Artificial Intelligence Used for the Interpretation of Combined Spectral Data. 3. Automated Generation of Interpretation Rules for Infrared Spectral Data

HENDRIK J. LUINGE,* GERARD J. KLEYWEGT, HENK A. VAN'T KLOOSTER, and JOHN H. VAN DER MAAS

Analytical Chemistry Laboratory, University of Utrecht, 3522 AD Utrecht, The Netherlands

Received December 30, 1986

The results of investigations on the development of an expert system for the elucidation of chemical structures from combined spectroscopic data (EXSPEC), in particular on the automated generation of knowledge necessary to interpret such data, are reported. A description of the program is given. Examples of automatically generated rules and interpretation results are presented and discussed.

INTRODUCTION

The use of artificial intelligence for chemical applications is currently enjoying a growing amount of attention.[1-7] This is partly due to the fact that new, high-level programming languages such as PROLOG[8] have become available that are capable of symbolic reasoning and list processing, important requirements for the application of artificial intelligence. Applications have been in such diverse areas as the design of organic chemical syntheses, the design of analytical procedures, and the interpretation of spectral data for structure elucidation of unknown compounds.

In the latter area, we are currently developing a microcomputer-based expert system (EXSPEC).[1,2] This system should ultimately be capable of handling various kinds of spectroscopic data (e.g., IR, MS, ^{13}C NMR, ^{1}H NMR, UV) and other relevant preinformation (e.g., chromatographic retention data, known structural fragments). The system consists of several units, each of which has a well-defined task. This is schematically depicted in Figure 1. Part I comprises the "translation" of spectral into structural data.[2] Part II concerns the generation of candidate structures for unknown compounds. Recently, a program has been completed that is capable of generating complete and irredundant sets of acyclic molecular structures.[1] The present investigations focus on part III of Figure 1, i.e., the automated generation of interpretation rules. Of course, other modules can be thought of (e.g., an explanation facility, a spectrum simulator), but either these are not yet available or they have been omitted from the illustration for the sake of simplicity. Human experts, in general, use two types of knowledge for the interpretation of spectra: correlation tablelike rules (e.g., "if there is a strong IR absorption near 1700 cm^{-1}, then carbonyl is likely/possible") and so-called "soft knowledge" (nonexplicit, heuristic expertise, largely based on experience, often pertaining to experimental conditions, patterns of peaks or exceptions to rules of the first type). Usually, both types of knowledge can be obtained by interrogating experts and transforming their reasoning into clear rules that can be handled by a computer. This job is often performed by a so-called knowledge engineer, who should ideally be a chemist and a psychologist as well as a computer scientist. Since such people are still rare (at the moment), it would be extremely useful if at least some of the expert's knowledge could be obtained in an automated fashion by a suitable computer program.

The aim of the present work is to develop a procedure for the generation of interpretation rules that (a) has a sound formal basis (using information theory and Bayes' statistics) and (b) is applicable to any kind of chemical functionality, thus making the availability of previously tabulated correlations a matter of strongly diminished relevance.

Research is in progress on the development of procedures to extract the second type of knowledge from human experts. In order to accomplish the goal mentioned above, two distinct problems have to be dealt with in order to reduce human effort to a minimum. First, the program has to be able to access a file of coded chemical structures and to select those compounds having structures that encompass a user-defined substructure (e.g., a secondary alcohol group or an amide linkage). Second, the program must produce a set of potential rules and select those that are least correlated and, hence, have the best classification characteristics.

To our knowledge, the first problem has only been tackled by using conventional programming languages (mostly FORTRAN), employing connectivity matrices for structure coding.[9-12] Algorithms to detect substructures are thus quite complex. However, employing the potential of symbolic reasoning, pattern matching, and backtracking inherent to PROLOG reduces this task to an almost trivial one.

The second problem has previously been addressed by Trulson and Munk[13] and Tomellini and co-workers.[14,15] In our approach we try to achieve three objectives (in decreasing order of importance): (1) using the interpretation rules generated, no substructure should be interpreted incorrectly as being absent, since this would make it more difficult to obtain complete sets of fragments for the structure generation stage; (2) the maximum amount of information should be extracted from the spectral data; and (3) as our system runs on a microcomputer, the number of rules should be as small as possible without any information loss. The first objective was achieved by generating only those spectral intervals in which every compound containing the functionality of interest shows an absorption. The second and third objectives were accomplished by explicity calculating the information content of each spectral interval and selecting only those intervals that were correlated to the least extent, resulting in a minimum number of rules with a maximum overall information content.

Since the spectral intervals generated in this way are generally larger than those obtained by Trulson and Munk[13] or Tomellini et al.,[14,15] our method obviously results in a higher number of false positives, but, in contrast to their methods, in the absence of false negatives. Furthermore, the information content gives direct insight into the interpretation capabilities of the generated rules. More specifically, the total number of correct interpretations is related to the relative amount of

0095-2338/87/1627-0095$01.50/0

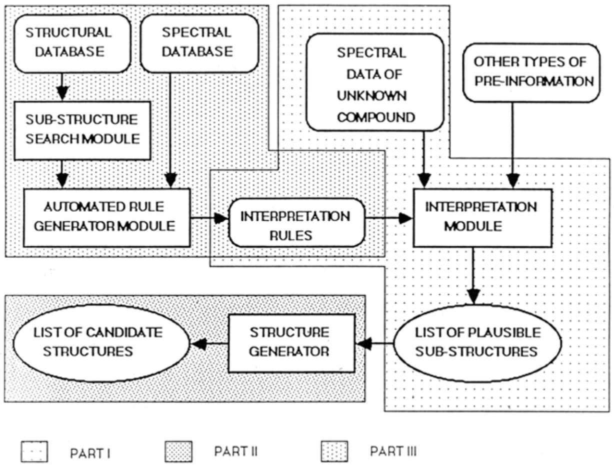

STRUCTURAL DATABASE
SPECTRAL DATABASE
SUB-STRUCTURE SEARCH MODULE
AUTOMATED RULE GENERATOR MODULE
INTERPRETATION RULES
SPECTRAL DATA OF UNKNOWN COMPOUND
OTHER TYPES OF PRE-INFORMATION
INTERPRETATION MODULE
LIST OF PLAUSIBLE SUB-STRUCTURES
STRUCTURE GENERATOR
LIST OF CANDIDATE STRUCTURES
PART I
PART II
PART III

Figure 1. Schematic representation of the EXSPEC system.

information described by the set of rules. Omission of intervals that are strongly correlated with those already selected finally leads to a smaller number of intervals without the loss of relevant information.

EXPERIMENTAL SECTION

For the generation of interpretation rules, use was made of a set of infrared spectra of 109 liquid alcohols and 141 liquid carbonyl compounds. The compounds, with molecular weights up to about 220, contained only carbon, hydrogen, and oxygen and were either commercial products or were obtained from the Laboratory of Organic Chemistry of the University of Utrecht. The purity of all compounds was over 98% (checked by means of GC).

The infrared spectra of the compounds were recorded as pure liquids on a Perkin-Elmer 180 spectrometer; the accuracy was ±2 cm^{-1} in the region 4000–2000 cm^{-1} and ±1 cm^{-1} in the region 2000–600 cm^{-1}. The base line was adjusted between 100 and 95% and the most intense band between 3 and 7% transmittance. For rule generation, use was made only of the wavenumbers of the peak maxima and the corresponding intensities (expressed as 100 minus transmittance in percent).

The rule generation program was written in LPA MacPROLOG (Standard syntax)[16] and runs on an Apple Macintosh Plus computer with 1 MB of internal memory.

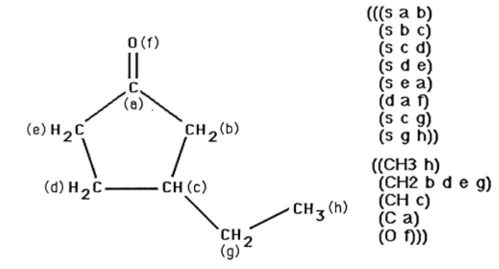
DESCRIPTION OF THE PROGRAM

The program consists of two main parts: a procedure for finding the compounds containing the structural fragment of interest and a procedure for the actual generation of rules.

The program is menu-driven, and substructure search and rule generation can be performed separately.

Search Module. For recognition of structural fragments in complete structures, a coding scheme has to be devised for representing these structures. This can be achieved by defining the units that are contained in the structure ("superatoms") together with the types of bonds connecting them. It is sufficient to define as superatoms all smallest possible combinations of non-hydrogen atoms with their attached hydrogen atoms (e.g., CH_3, CH_2, CH, C, OH, O). For reasons of efficiency, hydrogen atoms themselves are not defined as superatoms. Furthermore, four bond types are defined: s (single), d (double), t (triple), and a (aromatic).

Representing a molecular structure proceeds by assigning an arbitrary but unique symbol (e.g., a character) to each superatom present in the structure, defining the bond type between each pair of symbols, and adding a list that indicates the correspondence between symbols and superatoms. In this way, equivalent superatoms can be uniquely coded. All definitions are contained in a list structure as shown in Figure 2. Apart from structural information, any other information of importance can be stored in this list structure (e.g., name, formula, mp, bp, references).



```
(((s a b)
(s b c)
(s c d)
(s d e)
(s e a)
(d a f)
(s c g)
(s g h))
((CH3 h)
(CH2 b d e g)
(CH c)
(C a)
(O f)))
```

Figure 2. Structure coding for 3-ethylcyclopentanone.

Table I. Some Examples of Substructure Definitions

```
/*an ethyl group */
((substructure)
   (superatom CH3 __A)
   (superatom CH2 __B)
   (bond s __A __B))
/*carbonyl as a new basic unit*/
((carbonyl __A)
   (superatom C __A)
   (superatom O __B)
   (bond d __A __B))
/*an ester group*/
((substructure)
   (carbonyl __A)
   (superatom O __B)
   (bond s __A __B))
```

```
/*a saturated alcohol
   with a β-methyl group */
((substructure)
   (superatom OH __A)
   (superatom CH3 __B)
   (ON __X (CH2 C̅H̅ C))
   (superatom __X __C)
   (ON __Y (C̅H̅2̅ C̅H̅ C))
   (superatom __Y __D)
   (bond s __A __C)
   (bond s __C __D)
   (bond s __D __B))
```

For convenience, substructures to be searched for are defined in a somewhat different manner, but are converted into the above-mentioned list structure at run-time. Defining a substructure amounts to describing the superatoms and bond types of the unit of interest. Standard fragments (e.g., >C=O) can be defined from the basic units and used as new basic entities. Some examples of substructure definitions are given in Table I.

Having defined the substructure, a data base of structures can be scanned for those compounds that contain the unit of interest.

Rule Generator. Interpretation rules for infrared spectroscopy, in our approach, consist of characteristic absorption intervals (ω) for the functionality of interest (S) together with probabilities for the presence or absence of peaks in these intervals in the spectra of compounds that do or do not contain that functionality. Combination of different interpretation rules, i.e., combining the corresponding probabilities, ultimately yields an overall probability for each functionality as pointed out in Appendix 1.

Since the EXSPEC system is designed to generate molecular structures by combining structural fragments deduced from spectral data, it is important that no fragment present in the structure will be judged to be absent during the interpretation process. Therefore, we have chosen to generate selective interpretation intervals in which every spectrum of the class of compounds of interest shows an absorption.[17] Simultaneously, the number of interfering absorptions of compounds without the functionality of interest should be kept as low as possible.

In our approach, intervals are expanded around a startset of absorptions that may be correlated with a functionality S. This startset consists of all peaks encountered in the first spectrum that belongs to a compound containing S. Each absorption band is converted into an "interval" with zero width. Subsequently, the expansion of an interval occurs in case the spectrum of a compound with S has no absorption in the interval. The program then searches for the band "closest"

Table II. Example Output of the Rule-Generating Program for Primary Alcohols

```
***Start learning process***

Total number of compounds used: 109
Number of compounds with CH2–OH: 38
The maximum distance of an absorption to an interval:
  ±100 cm-1 and ± 100 (%T)
The weight of wavenumber and intensity in determining the distance:
  Weight (w) = 1 Weight (i) 1
There were 11 rules generated initially for CH2–OH!
Lowest allowed information content: 0
Minimum IC necessary for 100% correct identifications: 0.933

***Calculating information contents***

Interval (1): (3370 3295 92 53)  IC (1): 0.205
Interval (2): (1235 1116 63 11)  IC (1, 2, 3): 0.315
Interval (3): (2965 2915 93 40)  IC (1, 2): 0.385
Interval (4): ( 743  615 91 18)  IC (1, 2, 3): 0.453
Interval (5): (2938 2855 88 40)  IC (1, 2, 3, 4): 0.483
Interval (6): (1077 1002 93 45)  IC (1, 2, 3, 4, 5): 0.515

***Interpretation rules***

((CH2–OH IR (3370 3925 92 53)(38 0 30 41)(0.999 0.577 0.349)))
((CH2–OH IR (1235 1116 63 11)(38 0 15 56)(38 0 30 41)(0.999
  0.577 0.349)))
((CH2–OH IR (1235 1116 63 11)(38 0 15 56)(0.999 0.789 0.349)))
((CH2–OH IR (2965 2915 93 40)(38 0 13 58)(0.999 0.817 0.349)))
((CH2–OH IR ( 743 615 91 18)(38 0 16 55)(0.999 0.775 0.349)))
((CH2–OH IR (2938 2855 88 40)(38 0 11 60)(0.999 0.845 0.349)))
((CH2–OH IR (1077 1002 93 45)(38 0 26 45)(0.999 0.634 0.349)))
```

Table III. Results of Interpretations for 109 Alcohols and 141 Carbonyl Compounds with the EXSPEC Interpretation Module[a]

S	N(S)	N(r)	$\sum I$, bit	I_{rel}, %	TC, %
ArOH	14	2	0.553	100	100
ArCOR	19	5	0.544	100	100
RCHO	6	6	0.240	100	100
RR′R″COH	20	6	0.545	79	96
$RCOCH_3$	29	6	0.446	63	91
RCOOR′	74	5	0.584	58	90
RCH_2OH	38	6	0.515	55	86
RR′CHOH	38	6	0.424	45	81
RCOOH	19	4	0.215	40	78

[a] N(S) is the number of compounds with functionality S; N(r) is the number of rules generated; $\sum I$ is the total information content of all N(r) rules; I_{rel} is the relative information content (=100% × $\sum I/I_{100}$); TC is the total percentage of correct identifications (positive and negative).

to the interval, and expansion takes place to include it in the updated interval. A user-defined maximum value for the distance of the absorption to the interval prevents the intervals from becoming too large, since this would result in too many interfering absorptions. When no appropriate absorption is found, the corresponding interval is excluded from further consideration.

A spectrum is represented by two dimensions: a wavelength and some kind of intensity scale. For the determination of the band closest to an interval, at present, equal weights are assigned to both wavelength and intensity. The distance between a band and an interval can therefore be defined as the Euclidean distance in the space spanned by these scales. Other weight factors can, however, be imposed by the user.

Once selective intervals have been generated, conditional probabilities for the presence of an absorption in the interval ω, given the presence or absence of S, can be calculated. By use of Bayes' statistics, the probabilities for the presence of S can be computed. The formulas used to perform these calculations are given in Appendix 1.

For every interval generated, the information content can be determined as pointed out by Cleij and Dijkstra.[18,19] A threshold value for the information content can be imposed, below which intervals are excluded from further consideration. From the remaining intervals, the one with the highest information content is selected. Subsequently, for all combinations of this interval with the others the information contents are calculated, thereby implicitly taking into account correlations. In Appendix 2 the formulas to perform these computations have been summarized. In an iterative procedure, intervals that yield the highest increase in information content are added to those already selected until no further increase occurs. The total information content of the rules thus found can be used as a measure of their capability to interpret spectra for the functionality of interest.

RESULTS

The program was used to generate interpretation rules for several types of alcohols and carbonyl-containing compounds. Table II gives an example of the program output while generating rules for primary alcohols. Of 109 alcoholic compounds, 38 were found to contain a CH_2OH group. The spectra of these compounds were used to generate 11 intervals initially. Calculation of the information contents of these intervals, taking correlations implicitly into account, finally yielded six intervals in decreasing order of additional information content. These intervals were used to generate interpretation rules with the following format:

((name-of-functionality technique interval
counters probabilities))

Each interval consists of the maximum and minimum value of wavenumber and intensity. The counters are represented by $N(S,\omega)$, $N(S, \sim\omega)$, $N(\sim S, \sim\omega)$, and $N(\sim S, \omega)$, corresponding to the numbers of compounds with or without functionality S and with or without an absorption in interval ω ("~" stands for the logical "not"). The probabilities derived from these counters are given by $P(\omega|S)$, $P(\omega|\sim S)$, and $P(S)$. These are used by the interpretation module for calculating the probability of the presence or absence of S.

The performance of the rules was tested by using the EXSPEC interpretation module,[2] resulting in the scores shown in Table III. From the a priori probabilities for finding a functionality S, the information necessary for 100% correct identifications (I_{100}) can be calculated. The relative information content can then be defined as $I_{rel} = I/I_{100}$. As can be seen from the table, the percentage of correct identifications (TC) increases with increasing relative information content of the interpretation rules, and, for I_{rel} approaching 1.00, TC approaches 100%.

DISCUSSION

The above-described rule generator appears to give results that are comparable with spectrum–structure correlations found in tables in handbooks on infrared spectroscopy.[20] For example, the interval 3370–3295 cm^{-1} clearly corresponds with the O–H stretching vibration of alcohols and the interval 1077–1002 cm^{-1} with the C–O stretching vibration of primary alcohols. However, apart from these intervals, others are generated that seem less useful for interpretation purposes. Concerning this observation, two types of intervals can be distinguished: (a) large intervals with many interfering absorptions and (b) smaller intervals that do not seem to correspond with the functionality of interest. Occurrence of the first category is a direct consequence of the method of generating intervals (i.e., by expansion of existing intervals). However, as these intervals add only little to the overall information content, their occurrence can be prevented easily by raising the threshold value for the information content. The second category of intervals is harder to avoid as these intervals usually correspond to another functionality being incidentally present in all compounds containing the functionality of interest. Generation of such intervals clearly depends on the size and composition of the spectral data set used. As long as one adheres to specific types of compounds for rule generation and interpretation, these rules can be used quite satisfactorily. However, if one wants to use the interpretation rules for a wider variety of compounds, one should be careful about choosing a large training set for rule generation composed of compounds of widely different structure. (Since the program runs on a microcomputer, one should be aware of a corresponding increase in execution time.)

One might expect the rules generated to be dependent on the spectrum that is used to start the generation process. It appears, however, that approximately identical intervals are generated, irrespective of the choice of the first spectrum (small differences were found in the exact boundaries of the intervals and in the order in which they were added to the final set of rules). This surely is a direct consequence of the method used for rule generation, in which only selective intervals are generated, i.e., intervals in which every compound with the functionality of interest shows an absorption.

CONCLUSION

Automated generation of interpretation rules can relieve the developer of an expert system for interpretation purposes from much effort by extracting part of the knowledge necessary to perform interpretations from spectral and structural data with minimal human supervision. In a way, this program mimicks the learning behavior of an expert while working with a large amount of spectral data. In order to add a statistically sound basis to the process of rule generation, use was made of Bayes' statistics and information theory. This proves to be a useful approach and to a large extent relieves the system developer from extracting information from the literature and rather arbitrarily assigning certainties to interpretation rules.

Future research will focus on automated selection of substructures that show correlations with spectral data, extraction of the previously mentioned "soft" knowledge from human experts, and combination of the results of different spectroscopic techniques, which seems a viable way to obtain even better interpretation results.

APPENDIX 1. FORMULAS USED FOR THE CALCULATION OF CONDITIONAL PROBABILITIES

The probability of finding an absorption in spectral interval ω given a spectrum of a compound containing functionality S equals

$$p(\omega_i|\mathrm{S}_k) = N(\mathrm{S}_k,\omega_i)/N(\mathrm{S}_k)$$

Here, ω_i stands for the presence or absence of an absorption in spectral interval ω, S_k for the absence or presence of functionality S in the compound, and N for the corresponding number of spectra/compounds.

According to Bayes' theorem,[21] the probability of finding functionality S given an absorption ω equals

$$p(\mathrm{S}_k|\omega_i) = p(\mathrm{S}_k)p(\omega_i|\mathrm{S}_k)/p(\omega_i)$$

where

$$p(\omega_i) = \sum_{k=1}^{n} p(\mathrm{S}_k)p(\omega_i|\mathrm{S}_k)$$

Since we consider only two cases (S is present or absent), $n = 2$.

For the combination of more than one interpretation rule, it is assumed that one can state

$$p(\omega_{1i} \wedge \omega_{2i}|\mathrm{S}_k) \approx p(\omega_{1i}|\mathrm{S}_k)p(\omega_{2i}|\mathrm{S}_k)$$

The equality holds exactly when there is no correlation between the intervals ω_1 and ω_2. In our case, intervals are selected that are not highly correlated, so the equality will hold approximately.

APPENDIX 2. FORMULAS USED FOR THE CALCULATION OF THE INFORMATION CONTENTS OF SPECTRAL INTERVALS ω

According to Shannon,[22] the uncertainty about the presence or absence of a functionality S can be written as

$$H(\mathrm{S}) = -\sum_{k=1}^{n} p(\mathrm{S}_k) \text{ ld } p(\mathrm{S}_k) \qquad (\text{ld} = \log_2)$$

Analogously, the uncertainty about S after measuring a signal ω_i equals

$$H(\mathrm{S}|\omega_i) = -\sum_{k=1}^{n} p(\mathrm{S}_k|\omega_i) \text{ ld } p(\mathrm{S}_k|\omega_i)$$

The amount of information obtained in the case of an output signal ω_i is defined as the decrease of the uncertainty, i.e.

$$I(\mathrm{S}|\omega_i) = H(\mathrm{S}) - H(\mathrm{S}|\omega_i)$$

Finally, the expected value of the information or information content in bits is defined by

$$I = I(\mathrm{S}|\omega) = \sum_{i=1}^{m} p(\omega_i)I(\mathrm{S}|\omega_i)$$

Since there are only two possible signals, an absorption inside or outside a given interval, $m = 2$.

For reasons of computational efficiency, the information content is determined by calculating $I(\omega|\mathrm{S})$ instead of $I(\mathrm{S}|\omega)$. These can be shown to be equivalent.[18]

The computations for the information content of a combination of more than one interval are performed in an analogous way.

REFERENCES AND NOTES

(1) Kleywegt, G. J.; Luinge, H. J.; van't Klooster, H. A. "Artificial Intelligence Used for the Interpretation of Combined Spectral Data. Part II. A PROLOG Program for the Generation of Acyclic Molecular Structures". *Chemometrics Intelligent Lab. Syst.*, in press.

(2) Luinge, H. J.; van't Klooster, H. A. "Artificial Intelligence Used for the Interpretation of Combined Spectral Data". *Trends Anal. Chem.* **1985**, *4*, 242–243.

(3) Janssens, K.; van Espen, P. "Implementation of an Expert System for the Qualitative Interpretation of X-ray Fluorescence Spectra". *Anal. Chim. Acta* **1986**, *184*, 117–132.

(4) Gunasingham, H.; Srinivasan, B.; Ananda, A. L. "Design of a PROLOG-based Expert System for Planning Separations of Steroids by High-Performance Liquid Chromatography". *Anal. Chim. Acta* **1986**, *182*, 193–202.

(5) Gunasingham, H. "Heuristic Approaches to the Design of a Cybernetic Electroanalytical Instrument". *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 130–134.

(6) Hippe, Z. "Problems in the Application of Artificial Intelligence in Analytical Chemistry". *Anal. Chim. Acta* **1983**, *150*, 11–21.

(7) Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. *Applications of Artificial Intelligence for Organic Chemistry, The Dendral Project*; McGraw-Hill: New York, 1980.

(8) Clark, K. L.; McCabe, F. G. *micro-PROLOG: Programming in Logic*; Prentice-Hall: London, 1984. Clocksin, W. F.; Mellish, C. S. *Programming in PROLOG*; Springer Verlag: Berlin, 1984.

(9) Contreras, M. L.; Deliz, M.; Galaz, A.; Rozas, R.; Sepulveda, N. "A Microcomputer-Based System for Chemical Information and Molecular Structure Search". *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 105–108.

(10) Gillet, V. J.; Welford, S. M.; Lynch, M. F.; Willett, P.; Barnard, J. M.; Downs, G. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 7. Parallel Simulation of a Relaxation Algorithm for Chemical Substructure Search". *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 118–126.

(11) Synge, R. L. M. "Substructure Searching of Heterocycles by Computer Generation of Potential Aliphatic Precursors". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 50–55.

(12) Wipke, W. T.; Rogers, D. "Rapid Subgraph Search Using Parallelism". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 255–262.

(13) Trulson, M. O.; Munk, M. E. "Table-Driven Procedure for Infrared Spectrum Interpretation". *Anal. Chem.* **1983**, *55*, 2137–2142.

(14) Tomellini, S. A.; Hartwick, R. A.; Woodruff, H. B. "Automatic Tracing and Presentation of Interpretation Rules Used by PAIRS: Program for

the Analysis of IR Spectra". *Appl. Spectrosc.* **1985**, *39*, 331–333.
(15) Tomellini, S. A.; Hartwick, R. A.; Stevenson, J. M.; Woodruff, H. B. "Automated Rule Generation for the Program for the Analysis of Infrared Spectra (PAIRS)". *Anal. Chim. Acta* **1984**, *162*, 227–240.
(16) French, P.; Clark, K. L. *LPA MacPROLOG User Guide*; Logic Programming Associates: London, 1985.
(17) Visser, T.; van der Maas, J. H. "Systematic Computer-Aided Interpretation of Vibrational Spectra". *Anal. Chim. Acta* **1980**, *122*, 357–361.
(18) Cleij, P.; Dijkstra, A. "Information Theory Applied to Qualitative Analysis". *Fresenius' Z. Anal. Chem.* **1979**, *298*, 97–109.
(19) Dupuis, P. F.; Cleij, P.; van't Klooster, H. A.; Dijkstra, A. "Information Theory Applied to Feature Selection of Binary-Coded Infrared Spectra for Automated Interpretation by Retrieval of Reference Data". *Anal. Chim. Acta* **1979**, *112*, 83–93.
(20) Socrates, G. *Infrared Characteristic Group Frequencies*; Wiley: Chichester, UK, 1980.
(21) Weinberg, F. *Grundlagen der Wahrscheinlichkeitsrechnung und Statistik sowie Anwendungen in Operations Research*; Springer Verlag: Berlin, 1968.
(22) Shannon, E.; Weaver, W. *The Mathematical Theory of Information*; University of Illinois: Urbana, IL, 1947.

Description of Organic Reactions Based on Imaginary Transition Structures. 6. Classification and Enumeration of Two-String Reactions with One Common Node

SHINSAKU FUJITA

Research Laboratories, Ashigara, Fuji Photo Film Co., Ltd., Minami-Ashigara, Kanagawa, Japan 250-01

Received October 6, 1986

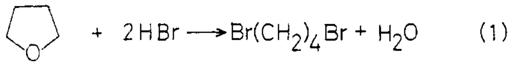
The number of reaction strings is a clue in the classification of organic reactions. Various two-string reactions are classified by their reaction graphs, each of which has two reaction strings sharing one node. The two-string reaction graphs are enumerated by Polya's theorem.

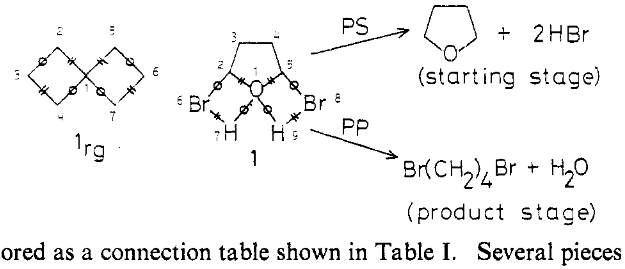
Classification of organic reactions is an important problem to be solved in order to construct computer systems for retrieval of organic reactions and for synthetic design. Many methods have been reported for this purpose and reviewed from various points of view.[1,2] We have proposed an imaginary transition structure (ITS) as a comprehensive representation of an individual organic reaction, which involves substrates and products as well as other components such as catalysts.[3] In the ITS approach, we have introduced three colored bonds, i.e., an out-bond (—╫—), an in-bond (—o—), and a par-bond (—).[4] The ITS is a kind of structural formula, in which all nodes are connected by the three colored bonds in accordance with structural change during a reaction.[5] From the ITS of an individual reaction, we have abstracted a reaction graph as a subgraph, which represents the corresponding reaction type. The reaction graph contains one or more reaction strings,[4] each of which has alternate in-bonds and out-bonds and can be modified by par-bonds.

In the previous papers,[3b,c] I have dealt with one-string reactions and enumerated trigonal, tetragonal, pentagonal, hexagonal, and octagonal reaction graphs. In this paper, I will describe two-string reactions and report that the present ITS approach has several advantages over other methods proposed for description of organic reactions.

ABSTRACTION OF REACTION STRINGS FROM AN ITS OR FROM A REACTION GRAPH

Formation of 1,4-dibromobutane by ring opening of tetrahydrofuran (entry 1)[7] is represented by ITS **1**, which can be





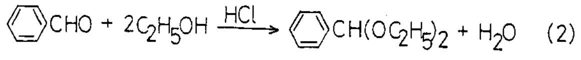
1_{rg}

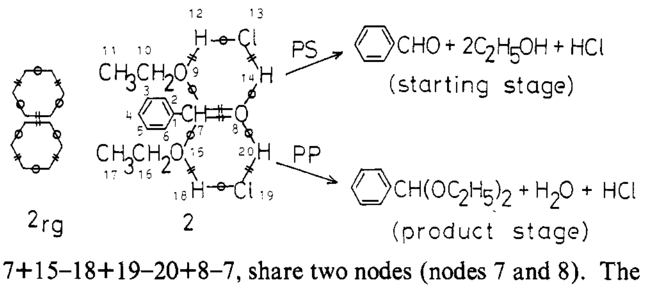
1

stored as a connection table shown in Table I. Several pieces of information concerned with this reaction can be abstracted in the light of newly defined operations upon the ITS or ITS connection table.[3a] For example, the starting and product stages are derived by PS and PP operations, respectively.[8] The corresponding reaction graph (**1**$_{rg}$) represents a generic reaction type involving this reaction.

Two reaction strings, 1–2+6–7+1 and 1–5+8–9+1, can be abstracted graphically from ITS **1** when – and + represent an out-bond and an in-bond, respectively.[9] These two reaction strings are stored by a connection table shown in Table II or by codes such as '1*O*'(1–1)2*C*(0+1)6*Br*(1–1)7*H*(0+1)'1*O*' and '1*O*'(1–1)5*C*(0+1)8*Br*(1–1)9*H*(0+1)'1*O*'. In the latter codes, the nodes and the ITS bonds[3] are taken up, and the common node(s) shared by two reaction strings is (are) indicated by the single quotes. The same reaction strings are involved in the reaction graph **1**$_{rg}$ in more abstract fashion.[10]

Acetalization of benzaldehyde (entry 2)[11] leads to ITS **2**, in which two reaction strings, i.e., 7+9–12+13–14+8–7 and 7+15–18+19–20+8–7, share two nodes (nodes 7 and 8). The reaction graph **2**$_{rg}$ abstracted from ITS **2** contains two hexagonal reaction strings sharing two nodes in a similar way.





2_{rg}

2

In the case of multistring reactions, there is some ambiguity in abstracting reaction strings. For example, in the case of entry 1, one can visualize a single reaction string, i.e., 1–2+6–7+1–5+8–9+1, in the manner of a picture drawn with a single stroke of the brush. However, this single string is to be forbidden. For the acetalization of entry 2, a single reaction string, 7–8+14–13+12–9+7–8+20–19+18–15+7, is possible but not permitted in the present method.

To prevent such ambiguity in the adoption of reaction strings, I have established a criterion that *the same colored bonds incident to a node are preferred to form different reaction strings from each other if possible.*[12] In other words,

0095-2338/87/1627-0099$01.50/0