomeric Structures found from Their
IR Spectra

Structures to be found	Solution found
$\begin{array}{c} \text{CH}_2 \text{ CH}_3 \\ \\ \text{HN} \\ \\ \text{CH}_2 \text{ CH}_3 \end{array}$	$\begin{array}{c} \text{CH}_2 \text{ CH}_2 \text{ CH}_3 \\ \\ \text{HN} \\ \\ \text{CH}_3 \end{array}$
$\text{CH}_3 \text{ CH}_2 \text{ CH}_2 \text{ CH}_2 \text{ C}(=\text{O})\text{OH}$	$\text{CH}_3 \text{ CH}_2 \text{ CH}_2 \text{ CH}_2 \text{ C}(=\text{O})\text{CH}_3$
$\text{H}_2\text{C}=\text{CHCH}_2-\text{O}-\text{C}(=\text{O})\text{CH}_3$	$\text{H}_2\text{C}=\text{CHCH}_2\text{C}(=\text{O})\text{OCH}_3$
$\begin{array}{c} \text{CH}_3 \\ \\ \text{N}-\text{CH}_2\text{CH}_2\text{C}\equiv\text{N} \\ \\ \text{CH}_3 \end{array}$	$\begin{array}{c} \text{CH}_3 \\ \\ \text{N}-\text{CH}-\text{C}\equiv\text{N} \\ \quad \\ \text{CH}_3 \text{ CH}_3 \end{array}$

TABLE 10 (continued)



ple, e.g., a doublet at 1380 cm^{-1} was accepted by STREC as evidence for an isopropyl or *t*-butyl group. Sometimes the block of logical equation solutions did not yield fragment sets, either because the empirical formula was overfilled or because the spectrum did not contain any characteristic frequencies in the basic IR file. In such situations, the program took an alternative decision: when the number of skeleton atoms in the empirical formula did not exceed seven or eight, the computer began automatically to synthesize all the isomers of a given composition, screening them during filtering. If the number of skeleton atoms in the empirical formula exceeds eight, the program required additional information to avoid excessive losses of computer time.

Several examples showing the effectiveness of individual blocks of STREC are given below.

Example 1

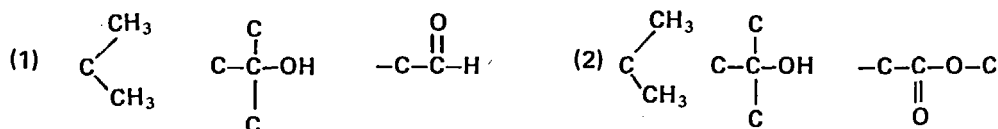
Empirical formula: $\text{C}_5\text{H}_{10}\text{O}_3$

IR: 970, 1150, 1200, 1270, 1370, 1450, 1740, 2980, 3500

NMR: 1.4 (1), * 3.7 (1)

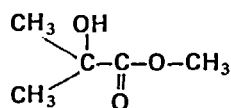
Mass: 27, 28, 29, 30, 31, 32, 33, 38, 39, 40, 41, 42, 43, 44, 45, 59, 61, 118

Selected sets of fragments:



Total structures examined: 116.

After the IR filter, three structures remained; after the NMR filter, one structure; and after the mass spectrum filter, one structure (which was consistent with the answer):



Example 2

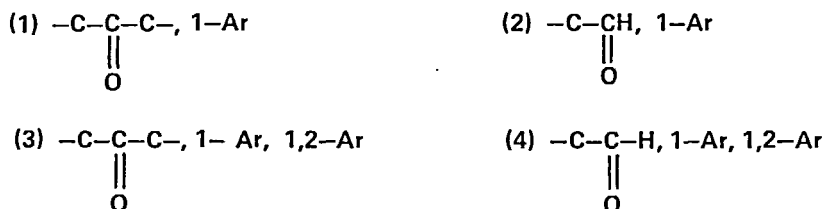
Empirical formula: $\text{C}_{15}\text{H}_{14}\text{O}$

IR: 695, 731, 752, 1056, 1120, 1333, 1447, 1500, 1610, 1725, 2970, 3040

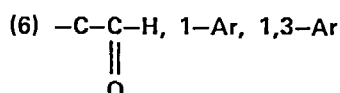
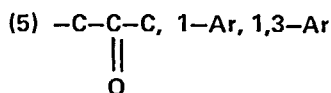
NMR: 3.55 (1), 7.1 (M)

Mass spectrum: 39, 41, 51, 63, 65, 91, 92, 118, 119, 210

Selected sets of fragments:

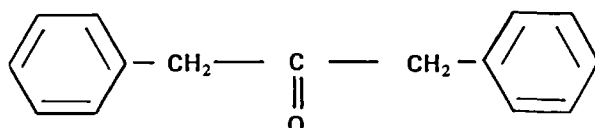


* The numbers in parenthesis designate multiplicities. The symbol (M) shows that the multiplicity was not established.



Total structures examined: 46

After the IR filter, five structures remained; after the NMR, one structure; and after the mass filter, one structure (which was consistent with the answer):



Example 3

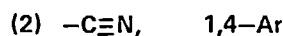
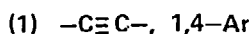
Empirical formula: $\text{C}_8\text{H}_7\text{N}$

IR: 703, 815, 950, 1022, 1041, 1120, 1180, 1292, 1385, 1456, 1506, 1608, 2230, 2940

NMR: 2.35 (1), 7.3 (M)

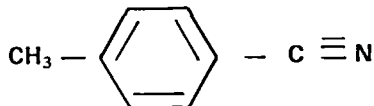
Mass: 27, 37, 38, 39, 41, 43, 44, 50, 51, 61, 62, 63, 64, 65, 75, 76, 89, 90, 91, 117

Selected sets of fragments:



Total structures examined: 5

After the IR filter, two structures remained; after the NMR filter, two structures; and after the mass filter, one structure (which was consistent with the answer):



Example 4

Empirical formula: $\text{C}_4\text{H}_8\text{O}$

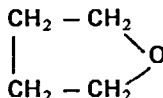
IR: 900, 1060, 1170, 1220, 1285, 1365, 1460, 2900

NMR: 1.79 (3), 3.63 (3)

No fragments were selected.

Total structures examined: 26.

After the IR filter, six structures remained; after the NMR filter, one structure (which was consistent with the answer):



Example 5

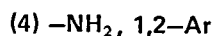
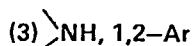
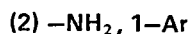
Empirical formula: $\text{C}_8\text{H}_{11}\text{N}$

IR: 700, 740, 830, 1030, 1070, 1370, 1435, 1605, 2960, 3020, 3390

NMR: 0.9 (1), 2.7 (M), 7.1 (M)

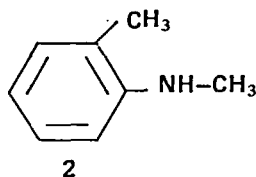
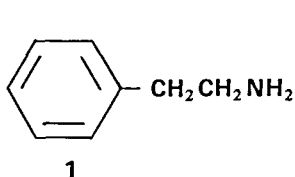
Mass: 28, 30, 32, 39, 41, 42, 50, 51, 52, 58, 63, 65, 77, 78, 84, 90, 91, 92, 103, 121

Selected sets of fragments:



Total structures examined: 16

After the IR filter, seven structures remained; after the NMR filter, two structures; and after the mass filter, two structures. The first (1) structure was consistent with the answer:



Example 6

Empirical formula: $C_9H_{13}N$

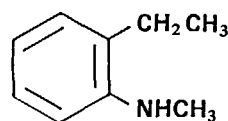
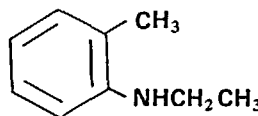
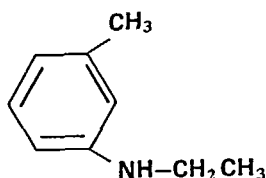
IR: 688, 770, 990, 1170, 1245, 1310, 1370, 1435, 1500, 1600, 2970, 3030, 3390

NMR: 1.1 (3), 2.2 (1), 2.9 to 3.1 (M), 6.1-7.1 (M)

Mass: 39, 51, 52, 53, 65, 70, 77, 91, 120, 135

Total structures examined: 56

After the IR filter, 14 structures remained; after the NMR filter, 8 structures; and after the mass filter, 4 structures. The first structure was consistent with the answer:



It can be seen that application of the recommended methods during the IR, NMR, and mass spectrum filtering, significantly increases the degree of selectivity of the answer.

As indicated above, when a problem cannot be solved unambiguously, the system provides for ranking the structures according to probability by calculating their vibrational spectra and checking each spectrum calculated for closeness to the experimental one. This approach appears promising, at least for nonbulky molecules.⁷¹

Overall, the results of STREC system tests are promising. The system should be particularly effective when oriented to solving problems of several special classes. It would then be expedient to use special-purpose libraries of standard fragments, consideration being given to the specific problems of a particular class. The library of standard fragments of the STREC system has been constructed on such principles so that it can easily be adapted to solve new types of problems.

F. Use of Computational Methods in the Identification of Molecules

The algorithm described above for identifying molecules is based on discrete spectral structural correlations and the existence of characteristic vibration frequencies. Such an algorithm may in principle prove to be unsatisfactory for the separation of isomers with an identical or close set of characteristic frequencies. In such cases, the information content of the vibrational spectrum can be enhanced, if in addition to the characteristic frequencies, consideration is also given to noncharacteristic frequencies corresponding to the relative motion of individual functional groups, which in consequence essentially depend on their combinations. This problem arises precisely when we are

interested in the identification of a compound which has no characteristic frequencies at all, and is made up, at least partly, of additive structural groups. Since each molecule has its own specific set of vibration frequencies, a unique solution can, in principle, be guaranteed in all cases, provided all the frequencies of vibrations of a system are taken into consideration in one way or another. It is useful here to make a preliminary evaluation of the amount of information contained in the characteristic and noncharacteristic spectral features.

Let N be the number of structural isomers with a given empirical formula and r be the number of structures derived from the solution of the identification problem. The magnitude of N can in principle, be found from a mathematical synthesis of all the isomers, using the appropriate algorithm.^{65,66} The amount of information inferred from the spectrum is thus expressed in the following form:

$$I = \log_2 N - \log_2 r$$

Here, I is the information that can only be obtained on the basis of spectral structural correlations and the logical combinatorial analysis described above. Assuming that unique identification of a molecule is achieved in a detailed analysis of the whole spectrum, we may evaluate the percentage of information that is obtained in a logical combinatorial solution of the corresponding inverse problem.

The estimate of information made on the basis of problems which we have solved, using the IR spectra only,* shows that about 40 to 100% of the information is inferred in this approach. Moreover, the percentage of a unique solution is quite high (see Table 11). This is indicative of the fact that this method is a fairly efficient one for solving the inverse problem. The information contained in the noncharacteristic frequencies is superfluous in this case. However, as already pointed out, certain isomer structures remain undistinguished. In such cases, the most promising route is that of constructing the complete molecular spectra on the basis of calculation methods and then comparing them with the experimental spectra.

As in the general case, the spectra have to be calculated for new compounds for which the force fields and geometry are not known, and the need arises to infer these parameters from those of the molecules already investigated in detail. Therefore, we can assert beforehand that the calculations have to be made in a very rough approximation, and success is not quite evident. In view of this situation, we have made an attempt to assess the prospects and potentialities of this method for identifying isomers on the basis of particular examples.

For this purpose, the frequencies and the shapes of 14 pairs of isomers of organic compounds were computed by the program⁷⁵ in the valence-field approximation. The values of the geometric parameters and the force constants were drawn directly from the data in the literature for molecules of similar structure. The degree of similarity between the calculated and the experimental spectra was evaluated by the following simple criterion.

Suppose it is required to compare two spectra (1 and 2) represented by the frequency sets, $\Omega_1(\nu_i)$, $i = 1, 2, \dots, d_1$, and $\Omega_2(\nu_j)$, $j = 1, 2, \dots, d_2$, respectively. Assume that the frequency $\nu_i \in \Omega_1$ corresponds to the frequency $\nu_j \in \Omega_2$ if

$$|\nu_i - \nu_j| \leq \delta \quad (9)$$

where δ is a certain given interval. We shall assume that, as a result of a comparison of all the elements $\nu_i \in \Omega_1$ with all the elements $\nu_j \in \Omega_2$, it is found that μ_{12} frequencies

* Compounds of such empirical formulas were chosen which did not need a long machine time to calculate N .

TABLE 11

Evaluation of Information Contained in
Vibrational Characteristic Frequencies


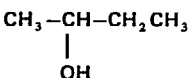
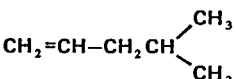
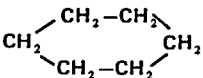
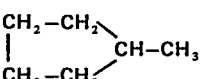
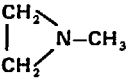
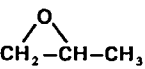
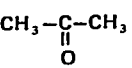
Structure to be found	Empirical formula	N	n	$I_0 = \log_2 N$	I	I/I_0 (%)
1	2	3	4	5	6	7
$(\text{CH}_3)_3\text{C}-\text{NH}_2$	$\text{C}_4\text{H}_{11}\text{N}$	8	1	3	3	100
$\text{CH}_3\text{CH}_2\text{NHCH}_2\text{CH}_3$	$\text{C}_4\text{H}_{11}\text{N}$	8	2	3	2	67
$\text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2\text{NH}_2$	$\text{C}_4\text{H}_{11}\text{N}$	8	2	3	2	67
	$\text{C}_2\text{H}_5\text{N}$	4	2	2	2	100
$(\text{CH}_3)_3\text{N}$	$\text{C}_3\text{H}_9\text{N}$	4	1	2	2	100
	$\text{C}_4\text{H}_{10}\text{O}$	7	1	2.807	2.807	100
$\text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2\text{OH}$	$\text{C}_4\text{H}_{10}\text{O}$	7	1	2.807	2.807	100
$(\text{CH}_3)_3\text{C}-\text{OH}$	$\text{C}_4\text{H}_{10}\text{O}$	7	1	2.807	2.807	100
	C_6H_{12}	25	1	4.644	4.644	100
	C_6H_{12}	25	5	4.644	2.322	50
	C_6H_{12}	25	5	4.644	2.322	50
$(\text{CH}_3)_3\text{COOH}$	$\text{C}_4\text{H}_{10}\text{O}_2$	27 ^a	2	4.755	3.755	80
	$\text{C}_3\text{H}_7\text{N}$	12	1	3.585	3.585	100
$\text{CH}_2=\text{CH}-\text{CH}_2\text{OH}$	$\text{C}_3\text{H}_6\text{O}$	9	1	3.170	3.170	100
	$\text{C}_3\text{H}_6\text{O}$	9	1	3.170	3.170	100
	$\text{C}_3\text{H}_6\text{O}$	9	1	3.170	3.170	100

TABLE 11 (continued)

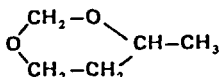
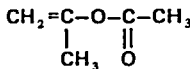
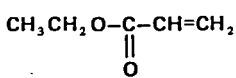
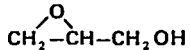
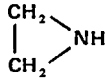
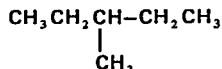
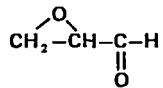
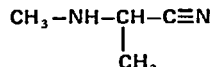
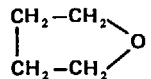
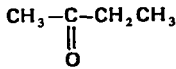
Structure to be found 1	Empirical formula 2	N 3	n 4	$I_o = \log_2 N$ 5	I 6	I/I_o 7
	$C_5H_{10}O_2$	266 ^a	35			
$CH_3CH_2O-CH_2CH_2OH$	$C_4H_{10}O_2$	27 ^a	6	4.755	2.170	46
	$C_5H_{10}O_2$	535 ^a	7	9.063	6.256	70
	$C_5H_{10}O_2$	535 ^a	4	9.063	7.063	78
$NH_2-(CH_2)_2-OH$	C_2H_7NO	8	1	3	3	100
	$C_3H_6O_2$	23	1	4.523	4.523	100
$(CH_3)_2N-C\equiv N$	$C_3H_6N_2$	136	1	7.087	7.087	100
	C_2H_5N	4	1	2	2	100
$(CH_3)_3N$	C_3H_9N	4	1	2	2	100
	C_6H_{14}	5	2	2.322	1.322	57
$NH_2CH_2CH_2CH_2OH$	C_3H_9NO	21	2	4.392	3.392	77
	$C_3H_4O_2$	36	1	5.170	5.170	100
	$C_4H_8N_2$	633	4	9.306	7.306	79
	C_4H_8O	26	3	4.70	3.115	66
	C_4H_8O	26	1	4.7	4.7	100

TABLE 11 (continued)

Evaluation of Information Contained in
Vibrational Characteristic Frequencies

Structure to be found	Empirical formula	N	n	$I_o = \log_2 N$	I	I/I_o
1	2	3	4	5	6	7
	$C_3H_6O_2$	34	1	5.087	5.087	100
	$C_3H_6O_2$	34	1	5.087	5.087	100
$CH_2=CH-CH_2CH_2CH=CH_2$	C_6H_{10}	77	2	6.267	5.267	56
$CH_3CH_2NO_2$	$C_2H_5NO_2$	86	1	6.426	6.426	100
	$C_4H_8O_2$	122	8	6.93	3.93	57
	$C_4H_8O_2$	122	2	6.93	5.93	86
$CH\equiv C-CH_2CH_2C\equiv CH$	C_6H_6	217	1	7.762	7.762	100
	$C_5H_{10}O_2$	266 ^a	2	8.055	7.055	75
	$C_5H_{10}O_2$	266 ^a	2	8.055	7.055	75
	$C_5H_{10}O_2$	266 ^a	2	8.055	7.055	75
	$C_5H_{10}O_2$	266 ^a	2	8.055	7.055	75
$CH_3CH_2CH_2CH_2C(=O)OH$	$C_5H_{10}O_2$	266 ^a	2	8.055	7.055	75
$CH_3CH_2CH_2NO_2$	$C_3H_7NO_2$	391	1	8.611	8.611	100
	$C_5H_8O_2$	535 ^a	1	9.063	9.063	100
	C_4H_9N	35	1	5.130	5.130	100
$CH_3-O-CH_2CH_2-C\equiv N$	C_4H_7NO	762	3	9.573	7.989	83
	$C_5H_{10}O_2$	266 ^a	10	8.055	4.734	59

^a The number of relatively stable isomers for the given empirical formula.

of the set Ω_1 satisfy the inequality (9). The relationship $\gamma_1 = \omega_{12}/\alpha_1$ then will in a certain sense characterize the probability that spectrum 1 corresponds to spectrum 2. Analogously, if we calculate $\gamma_2 = \mu_{21}/\alpha_2$, then the product

$$\epsilon_{12} = \gamma_1 \cdot \gamma_2 = \frac{\mu_{12} \mu_{21}}{\alpha_1 \cdot \alpha_2} \quad (10)$$

acquires the meaning of a coefficient for the mutual similarity of the two spectra. Obviously, in the general case, we have $\mu_{12} \neq \mu_{21}$.

As another similitude criterion, we used the distance between the spectra defined by Hemming⁷⁸ (Q_{12}) with due regard for the coincidence of the frequencies (inequality [9]). The magnitude of Q_{12} was determined as the total number of frequencies $\nu_i \in \Omega_1$ and $\nu_j \in \Omega_2$, none of which has a correspondence in the course of comparison of spectra

$$\rho_{12} = (\alpha_1 - \mu_{12}) + (\alpha_2 - \mu_{21}) \quad (11)$$

Let Ω_A denote the set of experimental frequencies of the sample under test; $\Omega_x, \Omega_y, \Omega_z, \dots$ stand for the calculated spectra of hypothetical structures. Evidently, by calculating $\epsilon_{Ax}, \epsilon_{Ay}, \epsilon_{Az}, \dots$ and Q_{Ax}, Q_{Ay}, \dots and then ranking them in the order of the increasing similarity factor (decreasing distance), a preferable structure may be identified.

In applying the ϵ and Q criteria, we took into consideration the fact that the mean distance between the frequencies of the experimental and calculated spectra is usually of the order of 30 to 40 cm^{-1} . In the inequality (9), we therefore took $\delta = 40 \text{ cm}^{-1}$.

Such isomers of close structures were combined into pairs that are difficult or impossible to distinguish from their IR spectra by means of empirical correlations. As our main aim was to check the validity of the approach in principle, molecules containing not more than 16 atoms were chosen as the model molecules. Their IR spectra were either specially recorded or taken from the works already published.

For a given isomer pair, each experimental spectrum was compared with both the theoretical spectra, i.e., the spectrum of the true compound and the spectrum of the competitor compound. Then the corresponding values of ϵ and Q were computed. As the spectral pattern of absorption due to valence vibrations of C-H bonds is unimportant from the viewpoint of information value for isomers of close structures, this spectral region was discarded in our consideration. The pairs of model compounds which we investigated (Table 12) had either no characteristic features or such features, if any, were identical. Hence, we may believe that these compounds are quite suitable for modeling the intricate situations that may arise in analytical practice.

By way of example, let us consider in detail the calculation of similarity coefficients and distances for *cis*- and *trans*-isomers of fluorobromomethylene. Their experimental and calculated spectra in the frequency range from 700 to 1700 cm^{-1} are shown in Figure 23. Frequencies which satisfy the inequality (9) are joined together by a broken line.

As is known,⁴⁹⁻⁵¹ the IR spectra of *trans*-olefines are, as a rule, distinct from the spectra of *cis*-olefines in that they contain a strong band at about 970 cm^{-1} . From Figure 23 it can be seen that there is no absorption in this region in both spectra, probably due to the influence of active substituents on the frequency of out-of-plane deformation vibrations of hydrogen atoms. By virtue of Equations 9 to 11, we get $\alpha_G = 5$, $\alpha_g = 6$, $\alpha_H = 10$, $\alpha_h = 7$, $\mu_{Gg} = \mu_{gG} = 5$, $\mu_{Gh} = \mu_{hG} = 5$, $\mu_{Hh} = 8$, $\mu_{hH} = 7$, $\mu_{gH} = \mu_{Hg} = 5$, $\epsilon_{Gg} = 0.84$, $Q_{Gg} = 1$, $\epsilon_{Hh} = 0.8$, $Q_{Hh} = 2$, $\epsilon_{Gh} = 0.71$, $Q_{Gh} = 2$, $\epsilon_{Hg} = 0.5$, and $Q_{Hg} = 6$.

A comparison of ϵ and Q shows that $\epsilon_{Gg} > \epsilon_{Gh}$, $Q_{Gg} < Q_{Gh}$, $\epsilon_{Hh} > \epsilon_{Hg}$, and $Q_{Hh} < Q_{Hg}$.

TABLE 12

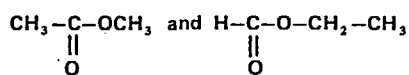
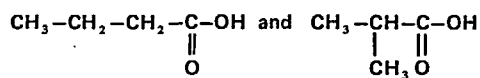
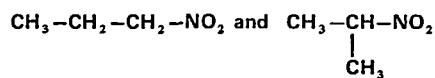
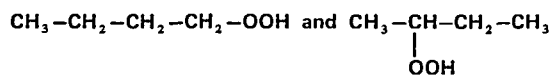
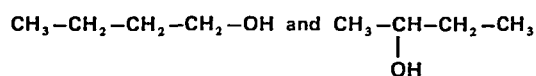
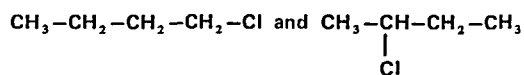
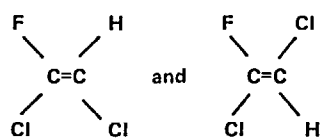
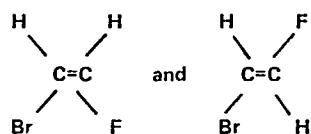
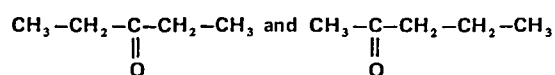
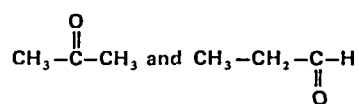
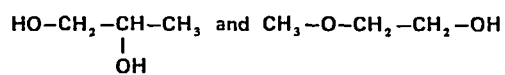
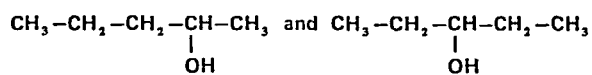
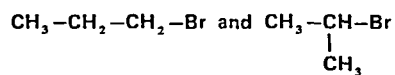
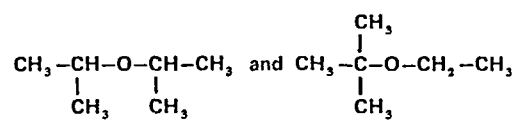


TABLE 12 (continued)



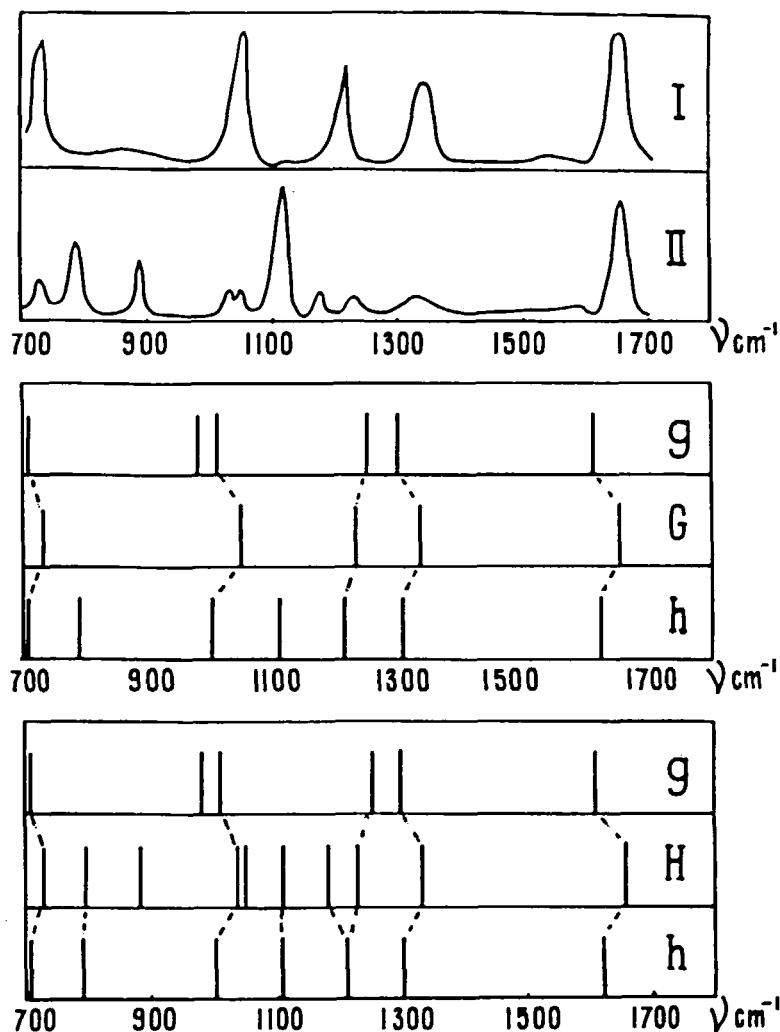


FIGURE 23. IR spectra of *cis*-fluorobromoethylene (I) and *trans*-fluorobromoethylene (II). Comparison of calculated (g, h) and experimental (G, H) frequencies in the spectra of compound I (G, g) and compound II (H, h).

Hence, we may conclude that in both cases the formal criteria gave the proper selection of probable structures. Similarity factors and distances obtained for the majority of the remaining isomers (28 cases in all were considered) have shown that in 20 out of 28 cases, even a very rough calculation of the vibrational spectra aided in proper selection of the most probable structures. Moreover, both criteria gave rise to identical results. Evidently, it would be almost impossible to arrive at a proper choice of the most probable structure by means of visual comparison of calculated and experimental spectra.

Thus, we may believe that the results obtained are quite satisfactory and the formal criteria proposed for the similarity of frequencies are highly effective. The experiments have likewise revealed the important role of the low-frequency part of the vibrational spectra in isomer identification: the probability of correct identification is enhanced if the IR spectrum is recorded starting at least from 200 cm^{-1} , because it is in this range that various kinds of noncharacteristic vibrations of atomic groups are readily exhib-

ited. The approach suggested in this paper is precisely the method which makes it possible to extract the structural information contained in the frequencies of these vibrations.

It is to be expected that if formal comparison criteria are worked out which also take account of the shape, intensity, and polarization of vibrations in experimental and calculated IR and Raman spectra, it may be possible to increase the reliability of identification based on calculation techniques of the molecular vibration theory.

From the above facts it follows that, although the inverse problem of the identification of molecules by their vibrational spectra is highly complicated, nonetheless, it readily yields to formalization. A computer-realizable algorithm can be designed that is adequately efficient and suitable for extensive application. This system is an integral part of a more sophisticated system which can be used not only for identification of molecules, but also for calculation of molecular vibrational and electron spectra, using the appropriate theories. This system is described in greater detail in Reference 73.

G. CONGEN Program

Of the works dealing with the application of artificial intelligence to the identification of molecular structures, special mention should be made of the studies conducted at Stanford University.⁷⁹⁻⁸⁵ In these works, initially oriented for the computer interpretation of mass spectra, an effective strategy has been developed for the first time for heuristic search (DENDRAL system) based on the generation of all the structural formulas which satisfy the constraints imposed and the empirical formula of a substance. The method used in this system for specifying the constraints in the form of desirable and forbidden structures (GOODLIST and BADLIST, respectively) has become an almost universal technique and is applied in several papers by other authors. The DENDRAL system is successfully being developed today for improving the recognition algorithms and programs for identifying complex molecules by their mass spectra.

Recently, Carhart et al.⁸⁶ designed a program called CONGEN which is an extremely universal tool for identification of molecular structures and which can be used to identify molecules with the use of various spectral methods. The basic idea underlying the CONGEN program is that of helping a chemist to analyze the information on the structure at his disposal and arriving at the correct structure by consecutive elimination of the wrong solutions.

The algorithm has two distinctive features pertaining to the method of generation of molecular structures and user-program interaction. Structures are generated from atoms and fragments (superatoms) by the "imbedding" technique. First, intermediate structures are generated in which the superatoms are included by their names (symbols). Each intermediate structure may represent a whole class of final structures. By giving the user an opportunity to interfere with the solution at this stage, the algorithm makes it possible to avoid the synthesis of an unduly large number of final structures by eliminating certain intermediate structures. Imbedding is a procedure whereby the meaning of the superatom symbol is uncovered and a mutually exclusive one-to-one correspondence is established between the symbol and the structural fragment. This technique is of considerable help in solving intricate problems by gradually building up a structure with due regard for the constraints.

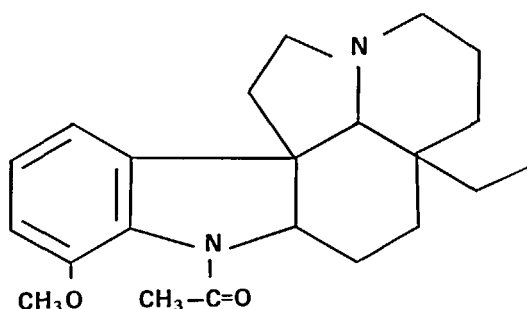
The possibility of realizing this procedure during solution helps in avoiding an uncontrollable combinatorial explosion, i.e., construction of an excessively large number of undesirable structures. A check of the intermediate structures is often very useful in detecting certain constraints that might have been disregarded in the initial stage.

An account of these limitations at the intermediate stage diminishes the complexity of the problem long before the next stage in computations.

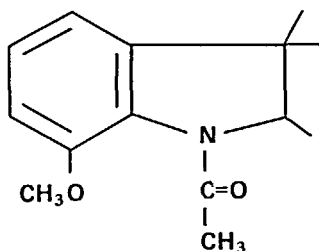
The program is based on limiting the structure generation process. In this respect, it includes both computational elements and blocks designed for automation of the solution strategy.

In principle, the general scheme of the method admits a few approaches to solving the problem and includes several levels. Moreover, at each stage the user can check the current structure, and if necessary, impose new constraints. For this purpose, several kinds of auxiliary operations have been included which can be realized at different stages in the solution. For example, the user can draw certain intermediate structures, store the results for subsequent use, restore the previous results, restart the solution afresh, print out the current commands, etc. These auxiliary functions are, incidentally, absolutely essential in any system where there is man-machine interaction. Therefore, considerable attention has been paid to this aspect of the program. The plotter program, for instance, can be connected to the standard computer terminal so that a remote user has access to the program and can put it to effective use. In this case, the user gets quite unambiguous representation of the structural formulas, although they may be rather unfamiliar to him (for example, the bonds may intersect).

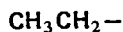
The method of solving structural problems is illustrated by the authors with examples taken from their own experience. Elucidation of the structure of the complex aspidospermine alkaloid is given as an example:



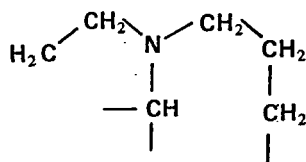
At the first stage, an initial set of superatoms is constructed from the data obtained by different methods. Three rather large atomic groups containing up to 14 atoms in the skeleton were chosen as such fragments and are denoted by the following formulas:



IND



ET



NP

It is clear that the valency of each superatom is equal to three, one, and four, respectively. Thereafter, the fragments obtained and the standard chemical atoms are entered in the COMPOSITION list in which the symbols and the number of each type of atom (superatoms) are shown. All the elements in the COMPOSITION list should in total give the empirical formula (for the example under consideration, $\text{C}_{22}\text{H}_{30}\text{O}_2$). If struc-

tures are generated without imposing any constraints, then an unduly large number of superfluous intermediate structures will be given (255 intermediate structures for the example under consideration). Therefore, the CONSTRAINTS list is used to eliminate the unnecessary structures.

The following constraints are introduced. The BADLIST and GOODLIST, as in the DENDRAL system, specify the forbidden and desirable fragments, respectively. The GOODRINGS list is used to specify the size and the number of the cycles which should be present in the structures generated. The forbidden ring fragments are entered in the BADRINGS list. The names of the fragments which determine the environment of hydrogen atoms and the corresponding number of hydrogen atoms are included in the PROTON list. The ISOPRENE list specifies the number of isoprene fragments and the method of their inclusion in the chain. The elements of the CONSTRAINTS list are formed by the EDITSTRUC program designed for editing the structures. Use of the constraint system and the introduction of every possible correction into this CONSTRAINTS list is permitted at all stages in the solution procedure.

In the case under consideration, on the basis of spectral data and calculated degree of unsaturation, EDITSTRUCT imposes a ban on the formation of three membered rings, methyl groups, and the addition of ethyl groups to the fragment NP. The requirement that four "two-membered" rings (multiple bonds) should necessarily be present is entered into the GOODRING list. This information is subsequently used at all CONGEN levels. A mathematical synthesis of the structures by a program developed by this research group⁷⁹⁻⁸⁵ earlier in this system is effected in the GENERATE phase. Synthesis is effected according to the valency specified beforehand for atoms and superatoms. Generation is effected using several strategies for the imposition of constraints. Strategy results from an analysis of the arguments put forward by the chemist.

Initially, the structural units are grouped in every possible way for the purpose of singling out the expected sets of atoms contained in the cyclic systems and acyclic chains ("partitions"). If the partition does not satisfy the constraints, it is discarded. For the purpose of verification, the partitions are divided into parts and then a rational sequence is established for scanning these parts. If, for example, the formation of a C=C bond is forbidden, all the paths leading to the synthesis of this bond are discarded.

Another strategy is that of building all the tree structures from each partition before attempting to construct ring systems from the partition parts. The constraints on the addition of a hydrogen atom to the other atoms are automatically imposed. For example, not a single hydrogen atom can be added to the superatom NP, and the program does not allow this.

For aspidospermine 59 intermediate structures were constructed in the first stage of generation. The most representative of these structures are shown in Figure 24.

The presence of three-membered cycles in these structures is not a violation of the constraints, because these cycles change their dimensions on exposing the internal structures of the fragments IND and NP. Elucidation is implemented by means of a subroutine called IMBED.

The IMBED routine restores the true linkage between the atoms without taking into consideration the molecular symmetry. The latter situation may lead to the appearance of duplicates in the intermediate structures. However, reduction of structures to their canonical forms avoids duplicates.

By imbedding the superatoms, the IMBED routine reveals the number of complete structural formulas that can be constructed from each intermediate structure. In the

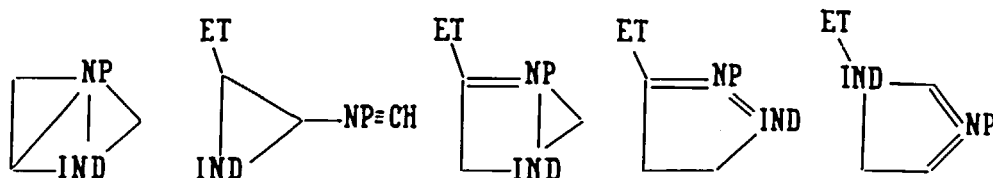


FIGURE 24. Examples of intermediate structures obtained in the GENERATE phase of the program. (from Cahart, R. E., Smith, D. H., Brown, H., and Djerassi, C., *J. Am. Chem. Soc.*, 97, 5755 (1975). With permission.)

course of the operation of the IMBED routine, the user can specify new constraints so that undesirable structures may be discarded at this stage.

For the given example, the IMBED routine restored 196 structures when applied to the fragment NP. Typical structures are shown in Figure 25. It can be seen that the superatoms IND and ET remain to be imbedded.

The PRUNE command is used to filter off all the undesirable structures. Pruning may be done automatically during generation and imbedding of the structural fragments or independently to apply new constraints. As a result of PRUNE, only ten structures survived in this example.

With new constraints imposed, IMBED and PRUNE, depending on the particular type of structures, resulted finally in 11 structures without any superatoms (Figure 26). At the final stage, PRUNE gave out one structure satisfying all the constraints and coinciding with the unknown compound (in Figure 26, structure 11).

In assessing the potentialities of the method, the authors point out that there is no need to impose the constraints from the very onset. They can be introduced step by step, depending on how they diminish the number of solutions. On the other hand, it is often more advantageous to introduce all the constraints during generation of the structures at the very beginning. IMBED may likewise be used at different stages, thus permitting flexibility of the system as a whole. CONGEN has a restricted capacity for storing information. Therefore, the intermediate structures should be entered in memory only when it is really essential. The authors mention a psychological nuance: most chemists feel a sense of despondency when they realize that under given constraints, hundreds of probable structures are given out. Therefore, the main strategy should lie in imposing effective constraints from the very beginning. Besides practical convenience, this also saves machine time.

The authors applied this method to various molecular structure problems and obtained good results. In this process, certain interesting situations came out which clearly demonstrate how difficult is it for man to foresee all the plausible structures in agreement with certain given constraints. Thus, under the conditions specified by Stoessl et al.,⁸⁷ 206 structural formulas correspond to the structure of lubimin, whereas CONGEN yielded only two of them as the most probable.

Although the spatial structure is not taken into account in the CONGEN program, a knowledge of molecular topology gives a lot of information. Elucidation of all the stereoisomers consistent with a given structural formula will considerably extend the potentialities of the program.

H. Various Strategies for Identification of Molecular Structure

Sasaki et al.⁸⁸ were among the first to attempt construction of an automatized system for elucidating molecular structure based on the use of PMR spectra. In their system, the initial data, i.e., PMR spectra introduced directly from the spectrometer, and the correlation tables were fed into the computer. The correlation tables which they com-

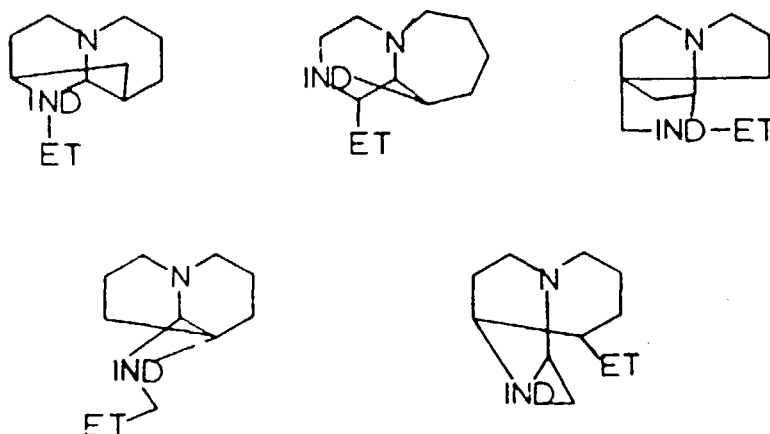


FIGURE 25. Some results of application of IMBED to NP.(from Cahart, R. E., Smith, D. H., Brown, H., and Djerassi, C., *J. Am. Chem. Soc.*, 97, 5755 (1975). With permission.)

piled from literature data included 150 fragments with the corresponding chemical shifts. The tables do not carry any information on the multiplicity of signals. The system is designed for identification of chemical compounds of the CHO class.

The authors made considerable efforts to elaborate a method of primary treatment of NMR spectra to be stored in the computer memory. The ordinates of the spectral curves were introduced with an interval of 0.2 Hz point by point. The program automatically detects the position of the maxima on the spectral curves, calculates the integral intensities, and the intensities of peak maxima. If OH groups are expected to be present on the basis of IR spectral data, then a preliminary analysis is made of the spectra at different temperatures and concentrations in order to identify the signals from the protons of the hydroxyl groups. The spectrum was divided into band groups, depending on the distances between the individual signals. Each group contained the signals covering a range of 20 Hz or more. Use of the integral intensities of each group of signals enabled the authors to assign a definite number of protons to each group, depending on the total number of hydrogen atoms consistent with the empirical formula. Then a structural-group analysis is made, using the spectral structural correlations. On the basis of the signal positions and signal groups present in the NMR spectrum, certain specific fragments are singled out for each group as the candidate solution. Here the data inferred from the IR spectra are partially used. Unfortunately, the structure and function of that part of the program pertaining to the use of the IR spectrum are not described in the paper.

The possibility that each of the candidate fragments is present is evaluated by comparing the number of protons in a fragment with the number of hydrogen atoms assigned to the corresponding signal group. On the basis of integral signal intensities, an estimate is made of the maximum possible number of each type of fragment which can make up the set explaining the spectrum. All possible structural formulas are then automatically synthesized from each set of fragments.

The authors note that no additional chemical information is utilized in the course of synthesis, and therefore there may be a large number of nonreal compounds, i.e., inconsistent with the requirements of organic chemistry. No provision is made in the system for the comparison of an experimental spectrum with the expected spectrum of each of the constructed structures. Evidently, such a block would considerably reduce the number of structures given in the answer.

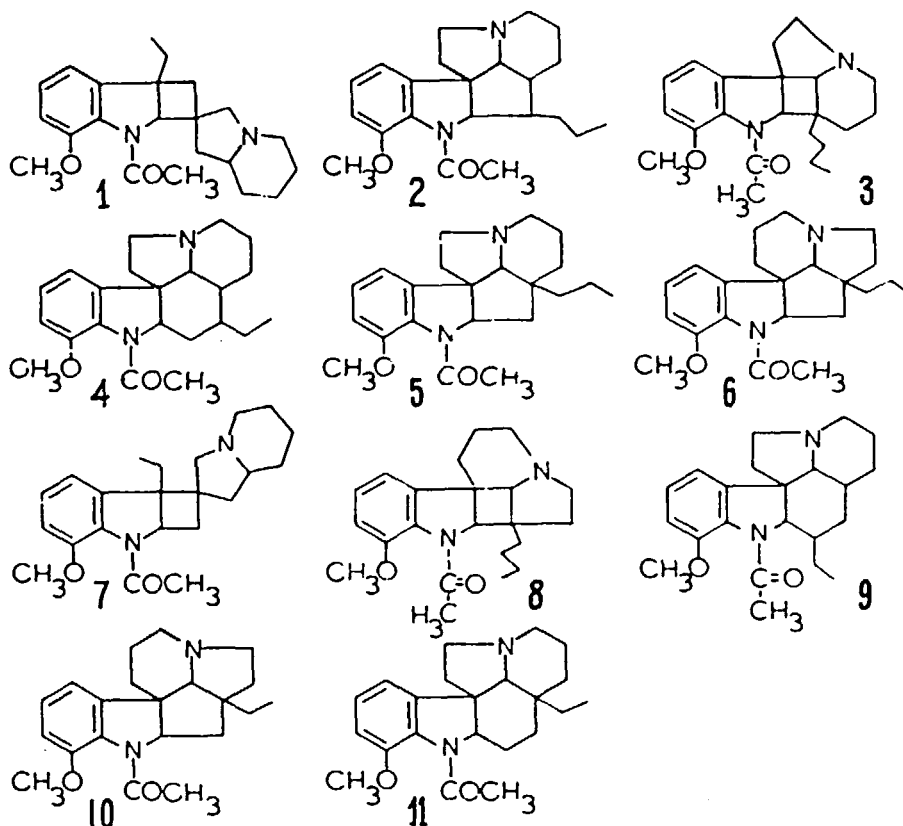


FIGURE 26. Final structures. (From Cahart, R. E., Smith, D. H., Brown, H., and Djerassi, C., *J. Am. Chem. Soc.*, 97, 5755 (1975). With permission.)

The function of the program is illustrated with reference to the identification of ethyl levulinate. This example clearly demonstrates the merits and demerits of the system. We shall therefore dwell on it in greater detail. As a result of structural group analysis of the spectrum of this compound, 21 fragments were chosen, out of which the program formed eight possible sets. The synthesis block gave the nine structural formulas shown in Figure 27.

From consideration of the example, it is obvious that the first structure (Figure 27) alone contains two carbonyl groups, to which the different frequencies in the IR spectrum should correspond. Furthermore, structures with three-membered cycles should likewise readily be detected from the IR spectrum, although in the NMR spectrum, a quite distinctive chemical shift corresponds to the cyclopropane ring. Therefore the investigator cannot accept this result as satisfactory. The potentialities of the system are fully illustrated in Table 13, which lists examples on structure identification.

On the whole, this work is a perceptible step forward in the automation of approaches to computer-aided elucidation of molecular structures. At the same time, this work clearly demonstrates the need for introduction of such constraints as GOODLIST and BADLIST⁷⁹⁻⁸⁵ into the synthesis of structural formulas and the advantage of comparing the expected spectra of generated structures with the experimental spectra. This work also shows that the IR spectral information has not been fully used.

In further developing the ideas and methods proposed in this work, Abe and Sasaki⁸⁹ elaborated on a system which they called the 1011 system. This program is useful in

identifying simple compounds of moderate size from their PMR and IR spectra, the candidate formulas being finally checked by the mass spectra. They assert that this program can elucidate the structure of compounds containing from six to ten carbon atoms and not more than one oxygen atom, the index of hydrogen deficiency being no greater than unity (i.e., the molecule may contain either one cycle or one multiple bond). An empirical formula has to be specified in order to solve the identification problem in this system. The block scheme of the 1011 system is shown in Figure 28.

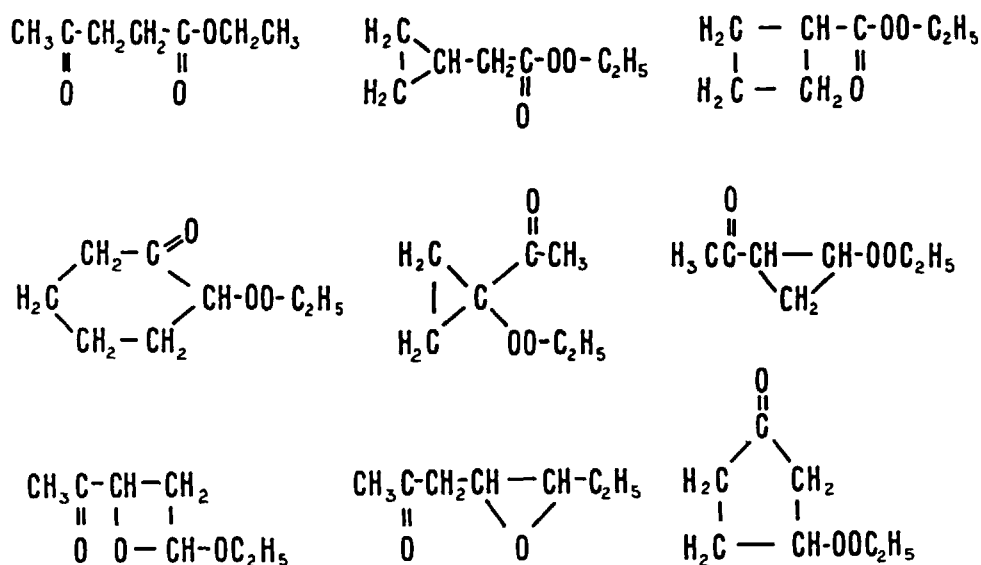


FIGURE 27. Structures identified by PMR spectrum of ethyl-levulinate. (From Sasaki, S., Kudo, Y., Ochiai, S., and Abe, H., *Microchim. Acta*, p. 726 (1971). With permission.)

TABLE 13

Examples of Recognized Structures

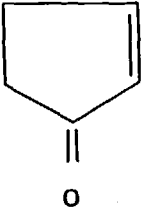
Compound tested	Answers
$\text{CH}_2=\text{CH}-\text{CH}_2-\text{OH}$	1
	1
$\text{CH}_2=\text{CH}-\text{CH}_2-\text{O}-\text{CH}_2-\text{CH}=\text{CH}_2$	1
$(\text{CH}_3\text{O})_2-\text{CH}-\text{CH}_2-\text{COCH}_3$	1
$\text{CH}_2=\text{CH}-\text{COOC}_2\text{H}_5$	1

Table 13 contd.

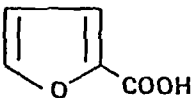
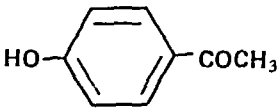
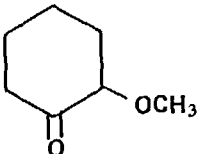
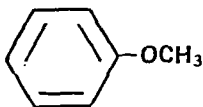
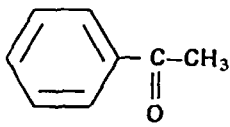
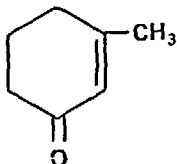
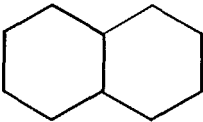
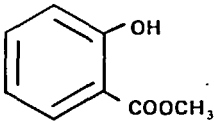
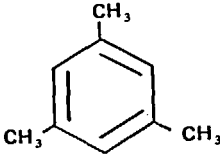
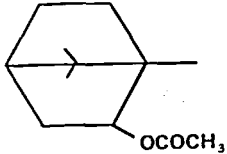
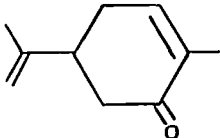
Compound tested	Answers
$C_6H_5CH=CH-COOC_2H_5$	1
$(C_2H_5O)_2-CH-COOC_2H_5$	1
$(CH_3)_3-C-COOH$	1
$CH_3-(CH_2)_5-C\equiv CH$	1
	3
	3
	3
$CH_3-CH=CH-CH=CH-COCH_3$	4
	4
	4
	7
$HOOC-CH-(CH_2)_5-C\equiv CH$ COOH	8

Table 13 contd.

Compounds tested	Answers
	21
	29
	37
	199
	221

From Sasaki, S., Kudo, Y., Ochiai, S., and Abe, H., *Mikrochim. Acta*, p. 726 (1971). With permission.

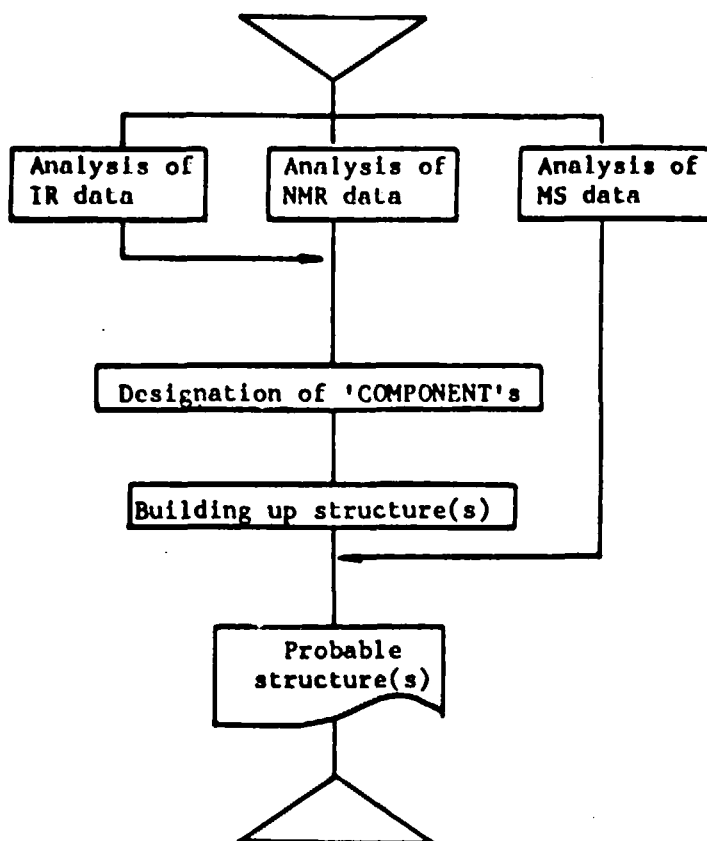


FIGURE 28. Block diagram of the structure determination program "1011". (From Abe, H. and Sasaki, S., *Sci. Rep. Tohoku Univ. Ser. 1*, 55, 63 (1972). With permission.)

The NMR, IR, and mass spectrometers are on-line to minicomputers which are built in within these spectrometers. The computers process the spectral information and then print out the NMR, IR, and mass spectra in digital form. Simultaneously, the mass spectra are normalized, and the spectral patterns of the NMR spectra are divided into groups, as in the previous work. The integral intensities of the signal groups and the peak heights are determined, and then for each signal group the possible number of protons responsible for absorption is established with the help of the correlation tables.

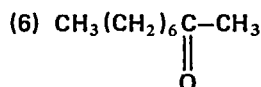
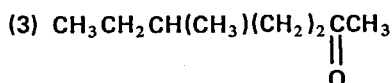
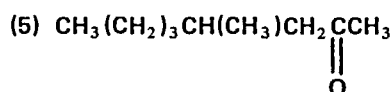
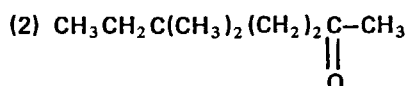
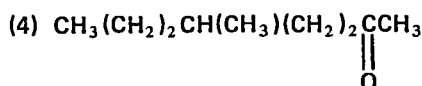
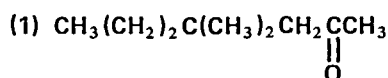
The list of discrete structural units, which the authors call the COMPONENT list, carry fragments with the nearest neighbors possible in a given system. A minimum empirical formula is assigned to each fragment. The list includes 83 components. Those components which have hydrogen atoms are entered into the table of spectral-structural correlations for NMR spectra. In this program, the IR spectra are used only to identify oxygen-containing groups and are not, therefore, needed to elucidate the structure of the carbon-containing compounds.

The whole frequency range of the IR spectrum is divided into seven intervals. Accordingly, seven strong bands are identified in the experimental spectrum. The most intensive band characterizing a given absorption range is selected in each interval. The most probable atomic groups are assigned to each of the seven intervals.

The first structural group analysis is carried out with the help of NMR and IR data.

In analogy with the system,⁸⁸ the fragments identified from the IR and NMR spectra are grouped into one set of components from which the subsets are subsequently constructed. These sets are so formed that each set contains atoms not exceeding the number specified in the empirical formula. Finally, a computer-assisted synthesis of structural formulas is implemented from all component sets.

The next stage consists in verifying the synthesized structures by their mass spectra. Verification is effected by means of several subprograms specially designed for each class of compounds (ketones, saturated esters, saturated alcohols, and unsaturated hydrocarbons). The lists of fragments carrying information on the mass numbers characteristic of each fragment and the peaks which, when they appear in the mass spectrum, contradict the presence of the corresponding fragment, are used as the data bank for the operation of mass spectral programs. The operation of the program is illustrated with reference to the elucidation of 2-nonanone. In identification of this compound, the program gave out the following three sets of fragments (Table 14). Six structural formulas were printed out:



Structures 1 and 2 were obtained from the first set, whereas structures 3, 4, and 5 were derived from the second set, and structure 6 (the true structure) from the third set.

This example shows that the use of three methods, provided the mass spectral programs are designed for seeking specific classes of chemical compounds, did not result in an unambiguous answer. Probably this is due not only to the difficulties involved in the extraction of spectral information, but also that the potentialities of the IR spectroscopy are not utilized to a fuller extent in the program.

Since there is only one band at the range 1380 cm^{-1} in the IR spectrum illustrated, it is doubtful whether structures 1 and 2 can be present in the specimen. Each of these structures contains the dimethyl group (it is known that a doublet in this range corresponds to the dimethyl group). Probably, the identification results could have been improved by the use of a more complete and comprehensive library of spectral structural correlations for IR spectra.

Absence of GOODLIST and BADLIST in this system in no way affects the analysis results because, in principle, program 1011 cannot give rise to structures forbidden in organic chemistry. It should be noted that both works^{88,89} do not give a description of the algorithm underlying the synthesis of all the structural formulas. Moreover, the technique used for identifying the fragments in structures during the verification of structural formulas by mass spectra is also obscure. The results listed in Table 15 to illustrate the operation of the program show that the number of structures detected is on the whole rather small. What is significant in this program is that the true structure is always found among the probable structures. Although the program is extremely limited in its capabilities, it can, nevertheless, identify molecules containing a few atoms falling within 11 different classes of chemical compounds.

TABLE 14
Possible Fragment Sets

Number of set	$\text{O}-\text{C}(\text{CH}_3)_3$	$\text{C}(\text{CH}_3)_3\text{C}$	$\text{CH}_3\text{CH}_2-\text{C}$	$\text{CH}_3\text{C}-\text{CHC}$	$\text{CH}_3\text{CO}-$	$-\text{CH}_2\text{CO}$	CCH_2C	$\overset{\text{C}}{\text{CCHC}}$	$\text{CH}_2=\text{C}$	$-\text{C}-$
1	0	1	1	0	1	1	2	0	0	0
2	0	0	1	1	1	1	3	1	0	0
3	0	0	1	0	1	1	5	0	0	0

From Abe, H. and Sasaki, S., *Sci. Rep. Tohoku Univ. Ser. I*, 55, 63 (1972). With permission.

TABLE 15

Results of Identification of Structures by means of the 1011 System

Compound	Molecular formula	Number of structures
1	2	3
$\begin{array}{c} \text{CH}_3-\text{CH}-\text{CH}_2-\text{CH}_2-\text{CH}_3 \\ \\ \text{CH}_3 \end{array}$	C_6H_{14}	3
$\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_3-\text{C}-\text{CH}_2-\text{CH}_3 \\ \\ \text{CH}_3 \end{array}$		1
$\begin{array}{c} \text{CH}_3-\text{CH}-\text{CH}-\text{CH}_3 \\ \quad \\ \text{CH}_3 \quad \text{CH}_3 \end{array}$		1
$\begin{array}{c} \text{CH}_3-\text{C}-\text{CH}_2-\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_3 \\ \\ \text{O} \end{array}$	C_7H_{14}	2
$\begin{array}{c} \text{CH}_3-\text{CH}_2-\text{C}-\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_3 \\ \\ \text{O} \end{array}$		1
$\begin{array}{c} \text{CH}_3-\text{CH}-\text{C}-\text{CH}-\text{CH}_3 \\ \quad \quad \\ \text{CH}_3 \quad \text{O} \quad \text{CH}_3 \end{array}$		1
$\begin{array}{c} \text{CH}_3-\text{CH}_2-\text{CH}_2-\text{C}-\text{CH}_2-\text{CH}_2\text{CH}_3 \\ \\ \text{O} \end{array}$		1
$\text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{OH}$	$\text{C}_7\text{H}_{16}\text{O}$	4
$\begin{array}{c} \text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2-\text{CH}-\text{CH}_3 \\ \\ \text{OH} \end{array}$		5
$\begin{array}{c} \text{CH}_3-\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}-\text{CH}_2\text{CH}_3 \\ \\ \text{OH} \end{array}$	$\text{C}_7\text{H}_{16}\text{O}$	3
$\begin{array}{c} \text{CH}_3-\text{CH}_2\text{CH}_2\text{CH}-\text{CH}_2\text{CH}_2\text{CH}_3 \\ \\ \text{OH} \end{array}$		2
$\begin{array}{c} \text{CH}_3-\text{CH}_2 \\ \\ \text{CH}_3-\text{CH}_2-\text{C}-\text{CH}_2\text{CH}_3 \\ \\ \text{OH} \end{array}$		2

TABLE 15 (continued)

Results of Identification of Structures by means of the 1011 System

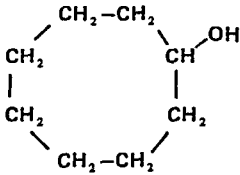
Compound	Molecular formula	Number of structures
1	2	3
$\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2-\text{C}-\text{OH} \\ \\ \text{CH}_3 \end{array}$		4
$\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_3-\text{CH}_2\text{CH}_2-\text{C}-\text{CH}_2\text{OH} \\ \\ \text{CH}_3 \end{array}$		3
$\text{CH}_2=\text{CH}-\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_3$		8
$\begin{array}{c} \text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}-\text{CHO} \\ \\ \text{CH}_3\text{CH}_2 \end{array}$	$\text{C}_8\text{H}_{16}\text{O}$	3
$\begin{array}{c} \text{CH}_3-\text{C}-\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_3 \\ \\ \text{O} \end{array}$		5
$\begin{array}{c} \text{CH}_3-\text{CH}_2-\text{C}-\text{CH}_2-\text{CH}-\text{CH}_2\text{CH}_3 \\ \quad \\ \text{O} \quad \text{CH}_3 \end{array}$		5
		1
$\text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2\text{OCH}_2\text{CH}_2\text{CH}_2\text{CH}_3$	$\text{C}_8\text{H}_{18}\text{O}$	2
$\text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{OH}$		
$\begin{array}{c} \text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}-\text{CH}_3 \\ \\ \text{OH} \end{array}$		7
$\begin{array}{c} \text{CH}_3-\text{CH}_2 \\ \\ \text{CH}_3-\text{CH}_2\text{CH}_2\text{CH}_2-\text{CH}-\text{CH}_2\text{OH} \end{array}$		5

TABLE 15 (continued)

Results of Identification of Structures by means of the 1011 System

Compound	Molecular formula	Number of structures
1	2	3
$\text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CHO}$	$\text{C}_9\text{H}_{18}\text{O}$	4
$\text{CH}_3-\underset{\text{O}}{\underset{\parallel}{\text{C}}}-\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_3$		6
$\text{CH}_3-\underset{\text{CH}_3}{\underset{ }{\text{CH}}}-\text{CH}_2-\underset{\text{O}}{\underset{\parallel}{\text{C}}}-\text{CH}_2-\underset{\text{CH}_3}{\underset{ }{\text{CH}}}-\text{CH}_3$		2
$\text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2-\underset{\text{O}}{\underset{\parallel}{\text{C}}}-\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_3$		6
$\text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{OH}$	$\text{C}_9\text{H}_{20}\text{O}$	21
$\text{CH}_3\text{CH}_2\underset{\text{OH}}{\underset{ }{\text{CH}}}-\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_3$		22
$\text{CH}_3\text{CH}_2\text{CH}_2\underset{\text{OH}}{\underset{ }{\text{CH}}}-\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_3$		45 45
$\text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_2\underset{\text{OH}}{\underset{ }{\text{CH}}}-\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_3$		6
$\text{CH}_3-\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2-\text{OCH}_2\text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_3$	$\text{C}_{10}\text{H}_{20}\text{O}$	9
$\text{CH}_3-\underset{\text{CH}_3}{\underset{ }{\text{CH}}}-\text{CH}_2\text{CH}_2-\text{O}-\text{CH}_2\text{CH}_2-\underset{\text{CH}_3}{\underset{ }{\text{CH}}}-\text{CH}_2$		4

From Abe, H. and Sasaki, S., *Sci. Rep. Tohoku Univ. Ser. 1*, 55, 63 (1972). With permission.

Of great interest is the program designed by Beech et al.³⁰ for the structural interpretation of first order proton magnetic resonance spectra. The program is based on an attempt to simulate the way of thinking of a specialist. Its distinctive features are that it makes use of not only chemical shifts of protons, but also their multiplicities, and there is no need to know the molecular formula of the compound to be identified. The latter fact is an important point because determination of the molecular formula is a separate and intricate problem.

The program, compiled in FORTRAN-IV is designed for use in off-line mode. A knowledge of the molecular ion is needed in this program for the final correction of the answer and for the elimination of the wrong structures. The block diagram of the program is shown in Figure 29.

The spectrum in the form of a sequence of line positions in ppm and the relative intensities of the peaks are fed into the program input. All the lines are divided into groups in such a manner that inside each group the distance between the lines does not exceed 10 Hz in the range of chemical shift up to 4 ppm and 20 Hz in the region of the weaker-field. In order to verify the multiplicity, the authors at first suggested certain formulas for estimating the probability of splitting shape. Thus, for example, the probability that a given signal may be assigned to a triplet is given by the expression:

$$P_T = \frac{S_{12} (I_1 + I_3)}{I_2 S_{23}}$$

where S_{12} and S_{23} are the distances between the first and the second, and between the second and the third lines, respectively; I_1 , I_2 , and I_3 are the integral intensities of lines. A similar formula is used to estimate the probability of a quartet:

$$P_Q = \frac{(S_{12} + S_{23} + S_{34}) (I_1 + I_4)}{S_{23} (I_2 + I_3)}$$

The value $P \geq 0.7$ shows that the choice of a multiplet is correct. Then the relative areas of the most probable multiplets are found by summing up the areas of separate peaks of each multiplet and then dividing them by the area of the least multiplet. The results of spectrum preliminary processing are fed for print-out.

The structural interpretation of spectra is effected with the help of machine correlation tables. The correlation tables contain 200 fragments, the limits of variation in the chemical shifts, and the multiplicity values being shown for each fragment. The multiplicity is given for the central group containing protons. The table also shows the possible environment of the central group. Thus, a central group with its environment forms a macrofragment. On the whole, the macrofragments are the different combinations of the 18 chemical groups listed in Table 16. The authors suggested the following two rules so that the selection of fragment may be consistent:

1. If a peripheral group does not contain protons and has a valency equal to n , it should occur at least $(n-1)$ times in the periphery parts of the macrofragments found from other multiplets.
2. If a peripheral group a of the fragment A is a proton-containing group, it should occur as a central group in one of the macrofragments derived from other multiplets; and if such a macrofragment is found, one of its peripheral groups is the central group a of the macrofragment A under consideration. The macrofragments which do not satisfy these rules are disregarded in the identification procedure.

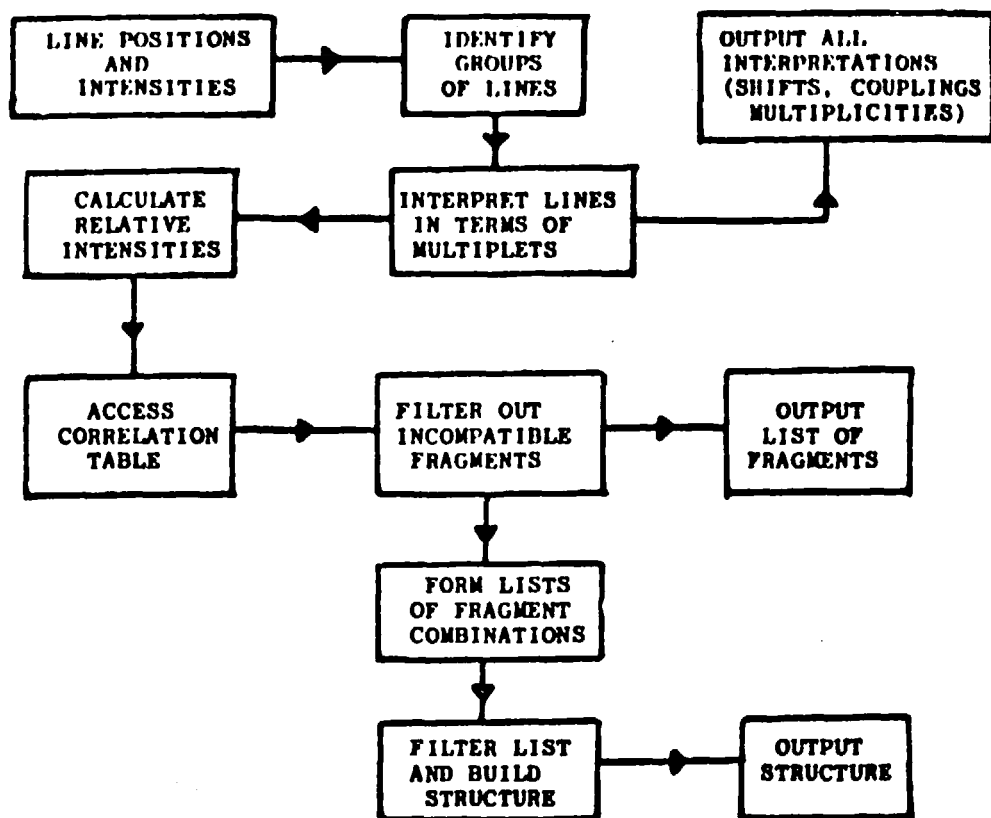


FIGURE 29. Schematic flow diagram of computer diagram. (Reprinted with permission from Beech, G., Jones, R. T., and Miller, K., *Anal. Chem.*, 46, 714 (1974). Copyright by the American Chemical Society.)

Moreover, an account is also taken of the fact that the fragments which are competitors to any one of the multiplets cannot be combined to form larger fragments because they are regarded as alternative candidates. Out of the fragments chosen that are consistent with these two rules, the program makes up sets which are fed to the input of a structural formula generator. If, as a result of verification, it is found that no structural formula can be constructed out of a given set in agreement with the experimental data, the set is discarded.

The method of building structures is based on matching the peripheral groups which form the macrofragments. The mass of the final structure is compared with the mass of the molecular ion. If they agree, the structure is given out as one that is consistent with the data. If a correct structure is not obtained at this stage, the relative areas of multiplets are doubled, and the program repeats its routine, beginning from the step where the fragments are generated. Thus, as a result, the so-called "dimer" molecules appear. This program was tested by solving 30 spectral analytical problems. The results are listed in Table 17. It is obvious from this table that the recognition results are quite satisfactory even in the absence of any information on the mass of molecular ions. It is noteworthy that the use of molecular ion masses resulted in unique answers in all cases. Nonetheless, the program demonstrated the limited possibility of the method based on the use of only one type of spectroscopy. It is capable of establishing the structural formulas only of chemical compounds to which the first order PMR spectra correspond. Therefore, several important classes of chemical compounds like olefines, molecules containing cycles, etc. are disregarded. The potentialities and the field of

TABLE 16
Basic Chemical Groups

Chemical group	CH ₃	CH ₂	CH	C ₆ H ₅	C ₆ H ₄	C
Valency	1	2	3	1	2	4
Chemical group	Cl	C=O	COO	N	O	OH
Valency	1	2	2	3	2	1
Chemical group	OOC	NO ₂	S	C≡C	CN	H
Valency	2	1	2	2	1	1

Reprinted with permission from Beech, G., Jones, R. T., and Miller, K., *Anal. Chem.*, 46, 714 (1974). Copyright by the American Chemical Society.

application of the method can be widened considerably by including, for instance, IR spectra in the program.

Another program which simulates man's thinking in elucidating the molecular structure from spectra is outlined in the paper by Gray.⁹¹ The initial data used in the construction of this program are protocols that describe the sequence of students' actions in solving certain spectral problems. The input information for the program consists of empirical formula, PMR, UV, IR, and mass spectra. The spectra are fed in the form of ordinates recorded every 10 nm in UV spectrum, every 10 cm⁻¹ in IR spectrum, and every 0.01 ppm in the NMR spectrum, etc. Information is fed into the computer in off-line mode. The empirical formula should be known beforehand for the program to be operated. The program is divided into three simple modules.

The first module carries out the preliminary processing of spectra and identifies the centers in the IR and UV bands, and also subdivides the NMR spectra into peak groups.

The second module, called the "Assign-fragments", consists of 20 routines, each of which corresponds to a set of operations essential for the elucidation of a specific functional group. The routines for a given particular problem are chosen by the program in accordance with the atomic composition of the unknown compound. For example, the routine for the identification of the carboxyl group is as follows: "if the IR spectrum has absorption in the range from 3200 to 2500 cm⁻¹, and strong broad band in the range 1670 to 1740 cm⁻¹, while a singlet is observed in the range 13 to 9 ppm in the PMR spectrum, then the spectral features should be taken as indicative of the presence of carboxyl groups, and constraints have to be imposed on the fragment's possible bonds."

If the presence of a fragment is confirmed by all types of spectra, the fragment is entered into the list of assigned fragments. After each fragment has been assigned, a check is made to find whether the list of possible routines can be simplified or shortened due to the elimination of subroutines which correspond to atoms and double bonds already used. It should be noted that the program selects only one list of probable fragments. There is no provision in the program for choosing fragments which are alternative to the fragments already chosen.

Each fragment has two valence descriptions. The first description determines the type of atom having free valency and contained in the composition of a given frag-

TABLE 17
Compounds Tested⁹⁰

Compound	Number of structures	
	With mass check	Without mass check
CH ₃ CH ₂ C ₆ H ₅	1	1
ClCH ₂ CH ₂ CO ₂ H	1	2
CH ₃ CH ₂ CH ₂ O ₂ CCH ₃	1	1
C ₆ H ₅ CH(CH ₃)O ₂ CCH ₃	1	1
(CH ₃ O) ₂ CHCH ₂ COCH ₃	1	5
CH(CO ₂ CH ₂ CH ₃) ₂	1	1
CH ₃ CH ₂ C(CH ₃)(CH ₂ CO ₂ H) ₂	1	1
CH ₃ CH ₂ CH(CO ₂ H)	1	1
C ₆ H ₅ CH ₂ O ₂ CCH ₂ CH ₃	1	1
CH ₃ O ₂ CCH ₂ CH ₂ CO ₂ H	1	4
CH ₃ OCH ₂ CO ₂ CH ₂ CH ₃	1	5
C ₆ H ₅ CH(OH)CH ₂ CH ₃	1	1
C ₆ H ₅ CH ₂ CH ₂ O ₂ CH	1	2
(CH ₃ CH ₂ O)CHCO ₂ CH ₂ CH ₃	1	1
CH ₃ CH(Cl)CO ₂ CH ₃	1	1
CH ₃ CH ₂ SCH ₂ C ₆ H ₅	1	2
ClCH ₂ CH ₂ CH ₂ C ₆ H ₅	1	3
CH ₃ CH ₂ CH(Cl)CO ₂ H	1	1
ClCH ₂ CH ₂ O ₂ CCH ₃	1	1
CH ₃ CH ₂ OCH ₂ CO ₂ CH ₂ CH ₃	1	1
CH ₃ OCH ₂ CN	1	3
CH ₃ CH(OH)COCH ₃	1	1
(CH ₃) ₂ NCH ₂ CO ₂ CH ₂ CH ₃	1	2
Cl ₂ CHCO ₂ CH ₂ CH ₃	1	1
C ₆ H ₅ CH ₂ CH ₂ OCH ₃	1	5
C ₆ H ₅ CH(CH ₃)NH ₂	1	1
CH ₃ C ₆ H ₄ CH(CH ₃) ₂	1	1
CH ₃ C ₆ H ₄ CH(CH ₃) ₂	1	3
C ₆ H ₅ CH ₂ N(H)CH ₂ CH ₃	1	1
C ₆ H ₅ CH ₂ N(CH ₂ CN) ₂	1	11

Reprinted with permission from Beech, G., Jones, R. T., and Miller, K., *Anal. Chem.*, 46, 714 (1974). Copyright by the American Chemical Society.

ment. The second description shows the possible valency of other fragments capable of forming bonds with the given fragments. The valence description makes it possible to introduce constraints on the structures synthesized.

The third module is called the "Solve-structure." This subprogram determines the connectivity matrix of the fragments. This matrix admits the formation of only simple bonds between the fragments. It is assumed that the connectivity matrix can be uniquely determined only by imposing constraints on the bonds linking the fragments.

All the possible bonds which each fragment may form with all the other fragments are shown in the initial connectivity matrix. Identification of a compound lies in sequential simplification of the initial connectivity matrix by elimination of those bonds, which are inconsistent with the information available on the molecule. In the first instance, those bonds are eliminated which lead to disconnected graphs. The possibility of addition of CH₂ and CH groups is verified by calculating the chemical shifts accord-

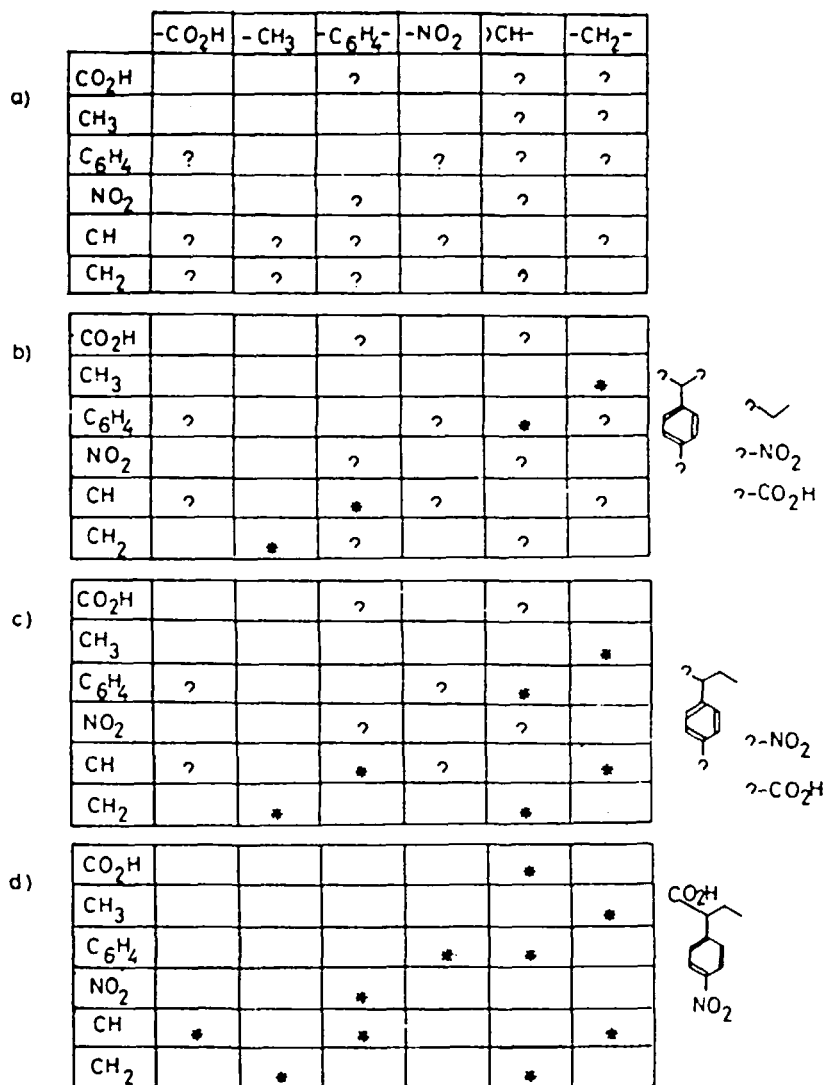


FIGURE 30. These different connectivity matrices illustrate the steps executed by the "solve structure" routine when combining the fragments assigned for the spectra. (Reprinted with permission from Beech, G., Jones, R. T., and Miller, K., *Anal. Chem.*, 46, 714 (1974). Copyright by the American Chemical Society.)

ing to the additivity law for the α -substituents. Here, account of multiplicity is taken only with regard to the methyl group. The mass spectrum can be used at this stage.

The corresponding program is based on inadequate empirical rules predicting the mass spectrum. Owing to the use of mass spectra, the program can make a proper choice between the possible structures. Moreover, the program includes certain heuristic routines. For example, there is a subroutine that implements the choice between five-membered, six-membered, or larger cyclic systems.

The logic of the program function is illustrated with reference to the example shown in Figure 30. The initial and succeeding connectivity matrices shown in this figure have been constructed for the set of fragments identified in the course of solution. The possible bonds between fragment pairs are marked with a question mark. After con-

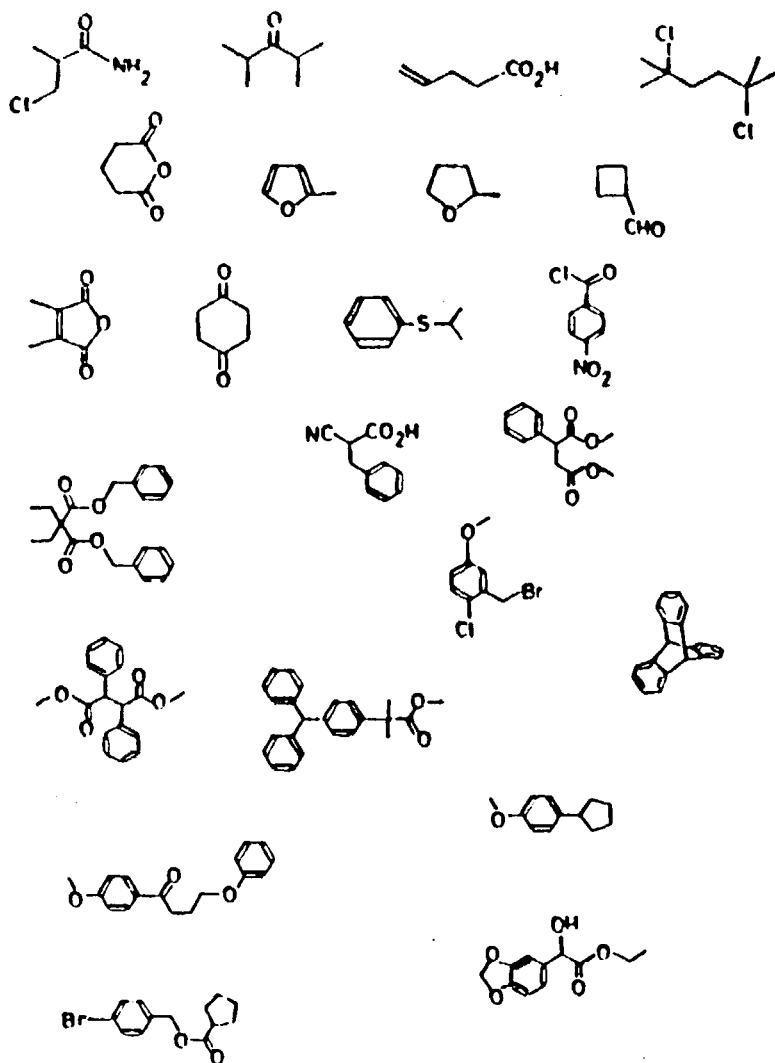


FIGURE 31. Examples of compounds correctly processed by the program. (Reprinted with permission from Gray, N. A., *Anal. Chem.*, 47, 2426 (1975). Copyright by the American Chemical Society.)

firming the presence of a certain bond, an asterisk is placed in the corresponding place instead of the question mark. The program was tested by solving spectral problems listed in the spectroscopy manual. It was found that the program confidently identifies quite complex structures (see Figure 31).

Gray, however, points out several demerits of his program. First, the result is affected by the fact that the program simulates the thinking of a student rather than that of a specialist. The program does not give an opportunity to utilize the data at the disposal of the synthesis chemist and does not include a provision for direct interaction with the user. In those cases where the specialist could easily have analyzed the shape of the NMR signal and arrived at the correct conclusions, the program was faced with insurmountable difficulties. There were cases where the program formed the wrong sets of fragments, although the structural group analysis was implemented correctly. In some rare cases, the correct structures were not derived from a good set of fragments.

Besides these demerits, the following should be mentioned. The requirement that only one set of fragments is to be chosen is a rather stringent condition. Although competing structural fragments, when taken into account, would lead to a large number of structures in many cases, nevertheless, the true structure may be established. Moreover, the program probably does not envisage a situation where it is not possible to select fragments from the experimental spectra that completely "absorb" all the atoms occurring in the empirical formula. Such situations are frequently encountered in practice and in order to deal with them, we should have a routine which might synthesize all the possible fragments from the remaining portions of the empirical formula by combining them with the fragments already established. The author does not mention whether the program ever failed to construct, from the fragment set selected, a structure resulting from the bans imposed by the structural chemistry. As a whole, Gray's paper is significant in that it raises several questions of great importance in the field of the automatic recognition system. This paper logically leads to an understanding of those requirements which should be met by a perfect recognition system. Probably, such a system should be able to simulate the thought of an experienced specialist in spectroscopy by implementing the logical combinatorial operations at a much faster rate with great precision. The requirements for the system should include such factors as the possibility of accounting for the origin of the sample, laws of structural chemistry, and the participation of the specialist in solving problems through interaction with the computer.

V. SUMMARY

In summing up this discussion, we shall deal once again with the three trends that were examined in detail in this review and assess their prospects.

The first trend is the creation of information retrieval systems. As is known, considerable advances have been made in this direction at several research centers. Information retrieval systems have been constructed which are capable of handling quite a large number of different problems. These systems are being run on an experimental basis, and in many cases have demonstrated their reliability.

Do such systems have any future? Will they develop? In what directions should they be developed? The answers to these questions are as follows.

Despite their inherent defects outlined elsewhere in this review, information retrieval systems should be developed in the future. Therefore, continuing efforts should be made to obtain more and more data about the spectra of pure substances, especially information about different kinds of spectral structural correlations. Such experiments should acquire ever increasing significance from year to year. Since a single solitary institution is not able to collect and store this information on a global scale, steps should be taken to invite international cooperation in this field, and methods have to be devised to publish the spectra of new compounds according to certain standards.

We can say that at present the mathematical tools needed in creating information retrieval systems and in processing the information stored in these systems have already been developed. Of course, the following complicated problem is still awaiting solution — how to store the spectra in the computer memory in the most complete form possible? This means that the spectrum should be stored in the form of a whole curve rather than in the form of certain information about some individual peaks or their integral or peak intensities. In turn, if the information about the compound to be identified is fed into the computer in the form of whole spectral curves, a serious problem arises in regard to the comparison of experimental and encoded spectra.²⁴

Nevertheless, this problem cannot be regarded as solved. In fact, the spectra may

differ significantly, depending on the recording conditions. To establish the similitude of the spectra is not a trivial problem. In complexity it is comparable to the problem of computer alpha scanning described in the literature. The methods of comparison of identified and atlas spectra in information retrieval systems will probably develop along these lines.

The second path of improving information retrieval systems is that of combining different kinds of spectra and constructing certain mathematical filters which deliver the final result after all the types of spectra have been filtered. No matter what advances are made in information retrieval systems, they will not be fully capable of solving the problems of elucidation of completely new compounds, at least as we now perceive the problems that are encountered.

Research centers engaged in the synthesis of new compounds are producing such a large number of these substances that it is almost impossible to have IRS which may contain fresh information about diverse molecules. The most powerful IRS in existence today contain about 100,000 spectra of the various classes of compounds. On the one hand, this is indeed quite a remarkable figure. Nevertheless, it should not be forgotten that thousands of new compounds are synthesized every year in different corners of the world. New compounds are being synthesized at an incredible rate and in future will probably supersede the capabilities of any information retrieval system. Since it is impossible to encompass all that is infinite, we have to hope for the appearance of a special IRS designed specifically for particular classes of compounds, for instance, organic compounds, organo-metallic compounds, inorganic compounds, or for the analysis of liquids, solids, gases, gas mixtures, etc.

The second group includes the methods based on pattern recognition. No essential advances have been made in recent years in this area. This is probably no accident. It is due to the fact that the pattern recognition method does not, as a rule, identify individual molecules, but only establishes whether a molecule belongs to one class or another. Therefore, it may not attract attention generally.

These methods are highly specialized in their goal, and any system based on them is capable of tackling only specific problems. There is general doubt whether it is possible on the basis of the pattern recognition method to design a sufficiently universal identification system capable of classifying a compound on its own. This does not, of course, imply that certain ideas and methods developed in this direction cannot be incorporated into the data processing programs based on other concepts. The point here is that the use of these methods as a separate entity seems inexpedient.

We believe that the most promising methods are those based on artificial intelligence, although they are extremely complicated to realize in practice. These methods are quite competitive with those based on information retrieval systems and, in addition, offer several advantages. For example, their data libraries and banks contain far less information bits than those which store the spectra of compounds directly or the spectra coded in the form of spectral features. These logical systems can easily be adapted to solve certain problems, and as such are invariably attractive in those laboratories which are engaged in the synthesis and study of new compounds.

Moreover, in these types of logical systems, it is immaterial which features are fed into the computer. They may, for instance, be spectral features, melting point, participation of compounds in specific reactions, the color of these compounds, smell, or any other suitable attribute. No specific constraints are imposed on the correlation between the features and the molecular structures. Any such suitable correlation can easily be introduced into the system in the course of analysis. All of these make this kind of system very flexible, convenient, and versatile to use. Nevertheless, it should be mentioned that data about spectral structural correlations in particular, are essential

for the operation of these systems. Such data are available today from printed matter, and as a rule they contain spectral structural correlations for average-sized fragments, individual bonds, three or four atom groups, etc.

The analysis of very large systems makes it necessary to elaborate these types of spectral structural correlations for large molecular fragments. Today such correlations are hardly available, and the task in the very near future is that of collecting them at least for the most important classes of organic and inorganic compounds. These correlations may, incidentally, prove to be more reliable than the structural spectral correlations for small-sized fragments because the probability of the presence of only the characteristic spectral features in large fragments is greatly enhanced. At the same time, it may happen that the large fragments will possess such a large number of spectral features that as a result, the spectrum of a molecule consisting of a few large parts may lose its discrete structure. Therefore, in analyzing large systems, recourse probably has to be taken to recording spectra under unusual conditions; for instance, at low temperatures or in liquids, gases, etc. In such cases, distinct spectra can be obtained, and hence the superposition of individual absorption bands is unimportant.

The compilation of new correlation tables for large systems and the drawing up of special tables, in turn, call for the elaboration and automatization of the appropriate mathematical techniques. Some progress has already been achieved in this respect. To solve this problem, we may use either the methods of symbolic logic,^{58,60,61} or the methods of statistical analysis of a large number of spectra in the atlas.⁶ What is important here is that the spectral-structural correlations may be ranked and the features specified both as certain or as probable. In turn, use of the ranked spectral-structural correlations calls for further refinement of the logical mathematical tools applied in solving the problem of group composition of the molecules in the first stage of molecular identification. As mentioned earlier in the body of the review, such spectral-structural correlations are already being used in solving this problem as fully reliable quantities, i.e., their real statistical weight is not taken into account.

The most important problem (if it can be called a problem), the next stage in the development of this area, is the analysis of a mixture of compounds. Indeed, although several highly effective techniques are available for the separation of mixtures into their components, of which the foremost are the diverse modifications of chromatographic analysis, electrophoresis, etc., we may, nevertheless, come upon a situation in which total separation is either impossible or extremely tedious. This point is hardly mentioned in the literature. It is not clear how the situation can be tackled in the case of multicomponent mixtures.

Perhaps, it is not the separation of mixtures into individual components that may be effective, but a promising method is that of evaluating the mean composition of these mixtures, elucidating the most significant or most important components with respect to certain criteria. For instance, the problem may arise in regard to the determination of biologically hazardous groups in one mixture or another. In this case, we can probably formulate the problem as the construction of a hypothetical molecule from a given spectrum that contains the most important fragments of all the molecules under study in the mixture.

This problem has not been tackled so far. Moreover, it is not clear how this problem should be formulated, and there can be no doubt that it is a pressing problem.

Thus, in conclusion, we stress that information retrieval systems need to be improved upon, thereby creating goal-oriented systems capable of functioning with coded spectra carrying detailed information.

Likewise, there is an urgent need to refine the methods based on artificial intelligence for the purpose of creating special-purpose systems.

It is not causal that man in his life and activities generally draws conclusions using logical reasoning. Certainly he cannot do without formal knowledge of various facts; however, the number of logical conclusions he constantly arrives at by far exceeds the number of facts kept in his mind. Nature, our best teacher, always selects the most rational ways. Having created man's brain it has clearly indicated that the combination of a limited "bank" of formally retained data with a powerful ability of logical thinking is the most economical way to provide unlimited possibilities for cognition. It is our strong conviction that systems designed to process spectroscopical data should be conceived following this model. Thus, there is no need to create comprehensive information retrieval systems. Rather, data banks should be a source of data for an efficient operation of logical programs. The combination of these two elements may be considered ideal.

Finally, it is quite essential to divert efforts toward elaborating general algorithms for the analysis of complicated mixtures made of both known and unknown compounds. We believe that the field described in this review will develop along these lines.

REFERENCES

1. Drobyshv, Yu. P., Nigmatullin, R. S., Lobanov, V. I., Korobeinicheva, I. K., Bochkarev, V. S., and Koptjug, V. A., *Vestn. Akad. Nauk SSSR*, 8, 75 (1970).
2. Drobyshv, Yu. P., Nigmatullin, R. S., Lobanov, V. I., Korobeinicheva, I. K., Bochkarev, V. S., and Koptjug, V. A., *Izv. Sib. Otd. Akad. Nauk SSSR Ser. Khim.*, 2, 108 (1972).
3. Barkhash, V. A., Sokolov, S. P., Sekerina, L. F., Drobyshv, Yu. P., and Koptjug, V. A., *Izv. Sib. Otd. Akad. Nauk SSSR Ser. Khim.*, 14(6), 111 (1974).
4. Bochkarev, V. S., Drobyshv, Yu. P., Koptjug, V. A., Korobeinicheva, I. K., Lobanov, V. I., and Nigmatullin, R. S., *Avotometriia*, 4, 124 (1972).
5. Smirnov, V. I., Nigmatullin, R. S., and Koptjug, V. A., *Zh. Prikl. Spektrosk.*, 22, 499 (1975).
6. Nigmatullin, R. S. and Smirnov, V. I., *Zh. Prikl. Spektrosk.*, 21, 307 (1974).
7. Koptjug, V. A., *Z. Chem.*, 15(2) 41 (1975).
8. Ul'yanov, G. P., Maslov, A. P., Piottukh-Peletsii, V. N., and Koptjug, V. A., Abstracts: 4-th all-union conference on the utilization of computers in spectroscopy of molecule, Novosibirsk, 1977, 78.
9. Erni, F. and Clerc, J. T., *Chimia*, 24, 388 (1970).
10. Erni, F. and Clerc, J. T., *Helv. Chim. Acta*, 55, 489 (1972).
11. Clerc, J. T. and Erni, F., *Computers in chemistry*, Springer-Verlag, Berlin, 1973, 91.
12. Naegeli, P. R. and Clerc, J. T., *Anal. Chem.*, 46, 739A (1974).
13. Penski, E. C., Padowski, D. A., and Bouck, J. B., *Anal. Chem.*, 46, 955 (1974).
14. Erley, D. S., *Anal. Chem.*, 40, 894 (1968).
15. Erley, D. S., *Appl. Spectrosc.*, 25, 200 (1971).
16. Rann, C. S., *Anal. Chem.*, 44, 1669 (1972).
17. Lytle, F. E., *Anal. Chem.*, 42, 355 (1970).
18. Kirby, E. M., Jones, R. N., and Cameron, D. G., *CODATA Bull.*, 21, 18 (1976).
19. Penca, M., Zupan, J., and Hadzi, D., *Anal. Chim. Acta Comp. Techn. Optimization*, 95, 3 (1977).
20. Miller, T. C. and Faulkner, L. R., *Anal. Chem.*, 48, 2083 (1977).
21. Sebesta, R. W. and Johnson, G. G., *Anal. Chem.*, 44, 260 (1972).
22. Fox, R. C., *Anal. Chem.*, 48, 717 (1976).
23. Zupan, J., Hadzi, D., and Penca, M., *Comp. Chem.*, 1, 71 (1976).
24. Tanabe, K. and Saeki, S., *Anal. Chem.*, 47, 118 (1975).
25. Vasil'ev, A. F., *Zavod. Lab.*, 40, 395 (1974).
26. Vasil'ev, A. F. and Pankova, M. B., *Zavod. Lab.*, 40, 1076 (1972).
27. Vasil'ev, A. F. and Pankova, M. B., *Zavod. Lab.*, 40, 1079 (1972).
28. Vasil'ev, A. F. and Aryutkina, N. L., *Zavod. Lab.*, 41, 339 (1975).

29. Aryutkina, N. L., Vasil'ev, A. F., and Kiseleva, A. A., *Avtometriya*, 2, 23 (1976).
30. Hirschfeld, T., *Anal. Chem.*, 48, 721 (1976).
31. Andrews, H. C., *Introduction to Mathematical Techniques in Pattern Recognition*, Interscience, New York, 1972.
32. Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1972.
33. Patrick, E. A., *Fundamentals of Pattern Recognition*, Prentice-Hall, Englewood Cliffs, N.J., 1972.
34. Vapnik, V. N., Ed., *Algorithms of Teaching of Pattern Recognition*, Vapnik, V. N., Ed., Soviet Radio, Moscow, 1973.
35. Jurs, P. C. and Isenhour, T. L., *Chemical Applications of Pattern Recognition*, John Wiley & Sons, New York 1975.
36. Kuznetsov, M. A., *Modern Problems of Organic Chemistry*, iss. 4, Leningrad State University, 1975, 5.
37. Jurs, P. C., Kowalski, B. R., and Isenhour, T. L., *Anal. Chem.*, 41, 690 (1969).
38. Kowalski, B. R., *Computers in Chemical and Biochemical Research*, Vol. 2, Academic Press, New York, 1974, 1.
39. Kozlova, S. I., Ph.D. thesis, Riga, 1970.
40. Kowalski, B. R., Jurs, P. C., and Isenhour, T. L., *Anal. Chem.*, 41, 1945 (1969).
41. Jurs, P. C., Kowalski, B. R., and Isenhour, T. L., *Anal. Chem.*, 41, 1949 (1969).
42. Preuss, D. R. and Jurs, P. C., *Anal. Chem.*, 46, 520 (1974).
43. Liddell, R. W. and Jurs, P. C., *Anal. Chem.*, 46, 2126 (1974).
44. Liddell, R. W. and Jurs, P. C., *Appl. Spectrosc.*, 27, 371 (1973).
45. Isenhour, T. L. and Jurs, P. C., *Anal. Chem.*, 43, 20A (1971).
46. Isenhour, T. L. and Jurs, P. C., *The Applications of Computer Techniques in Chemical Research*, Hepple, P., Ed., Institute of Petroleum, London, 1972, 189.
47. Jurs, P. C., *Anal. Chem.*, 43, 22 (1971).
48. Lowry, S. R., Woodruff, H. B., Ritter, G. L. and Isenhour, T. L., *Anal. Chem.*, 46, 1126 (1975).
49. Bellamy, L. J., *The Infra-red Spectra of Complex Molecules*, Methuen & Co., London, 1957.
50. Nakanishi, K., *Infrared Absorption Spectroscopy*, Holden-Day, San Francisco, 1962.
51. Colthup, N. B., Daly, L. H., and Wiberly, S. E., *Introduction to IR and Raman Spectroscopy*, Academic Press, New York, 1964.
52. Emsley, J. W., Feeney, J., and Sutcliffe, L. H., *High Resolution Nuclear Magnetic Resonance Spectroscopy*, Vol. 1 and 2, Pergamon Press, Oxford, 1965.
53. Bovey, F., *NMR Data Tables for Organic Compounds*, Vol. 1, Interscience, New York, 1967.
54. McLafferty, F. W., *Mass Spectral Correlations*, American Chemical Society, Washington, D. C., 1963.
55. Simon, W. and Clerc, T., *Strukturaufklärung Organischer Verbindungen Mit Spektroskopischen Methoden*, Springer, Frankfurt, 1967.
56. Gordon, A. J. and Ford, R. A., *The Chemist's Companion*, John Wiley & Sons, New York, 1972.
57. Elyashberg, M. E. and Gribov, L. A., *Zh. Prikl. Spektrosk.*, 8, 296 (1968).
58. Elyashberg, M. E., *Zh. Prikl. Spektrosk.*, 8, 648 (1968).
59. Elyashberg, M. E. and Moscovkina, L. A., *Zh. Prikl. Spektrosk.*, 8, 998, (1968).
60. Gribov, L. A. and Elyashberg, M. E., *J. Mol. Struct.*, 5, 179 (1970).
61. Gribov, L. A., Elyashberg, M. E., and Moscovkina, L. A., *J. Mol. Struct.*, 9, 357 (1971).
62. Ledley, R. S., *Digital Computer and Control Engineering*, McGraw-Hill, New York, 1960.
63. Harary, F., *Graph Theory*, Addison-Wesley, Reading, Mass., 1969.
64. Ore, O., *Theory of Graphs*, American Mathematics Society, Rhode Island, 1962.
65. Serov, V. V., Elyashberg, M. E., and Gribov, L. A., *Dokl. Akad. Nauk SSSR*, 224, 109 (1975).
66. Serov, V. V., Elyashberg, M. E., and Gribov, L. A., *J. Mol. Struct.*, 31, 381 (1976).
67. Serov, V. V., Elyashberg, M. E., and Gribov, L. A., *Zh. Strukt. Khim.*, 17, 1093 (1976).
68. Serov, V. V., Elyashberg, M. E., and Gribov, L. A., *Dokl. Akad. Nauk SSSR*, 232, 592 (1977).
69. Elyashberg, M. E. and Serov, V. V., *Theoretical Spectroscopy*, U.S.S.R. Academy of Sciences, Moscow, 1977, 102.
70. Elyashberg, M. E., Serov, V. V., and Gribov, L. A., *Zh. Prikl. Spektrosk.*, 26, 313 (1977).
71. Elyashberg, M. E. and Karasev, Yu. Z. *Zh. Strukt. Spektrosk.*, 26, 1047 (1977).
72. Gribov, L. A., Elyashberg, M. E. and Serov, V. V., *Anal. Chim. Acta Comp. Techn. and Optimiz.*, 95, 75 (1977).
73. Gribov, L. A., Dementyev, V. A., Elyashberg, M. E., and Yakupov, E. Z., *J. Mol. Struct.*, 22, 161 (1974).
74. Raznikov, V. V. and Talroze, V. L., *Zh. Prikl. Khim.*, 11, 357 (1970).
75. Gribov, L. A., *An Introduction to Molecular Spectroscopy*, Nauka, Moscow, 1976.
76. Beech, G., Jones, R. T., and Miller, K., *Anal. Chem.*, 46, 714, (1974).

77. Cheronis, N. D. and Ma, T. S., *Organic Functional Group Analysis by Micro and Semimicro Methods*, Interscience, New York, 1964.
78. Woodruff, H. B., Lowry, S. R., Ritter, G. L., and Isenhour, T. L., *Anal. Chem.*, **47**, 2027 (1975).
79. Lederberg, J., Sutherland, G. L., Buchanan, B. G., Feigenbaum, E. A., Robertson, A. V., Duffield, A. M., and Djerassi, C., *J. Am. Chem. Soc.*, **91**, 2973 (1969).
80. Duffield, A. M., Robertson, A. V., Djerassi, C., Buchanan, B. G., Sutherland, G. L., Feigenbaum, E. A., and Lederberg, J., *J. Am. Chem. Soc.*, **91**, 2977 (1969).
81. Schroll, G., Duffield, A. M., Djerassi, C., Buchanan, B. G., Sutherland, G. L., Feigenbaum, A. E., and Lederberg, J., *J. Am. Chem. Soc.*, **91**, 7440 (1969).
82. Buchs, A., Delfino, A. B., Duffield, A. M., Djerassi, C., Buchanan, B. G., Feigenbaum, E. A., and Lederberg, J., *Helv. Chim. Acta*, **53**, 1394 (1970).
83. Delfino, A. B. and Buchs, D. A., *Computers in Chemistry*, Springer-Verlag, Berlin, 1973, 109.
84. Sheikh, Y., Buchs, A., Delfino, A., Schroll, G., Duffield, A. M., Djerassi, C., Buchanan, B. G., Sutherland, G. L., Feigenbaum, E. A., and Lederberg, J., *Organic Mass Spectrometry*, Vol. 4, Heyden & Sons, London, 1970, 493.
85. Feigenbaum, E. A., Buchanan, B. G., and Lederberg, J., *Machine Intelligence*, Vol. 6, Elsevier, Edinburgh, 1971, 165.
86. Cahart, R. E., Smith, D. H., Brown, H., and Djerassi, C., *J. Am. Chem. Soc.*, **97**, 5755 (1975).
87. Stoessl, A., Stothers, J. B., and Ward, E. W., *J. Chem. Soc. Chem. Commun.*, p. 709 (1974).
88. Sasaki, S., Kudo, Y., Ochiai, S., and Abe, H., *Mikrochim. Acta*, p. 726 (1971).
89. Abe, H. and Sasaki, S., *Sci. Rep. Tohoku Univ. Ser. I*, **55**, 63 (1972).
90. Beech, G., Jones, R. T., and Miller, K., *Anal. Chem.*, **46**, 714 (1974).
91. Gray, N. A., *Anal. Chem.*, **47**, 2426 (1975).