ESSESA: An Expert System for Elucidation of Structures from Spectra. 1. Knowledge Base of Infrared Spectra and Analysis and Interpretation Programs

HONG HUIXIAO and XIN XINQUAN*

Department of Chemistry, Nanjing University, Nanjing 210008, People's Republic of China

Received November 9, 1989

This paper describes the logical structure of ESSESA, an Expert System for Structure Elucidation by Spectral Analysis, a knowledge base of infrared spectra and its inference logic. Logical representation and rules concerning analysis of IR spectra are discussed, as well as analysis tactics and inferential models useful in IR spectral analysis. The analysis program examines the unsaturation and the atomic composition of an unknown compound before developing the substructural constraints from the infrared spectrum. The interpretation program develops substructural constraints from its analysis of the infrared spectrum of a compound by inference from the knowledge base of spectral data. The knowledge base contains 229 substructures of 86 different structural types.

INTRODUCTION

Three distinct approaches have evolved for the analysis of spectra by computers. These are the search method, the pattern recognition method, and the artificial intelligence method. So-called "expert systems" represent a special case of the artificial intelligence approach. In database searching methods, the structure of the unknown compound is determined by comparing its spectrum with those in a spectral database. This method fails when the unknown compound, or its spectrum, is "new" to the database.[1,2] Pattern recognition current is used mainly to classify compounds on the basis of some property, such as their spectra; it is not useful in problems of structure determination.[3] The expert system method can elucidate the complete structure of an unknown compound, relying for this purpose upon inference from a "knowledge base" of spectra and spectral analysis.[4-12]

ESSESA is an expert system that aids the chemist in the interpretation of spectra. This paper describes ESSESA as it is applied to the analysis of infrared spectra. The program is designed to analyze infrared spectra and derive primary substructural constraints which can be used to drive a structure generation program.

ESSESA has been developed on an IBM PC-at using Turbo Prolog. Its output consists of substructural fragments inferred from the molecular formula, the spectrum, and other chemical information. An overview of the structure of ESSESA is provided in Figure 1.

KNOWLEDGE BASE OF INFRARED SPECTRAL ANALYSIS

The conventional approach taken by chemists to spectral interpretation is based upon models, often simplified, of the physical processes underlying resonance and the resulting spectral absorption. These physical models permit specific spectral signals to be related to particular structural components of the molecule. Usually, there are several factors which together determine the detailed characteristics of a spectral signal and which are often in some sort of a hierarchical relationship that defines their relative importance. The initial analysis of a spectral signal may identify the presence of a specific type of substructure in the unknown molecule, and more detailed analysis of the form of the signal may determine aspects of the larger environment of that substructure.

The chemists' knowledge of spectral analysis that has been incorporated into ESSESA is encoded in the form of spectral feature–substructure relationship rules, which comprise the knowledge base for infrared spectral analysis.

If a set of specific infrared absorption bands given by an unknown compound is W, such that

$$W = [W_i] \qquad i = 1, 2, 3, ..., n \tag{1}$$

and the set of substructures in the knowledge base is S:

$$S = [S_j] \qquad j = 1, 2, 3, ..., n \tag{2}$$

$$W_{sj} = W_{j1} \cup W_{j2} \cup W_{j3} ... \cup \mathrm{W}_{jk} \tag{3}$$

where W_{sj} is the set of specific absorption bands that correspond to substructure S_j. If W_j is a subset of set W and the specific absorption bands of substructure S_j are found in the spectrum, then the substructure S_j may be present in the structure of the compound. This idea may be represented by the rule

$$W_{sj} \subseteq W \rightarrow S_j \tag{4}$$

and this rule may be represented by the following Boolean function:

$$F(S_j) = f(W_{j1}) \wedge f(W_{j2}) \wedge f(W_{j3}) \wedge ... \wedge f(W_{jk}) \tag{5}$$

Equation 5 is equivalent to eqs 6 and 7.

$$f(W_{jr}) \subseteq 1 \ (r = 1, 2, 3, ..., k) \rightarrow F(S_j) \subseteq 1 \tag{6}$$

$$W_{jr} \subseteq W \rightarrow f(W_{jr}) \subseteq 1 \ (r = 1, 2, ...,k) \tag{7}$$

In order to interpret an infrared spectrum therefore, it is necessary to find the set of substructures S_{ir}:

$$S_{ir} = [S_{irt}] \ (t = 1, 2, 3, ..., p), S_{irt} \subseteq S \tag{8}$$

$$W_{sirt} = [W_{sirtu}] \ (u = 1, 2, 3, ..., o) \tag{9}$$

$$W_{sirtu} \subseteq W \tag{10}$$

According to this procedure, the construction of a knowledge base for infrared spectral analysis requires that the logical representation formula (eq 5) of the set of substructures S be found. In ESSESA, the information in eq 5 is expressed by the system rule—i.e., the spectral feature–substructure relationships. Given a sufficiently large number of spectra, or alternatively the appropriate information derived from published observations, it is possible to develop the ability to associate certain absorption lines with the corresponding functional groups. Thus a set of rules can be developed to determine the set S of substructures that are indicated by a specific infrared spectrum.

0095-2338/90/1630-0203$02.50/0 © 1990 American Chemical Society

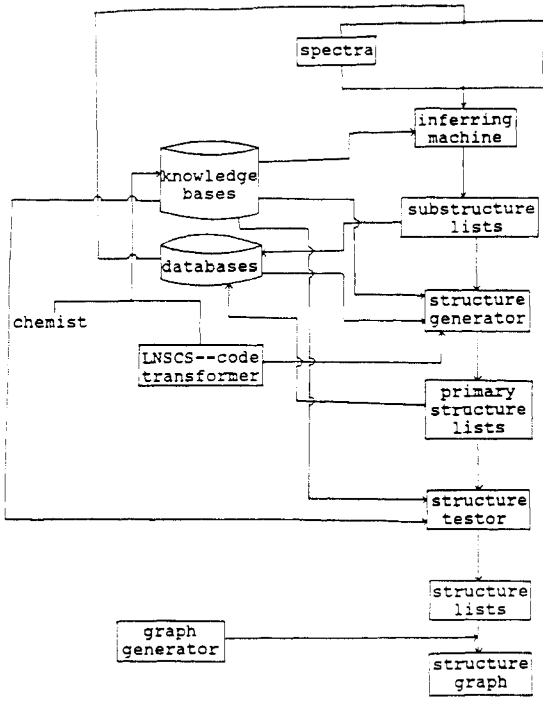


Figure 1. Overview of structure of ESSESA.

As an example, the rule that is derived for the identification of the aldehyde substructure (-CHO) is as follows:

IF in the infrared spectrum of the sample there are:
1740–1720 cm^{-1} (strong, sharp peak) **and**
2900–2700 cm^{-1} (two medium-weak sharp peaks) **and**
975–780 cm^{-1} (strong-medium, sharp peak) **and**
1400–1325 cm^{-1} (strong-medium sharp peak)
THEN the substructure -CHO may be present in the structure of the sample compound

The mathematical representation of this rule would be as follows:

$$F(\text{-CHO}) = f(W_1) \wedge f(W_2) \wedge f(W_3) \wedge f(W_4) \wedge f(W_5) \wedge f(W2 = W_3) \tag{11}$$

$$f(W_1) = 1,\ W_1 \subseteq W,\ W_1 \subseteq (1720\text{–}1740\ \text{cm}^{-1},\ \text{strong, sharp}) \tag{12}$$

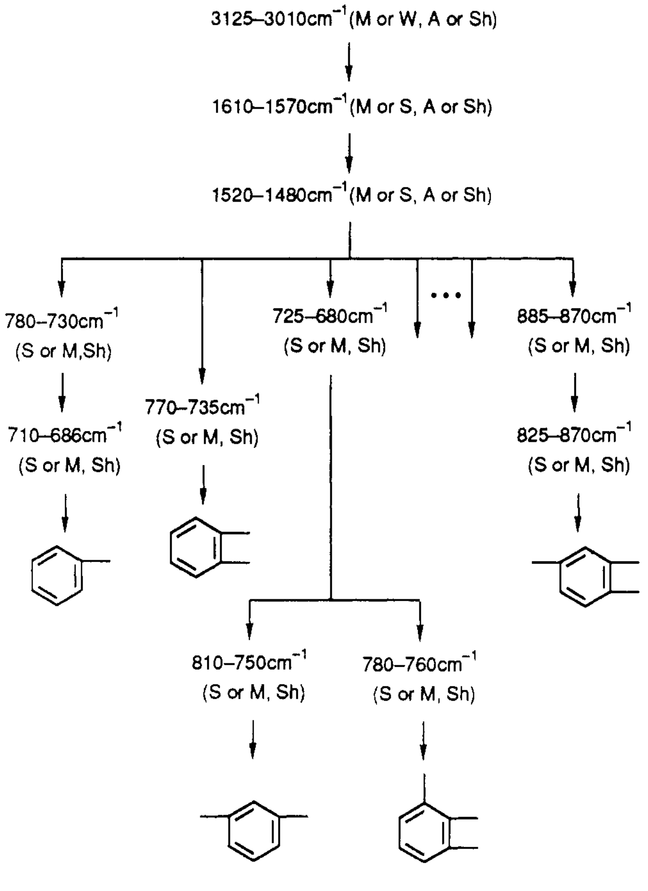
$$f(W_2) = 1,\ W_2 \subseteq W,\ W_2 \subseteq (2700\text{–}2900\ \text{cm}^{-1},\ \text{M-W, sharp}) \tag{13}$$

$$f(W_3) = 1,\ W_3 \subseteq W,\ W_3 \subseteq (2700\text{–}2900\ \text{cm}^{-1},\ \text{M-W, sharp}) \tag{14}$$

$$f(W_4) = 1,\ W_4 \subseteq W,\ W_4 \subseteq (780\text{–}975\ \text{cm}^{-1},\ \text{S-M, sharp}) \tag{15}$$

$$f(W_5) = 1,\ W_5 \subseteq W,\ W_5 \subseteq (1325\text{–}1440\ \text{cm}^{-1},\ \text{S-M, sharp}) \tag{16}$$

There are 229 structures of 86 different chemical types in the knowledge base used by ESSESA for infrared spectrum analysis. This exceeds the databases used by PAIRS,[13] EXSPE,[14] CHEMICS,[15] and CASE.[16] These 229 substructural fragments are listed in Table I. There is overlap within this list; the diene fragment $[>C{=}\overset{|}{C}{-}\underset{|}{C}{=}C<\ (Z)]$ for example, overlaps with —CH=CH— (*Z*), >C=CH—, and so on. In these cases, the structure generator detemines which fragments can be allowed in the full structure on the basis of information derived from the ^{1}H NMR and ^{13}C NMR spectra. The identification of mutually consistent set of fragments that are used by the structure generator can be achieved by means of a procedure that finds those combinations of permitted substructures that are compatible with the overall composition of the molecule and with the constraints derived from the spectral data. Each such combination of substructures then may be used to define a distinct problem which can be referred to a subsequent structure generation program.


3125–3010cm^{-1}(M or W, A or Sh)
1610–1570cm^{-1}(M or S, A or Sh)
1520–1480cm^{-1}(M or S, A or Sh)
780–730cm^{-1} (S or M,Sh)
710–686cm^{-1} (S or M, Sh)
770–735cm^{-1} (S or M, Sh)
725–680cm^{-1} (S or M, Sh)
...
885–870cm^{-1} (S or M, Sh)
825–870cm^{-1} (S or M, Sh)
810–750cm^{-1} (S or M, Sh)
780–760cm^{-1} (S or M, Sh)

S = strong; M = medium; W = weak; A = average; Sh = sharp

Figure 2. Partial search tree of benzene ring substructures.

Inference of substructures is achieved by depth-first searching of the tree-structured database. An example of the search tree is shown in Figure 2, which contains the partial search tree that is used by ESSESA in dealing with aromatic ring substructures. The C–H stretching region is examined first, and if a sharp or average peak is found between 3215 and 3010 cm^{-1}, then the region containing the aromatic ring skeletal vibration absorptions is examined. If no peak is found between 3125 and 3010 cm^{-1}, the search for aromatic substructures is terminated. If the search reveals peaks in both the C–H stretching region *and* the aromatic ring skeletal vibration region (1610–1570 cm^{-1} and 1520–1480 cm^{-1}; two medium or strong, sharp peaks), an aromatic fragment may be present in the full structure. Next, the branches of the tree are searched in order to decide which aromatic structures may be indicated. A sharp, strong peak in the 770–735 cm^{-1} region, for example, suggests the presence of a 1,2-disubstituted aromatic ring.

Search trees, frequently more complex than that shown in Figure 2, have been written in PROLOG for all of the structural fragments listed in Table I. Judgements must be made when building search trees of this sort, and accordingly, users are allowed to modify the rules to improve the interpretation procedures and the search trees. The interpretation rules are embedded in the knowledge base, which is quite

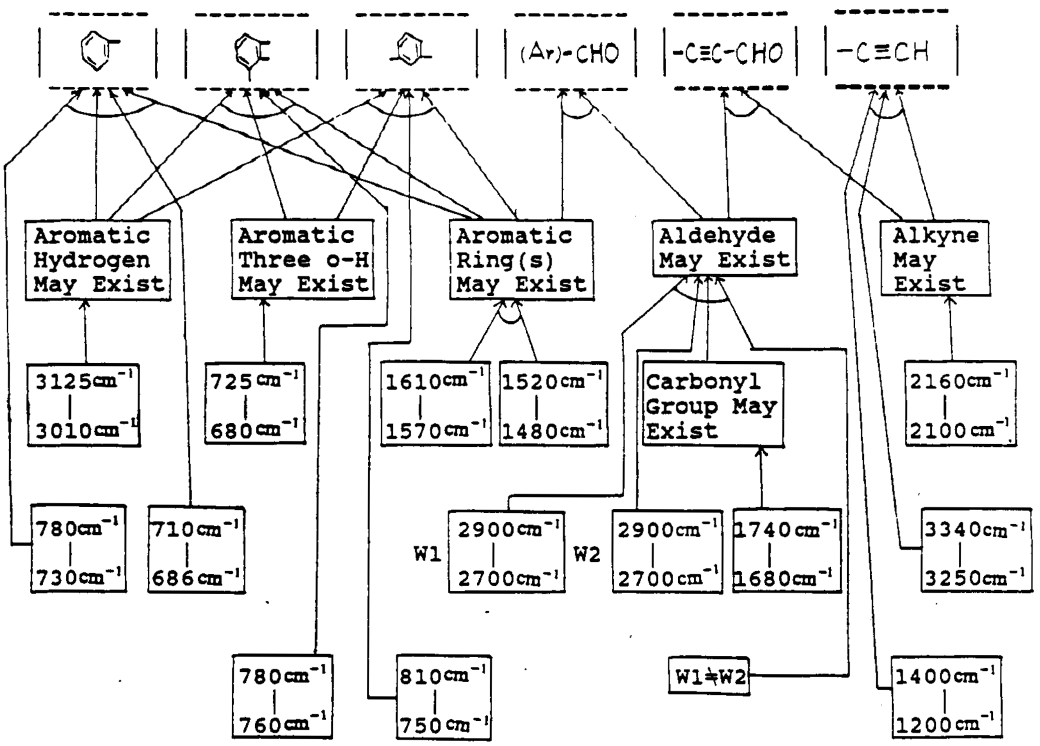


Figure 3. Part of the inferring network in the knowledge base of infrared spectra analysis of ESSESA.

distinct from the interpretation program in ESSESA. Modification of the rules does not cause any change in the interpretation program.

The search trees used by ESSESA for infrared spectral analysis are written in PROLOG as rules. The interpretation of infrared spectra is carried out by inference from these rules which, together with those in the knowledge base, constitute a very large inferential network, a small part of which is shown in Figure 3.

INTERPRETER PROGRAM

After the molecular formula of the compound and its digitized infrared spectrum have been entered, the computer identifies the set of specific absorption lines (W) and then, using formulas 7 and 8, begins to identify the various substructural fragments that may be present in the structure. This work is done by the interpreter program, which uses the knowledge base. This interpretation of an infrared spectrum leads to a group of structural fragments that are likely to be present in the unknown molecule. This fragment list is used in subsequent operations in ESSESA, such as generation of the structure of the unknown compound or interpretation of the ^{1}H NMR or ^{13}C NMR spectra. The structure of this program is shown in Figure 4.

First, the program analyzes the ring and double-bond characteristics of the entered molecular formula. The result from this step is passed to other programs, such as those that manage ^{1}H NMR interpretation and structure generation. Next, the atomic composition of the unknown compound is analyzed, and the number of atoms of each element is also passed forward to the same program. Finally, the interpretation of the infrared spectrum is started. The substructures that are deemed to be allowable are identified and stored, to be used as primary constraints in structure generation and the analysis of other spectra.

The interpretation program works with the set of 229 structural fragments. Each of these substructures is correlated with a defined pattern of infrared spectral absorption. The spectral features that characterize substructural fragments normally consist of spectral ranges within which specific types of peaks are expected. The initial set of substructures is, in effect, screened against the entered spectral data, and any substructure whose requisite spectral pattern is absent is discarded. The result of this analysis is a subset of the 229 fragments. Each of the members of the subset is related to absorption data, which is consistent with the infrared spectrum of the unknown compound.

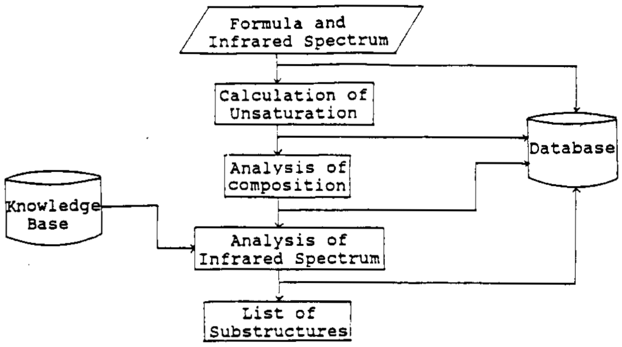

Formula and Infrared Spectrum
Calculation of Unsaturation
Database
Analysis of composition
Knowledge Base
Analysis of Infrared Spectrum
List of Substructures

Figure 4. Overview of the structure of the interpreter program.

In the interpretation, goal-driving inference tactics were used. In such an inference model, substructures from the knowledge base are used as the goals, and the program seeks spectral patterns which fulfill the premises of the goals, as defined by the production rules. If any single premise of a goal is not satisfied, that goal is determined to be false; that is to say, that particular substructure is not contained in the unknown structure. If all the premises of a goal are satisfied, then the goal is considered to be true, and the substructure corresponding to the goal may be embedded within the full structure of the unknown. The procedure is illustrated by the structure elucidation of (chloromethyl)phosphonodithioic acid *O*-ethyl *S*-*p*-tolyl ester, whose IR spectrum is shown in Figure 5.

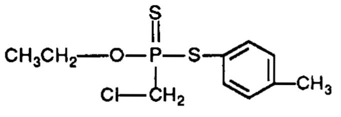


Table I. 229 Substructures in the Knowledge Base

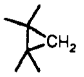
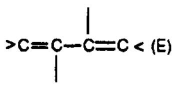
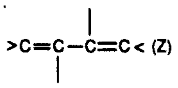
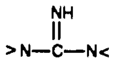
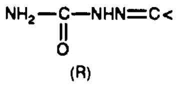
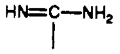
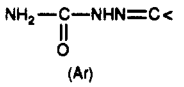
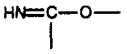
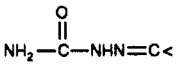
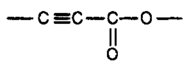
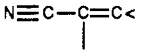
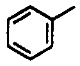
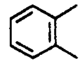
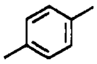
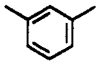
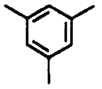
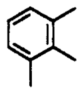
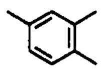
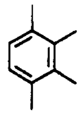
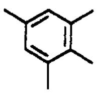
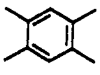
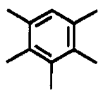
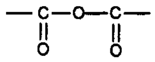
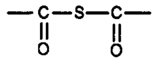
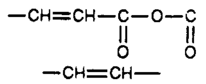
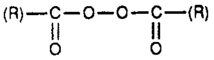
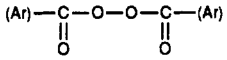
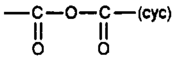
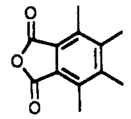
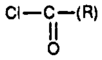
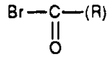
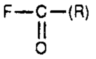
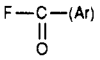
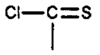
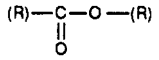
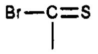
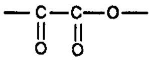
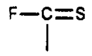
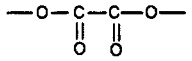
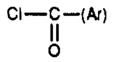
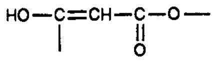
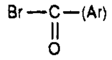
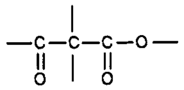
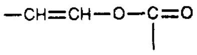
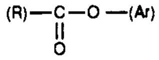
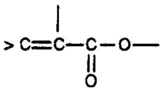
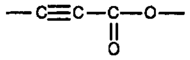
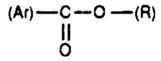
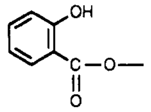
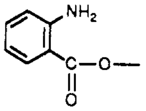
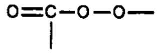
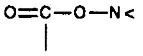
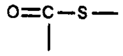
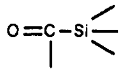
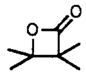
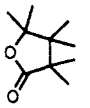
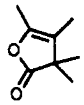
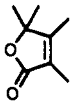
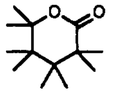
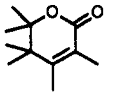
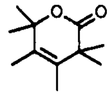
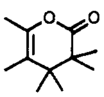
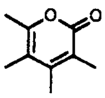
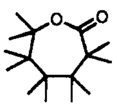
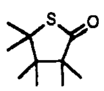
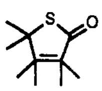
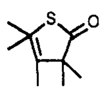
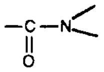
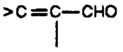
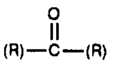
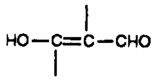
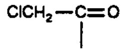
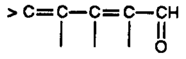
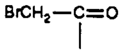
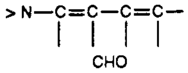
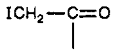
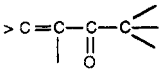
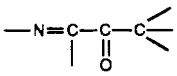
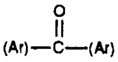
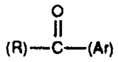
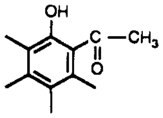
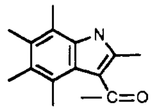
HO—(R) $NH(Ar)_2$ CH_3—N<	HO—(Ar) $\overset{+}{N}H_3(R)$ CH_3—O—	$NH_2(R)$ $\overset{+}{N}H_2(R)_2$ —$(CH_2)_3$—(n>4)	$NH(R)_2$ $\overset{+}{N}H(R)_3$ —CH_2—	$N(R)_3$ $\overset{+}{N}(R)_4$ —O—CH_2—O—	$NH_2(Ar)$ CH_3— 
—CH< >C=C<(Z)	CH_2=CH— >C=C<(E)	CH_2=C< 	—CH=CH—(Z) 	—CH=CH—(E) NH=CH—	>C=CH— >C=NH^+—
(R)—N=CH—(Ar) >C=N—N=C<	(Ar)—N=CH—(Ar) —CH=N—NH_2	 	 	 	HO—N=C< —N=N—
—NH—NH—	CH≡C—	—C≡C—	—C≡C—CH=C<	—C≡C—C≡C—	
N≡C—(R)	N≡C—(Ar)		C≡N—(R)	C≡N—(Ar)	>C≡C≡C<
O=C=N— N^-=N^+=N—	S=C=N— N^-=N^+=CH—	N≡C—O— N^-=N^+=C<	N≡C—S—(R) N^+≡N—(Ar)	N≡C—S—(Ar) 	—N=C=N— 
					
					
					
 	 	 	 	 	 
					
					
					
					
 —C≡C—CHO	(R)—CHO (Ar)—CHO	 	 	 	 
					

Table I (Continued)

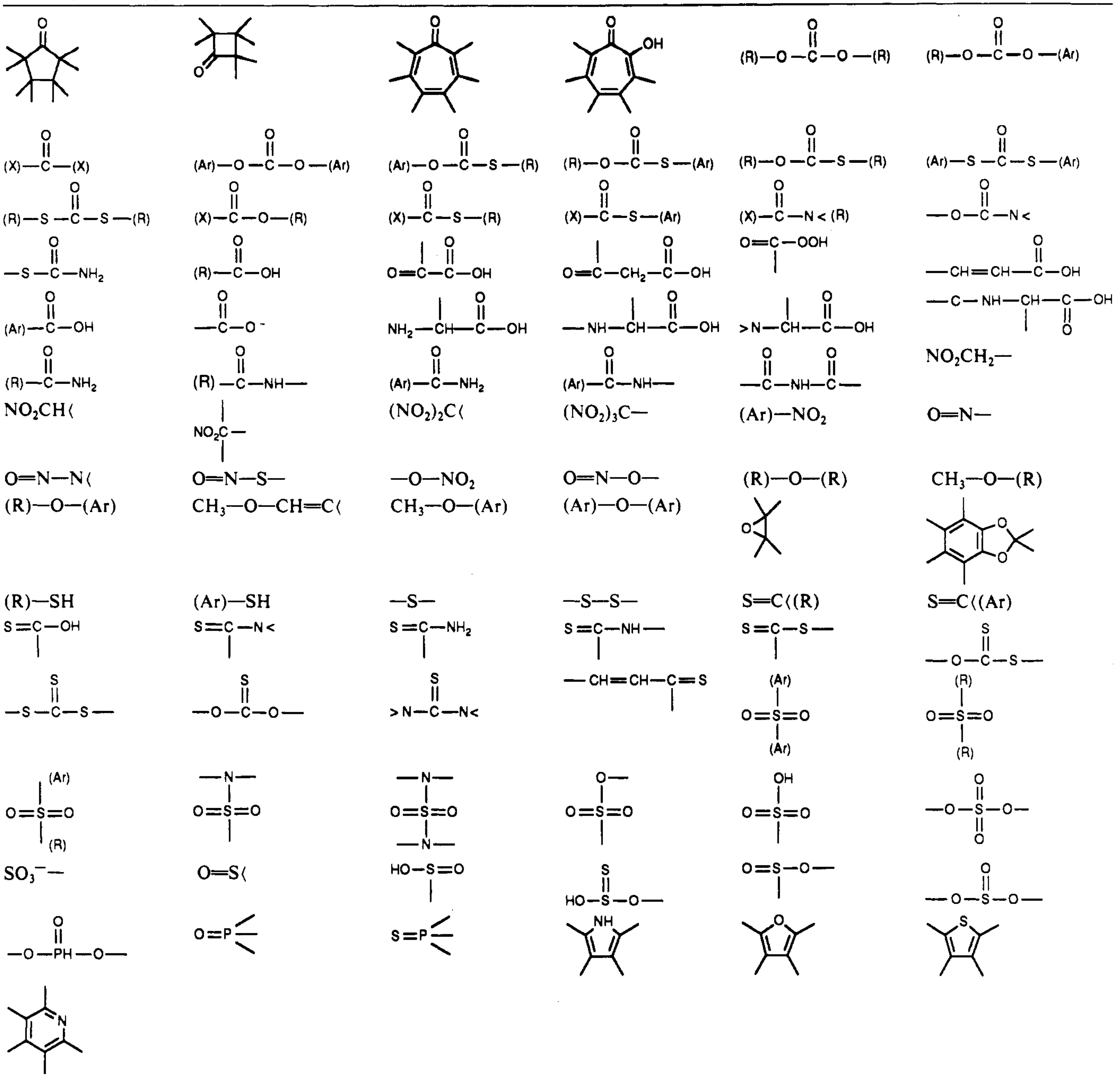


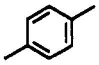
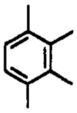
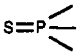
Table II. Digitized Spectrum

location	intensity	shape	location	intensity	shape
3010 cm^{-1}	4	2	1450 cm^{-1}	5	1
2980 cm^{-1}	6	1	1390 cm^{-1}	6	1
2920 cm^{-1}	5	2	1020 cm^{-1}	9	2
2860 cm^{-1}	4	2	955 cm^{-1}	8	1
1592 cm^{-1}	7	1	830 cm^{-1}	8	1
1490 cm^{-1}	4	2	810 cm^{-1}	8	1

First, the molecular formula, $C_{10}H_{14}ClOPS_2$, and the infrared spectrum are entered. The spectral data must be presented in digital form, but the means of digitization is unimportant, as long as it is accurate. ESSESA accepts the following frequency ranges for the digitized data: peak locations between 4000 and 600 cm^{-1}; peak intensities of 1–9 (1 very weak, 1–3 weak, 4–6 medium, 7–9 strong, and 9 very strong); peak shape 1 sharp, 2 average, 3 broad. The digital data from the spectrum in Figure 5 are given in Table II.

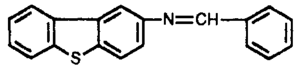
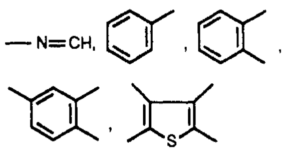
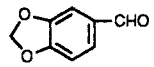
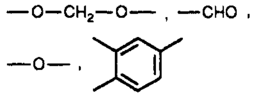
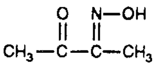
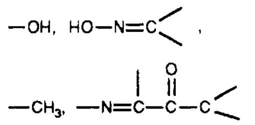
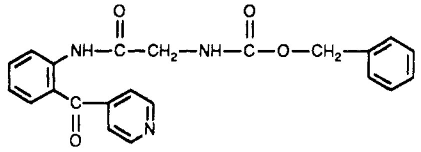
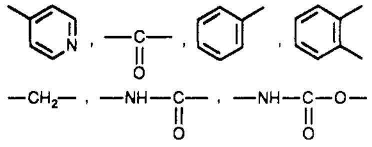
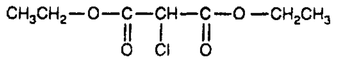
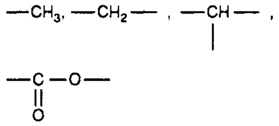
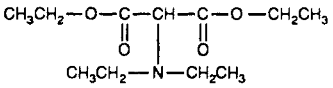
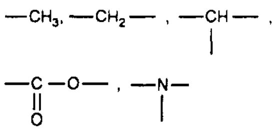
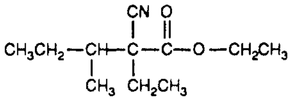
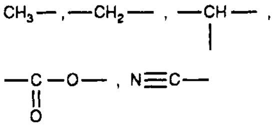
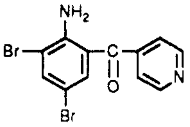
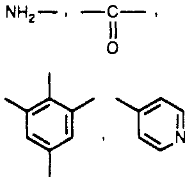
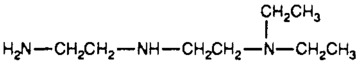
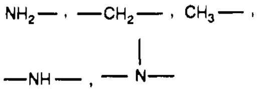
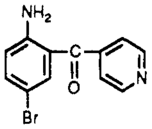
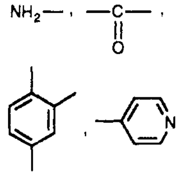
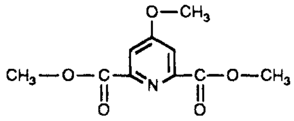
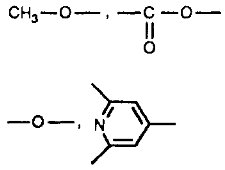
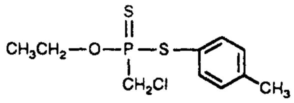
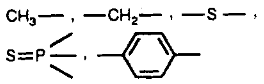
As the interpreter acquires the digital spectral data, it makes use of the rules in the knowledge base to compare the stored

Table III. Analysis Result of Infrared Spectrum of $C_{10}H_{14}ClOPS_2$

substructures			chemical structures
no.	LNSCS code[a]	environment[b]	
1	C.	c	CH_3-
2	C.	O	CH_3-O-
3	C.2	c	-CH_2-
4	Cl = CC. = CC = Cl.	c	
5	Cl = CC. = C.C. = Cl.	c	
6	O.2	c	-O-
7	S.2	c	-S-
8	S = P.3	c	

[a] LNSCS is the linear code used by ESSESA, designed by authors. [b] Symbols of ESSESA: C, saturated carbon; Ar, aromatic carbon; c, arbitrary; and so on.

Table IV. Testing Details of 12 Infrared Spectra

structure formula (molecular formula)	peaks input	substructures found	time (s)	substructures exist in molecule
	19	9	6.0	
	19	11	7.0	
	14	7	3.5	
	32	15	9.5	
	16	8	4.5	
	14	10	4.5	
	12	9	4.0	
	25	9	6.5	
	12	7	2.5	
	22	9	5.5	
	25	10	7.5	
	19	8	5.0	

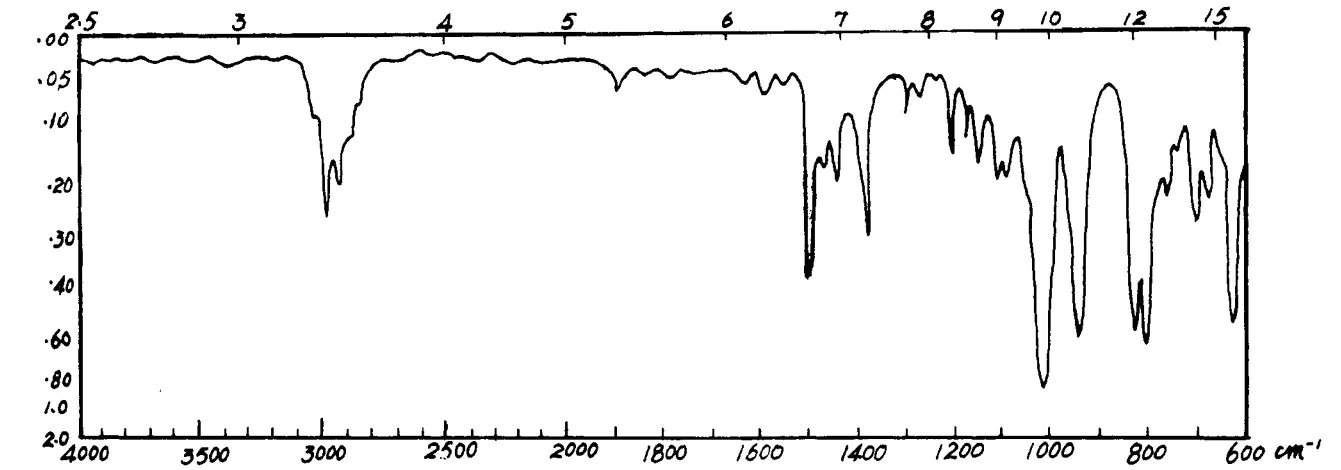


Figure 5. Infrared spectrum of $C_{10}H_{14}ClOPS_2$.

spectral patterns with the digital experimental data to identify the substructures that might be contributing to the $C_{10}H_{14}$-$ClOPS_2$ molecular formula.

The structure of the (chloromethyl)phosphonodithioic acid *O*-ethyl *S*-*p*-tolyl ester above can be compared with the fragments shown in Table III, which ESSESA decided should be present in the molecule, based upon the IR spectrum. Fragments 1, 3, 4, 6, 7, and 8 are present in the molecule, but fragments 2 and 5 are not. These erroneous fragments are eliminated by consideration of the ^1H NMR and ^{13}C NMR spectra and the subsequent structure generation. This program does not infer the presence of the chlorine atom, but the structure generator can proceed without such an inference. If the inference of the chlorine does not follow from the analysis of the ^1H NMR and ^{13}C NMR spectra, the structure generator, which has access to the molecular formula, will automatically produce structures which contains the six correct fragments and a chlorine, in any of the permitted positions.

RESULTS AND DISCUSSION

The interpretation program and the knowledge base were tested with 48 infrared spectra selected from Nakanishi and Solomon[17] and from the Sadtler Infrared Spectral Database. All the structure fragments in these 48 compounds were correctly found by the program. The results obtained from 12 of the spectra are shown in Table IV.

The vibrational frequencies of chemical bonds in a molecule are affected by the overall structural environment, and different structural fragments can absorb infrared radiation in the same frequency range. As a result, the infrared spectrum can be very complex, and it is usually difficult or impossible to determine the structure of a compound with only an infrared spectrum. A structure elucidation system that relies solely upon IR spectral data will obviously be limited to bonds that give signals in the infrared spectrum, but since many bonds, particularly skeletal bonds in a molecule, fail to exhibit clearly characteristic and unique peaks in the infrared spectrum, no complete description of the structure will be derivable from the spectrum. The function of the infrared spectrum analysis component of ESSESA is to identify structural fragments that, in the light of the IR spectral data, are likely to be present in the molecule. The presence or absence in the molecule of specific fragments is then used as a constraint in subsequent ESSESA modules, ^1H NMR and ^{13}C NMR spectral interpretation, and structure generation. Fragments that are considered to be absent are not considered in these subsequent steps.

The time of analysis and the number of fragments reported to be present both depend upon the number of peaks in the spectrum that is entererd. As more peaks are entered, the analysis time and the number of fragments present both increase. If too few peaks are entered however, some important substructures may be missed and this makes the structure generation more difficult. In general, an attempt should be made to enter, at a minimum, the more characteristic peaks from the spectrum.

The correct molecular formula must be provided before ESSESA begins the spectral analysis. This information is useful to avoid some nonessential analysis and invalid searches. As an example, if the molecular formula shows the presence of only one oxygen atom, then an absorption peak in the 1350–1290-cm^{-1} range will not be interpreted as diagnostic of a sulfone moiety, which would require two oxygens. The appearance of such a combination of elemental composition and spectral absorption would in fact cast doubt upon the premise that an absorption in this range indicates a sulfone.

After the molecular formula has been obtained, the number of double bonds that may be in the molecule is calculated and is itself used as a constraint. If only three rings/double bonds are present in the molecule for example, the presence of a benzene ring is precluded and any absorption lines between 1610 and 1570 cm^{-1} or between 1520 and 1480 cm^{-1} must be explained otherwise. The calculation of the elemental composition and the degree of unsaturation of the molecule before the spectral analysis begins not only avoid invalid structures but also decrease the time required for the analysis.

ACKNOWLEDGMENT

We thank Professor Shi Yaozhen and Mrs. Zhang Haiwan for helpful discussions of infrared spectral analysis. We thank the Editor, Dr. G. W. A. Milne, for redrafting our manuscript.

REFERENCES AND NOTES

(1) Clerc, J. T.; Zupan, J. *Pure Appl. Chem.* **1977**, *49*, 1827.
(2) Lowry, S. R.; Huppler, D. A. *Anal. Chem.* **1981**, *53*, 889.
(3) Jurs, P. C.; Isenhour, T. L. *Chemical Applications of Pattern Recognition*; Wiley: New York, 1975.
(4) Sasaki, S.; Abe, H.; Ouki, T.; Sakamoto, M.; Ochiai, S. *Anal. Chem.* **1968**, *40*, 2220.
(5) Abe, H.; Fujiwara, I.; Nishimura, T.; Okayama, T.; Kida, T.; Sasaki, S. *Comput. Enhanced Spectrosc.* **1983**, *1*, 55.
(6) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. *J. Am. Chem. Soc.* **1975**, *97*, 5755.
(7) Gribov, L. A. *Anal. Chim. Acta* **1980**, *122*, 249.
(8) Shelley, C. A.; Munk, M. E. *Anal. Chim. Acta* **1981**, *133*, 499.
(9) Debska, B.; Duliban, J.; Guzowska, B.; Hippe, Z. *Anal. Chim. Acta* **1981**, *133*, 303.
(10) Funatsu, K.; Susuta, Y.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 6.
(11) Lipkind, G. M.; Shashkov, A. S.; Knirel, Yu. A.; Vinogradov, E. V.; Kochetkov, N. K. *Carbohydr. Res.* **1988**, *175*, 59.

(12) Munk, M. E.; Farkas, M.; Lipkis, A. H.; Christie, B. D. *Mikrochim. Acta* **1986**, *2*, 199.
(13) Woodruff, H. B.; Smith, G. M. *Anal. Chem.* **1980**, *52*, 2321.
(14) Luinge, H. J.; Kleywegt, G. T.; Van't Klooster, H. A.; Van Der Mass, J. H. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 95.
(15) Miyashita, Y.; Ochiai, S.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 228.
(16) Woodruff, H. B.; Munk, M. E. *J. Org. Chem.* **1977**, *42*, 1761.
(17) Nakanishi, K.; Solomon, P. H. *Infrared Absorption Spectroscopy*; Holden-Day: New York, 1977.

Enumeration and Classification of Coronoid Hydrocarbons. 10. Double Coronoids

S. J. CYVIN,* J. BRUNVOLL, and B. N. CYVIN

Division of Physical Chemistry, The University of Trondheim, N-7034 Trondheim-NTH, Norway

Received December 1, 1989

A multiple coronoid is a polyhex with more than two corona holes. Double coronoids with two holes are treated most extensively. They are classified into (a) those obtained by additions of hexagons, (b) primitive, (c) nonprimitive basic, and (d) nonbasic extras. This classification is important under the computer-aided specific generation of the double coronoids. A full account is given for the numbers of nonisomorphic double coronoids with the number of hexagons (h) up to 16. Among the 1618 systems with $h = 16$, the 64 systems under the categories (b), (c), and (d) are described extensively, and their forms are depicted. Machine-computed Kekulē structure counts are reported for many double coronoids, and for some of them also an analytical solution according to the method of fragmentation is shown. Some features of triple coronoids (three holes) are included.

INTRODUCTION

A polyhex is a geometrical object consisting of congruent regular hexagons in a plane. According to a strict definition this system of hexagons is connected, and any pair of hexagons in it either share exactly one edge, or they are disjoint. A coronoid (system) is a polyhex with a (corona) hole of the size of more than one hexagon. For general references to treatments of benzenoids (polyhexes without holes) and coronoids we cite some monographs.[1-6]

Benzenoids and coronoids have chemical counterparts in a subset of polycyclic conjugated hydrocarbons, that is in polyhex hydrocarbons. Two of the molecules corresponding to coronoids have been synthesized, viz., cyclo[*d.e.d.e.d.e.d.e.d.e.d.e*]dodecakisbenzene or kekulene[7] and cyclo[*d.e.d.e.e.d.e.d.e.e*]decakisbenzene.[8] They both belong to the class of cycloarenes.[9] The corresponding coronoid systems are classified as primitive coronoids.[10] The most famous of these compounds is $C_{48}H_{24}$, kekulene; the same name is used also about the primitive coronoid system, which is depicted in Figure 1 (left).

Kekulene is a rather stable chemical compound, characterized as "greenish-yellow microcrystals", "with its extreme insolubility in solvents of all kinds".[7] It should be noted that the molecule actually has a cavity, into which carbon-hydrogen bonds are pointing, this being a characteristic feature of cycloarenes. Yet the steric hindrances[11] are not serious enough to prevent the molecule from being basically planar.[12,13] It can be predicted that the regular hexagonal hexabenzokekulene ($C_{72}H_{36}$), as is depicted in Figure 1 (right), would be extraordinarily stable chemically.[14,15] The coronoid system belongs to all-coronoids, defined in the same way as all-benzenoids[16] (or fully benzenoids[17]). In our case (Figure 1) the all-coronoid has $2^{12} = 4096$ Kekulē structures with 12 aromatic sextets[17] each out of the total number of 7776.

Theoretically there should be no objection against the possibility to synthesize the "double kekulene" $C_{92}H_{38}$ and its extension to an all-coronoid $C_{124}H_{54}$; see Figure 2. These two systems, having two corona holes each, are called double coronoids. A coronoid with one hole should then be termed more strictly a single coronoid.

The smallest (single) coronoid has eight hexagons, while the smallest double coronoid has 13. These two systems are shown in Figure 3 together with their extensions to all-coronoids. However, one should not be misled to believe that every coronoid system can be converted to an all-coronoid in this way; the cases of Figures 1-3 are rather exceptional.

We shall (as usual) identify the number of hexagons of a polyhex (benzenoid or coronoid) by the symbol h. The Kekulē structure count (or number of Kekulē structures) is designated by K. Hence we also speak about the K number.

The $h = 8$ system of Figure 3 was depicted, perhaps for the first time, by Ege and Vogler,[18,19] and later by several others in connection with different theoretical works.[10,20-27] The $h = 12$ system of Figure 3 is the smallest all-coronoid. The possibility to construct what we call all-coronoids was pointed out already by Polansky and Rouvray,[28] while the particular smallest system of this kind was depicted by Bergan et al.,[29] who also gave its Kekulē structure count (see Figure 3). Hydrocarbons corresponding to the $h = 8$ and $h = 12$ coronoids referred to above are not likely to be synthesized because of severe steric hindrances inside the cavities of these (hypothetic) molecules. However, it should be clear after the above description that these systems are of great importance in theoretical and mathematical chemistry. The same can be said about the two double coronoids in Figure 3. The bottom-left ($h = 13$) system is the absolutely smallest double coronoid, while the bottom-right ($h = 17$) system is the smallest double all-coronoid. The former ($h = 13$) of these systems was mentioned, probably for the first time, by Dias,[30] and later by Brunvoll et al.,[10] who also gave the K number of this coronoid.

A large amount of work has been published on the single coronoids. A complete list of references would be too voluminous to be cited here, but the reader is referred to some additional references in the following. With regard to the double coronoids, however, the amount of work done is much more modest. In a recent theoretical work of considerable significance, Hall[26] made allowance for more than one corona hole of a polyhex and referred to the number of holes as the

0095-2338/90/1630-0210$02.50/0