gramming language INTERLISP. Although we will provide copies and listings of the program to interested persons, this information will only be useful to a facility which maintains the INTERLISP language. The Stanford University Medical Experimental (SUMEX) computer facility has been established to encourage sharing of such programs among a collaborative community of users via a nationwide computer network. For additional information, write to the author, or to Professor Joshua Lederberg, Principal Investigator, SUMEX Project, Stanford University Medical School, Stanford, California 94305. Structures of isomers, including those presented here, with or without constraints, are available via this mechanism.

ACKNOWLEDGMENTS

I wish to thank the National Institutes of Health for their direct support of this work (RR 00612-05A1) and for their support of the SUMEX resource (RR 00785-02) which enabled the computations to be carried out. The contributions of my coworkers in developing the structure generator and its refinements, and the support of the SUMEX staff were invaluable.

REFERENCES AND NOTES

(1) Applications of Artificial Intelligence for Chemical Inference. XVIII. For part XVII, see ref. 10.
(2) Rouvray, D. H., *Chem. Brit.*, **10**, 11 (1974).
(3) (a) 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); (b) Buchanan, B. G., Duffield, A. M., and Robertson, A. V., "Mass Spectrometry: Techniques and Applications", G. W. A. Milne, Ed., Wiley, New York, N. Y., 1971, p 121.
(4) Sasaki, S., Kudo, Y., Ochiai, S., and Abe, H., *Mikrochim. Acta*, 726 (1971).
(5) Balaban, A. T., *Rev. Roum. Chim.*, **18**, 635 (1973).
(6) Polya, G., *Acta Math.*, **68**, 145 (1937).
(7) Brown, H., and Masinter, L. M., *Discrete Math.*, **8**, 227 (1974).
(8) Masinter, L. M., Sridharan, N. S., Lederberg, J., and Smith, D. H., *J. Am. Chem. Soc.*, **96**, 7702 (1974).
(9) Masinter, L. M., unpublished results.
(10) Carhart, R. E., Smith, D. H., Brown, H., and Djerassi, C., *J. Am. Chem. Soc.*, in press.
(11) Rouvray[2] uses the term "enumeration" in reference to both the constructive approach of Lederberg[3] and the isomer counting technique based on Polya's theorem.[6] The former technique yields the identity of each isomer, information we deem critical to chemists; the latter technique does not. Thus, we refer to structure generation, or construction, with reference to methods which yield the actual structures.
(12) This set of problems is restricted for several reasons. Selected isomer counts are presented as they are important to the discussion. Practicality dictates against other, larger problems; they would represent only extensions to the discussion. The program is not restricted to the problems presented. Structures from these problems and others are available (see Conclusions and Experimental).
(13) Disagreements between the first column (U = 0) of Table I and Table III of reference 3a arise from the fact that some undesired substructures (BADLIST[3]) were not considered in the latter tabulation.
(14) We use the terms primary, secondary, . . . , in reference to a broad definition of the word "degree". We consider for the purposes of this discussion that a tertiary carbon atom possesses three bonds to other, nonhydrogen atoms, independent of the multiplicity of these bonds. Although exception might be taken to this broad a definition, it is consistent with the structure generator's topological view of chemistry and saves having to introduce the textual complexities of differing hybridizations. We only note that the different hybridizations which, for example, a quaternary (by our definition) carbon may possess lead also to greater variety, which is responsible for there being more isomers of C_6U_7 (no hydrogens) than C_6U_0. Although the latter composition list can support carbon atoms with several degree values (e.g., 2,2-dimethylbutane has carbon atoms of degree 1, 2, and 4), they are all sp^3 carbons. The former composition list has only quaternary carbon atoms, but there are several different hybridizations possible.

A Text Search System Using Boolean Strategies for the Identification of Infrared Spectra†

H. B. WOODRUFF, S. R. LOWRY, and T. L. ISENHOUR*

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

Received August 11, 1975

A text searching algorithm, which operates on a tape based minicomputer, has been developed in this laboratory and applied to *Chemical Condensates*. This search is general in nature, and thus is applicable to any type of data set following an appropriate preprocessing step. This paper presents the application of general text searching concepts (including truncation and Boolean logic) to searching the ASTM file of 91,875 infrared spectra. A text-oriented search allows information supplemental to peak positions and intensities to be included in the search input. Some examples include the compound name or fragments thereof, molecular formula information, and chemical functionality information. In addition, the system is useful for studies other than the conventional technique of searching a library file for the closest match to an unknown spectrum. One can readily obtain a specific subset of spectra for further study by pattern recognition or statistical techniques using this system.

Recent work in this laboratory has resulted in the development of a rapid and efficient generalized minicomputer text searching system.[1] The system has been applied to *Chemical Condensates* of the Chemical Abstracts Service. Complete Boolean algebraic search strategy expressions may be used as direct entries and all forms of truncation are automatically processed. Currently the system is operating on 72 complex literature profiles developed by vari-

† Presented, in part, at the 26th Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy, Cleveland, Ohio, March 4, 1975.

ous members of the Chemistry Department at the University of North Carolina. The entire search system is independent of the data base used; only the preprocessing program which prepares the initial search tape (library tape) is peculiar to the Chemical Abstracts data base. Therefore, this system can be applied to other literature bases which may be beneficially searched in this fashion.

This paper reports an investigation designed to determine the utility of the text searching algorithm for searching an appropriately preprocessed spectral data base. In particular, the data base employed is the collection of 91,875 infrared spectra compiled by the American Society for Testing and Materials (ASTM). The text-oriented search is slower than many of the systems previously reported for infrared data, but has a number of unique capabilities. These capabilities become especially apparent when information from other sources (e.g., mass spectra or NMR spectra) is included.

BACKGROUND

Sorting punched cards with an electric sorting machine was the first technique used to search rapidly a file of infrared spectra.[2,3] Now that the ASTM file has increased to over 100,000 entries, card sorting is highly inefficient. A number of computerized techniques have been reported for searching the file with a magnetic tape or disk input.[4-9] In an effort to decrease search times, principles such as spectral compression,[10-12] file inversion,[13] and hash coding[14] have been used.

The ASTM file is coded in a binary format. The only absorption information retained is whether or not a peak maximum occurs in a given 0.1-μm interval. This results in a considerable savings of storage space over compilations retaining intensity information. However, Penski, Padowski, and Bouck[15] contend that unless additional information is used in conjunction with the peak presence or absence information of the ASTM file, a search will result in an excessive number of matches. As an alternative, they advocate a search which utilizes both peak intensity and peak shape information. Inclusion of this type of information, however, may greatly increase the size of the data base and the corresponding search times. Several other workers have reported retrieval systems using peak intensity information.[16-20] Finally algorithms designed to retrieve the components of an unknown mixture by means of a search have been presented.[21-22]

Penski et al. admit that the ASTM file supplies considerable chemical data for each spectrum, thereby providing a means of further discrimination. However, they find that the coding techniques for these chemical data are difficult to adapt to standard searching methods.[15] It is because of this difficulty that the nonstandard approach of a text-oriented search is useful. The coding and searching of information supplemental to peak positions is not at all difficult with the system presented in this paper.

GENERATION OF THE LIBRARY FILE

The ASTM infrared file is coded in a peak/no peak format, with each encoded value corresponding to a 0.1-μm interval.[23] In addition, there is an entry which corresponds to the presence or absence of a "strong" peak within a particular 1.0-μm interval. Melting point or boiling point values are available for some of the spectra. Each entry also contains a serial number and an abundance of chemical classification data.

The ASTM file was purchased by the R. J. Reynolds Tobacco Company and is made accessible on disk at the Triangle Universities Computation Center (TUCC) in North Carolina. Also included on this disk are the name and a portion of the molecular formula (the number of C, H, N, O, S, F, Cl, Br, I, and Si atoms) for each compound.

```
166CA*** CCCCCC O ***      CYCLOHEXANONE*** 5
*** 3.4 3.5 5.8 6.9 7.0 7.4 7.6 8.2 9.0 9.5 9.8 11.0
11.6 13.3 *** 1 26 35 50 62 74 124
```

Figure 1.

Processing the data available on the TUCC disk to make it compatible with the existing minicomputer search system required a fairly complex PL/I program. The available data are in two different files (spectral file and name file) and are written as bit strings (peak positions, "strong" peak positions, and chemical classification numbers), character strings (name and serial number), and numbers (molecular formula information). All of this information was converted into one character string for each spectrum. The resulting character strings were written on magnetic tape, and, to be compatible with the existing minicomputer, an EBCDIC to ASCII conversion was required.

Each character string is subdivided into six different keystring types. Preceding each keystring are three bytes of pointer information, which include the type and length of the subsequent keystring.

The six keystring types are listed below along with a brief description. The descriptions refer to a sample character string shown in Figure 1. Note that the three bytes of pointer information preceding each keystring are denoted by "***" in Figure 1.

1. **Serial Number.** The serial number consists of ten characters as given by the ASTM file. The ninth and tenth characters are letters, which indicate the source of the spectrum and the spectral region, respectively. In Figure 1, "C" indicates the source is the Sadtler Research Laboratories and "A" indicates the spectral region is the infrared (as opposed to the far-infrared, which is indicated by a "G").

2. **Partial Molecular Formula.** The number of C, N, O, S, F, Cl, Br, I, and Si atoms present in the molecule are indicated here. The notation of "CCCCCC" for six carbon atoms appears cumbersome, but the advantage realized with the truncation capability of the search algorithm, which will be discussed later, is substantial.

3. **Compound Name.** The compound name is recorded exactly as found on the TUCC disk. As a result, often there is considerable wasted space due to excess blanks. The example in Figure 1 has six unnecessary blanks preceding the name cyclohexanone.

4. **Intervals of "Strong" Absorption.** The utility of this information is hampered by the fact that the rules for deciding what constitutes "strong" absorption are somewhat nebulous. Cyclohexanone has a "strong" peak in the 5–6-μm region.

5. **Peak Positions.** The one-tenth micrometer intervals containing peaks are indicated by this keystring.

6. **Chemical Classification Data.** These numbers range from 1 to 312, corresponding to the 312 possible classifications represented by columns 32 to 57 on the Wyandotte–ASTM punched cards.[23] In Figure 1, the 1 means cyclohexanone contains oxygen, the 124 means it is a ketone, etc.

The conversion program was run on the TUCC IBM 370/165 teleprocessing with the University of North Carolina Computation Center IBM 360/75. All 91,875 spectra require approximately 2200 feet of magnetic tape written in an 800 bpi density.

GENERATION OF A PROFILE

In order to search the library, the user must generate a profile, a sample of which is shown in Figure 2, containing specific information about the unknown. Each profile is initiated by an operator-assigned job number; thus many

```
/ SAMPLE PROFILE #1
PP,A=5.9
PP,B=6.2
PP,C=6.6
PP,D1=7.0
PP,D2=7.1
D=D1:D2
PP,E1=7.3
PP,E2=7.4
E=E1:E2
PP,F=7.9
PP,G=8.1
PP,H1=8.6
PP,H2=8.7
H=H1:H2
PP,I=10.4
PP,J1=11.8
PP,J2=11.9
PP,J3=12.0
J=J1:J2:J3
PP,K=4.*
IS,L=27
IS,M=97
AU,N=*O
AU,P=CCCCCCCCCCC*
**=A&B&C&D&E&F&G&H&I&J&-K&L&M&N&-P
```

Figure 2.

profiles can be searched with one pass through the library tape. The user supplies three types of profile cards: comment cards, keystring cards, and predicate definition cards.

Comment cards begin with "/" and allow any extent of comments that the user desires. The keystring definition cards consist of three components. The first two characters on the card are the keystring category identifier. (Since this investigation is designed to utilize a previously existing text search system, the six identifiers listed in Table I are more applicable to the *Chemical Condensates* search. While CN is a good mnemonic for *Chemical Abstracts* number and AU for author, one might be more inclined to select SN for serial number and FO for the molecular formula keystring in a search designed exclusively for spectral data.) The identifier characters are followed by a comma and a one- or two-character variable. This variable is assigned by the user and allows reference to the keystring definition in the subsequent Boolean search logic expressions. The third component is the keystring definition itself. This follows the equal sign and is a text fragment that is to be detected in the text search. It consists of from 2 to 795 alphanumeric characters (including blanks).

From Figure 2, the first keystring definition card is

$$\text{PP,A=5.9.} \qquad (1)$$

The PP is the identifier for peak position; the A is the user assigned variable; and the 5.9 is the keystring definition. Thus, the search will be looking for spectra with a peak at 5.9 μm. Similarly, IS,L=27 indicates the search will be checking whether classification number 27 is present. (Classification number 27 means the compound is aromatic and number 97, which also appears in the sample profile, indicates methyl group presence.)

The presence of an asterisk in a keystring definition has special significance and indicates truncation. A keystring definition beginning with an asterisk may have any prefix (including a blank) and still be recorded as a match. Likewise, a keystring definition ending with an asterisk may have any suffix. A non-left-truncated keystring (*i.e.*, no asterisk at the beginning) must be preceded by a blank when encountered in text search in order to be recognized. A typical keystring definition statement is shown below just as it might appear in a user profile.

$$\text{PU,AA=*CYCLOHEXANONE*} \qquad (2)$$

Table I. Identifiers for the Six Different Keystring Types

Identifier	Keystring type
CN	SERIAL #
AU	MOLECULAR FORMULA
PU	COMPOUND NAME
KT	"STRONG" PEAKS
PP	PEAK POSITIONS
IS	CLASSIFICATION DATA

The keystring category identifier (PU) indicates that the keystring definition is a compound name. (The actual keystring definition is both left- and right-truncated.) Therefore, all three of the compound names listed below would be recorded as hits.

CYCLOHEXANONE

HYDRAZONE, 2,4-DINITROPHENYL-, CYCLOHEXANONE (3)

CYCLOHEXANONE OXIME

In addition, if CYCLOHEXANONE were in any other environment, it would be detected.

In the example in Figure 2, three keystring definitions include truncation.

$$\begin{aligned}&\text{PP,K=4.*}\\&\text{AU,N=*O}\\&\text{AU,P=CCCCCCCCCCC*}\end{aligned} \qquad (4)$$

User variable K is set to "true" any time a peak is present between 4.0 and 4.9 μm, inclusive. Variable N is set to "true" whenever one or more O is present in the molecular formula keystring, and P is "true" when eleven or more C's are present.

Finally, the user supplies predicate definition cards. Predicate definition cards contain two components. The first is a one- or two-character logic variable which is assigned by the user, and to which is assigned the result of a Boolean algebra expression. The second component is the Boolean algebra expression, consisting of previously defined variables connected with the usual Boolean logic connectives: "and" (&), "or" (:), and "not" (-). In addition, parentheses may be used freely to denote the desired sequence of expression evaluation. In the absence of parentheses, evaluation proceeds from left to right. The predicate definition cards may be interspersed throughout the profile, provided all variables in the Boolean expression are previously defined. E=E1:E2 is a predicate definition statement from Figure 2. Predicate definition cards within the profile are optional. Only the terminal predicate (indicated by the symbol "**" in columns 1 and 2) is required. If and only if this Boolean expression evaluates as "true" would the spectrum (citation) under consideration be listed as a hit.

SEARCH HARDWARE

The minicomputer and peripherals employed in this search system have been in use in this research group for over three years. Thus, the search algorithm is designed to use existing equipment, rather than the equipment being purchased exclusively to perform searches. The system includes a 64K-byte, 1.0-μsec cycle time, Raytheon 704 computer, equipped with two Peripheral Equipment Corp. 800 bpi, 25 ips, IBM compatible magnetic tape drives, a 500-cpm card reader, and a 300-lpm line printer. The total equipment investment is about $33,000.

SEARCH ALGORITHM

The system requires use of the library tape created on the TUCC computer and three programs developed on the

```
     19654CA*** CCCCCCCC F 0 ***      ACETOPHENONE,
4PR-FLUORO-*** 5 6 7 8 11 *** 5.9 6.2 6.6 7.1 7.3 7.7
 7.9 8.1 8.6 9.1 9.8 10.4 11.9 12.2 *** 1 4 25 27 39
51 62 65 74 97 124 143

     17617EA*** CCCCCCCC F 0 ***      ACETOPHENONE,
P-FLUORO-*** 5 7 8 11 *** 5.9 6.2 6.6 7.1 7.3 7.9 8.1
 8.6 9.1 9.9 10.4 11.9 12.2 *** 1 4 25 27 39 51 62 65
 74 97 124 143
```

Figure 3.

Raytheon 704 for the *Chemical Condensates* search.[1] The purpose of the programs is as follows:

1. WARPSET. This program compiles the input data (the search profiles) into a set designed for sequential and canonical searching of the data base.
2. WARP-8. This program is a multiprofile search. The program makes all matches, solves Boolean logic expressions defining search logic, and outputs spectral hits with labels according to the search profiles scoring the hits.
3. SORPRINT. This program sorts the output tape in the order of profile numbers and prints the individual outputs.

A more detailed description of the searching algorithm is available in reference 1.

RESULTS AND DISCUSSION

As with any search involving infrared data, one is forced to decide between a very specific profile which might result in missing the desired compound, and an overly vague profile which results in far too many matches. Usually, a compromise must be made. The truncation capabilities available in this search system aid in achieving this compromise. If one is certain of only a small portion of the molecular formula or the name, he can easily include this information. If one knows the molecular weight of a compound is 118, but has no molecular formula information, he can still include the following term in his profile.

```
AU,A=CCCCCCCCCC*
      .
      .
      .
**=-A&(···)
```

This profile would only result in hits for compounds containing fewer than ten carbon atoms, a discrimination step that could substantially reduce the number of hits.

Several examples are included in this paper to demonstrate some of the capabilities of the text search system. The first example uses the profile already presented in Figure 2. This profile is created following a fairly cursory examination of one of the unknown sets of spectra (IR, MS, NMR) from Silverstein and Bassler's book (Compound 8-10).[24] Selection of the infrared peak positions is straightforward, with the results being shown in all the PP terms in Figure 2. Those peaks for which uncertainty exists have two or three peak positions connected by "or" statements (e.g., D1:D2). No peaks appear between 4.0 and 5.0 μm, so a "not" 4.* operation is included. Several other obvious facts are included in the profile. There is a frequent occurrence of a loss of 15 mass units from the parent peak in the mass spectrum. In addition, the NMR spectrum has a peak at 2.5 ppm. Hence, a methyl group should be on the molecule (classification #97). The NMR output also demonstrates the presence of a phenyl group (classification #27 for aromatics). The obvious carbonyl peak in the IR spectrum requires that at least one oxygen atom be in the molecular formula (AU,N=*O). Finally, the molecular weight of 138 means that, since at least one oxygen is present, there can be no more than ten carbon atoms.

When the library file is searched for entries which fulfill all these requirements, only two hits are recorded—both the correct compound. The output is shown in Figure 3. While it is quite satisfying to be this successful, one can seldom specify the infrared peak positions with the accuracy found in the sample profile. Most infrared search systems reported previously incorporate a "wiggle" option. With this option, an entry of 3.4 μm for a peak is automatically wiggled to include 3.3 and 3.5 μm. Thus, any library entries having a peak at any of these three positions will result in a favorable response to this particular query. The "or" capability of the text search allows such a wiggle to be employed, if desired, but it has the added advantage of being more specific. With most existing search systems, the wiggle option requires that either all or none of the peaks be wiggled. With the text search, not all peaks need to be wiggled at the same time or to the same degree.

A clearer example of the discrimination capabilities of a text search follows. The ASTM data set at TUCC has been used to implement the "Infrared Spectral Information Service" (ISIS).[25] An example in the ISIS manual is a search for cyclohexanone. Their input includes the following information:

Peaks present: 3.4, 5.8, 6.9, 7.6, 8.2, 8.9 μm
Intervals with no peaks: 3.7–5.6, 6.1–6.7, 11.8–13.1 μm (6)

A wiggle option is included. The example has a cutoff limit of 50 hits. When that limit is removed, the result of their search is a listing of 301 serial numbers. If a profile is developed which includes exactly the same information and a text search is performed, the result is the same 301 hits. Of the 301 hits, 10 are cyclohexanone. In fact, the data set includes 16 cyclohexanone entries. Clearly, this is a case in which the profile is too vague, as far too many hits are recorded to be of any value. Further discrimination is necessary.

The procedure used in the ISIS manual is to specify an increased number of intervals containing no peaks. The peaks present are the same as before, but the intervals with no peaks become:

3.7–5.6, 6.0–6.7, 8.4–8.7, 10.1–10.8, 12.0–14.9 μm (7)

Now the result of their search is a listing of 16 serial numbers, only one of which is cyclohexanone.

As mentioned previously, the text searching concept is especially useful when data other than strictly the IR spectrum are available. With the cyclohexanone unknown, suppose a low resolution mass spectrum is obtained. Even if the only supplemental information is the molecular weight, discrimination is greatly enhanced. To the profile which generated 301 hits, the following information is added: (1) an obvious carbonyl peak in the IR means at least one oxygen is present; and (2) the molecular weight is such that with at least one oxygen present, there can be no more than six carbon atoms. The following statements are added to the profile containing peak presence and peak absence information.

```
AU,E6=*O
AU,E7=CCCCCCC*
      .
      .
      .
**=···&E6&-E7
```
(8)

The result of a text search using this profile is 32 hits, but

Table II. Some Sample Search Speeds

No. of profiles	Search time (min)	No. of spectra/ profile/min
1	18.7	4900
2	20.9	8800
5	21.3	22000
10	28.8	32000
50	117	39000

10 of them are cyclohexanone. Thus, about 1 out of every 3 hits is the correct answer.

A text search system for spectral data has uses other than just the conventional search and compare of an unknown to a library. One can easily obtain a highly selective subset of the data file for subsequent investigation by pattern recognition or statistical techniques. In addition, work is in progress to try to develop systematic approaches to using a text search system to find data set inconsistencies. As an example, the following profile results in hits for all entries containing the word cyclohexanone in the compound name keystring.

/PROFILE TO SELECT ONLY THE CYCLOHEXANONE ENTRIES

PU,A=CYCLOHEXANONE (9)

**=A

Sixteen of the hits are entries for the compound cyclohexanone. It is evident from the 16 cyclohexanone hits that no data compilation is likely to be perfect. The number of peaks encoded ranges from a low of 7 to a high of 27. Also, errors exist in the chemical classification data, as one entry does not record cyclohexanone as containing a six-member ring, and a second entry does not include the ketone classification number.

The presence of the compound name is both beneficial and detrimental. Including the name causes search times to be 15 to 20% longer than they would otherwise be. However, having a printout returned as in Figures 1 and 3 in which the name is present is much better than simply receiving a list of serial numbers which must subsequently be looked up, as is the case with most other search systems. Also, one can generate useful profiles using just the compound name or a fragment thereof, as is demonstrated in the previous example.

Finally, some search speed information is shown in Table II. It is evident that the ability to stack profiles (up to a maximum of 224) for simultaneous searching during a single pass through the library tape is extremely beneficial. On a basis of spectra searched per profile in one minute, the stacking of 50 different profiles increases the search speed by nearly a factor of 8 over a run using a single profile.

As previously mentioned, the minicomputer system employed in this investigation was not purchased specifically for searching. An optimized system, which would cost about $34,000 and would increase search speeds by at least a factor of 5, is described in ref 1. Using this system, a search of all 91,875 spectra for a single profile would require about 3 min.

CONCLUSIONS

The feasibility of using a text search system on a spectral data set has been demonstrated. For problems requiring the matching of an unknown to a library using only the positions of infrared absorption, this system is the equal of other reported search systems in all aspects but one, speed. A text search is superior to other systems when additional chemical information is available. The ease of encoding functional group information, the presence of molecular formula and name information, the relatively unique "or" capabilities, and the truncation capabilities, all contribute to this superiority over conventional infrared search systems. Finally, the diversity of the system, which enables investigations other than simple search and compare schemes to be performed and allows specific subsets of the data set to be selected for future study, is beneficial and unique.

ACKNOWLEDGMENT

The authors are indebted to W. S. Woodward for many helpful discussions. The financial support of the National Science Foundation is gratefully acknowledged. T. L. Isenhour is an Alfred P. Sloan Research Fellow, 1971–1975.

LITERATURE CITED

(1) Isenhour, T. L., Woodward, W. S., and Lowry, S. R., "A Rapid Minicomputer Text Search System Incorporating Algebraic Entry of Boolean Strategies", *J. Chem. Inf. Comput. Sci.*, **15**, 115–9 (1975).

(2) Kuentzel, L. E., "New Codes for Hollerith-Type Punched Cards", *Anal. Chem.*, **23**, 1413–8 (1951).

(3) Baker, A. W., Wright, N., and Opler, A., "Automatic Infrared Punched-Card Identification of Mixtures", *Anal. Chem.*, **25**, 1457–60 (1953).

(4) Sparks, R. A., "Storage and Retrieval of Wyandotte-ASTM Infrared Spectral Data Using an IBM 1401 Computer", ASTM, Philadelphia, Pa., 1964.

(5) Smithson, L. D., Fall, L. B., Pitts, F. D., and Bauer, F. W., "Storage and Retrieval of Wyandotte-ASTM Infrared Spectral Data Using a 7090 Computer", Tech. Doc. Rept. No. RTD-TDR-63-4265, Research and Technology Division, Wright-Patterson Air Force Base, Ohio, 1964.

(6) Entzminger, T. A., and Diephaus, E. A., "Storage and Retrieval of Wyandotte-ASTM Infrared Spectral Data Using a Honeywell 400 Computer", U.S. Public Health Service, Robert A. Taft Sanitary Engineering Center, Cincinnati, Ohio, 1964.

(7) Anderson, D. H., and Covert, G. L., "Computer Search System for Retrieval of Infrared Data", *Anal. Chem.*, **39**, 1288–93 (1967).

(8) Cross, L. H., Haw, J., and Shields, D. J., in "Molecular Spectroscopy, Proceedings of a Conference Held at Brighton, England, 17–19 April 1968", P. Hepple, Ed., The Institute of Petroleum, London, 1968, p 189.

(9) Massios, G. A., "Infrared Spectral Data Retrieval by Computer", *Am. Lab.*, **3**, 55–62 (Sept 1971).

(10) Erley, D. S., "Fast Searching for the ASTM Infrared Data File", *Anal. Chem.*, **40**, 894–8 (1968).

(11) Lytle, F. E., and Brazie, T. L., "Effects of Data Compression on Computer Searchable Files", *Anal. Chem.*, **42**, 1532–5 (1970).

(12) Erley, D. S., "A Quantitative Evaluation of Several Infrared Searching Systems", *Appl. Spectrosc.*, **25**, 200–2 (1971).

(13) Lytle, F. E., "Computerized Searching of Inverted Files", *Anal. Chem.*, **42**, 355–7 (1970).

(14) Jurs, P. C., "Near Optimum Computer Searching of Information Files Using Hash Coding", *Anal. Chem.*, **43**, 364–7 (1971).

(15) Penski, E. C., Padowski, D. A., and Bouck, J. B., "Computer Storage and Search System for Infrared Spectra Including Peak Width and Intensity", *Anal. Chem.*, **46**, 955–7 (1974).

(16) Drobyshev, Yu. P., Nigmatullin, R. S., Lobanov, V. I., Korobeinicheva, I. K., Bochkarev, V. S., and Koptyug, V. A., "Use of Computers to Identify Chemical Compounds According to Spectra Characteristics (Retrieval System of the IR Spectra of Compounds)", *Vestn. Akad. Nauk SSSR*, **40** (8), 75–83 (1970); *Chem. Abstr.*, **74**, 69g (1971).

(17) "Computer Systems for Identification of Chemical Compounds from Their Spectral Characteristics. II. Computer Search System for Retrieval of Sadtler Standard Infrared Spectra", *Izv. Sib. Otd. Akad. Nauk SSSR, Ser. Khim. Nauk*, 108–14 (1972); *Chem. Abstr.*, **77**, 96627y (1972).

(18) Rann, C. S., "Automatic Sorting of Infrared Spectra", *Anal. Chem.*, **44**, 1669–72 (1972).

(19) Schaarschmidt, K., Reimer, R., and Steger, E., "Machine Oriented Searching System for Infrared Spectra", *Z. Chem.*, **14**, 374–5 (1974).

(20) Tanabe, K., and Saëki, S., "Computer Retrieval of Infrared Spectra by a Correlation Coefficient Method", *Anal. Chem.*, **47**, 118–22 (1975).

(21) Sebesta, R. W., and Johnson, G. G., "New Computerized Infrared Substance Identification System", *Anal. Chem.*, **44**, 260–5 (1972).

(22) Isenhour, T. L., "Rapid Memory-Conserving, Compiler-Level Search Algorithm for Mixture Spectra", *Anal. Chem.*, **45**, 2153–4 (1973).
(23) "Codes and Instructions for Wyandotte-ASTM Punched Cards, Indexing Spectral Absorption Data", American Society for Testing and Materials, Philadelphia, Pa., 1964.
(24) Silverstein, R. M., and Bassler, G. C., , "Spectrometric Identification of Organic Compounds", 2nd ed, Wiley, New York, N.Y., 1967, p 237.
(25) Denk, J. R., and Gunn, J., "ISIS-Infrared Spectral Information System—User's Manual", Triangle Universities Computation Center Document No. LSR-98, Research Triangle Park, N.C., 1970.

A Feature Selection Technique for Binary Infrared Spectra

S. R. LOWRY and T. L. ISENHOUR*

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

Received June 18, 1975

Feature selection is frequently employed in the computer classification of spectral data to overcome storage limitations and to decrease computation times during the development of discriminant functions. This paper describes a method of choosing a subset of the important features in binary data based on the a posteriori probabilities as determined by Bayes rule. When this technique was applied to 2600 infrared spectra taken from the ASTM file, the number of 0.1-μm intervals was reduced from 139 to 32 with an overall loss of about 2% in classification ability. The simplicity of this feature selection technique and its success with binary infrared spectra demonstrate that it should be considered in situations involving binary data.

INTRODUCTION

In computer classification of chemical infrared spectral data, a training set of spectra whose classifications are known is frequently employed to empirically develop discriminant functions.[1-4] These discriminant functions can then be applied to an unknown spectrum to give information concerning molecular structure.

Because of the limited memory size of computers and the large amounts of time often required to develop discriminant functions, rarely are all possible spectra incorporated into the training set. In order to overcome storage limitations and to decrease computation times some form of information compression is needed. One method is to optimally pack each spectra as a series of bits, each representing a peak maximum in the infrared spectra. An alternate method is to retain only the information necessary to differentiate among classes. The number of peaks important in a particular classification problem is normally much smaller than the total number of peaks present in an infrared spectrum. If some method is used to select only the important peaks, the "size" of the stored spectra can be greatly reduced. While sophisticated feature selection techniques have been reported in the pattern recognition literature[5,6] and several have been applied to chemical data containing intensity information,[7,8] very little has been reported concerning feature selection of binary data (peak–no peak).[9] This problem is quite significant when one tries to develop discriminant functions using the ASTM file of 91,875 infrared spectra. In this file, each spectrum is stored as a string of ones and zeros. A one indicates that a peak maximum is present in the corresponding wavelength interval of the spectrum. If the number of intervals (features) can be reduced from 100 to 25, the size of the training set could be increased by a factor of 4. Correspondingly, the number of arithmetic operations occurring in the computations could be reduced by as much as a factor of 4, thereby greatly increasing the speed of calculating a discriminant function. However, if considerable interclass information is lost in the feature selection process, the reduced spectra will be less useful as a training set for any type of pattern recognition technique.

To select a set of features that retain most of the discriminating information, we have taken a training set of binary infrared spectra and progressively deleted those wavelength intervals from each spectra which contain the least interclass information. By comparing the classifying ability of discriminant functions developed on the reduced spectra, we obtain a measure of goodness for the particular reduced set of features. This paper describes a method to evaluate the usefulness of each feature in classifying unknown compounds and to assign a relative order of goodness to the features. By selecting the best 32 features out of a total of 139 features, the predictive ability of several types of discriminant functions approached values found using all 139 features.

DATA SET

The data for this study were obtained from the file of 91,875 binary infrared spectra assembled by the American Society for Testing and Materials and made accessible by the Triangle Universities Computation Center (TUCC), the North Carolina Educational Computing Service, and the R. J. Reynolds Tobacco Company. Thirteen mutually exclusive classes were chosen with the main criterion for their selection being that they were similar to ones reported in previous work,[10] thus simplifying the comparison of results. Compounds containing C, H, O, and N atoms exclusively and with a carbon content ranging from C_1 to C_{15} were the only ones selected for this study. From those spectra belonging strictly to each of the 13 classes, 200 were randomly selected from each class, resulting in a data set of 2600 spectra. The range 2.0 to 15.9 μm was divided into 139 intervals of 0.1 μm each. Computations were done on the TUCC IBM 370/165 teleprocessing with the University of North Carolina Computation Center IBM 360/75 using Fortran IV and PL/I computer programs.

* Author to whom correspondence should be addressed; Alfred P. Sloan Fellow, 1971–1975.