(37) D. F. Meigh, *Nature (London)*, 579 (1952).
(38) F. Sanger, *Biochem. J.*, **44**, 126 (1949).
(39) S. Fittkau, *Naturwissenschaften*, **15**, 522 (1963).
(40) C. R. Hauser and W. B. Renfrow, Jr., "Organic Syntheses", Collect. Vol. II, Wiley, New York, N.Y., 1943, p 67.
(41) J. Chapman, *J. Chem. Soc.*, 507, 511 (1931).
(42) H. Woelsch, *Methods Enzymol.*, **3**, 570 (1957).
(43) W. J. Maimind, K. M. Ermoljajew, and M. M. Shemjakin, *Zh. Obshch. Khim.*, **26**, 2313 (1956).
(44) Aspartic acid residues were cyclized in a similar manner.
(45) Private communication from Dr. Francis Cardinaux.
(46) Although our analytical facilities are limited to a few tenths of a micromole or above, our procedures seem to work best at the lowest peptide levels attempted. There seems to be no reason why nanomole-level degradations using the procedure described here and elsewhere[1] should not be successful.
(47) M. Caplow, *J. Am. Chem. Soc.*, **90**, 6795 (1968).
(48) A. Williams and W. P. Jencks, *J. Chem. Soc., Perkin Trans. 2*, 1753 (1974).
(49) G. R. Stark, *Biochemistry*, **7**, 1796 (1968).
(50) (a) F. Weygand, D. Hffmann, and E. Wünsch, *Z. Naturforsch.*, **216**, 426 (1966); (b) E. Wünsch and F. Drees, *Chem. Ber.*, **99**, 110 (1966).
(51) H. T. Clark and L. D. Behr, "Organic Syntheses", Collect. Vol. II, Wiley, New York, N.Y., 1943, p 19.

A Computerized Infrared Spectral Interpreter as a Tool in Structure Elucidation of Natural Products[1]

Hugh B. Woodruff and Morton E. Munk*

Department of Chemistry, Arizona State University, Tempe, Arizona 85281

Received September 29, 1976

The process of structure elucidation as practiced by the natural products chemist requires the determination of structural constraints (the polyatomic fragments known to be present in the molecule and others known to be absent). The structural constraints may be chemist derived or computer derived. This paper decscribes an interactive computer program that interprets infrared spectra in order to help the chemist determine structural constraints. The program attempts to parallel the reasoning a chemist uses in interpreting a spectrum. If the program is to be of value to the chemist, it must be able to interpret the spectra of relatively complex molecules and make decisions concerning the presence or absence of a large number of functional groups. This program makes decisions concerning 159 different classes of compounds and has been tested on a wide variety of sample spectra.

Chemists in general, and natural products chemists in particular, are frequently faced with the task of piecing together results from chemical experimentation and spectroscopic work to deduce the molecular structure of an unknown compound. The chemist interprets these data, expresses the result in terms of a partial structure (structural fragments plus unaccounted for atoms), and intuitively attempts to reduce the partial structure to all molecular structures consistent with the available evidence. While intuition is a valuable asset to the chemist, an asset that cannot be adequately programmed into a computer model of the structure elucidation process, the chemist may frequently overlook a valid combination pathway, especially when dealing with the relatively complex molecules of nature. To preclude such an occurrence, and to relieve the chemist of the tedious task of manually assembling molecular structures, several computerized structure generators have been developed that ensure that *all* chemically feasible molecules are considered.[2-9]

For all but the simplest molecular formula, the number of valid structures generated without structural constraints (the polyatomic fragments known to be present and others known to be absent) is unmanageably large. As structural constraints are imposed on the process, the number of structures generated is diminished, ultimately to one. The structural constraints may be chemist derived or computer derived. Clearly, an essential component of any computerized structure elucidation package is an effective spectral interpretation procedure.

This paper describes an artificial intelligence program that aids the chemist in the interpretation of infrared spectra. The program is designed to present the chemist wit the most logical interpretation of a spectrum, not to interpret it definitively and pass the results directly to the structure generator, bypassing the chemist entirely. The chemist is an integral part of the decision making process and, for that reason, the program is interactive in nature.

When the task is to identify an unknown compound from its spectral data, and it is suspected that the unknown might be included in an accessible large library of spectra, a search and compare scheme is probably the best approach.[10,11] Pattern recognition procedures have enjoyed some success in predicting the functional groups that are present using infrared spectra.[12-14] A purely empirical approach for interpreting spectra has recently been described by Gray.[15] Like the method of Gray, the program described in this paper is completely empirical. It attempts to parallel the chemist's reasoning in interpreting an infrared spectrum as much as possible. The advantage in using an interpretation program such as this one in a structure elucidation package is evident when one considers that it is imperative that no information passed on to the structure generator can be incorrect. When an error is found, this program can be altered to correct the error, a capability that does not exist with pattern recognition programs.

General Approach

The chemist uses an empirical approach to interpret infrared spectra. After observing a sufficient number of spectra, or alternatively reading textbooks and learning from the observations of others, the chemist develops the ability to associate certain absorption peaks with their corresponding functional groups. In other words, a set of "rules" for interpreting infrared spectra is learned. For example, the "rules" for identifying a carboxylic acid might be to look for a broad, medium to strong peak centered around 3000 cm^{-1}, a strong peak near 1720 cm^{-1}, and a broad, medium intensity band around 920 cm^{-1}. It is this type of reasoning that must be programmed in order to develop a successful spectral interpreter. One of the first problems encountered is the need to digitize the spectra. It must be decided how much peak shape and intensity information should be retained. Many search systems and some pattern recognition procedures have been

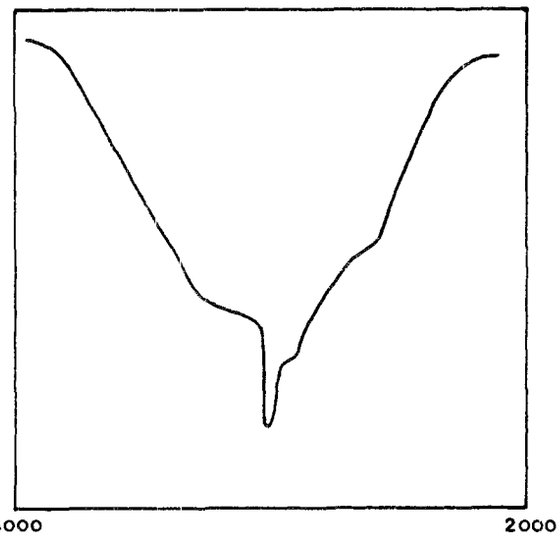


Figure 1. Partial spectra A and B and their peak/no peak representation.

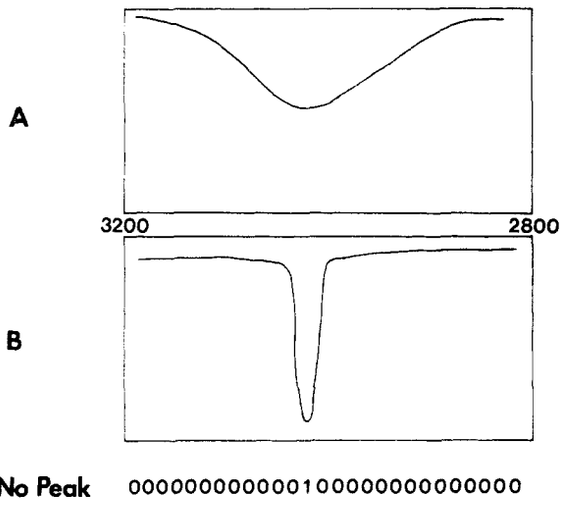


Figure 2. Portion of the spectrum of 4-methylpentanoic acid. Original-data from Silverstein and Bassler.[16]

successful while only employing peak positions and discarding all intensity and shape information. Figure 1 demonstrates that a chemist would not fare too well with only peak/no peak data. The region between 3200 and 2800 cm^{-1} is shown for two spectra labeled A and B. If these two spectra were digitized in a peak/no peak fashion, any interval that contained a peak maximum would be represented by a "1" and those intervals without a peak maximum would contain a "0". Thus, both A and B would be represented by the identical string of numbers, also shown in Figure 1, since their peak maxima coincide. Yet the chemist would have no difficulty distinguishing between the two nondigitized peaks.

A second approach to digitizing spectra is to break the spectrum into intervals (e.g., 0.1-μm or 10-cm^{-1} intervals) as in the peak/no peak approach, but then to record the percent transmission for each interval. This method is used by Gray[15] and results in a fairly accurate representation of the spectrum. However, it is not the approach a chemist would employ. Rather, a chemist notes the positions of peaks as well as their intensities and shapes. He does not break the spectrum into intervals. Thus, the approach used in this program is similar to the one reported by Penski, Padowski, and Bouck.[11] For each peak, a code number indicating intensity and shape information is recorded along with the peak position. The peak positions may be encoded in units of either cm^{-1} or μm. There are ten possible code numbers ranging from 0 to 9. Code number 0 indicates that the peak is a shoulder. Numbers 1, 2, and 3 are used for weak peaks; 4, 5, and 6 for medium peaks; and 7, 8, and 9 for strong peaks. Additionally, 1, 4, and 7 represent sharp peaks; 2, 5, and 8 are for peaks of average width; and 3, 6, and 9 are for broad peaks. With this system, peak A in Figure 1 would be encoded as 3000, 6, indicating that the peak position is 3000 cm^{-1} and it is a broad, medium intensity peak. Similarly, peak B would be encoded as 3000, 8, i.e., a strong peak at 3000 cm^{-1} and of average width. Of course, a chemist also relies very heavily on peak absence information when interpreting an infrared spectrum. As will be discussed later, this program utilizes peak absence information as well as peak presence information.

At this stage of development of the program it is conceivable that different chemists might encode a spectrum differently using the ten codes described above. For this reason some latitude has been provided in many of the subroutines. In the future, we plan to digitize spectra directly as they are recorded using an A/D converter. A preprocessor program would be written to reduce the digitized spectra to the codes required by the infrared interpreter. This approach requires a thoroughly tested set of rules to define "broad", "average", and "sharp".

In most cases, the method of digitizing spectra used for this program has been found to be quite satisfactory. However, there are still a few instances in which a chemist would have little difficulty associating an absorption pattern with a functional group, but the program does have difficulty. The value of the interactive nature of the program becomes apparent in those cases. As an example, a portion of a spectrum of 4-methylpentanoic acid is drawn in Figure 2. To the chemist, it would be fairly obvious that the absorption pattern shown results from a strong C–H stretch peak being superimposed on the broad O–H stretch peak characteristic of a carboxyl group. However, if asked to digitize the same absorption pattern, the result would probably be 3020, 0; 2920, 8; 2880, 0; 2650, 0. Since the "rules" state that a carboxylic acid should have a broad peak in the O–H stretching region, yet no broad peak is indicated by the digitized data, the program understandably has some difficulty. The carbonyl peak and the peak around 920 cm^{-1} due to out-of-plane bending of the bonded O–H are both present and indicative of a carboxyl group, but the O–H stretching region causes ambiguity. Rather than simply instructing the program to venture a guess, the interactive nature of the program is exploited. The user is confronted with the following question. "You input a strong and average peak between 3150 and 2850 cm^{-1}. Is this peak part of a broad envelope of peaks in that region?" Based on the user's response, a more intelligent decision can be made regarding the presence or absence of a carboxyl group.

Discussion

Part of what makes this interpretation program novel is that it is a very ambitious undertaking. Nearly all statistical pattern classification procedures involving infrared data attempt to discriminate between fewer than 15 functional groups. In his paper, Gray[15] does not indicate the number of classes for which his program tests. The information returned by the program described in this paper is categorized among the classes shown in Table I. Discrimination between 25 major classes is attempted, a number that is not too much larger than the number of classes investigated in some previous works.[12-14] However, this program attempts considerably more discrimination. Nearly all of the major classes are divided into several subclasses that deal with the environment of the functional group (substitution patterns, α,β unsaturation, etc.). Thus, counting both major classes and subclasses,

Table I. Major Classes and Their Associated Subclasses

1 ACETAL or KETAL
2 ACID, CARBOXYLIC
- a. saturated
- b. α,β-unsaturated
- c. α-electronegative group
- d. pyridinecarboxylic acid (or similar to one)

3 ACID HALIDE
4 ALCOHOL
- a. phenol
- b. primary
- c. secondary
- d. tertiary
- e. primary-α-unsatn or branch[a]
- f. secondary-α-unsatn or branch
- g. tertiary-α-unsatn or branch
- h. secondary-α,α'-unsatn and/or branch
- i. tertiary-α,α'-unsatn and/or branch
- j. tertiary-α,α',α''-unsatn and/or branch
- k. secondary contained in ring
- l. tertiary contained in ring

5 ALDEHYDE
- a. saturated
- b. α,β-unsaturated

6 AMIDE
- a. primary amide
- b. secondary amide
- c. tertiary amide
- d. lactam
  - (1) 4 members
  - (2) 5 members
  - (3) 6 or more members
- e. carbamate
  - (1) primary carbamate
  - (2) secondary carbamate
  - (3) tertiary carbamate

7 AMINE
- a. primary amine
- b. secondary amine
- c. tertiary amine
- d. aromatic amine (primary, secondary, or tertiary)
- e. $C{=}CNR_2$
- f. NH_3^+
- g. NH_2^+
- h. NH^+

8 ANHYDRIDE (of carboxylic acid)
- a. linear, saturated
- b. linear, α,β-unsaturated
- c. cyclic, 5 members
- d. cyclic, 6 members
- e. diacyl peroxide

9 AROMATIC
- a. thiophene
- b. furan
- c. pyrrole
- d. other aromatic
  - (1) monosubstituted benzene
  - (2) 1,2 disubstituted
  - (3) 1,3 disubstituted
  - (4) 1,4 disubstituted
  - (5) 1,2,3 trisubstituted
  - (6) 1,2,4 trisubstituted
  - (7) 1,3,5 trisubstituted
  - (8) 1,2,3,4 tetrasubstituted
  - (9) 1,2,3,5 tetrasubstituted
  - (10) 1,2,4,5 tetrasubstituted
  - (11) pentasubstituted
  - (12) α-naphthalene
  - (13) β-naphthalene
  - (14) pyridine
  - (15) pyrazine
  - (16) pyrimidine
  - (17) purine

10 AZO COMPOUNDS
11 C=C (nonaromatic)
- a. $CHR{=}CH_2$
- b. $CHR{=}CR_2$
- c. $CH_2{=}CR_2$
- d. *trans*-CHR=CHR
- e. *cis*-CHR=CHR

12 CARBOXYLATE ANION
- a. metal salt
- b. ammonium salt
- c. amino acid zwitterion

13 ESTER (of carboxylic acid)
- a. saturated[b,c]
- b. saturated acetate
- c. saturated formate
- d. linear with unsatn α to carbonyl
- e. benzoate
- f. linear[b] with unsatn α to oxygen
- g. acetate with unsatn α to oxygen
- h. benzoate with unsatn α to oxygen
- i. linear with α-unsatn on both sides
- j. 6 (or greater) member lactone
  - (1) saturated
  - (2) unsatn α to C=O
  - (3) unsatn α to O
- k. 5 member lactone
  - (1) saturated
  - (2) unsatn α to C=O
  - (3) unsatn α to O

14 ETHER
- a. saturated
- b. α,β-unsaturated
- c. epoxide

15 HYDROXYLAMINE
16 IMIDE
- a. linear
- b. cyclic, 5 member
- c. cyclic, 6 member
- d. cyclic, 5 member with α,β-unsatn
- e. cyclic, 6 member with α,β-unsatn

17 IMINE
18 KETONE
- a. saturated (inc. 6 member ring)
- b. PhCOR
- c. PhCOPh
- d. α,β-unsaturated
- e. α,β-γ,δ or α,β-α',β'-unsatd
- f. 5 member ring
- g. 4 member ring
- h. α-diketone
- i. quinone (1,2 or 1,4)
- j. chelate

19 MERCAPTAN
20 METHYL
- a. *gem*-dimethyl

21 NITRO or NITROSO COMPOUNDS
- a. nitrite
- b. nitrate
- c. nitramine
- d. nitro, saturated
- e. nitro, α,β-unsaturated
- f. nitro, aromatic
- g. nitro, α-electronegative group

22 OXIME
23 SULFUR-OXYGEN COMPOUNDS
- a. sulfinate
- b. sulfonate
- c. sulfonamide
- d. sulfone
- e. sulfonic acid
- f. sulfoxide

24 THIOCARBONYL COMPOUNDS
25 TRIPLE BONDS or CUMULATIVE DOUBLE BONDS
- a. internal acetylene
- b. terminal acetylene
- c. saturated nitrile
- d. α,β-unsaturated nitrile
- e. saturated isonitrile
- f. α,β-unsaturated isonitrile
- g. thiocyanate
- h. isocyanate
- i. isothiocyanate
- j. azide
- k. carbodiimide
- l. diazo compounds
- m. ketene
- n. allene

[a] There is either α,β unsaturation or the carbon atom α to the alcohol group is branched. [b] Except acetate. [c] Except formate.

decisions concerning 159 classes must be made. It is most important for the program to make correct decisions concerning the major classes. For example, the program must be able to identify correctly that the compound is aromatic. Should it also be able to determine that the compound is a monosubstituted benzene, that information is considered as a bonus.

Discrimination between all 159 classes in Table I using only infrared data is virtually impossible for many compounds. For example, the standard positions for the C–O stretching bands attributable to primary and secondary alcohols are 1050 and 1100 cm^{-1}, respectively. If the secondary alcohol is part of a ring, the C–O stretching band is lowered by 50 cm^{-1}. Thus, both primary alcohols and secondary alcohols in a ring should have a band near 1050 cm^{-1} and discrimination between them may not be possible, assuming that the molecular formula does not preclude the presence of a ring. Yet each type of alcohol is listed as a separate subclass in Table I (alcohol subclasses b and k). Remembering the purpose of this program, no inconsistency exists. The program is designed to assist the chemist by making him aware of the various possibilities. To this end, it is best if the program returns information to the chemist pertaining to as large a number of classes as possible, even if many times it is not possible to distinguish between some of the classes.

As stated earlier, the program attempts to simulate the reasoning employed by a chemist in interpreting an infrared spectrum. Unfortunately, a chemist has difficulty verbalizing the reasoning that he uses. When he looks at a spectrum, the chemist instantaneously makes decisions concerning the presence or absence of certain more obvious functional groups based upon the presence or absence of the patterns he has come to associate with those functional groups. However, the computer must be told in what order to consider the classes. If several chemists were asked to specify the order in which they consider the functional groups, each might give a different response. With certain exceptions, the order in which the classes are considered is probably not overly crucial. If one chemist chooses to consider acids first and aromatics second, while another chemist chooses the reverse order, both chemists should probably reach the same conclusions concerning the spectrum. However, it would be unwise to consider the ether class first, as the only prominent bands in ethers are caused by C–O stretching, a phenomenon that occurs in several other classes that are more readily identified (e.g., alcohols and esters). Thus, one should consider the ether class subsequent to esters and alcohols.

The program consists of 25 hard-coded subroutines, one for each of the major classes. Each functional group is considered in turn, and its presence or absence is determined by the patterns the program has been instructed to recognize. Some interdependence among the subroutines does exist. As an example, if the program determines that the compound is an alcohol, but does not contain a benzene ring, there is no point in considering the phenol subclass. Likewise, if the compound contains only two oxygen atoms and the program determines from the spectrum that the compound is definitely an ester, no further testing of oxygen containing functional groups is performed.

All routines are coded in FORTRAN IV. The program requires approximately 30K words of core on the Arizona State University Univac 1110 computer.

Results and Examples

The development and improvement of an artificial intelligence interpretation program follows a cyclic pathway. Following the initial writing of the program, it is tested using sample spectra. Based on any erroneous results, improvements are made in the program. The revised version is then tested, with additional changes resulting from any errors that are found. The cycle continues and never actually terminates with this type of program. As more samples are tested, new errors might appear, necessitating new improvements in the program. Once the program has been debugged, any changes may be considered as adding new "rules" to the program's collection (or possibly an exception to an existing "rule").

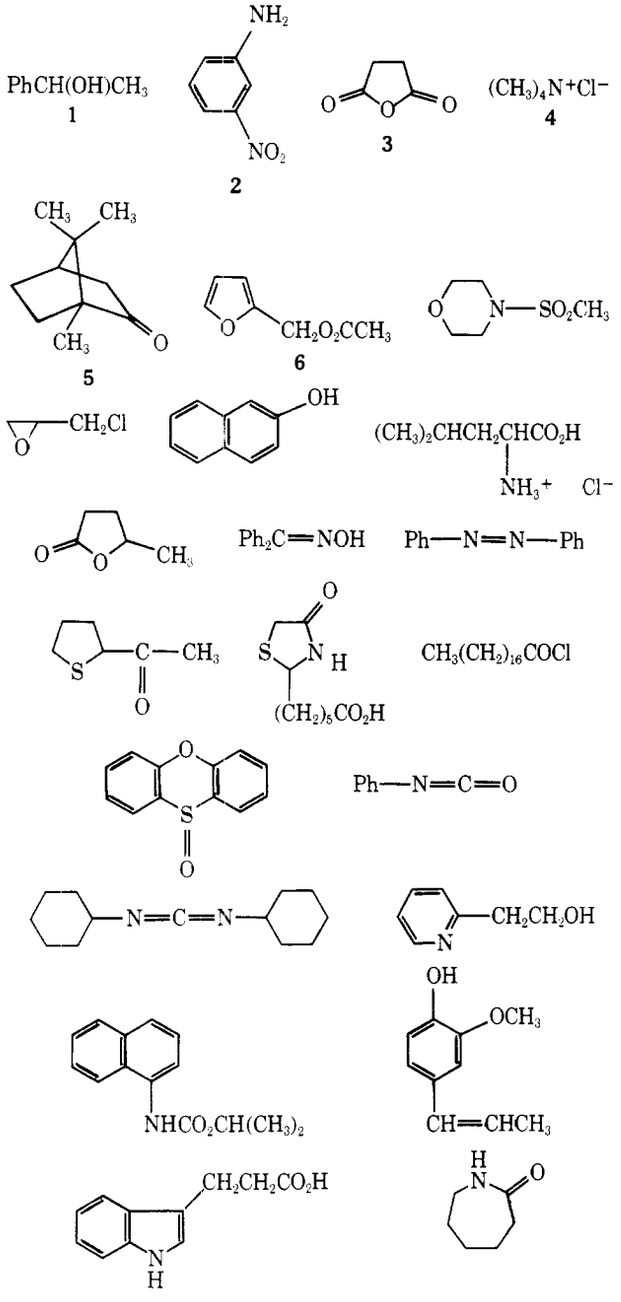
Two examples will serve to illustrate this cyclic development process. The original data are from Nakanishi,[17] and the results described are obtained using an earlier version of this program. Spectrum C includes a strong peak at 1710 cm^{-1} and two broad, medium intensity peaks, one at 3000 cm^{-1} and one at 935 cm^{-1}. Not surprisingly, the earlier version of the program found that compound C is a carboxylic acid, which it is (propionic acid). Spectrum D has a strong peak at 1703 cm^{-1} and a weak peak at 940 cm^{-1}. While these two peaks might possibly originate from a carboxylic group, especially an α,β-unsaturated carboxylic acid, no peak appears in the 3000-cm^{-1} region. Thus, the earlier version of the program understandably found compound D to be a nonacid, when in fact, D is 2-methylpyridine-5-carboxylic acid. This compound illustrates the need for an exception to the "rules" for categorizing carboxylic acids. Pyridinecarboxylic acids, unlike conventional dimeric acids, produce broad peaks near 2450 and 1900 cm^{-1} rather than the usual peak near 3000 cm^{-1}.[17] So the present version of the program considers this additional possibility, and correctly suggests that compound D might be a pyridinecarboxylic acid.

The program has been developed using 243 different spectra, originating from a variety of sources. All spectra for which their compounds are shown (containing C, H, O, N, S, halogen, or metal atoms) from the books by Silverstein and Bassler,[16] Nakanishi,[17] and Pasto and Johnson[18] are included in the test set. Three additional sources of spectra are the Sadtler collection,[19] Umezawa's index of antibiotics,[20] and an article by Hayden et al.[21]

It would be unrealistic to expect a program to select only the correct classes from the 159 possible classes every time. A chemist could not perform with absolute accuracy, and since the program parallels the reasoning of a chemist, it should be expected to perform nearly as well as the chemist, but no better. The most important function of the program is to suggest logical possibilities to the chemist, not to make the final decisions for him. With this function in mind, the program is designed to return one of five confidence values for each of the 159 classes, with selection of the confidence levels in any given example being based upon how well the "rules" for each class are obeyed. The values range from 0 to 4, and expressed in words, their meanings may be considered as: 0, definitely absent; 1, low probability of being present; 2, medium probability of being present; 3, high probability of being present; 4, definitely present. Obviously, the most grievous errors are cases in which the program returns 0 for a functional group that is present or 4 for a group that is absent.

Space limitations prevent discussion of all the examples. To aid in describing the achievements of this program, the 243 compounds are divided into four groups (A, B, C, and D). Group A, which consists of 177 compounds or 72.8% of the test set, contains those spectra for which no incorrect answer has a higher probability of being present than a correct answer. Thus, if three of the 25 major classes are present in the compound, the program returns confidence values for those three classes that are greater than or equal to the values for any of the other major classes. Likewise, with the subclasses associated with any major class that is present, the subclass that is correct has the highest (or tied for the highest) probability of all the remaining subclasses. A few brief examples from the test set should clarify the situation. Chart I shows a representative sample from group A. Structure 1 from that figure,

Chart I. Representative Set of Group A Compounds



1-phenylethanol, contains three major functional groups: aromatic, alcohol, and methyl. The program correctly identifies those three groups with confidence values of 4, 4, and 3, respectively. The output for all the remaining major classes equals 0. The program determines that four of the aromatic subclasses are plausible for this case, but the one with the largest confidence value is the monosubstituted benzene class. Finally, the program must consider the alcohol environment. Three alcohol subclasses have nonzero values, one of which is the secondary alcohol with α-unsaturation subclass. In this instance, all three alcohol subclasses have a value of 2; thus the correct answer is tied for the highest probability.

The complete output, including the confidence values, for five other compounds from Chart I is shown below. Major classes are indicated by capital letters.

2 *m*-Nitroaniline: AROMATIC, 4; NITRO or NITROSO COMPOUND, 4; aromatic nitro, 4; AMINE, 4; primary amine, 3; aromatic amine, 2

3 Succinic anhydride: ANHYDRIDE, 4; five-membered cyclic anhydride, 3

4 Tetramethylammonium chloride: METHYL, 2

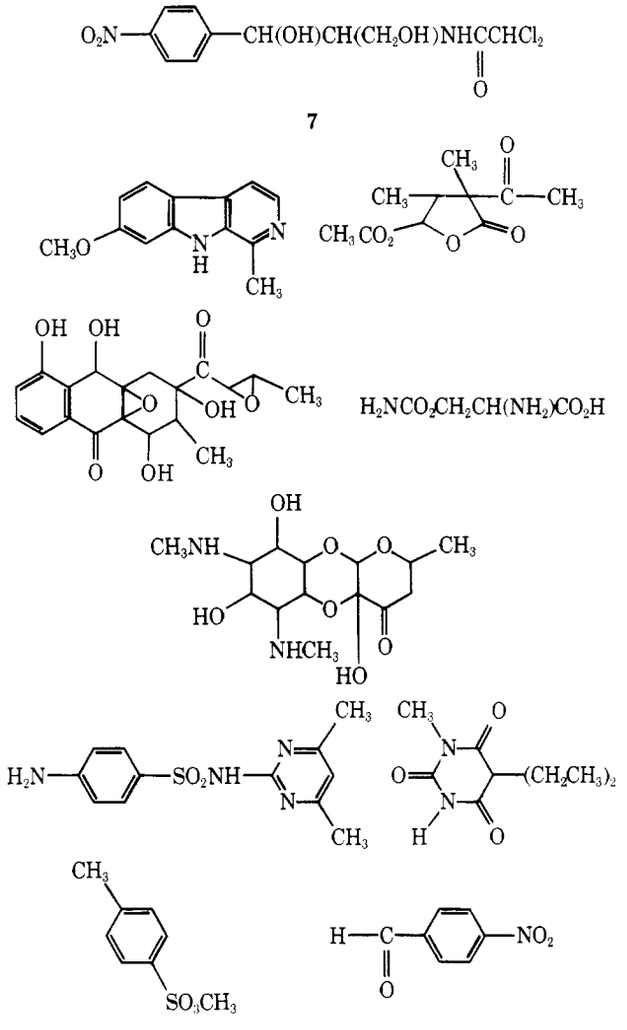
5 Camphor: METHYL, 3; *gem*-dimethyl, 3; KETONE, 3; five-membered ketone, 2; C=C (NONAROMATIC), 1; *cis*-CHR=CHR, 1; $CHR{=}CR_2$, 1

6 Furfuryl acetate: ESTER, 4; acetate, 4; METHYL, 3; AROMATIC, 2; furan, 2; ETHER, 2; unsaturated ether, 2; epoxide, 2; C=C (NONAROMATIC), 2; $CHR{=}CR_2$, 2; *cis*-CHR=CHR, 2; $CHR{=}CH_2$, 2; *trans*-CHR=CHR, 1; $CH_2{=}CR_2$, 1; KETONE, 1; unsaturated ketone, 1; five-membered ketone, 1

The spectrum of *m*-nitroaniline was run in chloroform solution; thus one cannot determine the benzene substitution pattern. Also, quaternary ammonium salts produce no characteristic bands in the infrared,[18] so the methyl class is the only correct response. Once again, all 177 members of group A have the correct major classes and subclasses returned with the highest probability or tied for the highest probability.

Several of the 35 members of group B are shown in Chart II. To qualify for group B, the correct major classes must once again be at least tied for the highest probability; however, at least one incorrect subclass has a higher confidence value than a correct subclass. For example, the spectrum of chloramphenicol (**7**) results in the following output for the major classes.

Chart II. Representative Set of Group B Compounds



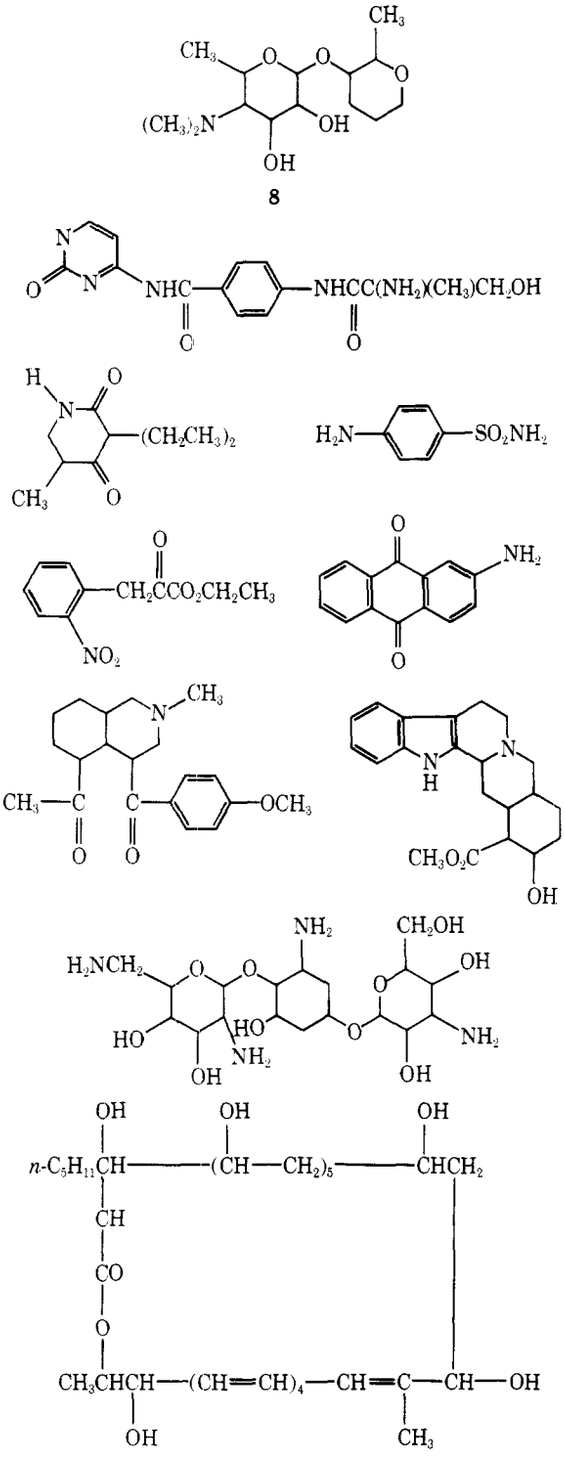
Amide, 4; nitro or nitroso, 4; aromatic, 3; alcohol, 3; nonaromatic C=C, 2; ether, 1; methyl, 1; ketone, 1

The four correct major classes have the largest confidence values. However, among the subclass output, the program finds a 1,2,4-trisubstituted benzene to be more likely than a 1,4-disubstituted benzene. So the 35 group B members have

some problem in the subclass output. Since it was already stated that correct subclass information is considered a bonus, the members of group B should certainly be deemed correct answers.

Group C contains the remaining compounds for which all correct classes have nonzero confidence values, and no incorrect class is claimed definitely to be present. For all 24 group C compounds, at least one incorrect major class has a greater confidence value than one or more of the correct major classes. Amicetin, structure 8 (Chart III), provides an example.

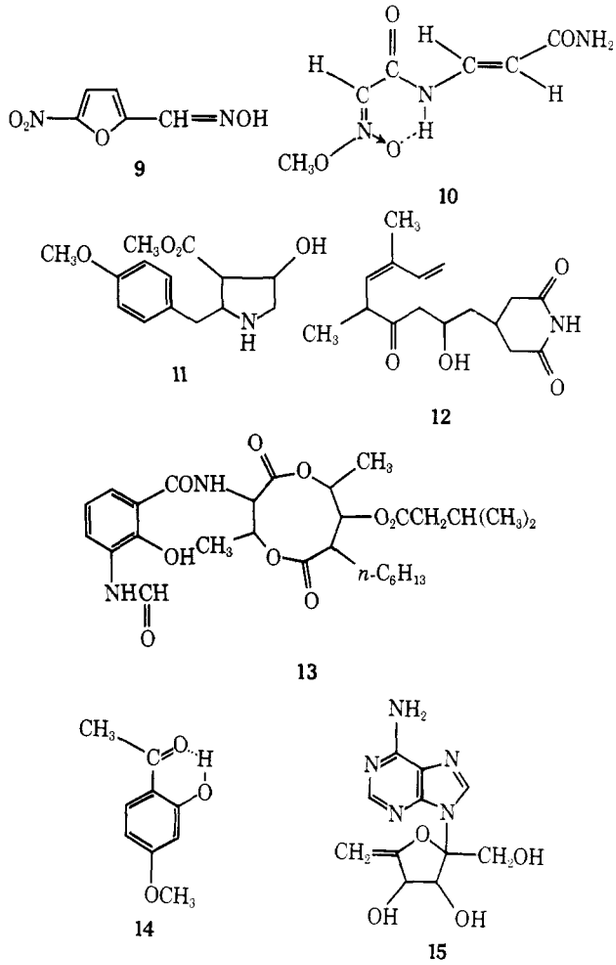
Chart III. Representative Set of Group C Compounds



The program finds that amicetin might be aromatic (confidence = 3), an amide (3), an amine (3), an ether (3), an alcohol (1), and an acetal or ketal (1), all of which are correct, but it also finds that other possibilities are a ketone (3), an azo compound (2), and a nitro or nitroso compound (3), all of which are incorrect and have a larger confidence value than two of the correct classes. So for group C, the correct classes are all nonzero, but not all the correct classes are the most probable.

Only 7 of the 243 test set members (2.9%) produce errors, either a response of 4 for a group that is not present in the molecule, or 0 for a group that is present. Since a major advantage of this type of programming is the capability of altering the program to eliminate the mistakes, it would be beneficial to examine each of the seven group D compounds more closely (Chart IV). Nifuroxime (**9**) is said to contain an

Chart IV. The Seven Members of Group D



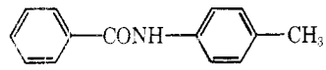
aromatic nitro (4), an oxime (3), and a furan (2), but the program also finds sufficient evidence to believe that an aliphatic nitro group is definitely present. It realizes one of the nitro responses is incorrect, but leaves it to the chemist to determine which one, a relatively easy task in this example.

Enteromycin carboxamide (**10**) again produces all the correct classes, *trans*-CHR=CHR and primary and secondary amide, but also results in a confidence of 4 for a primary amine. In the spectra of anisomycin (**11**) and streptimidone (**12**), the O–H stretching band of the alcohol is not evident. The band position and intensity are very likely altered by hydrogen bonding. Similarly, in the spectra of antimycin A_1 (**13**) and 2-hydroxy-4-methoxyacetophenone (**14**), the O–H stretch is not apparent. In the case of the chelate (**14**), the absence of the O–H band is expected, and since the program returns a confidence of 2 that the compound is in fact a chelate, the chemist should have little difficulty rectifying the situation. The carbonyl band due to the amide group in structure **13** is likewise not evident. Finally, the spectrum of

angustmycin A (**15**) has a strong peak at 1652 cm^{-1} (probably due to the C=C adjacent to an oxygen) and a shoulder at 1612 cm^{-1}. The program mistakenly identifies these peaks as the amide I and amide II bands of a secondary amide, and thus returns a value of 4 for the amide class. The program does inform the chemist that an unsaturated ether is a possibility, so the chemist is alerted to the correct functional group.

Of the seven errors, it is evident that the most common problem is the lack of the expected O–H stretching band in compounds with strong hydrogen bonding. Thus, the chemist would be wise to verify the alcohol results by some supplementary experimental evidence, such as an NMR spectrum.[22]

Finally, an example is presented for which the approach used by the program is described, without going into programming details. The compound is *N*-benzoyl-*p*-toluidine, and Nakanishi[17] supplies the original data.



As input, the program asks whether the spectrum was obtained from a solution (no), asks for the molecular formula ($C_{14}H_{13}ON$), and asks for the peak positions and intensity and shape code for each peak (3298 8, 3057 2, 2910 2, 2849 2, 1649 8, 1602 5, 1582 5, 1522 8, 1494 5, 1450 2, 1407 5, 1322 5, 1302 5, 1271 5, 815 7, 692 5). Since the formula indicates a sufficient amount of unsaturation (nine double bond equivalents), the program tests the aromatic class first. Similar to the approach a chemist would use, the program checks for peaks that it has been instructed to associate with aromatic groups. The peaks at 3057, 1602, 1494, 815, and 692 cm^{-1} are all at positions required by the "rules" for aromatic presence. The program finds sufficient evidence to return the highest confidence value for the aromatic class. The strong peak at 815 cm^{-1} suggests a para-substituted benzene as one possibility, and the program finds the three most probable substitution patterns to be para, 1,2,4-, and mono. Other possibilities are suggested, including pyridine. However, the final output will exclude pyridine, a fact that will be discussed later.

Since the program is satisfied that the compound contains an aromatic ring, four double bond equivalents are subtracted from the original total, three for the double bonds and one for the ring. Any time a confidence of four is found for a class, the appropriate atoms and double bond equivalents are subtracted from the formula.

Next the program checks for a possible carbonyl peak. Finding one, it checks several classes. The carboxylic acid, acid halide, acid anhydride, carboxylate anion, ester, and imide classes are all excluded from consideration owing to insufficient atoms of one type or another in the formula. After testing the aldehyde class "rules" and finding no evidence for an aldehyde, the program tests the amide class. The 3298, 1649, and 1522 cm^{-1} peaks all closely obey the "rules" for a secondary amide and the program determines that the compound is definitely an amide. Thus, one additional unsaturation site is accounted for as well as the oxygen and nitrogen atoms. No subsequent classes that require any atoms other than carbon or hydrogen are tested. The ensuing tests find evidence to suspect methyl group presence. In addition, following the first check of all the classes, any classes with nonzero confidence values are rechecked to ascertain whether sufficient atoms of the proper type still remain. During the rechecking phase, the program realizes that since no nitrogens remain unaccounted for, it is not possible for the compound to contain a pyridine group and the confidence value for that subclass is changed to zero.

This example has ignored many of the programming details and has greatly simplified the "rules" employed for the various classes. However, it does illustrate the approach used and that the program attempts to utilize all the information at its disposal in order to return only logical interpretation results to the chemist.

Conclusions

As stated earlier, the developmental stage of this type of interpretation program never actually reaches completion. If a chemist were able to do a better job of interpreting any of the test spectra than the program is able to do, then it should be possible to include whatever reasoning the chemist used in a new version of the program. This new version would then be expected to return the correct interpretation. This type of program can only be expected to approach the abilities of the chemist, and not to surpass them. However, it does perform its interpretation much more quickly than a chemist. It must be remembered that the goal of this work is only to develop a tool to aid the chemist and not to replace the chemist. Considering that the program attempts to test far more classes than has ever been attempted previously, obtaining an actual wrong response for only 2.9% of the test spectra should be evaluated as a definite fulfillment of that goal.

Obviously, infrared spectroscopy has some limitations in the field of structure elucidation. For this reason, the chemist performs additional experiments to aid in the decision making process. Thus, it would be unrealistic to expect an infrared spectral interpretation program to solve all problems. Sometimes information from other sources is needed. While this one program by itself would be a valuable asset to any structure elucidation scheme, information gained from mass spectral data and NMR data would enhance the scheme even more. To this end, this research group is in the process of developing similar interactive interpretation programs to analyze other types of spectra.

References and Notes

(1) Support by the National Institute of General Medical Sciences (NIH Grant GM 21703) is gratefully acknowledged.
(2) M. E. Munk, C. S. Sodano, R. L. McLean, and T. H. Haskell, *J. Am. Chem. Soc.*, **89**, 4158 (1967).
(3) D. B. Nelson, M. E. Munk, K. B. Gash, and D. L. Herald, Jr., *J. Org. Chem.*, **34**, 3800 (1969).
(4) B. D. Cox, Ph.D. Thesis, Department of Chemistry, Arizona State University, 1973.
(5) H. Abe and S. Sasaki, *Sci. Rep. Tohoku Imp. Univ., Ser. 1*, **55**, 63 (1972).
(6) L. A. Gribov, V. A. Dementyev, M. E. Elyashberg, and E. Z. Yakupov, *J. Mol. Struct.*, **22**, 161 (1974).
(7) H. Abe and P. C. Jurs, *Anal. Chem.*, **47**, 1829 (1975).
(8) R. E. Carhart, D. H. Smith, H. Brown, and C. Djerassi, *J. Am. Chem. Soc.*, **97**, 5755 (1975).
(9) C. A. Shelley and M. E. Munk, in preparation.
(10) D. H. Anderson and G. L. Covert, *Anal. Chem.*, **39**, 1288 (1967).
(11) E. C. Penski, D. A. Padowski, and J. B. Bouck, *Anal. Chem.*, **46**, 955 (1974).
(12) B. R. Kowalski, P. C. Jurs, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **41**, 1945 (1969).
(13) R. W. Liddell III and P. C. Jurs, *Appl. Spectrosc.*, **27**, 371 (1973).
(14) H. B. Woodruff, G. L. Ritter, S. R. Lowry, and T. L. Isenhour, *Appl. Spectrosc.*, **30**, 213 (1976).
(15) N. A. B. Gray, *Anal. Chem.*, **47**, 2426 (1975).
(16) R. M. Silverstein and G. C. Bassler, "Spectrometric Identification of Organic Compounds", 2nd ed, Wiley, New York, N.Y., 1967.
(17) K. Nakanishi, "Infrared Absorption Spectroscopy", Holden-Day, San Francisco, Calif., 1962.
(18) D. J. Pasto and C. R. Johnson, "Organic Structure Determination", Prentice-Hall, Englewood Cliffs, N.J., 1969.
(19) "Sadtler Standard Infrared Spectra", Sadtler Research Laboratories, Philadelphia, Pa.
(20) H. Umezawa, "Index of Antibiotics from Actinomycetes", University Park Press, State College, Pa., 1967.
(21) A. L. Hayden, O. R. Sammul, G. B. Selzer, and J. Carol, *J. Assoc. Off. Agric. Chem.*, **45**, 797 (1962).
(22) O. L. Chapman and R. W. King, *J. Am. Chem. Soc.*, **86**, 1256 (1964).