

## Simulation of Infrared Spectra: An Infrared Spectral Simulation Program (SIRS) Which Uses DARC Topological Substructures

J. E. DUBOIS,\* G. MATHIEU, P. PEGUET, A. PANAYE, and J. P. DOUCET

Institut de Topologie et de Dynamique des Systèmes de l'Université Paris 7, associé au CNRS, 1 rue Guy de la Brosse, 75005 Paris, France

Received November 16, 1989

The infrared spectrum of a molecule can be simulated with a computer program called DARC-SIRS. This program recognizes the substructures which characterize the molecule and searches for characteristic substructure/spectra dual relations in a database of nearly 700 fragments. These correspond to "bond FREL" fragments with varying depth and precision and to complex foci, or FOXE, organized by filiation. The spectrum simulator produces spectral bands with frequency, intensity, and width. The similarity of the simulated spectrum in comparison to the experimentally determined spectrum is evaluated with several similarity indices. The implementation of the data files and the system is described, as is the interaction between FREL, FOXE, and version 02 of SIRS.

### INTRODUCTION

Simulation or reconstruction of a spectrum for a given structure is a process that is used in situations of varying complexity. In simple cases, simulation can be used for rapid verification of the structure of a synthetic material. In such cases, the chemist can decide which of several potential reaction products is correct by a comparison of the experimental spectrum with the simulated spectrum. In structure elucidation, this process, which is usually limited to partial spectral data, is used to determine part or all of the structure of the molecule and to screen candidate structures. Although infrared (IR) spectroscopy is used to detect characteristic structural elements, it has not generally been used in computer-assisted structure elucidation systems<sup>1-11</sup> because of the difficulty involved in assigning all the absorption bands. For example, the DENDRAL,<sup>2</sup> CASE,<sup>3,4d</sup> STREC,<sup>5</sup> and SPEK-TREN<sup>6</sup> systems do not use the representative IR curves in their successive comparisons. It should be noted in particular that the PAIRS<sup>7</sup> system starts from an IR spectrum and proceeds through decision trees to determine the presence or absence of the usual functional groups found in organic compounds. The IDIOTS<sup>12</sup> system proceeds from subspectra (ssp) to substructural fragments (SS) by using concentric structures to determine, by an ssp  $\rightarrow$  SS process, the existence in the unknown compound's IR spectrum of absorption bands associated with specific chemical groups. In all of these systems, the aim is to identify, more or less precisely, substructural parts of the molecule on the basis of its IR spectral behavior. Such sp  $\rightarrow$  S characterization provides local structural information but is rarely exhaustive. Local assignments can be used to screen potential structural solutions by creating constraints at the level of the generators of isomeric candidate formulas in, for example, CONGEN,<sup>4d,13</sup> GENOA,<sup>14</sup> CHEMICS,<sup>4</sup> and GENIAC.<sup>15</sup> When predicting an IR spectrum from a structure (S  $\rightarrow$  sp), there are fewer difficulties and constraints because the structural information is not all reflected in the spectrum. Outside the "fingerprint region" IR spectral peaks can generally be related to certain specific structural fragments. When proceeding from the structure to the spectrum, the paucity of known chromophores is no longer a grave handicap because the detail and quality of the simulated spectrum can be greatly improved by the use of a database rich in structure-spectra (SS/ssp) correlates and sound models bearing on the functional groups, their interactions, and those of their environments (ordered topology).

**Objectives of the SIRS System.** By using methods which combine local spectral and structural information, efficient tools for the simulation of the spectra of complex molecules

have been developed. The breadth and diversity of many of the problems examined led to the development of the SIRS system, which consists of the necessary computer programs together with correlation and validation files. Prediction of the spectrum of a simple molecule requires only application of the usual theories of spectroscopy.<sup>16</sup> In complex molecules, however, such procedures are delicate, sometimes ambiguous, and often impossible. In these cases, it has been customary to reconstruct part of a more or less empirical spectrum from the spectral properties associated with specific structural fragments. These simple additive processes can be improved by correlations dealing with various types of internal or external interactions. With computer-managed databases, it has been possible to enrich the set of substructure-subspectral pairs, considered as primitives, and to master fairly complex combinations of these primitives. With *simple systems*, progress has resulted from the use of learning procedures to optimize substructure/subspectra correlations. In this way, ssp/SS correlations have been refined, and their number has been increased; there are now nearly 700 such pairs in version 02 of SIRS. A search for the size of structural primitives required specification of the objectives. These could be either highly refined applications in very narrow sectors or more general systems grouping very diverse molecules. The compromise reached here is based upon a fine-tuned formalization of the definition of substructures, and also on a broad perspective with respect to categories. This is made possible by the choice of *complex reference foci*, polycyclic systems or chemical families, for example, which are treated as open lists.

The characteristics of such a system, which combines broad openness with good SS/ssp correlation, were defined according to the DARC/SIRS applications. To this end, the *analysis of spectral data* and an *efficient organization of the files* were objectives. Further, the tools discriminate sufficiently to guide choices between various structural assignments, either *directly* or as a supplementary contribution in the final stages of *elucidation*. These are mixed strategies, S  $\rightarrow$  sp and sp  $\rightarrow$  S, combining various spectroscopic methods.

**Bases of the SIRS System: Structure-Spectra Assignments.** Prediction of the IR spectrum that will be given by a specific structure can be accomplished by development of a molecular model and by determination of the normal vibration modes from empirical force fields using the usual procedures of classical mechanics.<sup>17-19</sup> In the quantum chemical approach, one proceeds by calculating the potential energy of the system, using commonly available programs such as MNDO or GAUSSIAN.<sup>20</sup> This approach is computationally intensive and, consequently, is not used commonly and is applied most fre-

quently to simple molecules. This explains why software that is available for use on microcomputers tends to be oriented toward simulation of experimental conditions for a spectrometer (NMR SIMULATOR and IR SIMULATOR)<sup>21</sup> or toward learning characteristic frequencies (EXP'AIR)<sup>16</sup> for didactic purposes.

This paper describes the automatic simulation of the IR spectrum of a compound by a computer program called SIRS (Simulated InfraRed Spectra).<sup>22</sup> The program works with topologies and has been developed within the framework of the DARC system. It reconstructs the characteristic elements of the spectrum of a compound whose structure is known. By a process similar to that ordinarily used by chemists for spectral analysis, the SIRS software identifies the relevant structural fragments of the compound and generates the spectrum from the appropriate spectral elements<sup>23</sup> that have been organized in the form of subspectra/substructure pairs. It is first necessary to describe the generation of the four-level hierarchical database, and then spectral generation will be considered both qualitatively and quantitatively. Finally, the quality of the simulated spectra is evaluated by means of similarity procedures and indices which are contained within the SIRS software.

### CREATION OF THE DATABASES

In spectral analysis, the chemists usually relies upon the identification of recognized characteristics of spectral elements provided by specific structural moieties. This expert knowledge is expressed in terms of substructure/subspectra relationships. In computer-assisted spectral simulation such relationships are best defined by building a database of substructures and one of spectral features and by establishing links between them. Such a SS  $\rightarrow$  sp pair is termed here a "primitive couple".

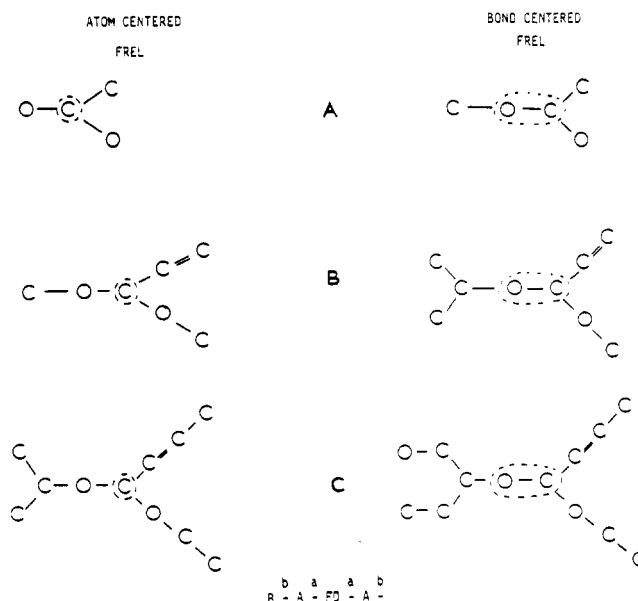
**Subspectral Database.** The validity of the model is determined by the relevance of the substructure/subspectra relationships. These can be retrieved from databases with many spectra, but in IR spectroscopy the information that has been derived from large numbers of spectra has enabled the production of secondary information in the form of correlation charts which allow the association of spectral bands with specific structural fragments or functional groups.<sup>24-29</sup> The current version of SIRS makes use of this level of information. The spectral database consists of the characteristic group frequencies as they are found in most standard textbooks. For uncommon structural fragments which are frequently absent from such lists, and for functional groups situated in an unusual environment (e.g., very hindered systems), it is still necessary to use primary information. The databases used by SIRS will evolve as they are updated on a regular basis.

For purpose of spectral simulation, SIRS-02 uses only those frequencies which unambiguously show the presence of a structural grouping. Certain bands whose frequencies are highly dependent upon skeletal structure (e.g., C-C and C-N stretching) and are not shown in the correlation charts have been eliminated. As a consequence, the quality of the simulation in the fingerprint region is diminished.

For each absorption band in the database, the following characteristics are retained:

- the frequency range where the signal is found
- the peak intensity—five levels are used: weak, medium, strong, very strong, and variable
- the line shapes, identified in terms of the half-height width and an identifier specifying type (Gaussian, Lorentzian, and so on)

Finally, when necessary, a comment identifies the attributes (valency, deformation, harmonic vibration, combination band, and so on). This information is not used by the current SIRS version but it has been incorporated into the database and will



**Figure 1.** Concentric substructures: FRELs—Fragment Reduced to an Environment that is Limited. A capital letter A, B, C ... designates the topological radial distance of the environment around the focus which may be an atom or a bond. Lower case letters a, b, c ... are associated with the anterior bonds. The FREL can be identified as FREL-a (explicit bonds and nonidentified atoms), FREL-A (row A explicit), FREL-b (row A, b bonds explicit and B atoms nonidentified), or FREL-B (rows A and B explicit).

be used with future SIRS versions that will allow for the influence of symmetry leading to the attenuation of some stretching vibrations. It will also be used for the generation of programs which can didactically learn characteristic frequencies. This comment field will eventually supply information concerning changes in the physical state. Display of the associated comments is optional and will facilitate interactive modification of the spectral data.

**Substructure Database.** The information in IR spectra is linked intrinsically to bonds, and an algorithmic fragmentation logic can be used. In about 10% of the cases, however, the IR signal cannot be linked to a single bond, and in these cases a whole fragment must be considered. This necessitates the formation of a database of identified fragments.

(A) *Algorithmic Concentric Substructures: FRELs.* Spectral information that is linked to a bond depends upon the bond's molecular environment. The active environment of a bond is identified in a particularly efficient way by a topological concentric structural exploration. This idea of concentricity lies at the heart of the DARC system. Such a topological approach to the definition of substructures allows expression of their formal links. The relevance of the SIRS fragmentary model is based upon the complexity of the various theoretical filiations that can be established between concentric substructures.

**FREL Concept.** The FREL concept implies a topologically limited and organized substructure. Where substructures are concerned, the concentric nature of the environment of a site is expressed by the concept of the Fragment Reduced to an Environment that is Limited or FREL<sup>30-34</sup> defined about a focus. The focus can be the atom at the center of the FREL (as in substructure query language of compounds<sup>35</sup> or the DARC EPIOS <sup>13</sup>C NMR elucidation system<sup>36</sup>), a bond and the two atoms it connects (as in the substructure query language of reactions<sup>37</sup> or the IR simulation system described here), or a group of atoms or bonds (as in most of the DARC PELCO correlations<sup>38,39</sup>). In the last situation, this complex focus is termed a FOXE.

A typical FREL is shown in Figure 1 and is defined by:

- its focus.
- its environment described in terms of concentric organization. The scope of the environment is symbolized by the letter A, B, ... that designates the topological distance between the focus atom and further atoms.
- the chromatism authorized at the different sites. This describes the nature of the atoms and the multiplicity of the bonds. The nature of the sites can be clearly displayed with color graphics, hence the use of the word *chromatism*.

An order is imposed upon the environment and the formal generation of an ordered group representing the FREL creates filiation relationships between the products of successive stages of the FREL. Thus a FREL of row  $n$  must include the FREL from row  $n - 1$ , which can be regarded as its son. The filiations that can be so defined between substructures can be used in spectrum similarity searches for members of a single chemical family. This original DARC concept, linking structures to one another in controlled filiations, is sometimes referred to as a heredity function (father FREL-son FREL) imposed upon the descendants in a graph structure of a space of states.<sup>40</sup>

**Chromatism Used.** The characteristic substructures in the database contain the more frequently occurring elements in the full CAS file; H, C, N, O, F, Cl, Br, and I. The bonds are single, double, triple, or aromatic. Tautomeric bonds are not allowed. This requires that the individualization of the spectra in terms of each tautomeric species must be handled separately.

Three additional characteristics, defined as *secondary chromatism*, are used in the description:

- the charge that is linked to the indication of an unusual valency. This allows description of the whole set of charged atoms, according to the DARC rules.
- the atom hybridization. Hybridization character is linked to interactions involving conjugation and homoconjugation effects that significantly modify the position of stretching bands. These effects rely mainly on the multiplicity of the bond originating from the neighbor A (conjugation) or B (homoconjugation). Two types of hybridization ( $sp^3$  and non- $sp^3$ ) are noted. No distinction is made between  $sp^2$  and  $sp$  hybridizations.
- the cyclic nature of bonds. Constraints resulting from the presence of a bond in a ring often modify the IR frequencies of the various fragments in different ways. Cyclization chromatism allows specification of the cyclic or acyclic nature of a bond and also the number of atoms of the ring concerned. Bonds in rings with 6 or more atoms are assimilated with acyclic bonds because the resulting stretching frequencies are usually close. For bonds at ring junctions, only the size of the smaller ring is considered because it is this ring that provides the more significant frequency constraints.

**Levels of Knowledge: Scope and Chromatic Fuzziness.** The effects of the environment on the vibrational frequencies associated with each bond are described by limiting the environment to row B, i.e., two atoms from the focus. This choice leads to a good description of most of the usual functions which is consistent with the very weak influence on the vibration frequency of the focus of atoms in the C row and beyond, except in particular cases that can be dealt with either by secondary hybridization criteria, cyclization, or specific fragments considered as FRELs reduced to complex foci or FOXEs (see below).

**Table I.** Atom Chromatism for the Four FREL Levels<sup>a</sup>

FREL	positions		
	$F_0$	$A_i$	$B_{ij}$
level 1	H, C, N, O, F, Cl, Br, I	0	0
level 2	H, C, N, O, F, Cl, Br, I	H, C, N, O, X, F	0
level 3	H, C, N, O, F, Cl, Br, I	H, C, N, O, X, F	R, N, O, X, F
level 4	H, C, N, O, F, Cl, Br, I	H, C, N, O, X, F	H, C, N, O, X, F

<sup>a</sup>R = (C, H). X = (Cl, Br, I). C for  $sp^3$  C and \*C =  $sp^2$  C or  $sp$  C.

In organizing the substructure database, the fragments must be arranged in such a way as to avoid redundancy. It is important however that when the required substructure is absent from the database it is possible to make use of the information associated with neighboring fragments which are present. This is made possible by the use of *four levels of precision* and by managing their filiation efficiently. For this purpose, two parameters are used: variable depth of substructures and fuzzy chromatism. Atoms that have a similar influence on the spectral features of the fragments are not distinguished. This reasoning is analogous to that which in chemistry uses the generic symbol X to denote any of the four halogen atoms F, Cl, Br, or I. The reduction of fuzziness and the increase in the substructure dimensions help to refine spectral information, as is shown in Figures 2 and 3.

Fuzzy chromatism creates equivalent classes in the substructure/subspectra relationship. This grouping considers both the nature of the atoms and their distance from the focus in order to master the equivalence between the spectral precision and the structural precision. In this way, all halogen atoms except F can be merged on A positions and even more so on B positions. Similarly, hydrogen and  $sp^3$  carbon can also be merged on the B row for the third level. For the most precise level, the influence of alkyl ramification leads to distinction between these H and ( $sp^3$ ) C atoms. This is seen in the chromatism data given in Table I.

Given a maximum width of two rows of atoms, four bond chromatisms, eight atom chromatisms, and secondary characteristics such as hybridization, charge, and cyclization, a very large number of FRELs are theoretically possible. This is a reason for an architecture which has four levels of knowledge, in which the localization of available information avoids all redundancy and does not lead to useless searching.

Once the vibrational groups (FREL and Focus) that are to be considered are chosen, it is necessary to eliminate those for which there is little spectral data or which do not seem to possess characteristic frequencies. Also eliminated are those groups which have a very low occurrence frequency or those which, like aromatic bonds, are handled by complex focus fragments (FOXEs). This elimination process leaves 43 Focus FRELs at the primary level (level 1).

The 43 Focus-FRELs of the first level give access to the 306 FRELs-A retained on level 2 as shown in Table II. This second level defines the neighbors around the focus.

For FRELs-B (focus augmented by two rows of atoms) of higher levels 3 and 4, and differing in their chromatic fuzziness, only four well-defined focuses are retained. These are shown in Table III.

This type of description, which exploits structural filiation, is very open. It permits integration of additional refinements for specific vibrating groups into one of the four levels and so facilitates more specialized studies. [In Tables II and III, following conventional structural adaptation rules,<sup>41</sup> the donor bond between nitrogen and oxygen is shown as a double bond, indicating a pentavalent nitrogen, while the donor bond between nitrogen and carbon (present in isonitriles, for example) is incorporated into the database as a single bond joining a tetravalent nitrogen to a trivalent carbon.]

(B) *Fragments Reduced to Complex Foci: FOXEs.* Sometimes, a description involving two rows of atoms around

**Table II.** Selected IR Frel-Focus with Stretching Vibration Range and Number of Related FREL A

no.	$F_0^a$	no. of FREL A	max <sup>b</sup>	min <sup>b</sup>	range
1	C-H	20	3100	2710	390
2	*C-H	9	3390	2650	690
3	C=C	35	2030	1580	450
4	C <sub>3</sub> =C	1	1900	1865	35
5	C <sub>4</sub> =C	2	1695	1555	140
6	C <sub>5</sub> =C	2	1685	1600	85
7	C <sub>6</sub> =C	3	1690	1580	110
8	C≡C	8	2310	2100	210
9	N-H	8	3550	2800	750
10	*N-H	1	3400	3300	100
11	+N-H	4	3390	1800	1590
12	**N-H	2	2700	2250	450
13	N=C	34	2285	1530	755
14	+N=C	7	2100	1550	550
15	N=N	3	1575	1410	165
16	N=N <sup>+</sup>	4	1530	1285	245
17	+N=N <sup>-</sup>	2	2075	2000	75
18	V <sub>3</sub> N=NV <sub>5</sub>	2	1425	1175	250
19	C=N	3	2305	2175	130
20	C≡NV <sub>5</sub>	1	2305	2285	20
21	V <sub>4</sub> N=CV <sub>3</sub>	2	2175	2115	60
22	O-H	6	3670	2500	1170
23	O-C	18	1275	900	375
24	O <sub>3</sub> -C	3	1260	1240	20
25	O <sub>4</sub> -C	1	1040	970	70
26	O <sub>5</sub> -C	9	1265	1040	235
27	O <sub>6</sub> -C	3	1270	1030	240
28	O-C*	13	1320	980	340
29	O <sub>5</sub> -C*	3	1370	895	475
30	O-O	4	900	830	70
31	O <sub>5</sub> -O	1	1005	1040	25
32	O <sub>6</sub> -O	3	900	830	70
33	O=C	45	2285	1550	735
34	O=N	4	1680	1435	245
35	O=NV <sub>5</sub>	9	1640	950	690
36	F-C	6	1365	1000	365
37	F-C*	3	1340	1100	240
38	Cl-C	6	800	505	295
39	Cl-C*	5	890	370	520
40	Br-C	3	650	485	165
41	Br-C*	3	650	485	165
42	I-C	3	600	200	400
43	I-C*	2	600	185	415
total A: 306					

<sup>a</sup>Secondary atom chromatisms are indicated by V\* (see text). <sup>b</sup>max and min (cm<sup>-1</sup>) refer to the upper and lower values of stretching vibration frequency.

**Table III.** FREL A, FREL b, and FREL B Occurrences in the Databank for Various Focuses

Focus $F_0^a$	FREL A	FREL b	FREL B
C=O	45	131	64
C≡N	3	9	4
V <sub>3</sub> C≡N V <sup>*</sup>	2	2	0
C≡N V <sup>5</sup>	1	2	0

<sup>a</sup>Secondary atom chromatisms are indicated by V\* (see text).

the focus does not sufficiently reflect spectral reality. Situations of this sort include delocalized vibrations on a set of bonds (e.g., aromatic systems) and vibrations whose category was not previously retained. An example of the latter is given by the deformation vibrations of the isopropyl group, which give rise to characteristic frequencies, while single C-C bonds provide little information and are usually ignored in SIRS. Another type of situation is seen with structural moieties

**Table IV.** Tricyclic Complex Focus (X = Free Site): FOXEs

no.	substructure
1	
2	
3	

between which differentiation is not possible at level 4—the most precise level retained. An example is provided by the anhydride structure, which requires three rows of atoms around the focus for a complete definition. Such difficulties can usually be resolved by including an extra row of atoms, but the complexity of managing an additional level of precision is not justified given the small number of cases in which this difficulty arises.

Instead of refining relationships between the various FREL categories, it appeared to be preferable to gather together these diverse situations and provide for them a homogeneous computer solution based on the structural moieties handled as fragments reduced to complex foci, or FOXEs. Ninety-four fragments were selected. They varied in size from 3 non-hydrogens (isopropylene) to 14 (anthracene). Their introduction requires assurance that when consulting the database a bond will not be handled several times, once as a FREL, again as a FOXE, and yet again within two FOXEs.

Among the 69 cyclic fragments, SIR-02 identifies the following:

- 3 tricyclic compounds where the eventual existence of substituents is ignored (Table IV).
- 16 bicyclic compounds in which only substitution on a single bond is considered. This substitution is standardized as Z = C, O, F, Cl, Br, I; the other ring preserves its free sites (Table V).
- 37 monocyclic heterocycles with 5 or 6 atoms at which free sites are allowed (Table VI).
- 13 diversely substituted benzene derivatives differentiated by the number and localization of these substituents (Table VII).

To these must be added 25 acyclic fragments that permit differentiation of amines according to their carbon environment or specification of some aliphatic fragments such as *t*-Bu, *i*-Pr, Et<sub>2</sub>C, and acid anhydrides or peranhydrides (Table VII).

## SIMULATING AN IR SPECTRUM

The procedure for simulation of an IR spectrum by SIRS is executed by the SIMULIR software, which is composed of three modules of diverse complexity and organization. Given a target molecule defined by its structural formula, the programs operate in turn to:

- determine structural characteristics by identifying complex foci and searching for bond FRELs in a bond-by-bond exploration of the structural formula
- retrieve the relevant absorption frequencies from the database
- edit the spectrum that results as the sum of the extracted frequencies

These modules are discussed in more detail in this section.

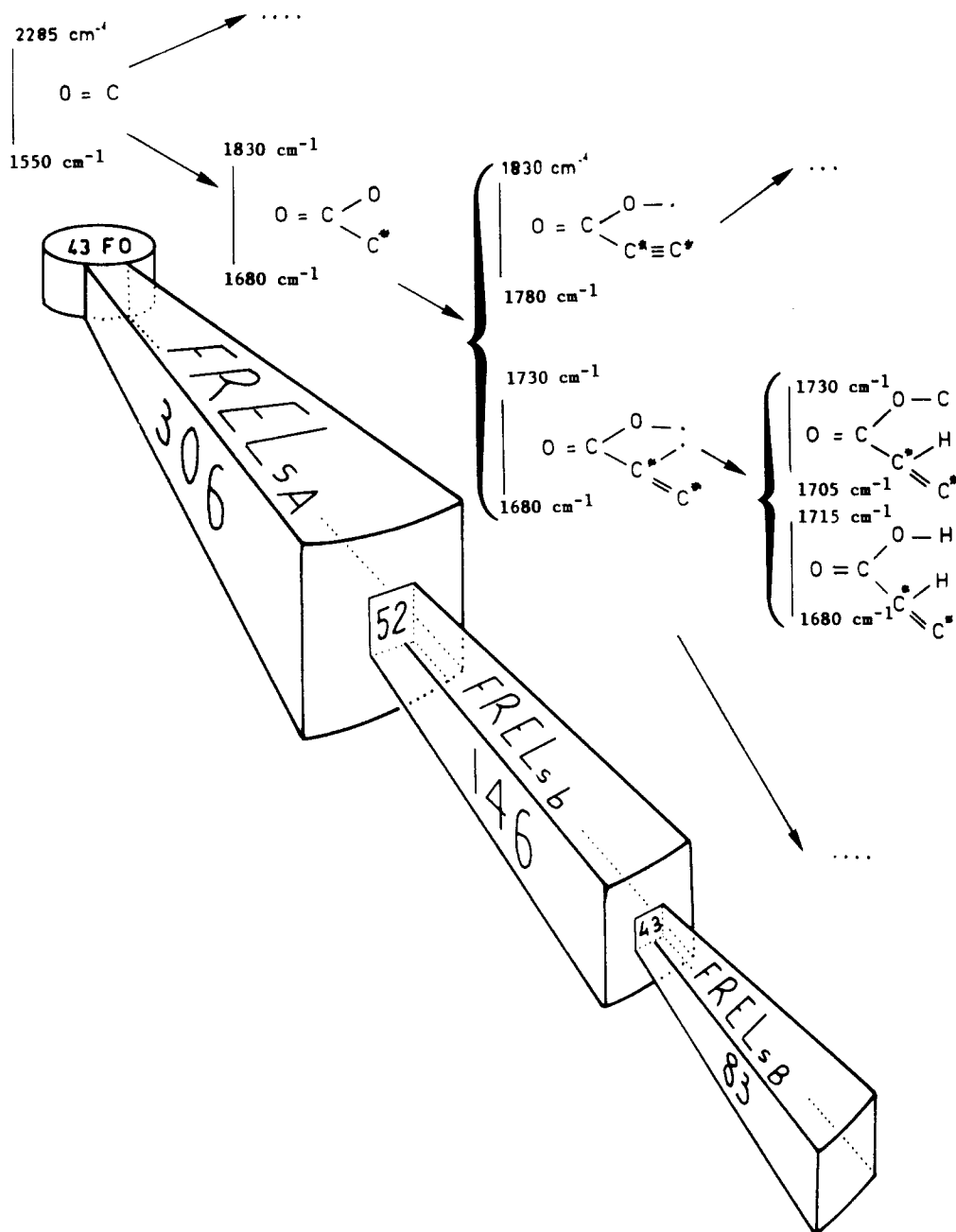


Figure 2. Four levels of knowledge: spectral ranges decrease with structural refinements (variable depth and fuzzy chromatism).

**Module I—Acquisition of the Structure and Identification of the Structural Information.** A high degree of flexibility is desirable during input of the structure, and identification of the relevant substructures must be carried out in a manner that is transparent to the user.

**Structure Input.** Input of the compound's structural formula can be done either graphically or alphanumerically and is done in the same way as the database structures were entered. A new structure may also be extracted from an existing file.

**Determining Structural Characteristics.** Certain structural facts, which can be deduced from the structural formula and which are associated with the chromatism of the atoms or bonds, are established by this program. Determinations are made successively of the hybridization of each atom ( $sp^3$  or not) from the multiplicity of the related bonds and the number of hydrogens at each atom. Hydrogen is implicit in the topological description because its presence or absence can be determined from atomic valencies.

By successively eliminating nodes of degree 1 (atoms with only one non-hydrogen neighbor), it is possible to differentiate

acyclic from cyclic atoms and then seek the size of the smallest ring to which the latter belong.

**FOX Identification.** A homomorphism procedure detects the presence of complex foci. A distinction is drawn between substructures (SS) that can or cannot be structurally included, i.e., included in the more important SS. For those that cannot, such as those involving acyclic or benzene substructures, it is sufficient to identify their presence. It is more difficult to handle polycyclic and monocyclic moieties where numerous free sites are possible. Thus, the quinoline nucleus subsumes the pyridine nucleus, and both will be among the substructures that are retained. When a substructure is totally embedded in another, as in this case, the best solution is to retain the larger of the two. On the other hand, two substructures may overlap only partially; as an example, phenanthridine contains a quinoline and an isoquinoline nucleus. In such cases, the various possibilities must all be retained, even though this will lead to some redundancy in the spectral information. After the two families of complex foci are treated, all the bonds involved are kept in memory, to avoid redundancy of spectral information during the processing.

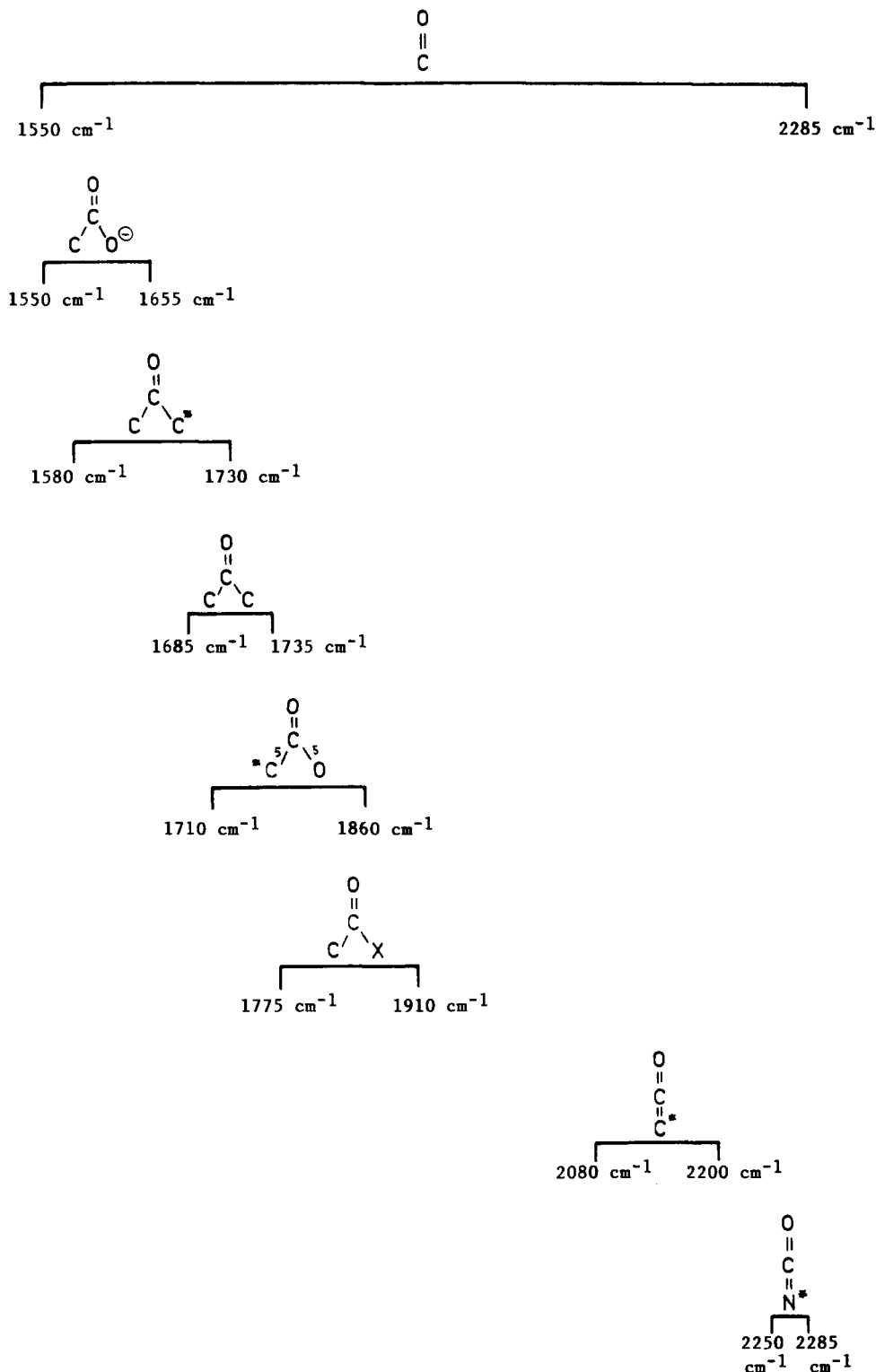


Figure 3. Splitting C=O frequencies in various structural environments by taking account of only the first neighbors (FREL-A level).

**FRELS Output.** Examination of the structure on a bond-by-bond basis allows the extraction of concentric substructures with a maximum of two rows of atoms, or FRELS-B, around each of them, with the exception of bonds that belong to particular fragments and those that are a priori ignored (e.g.,  $\text{sp}^3 \text{C} - \text{sp}^3 \text{C}$ ). Canonical ordering and identity searching allows elimination of specific redundant FRELS-B.

**Module 2—Exploiting the Data, Search for Correlated Subspectra.** The two structural subsets, FREL and FOXE, are next organized to permit searching.

**Identification of Compound FRELS and Base FRELS.** FRELS extracted from the proposed structure are searched on the basis of homomorphism in the database files. Regis-

tered filiations between FRELS of different levels of knowledge lead to a tree structure which can be read from top to bottom or in the other direction. For example, a FREL-B [ $\text{O}=\text{C}(\text{OH})(\text{CH}=\text{C})$ ] (see Figure 2) extracted from a structure can be searched for by first identifying the FREL-FO [ $\text{O}=\text{C}$ ] and then the FREL-A [ $\text{O}=\text{C}(\text{O})(\text{C})$ ] from among the sons of the FREL-FO and so on. Alternatively, the FREL-B itself can be searched for directly in level 4; if such a search fails, the corresponding FREL-b [ $\text{O}=\text{C}(\text{O})(\text{C}=\text{C})$ ] is sought for in level 3 and so on. It is found that the number of hits that must be managed is minimized when the direction goes from the least to the most precise, i.e., FREL-focus  $\rightarrow$  FREL-A  $\rightarrow$  fuzzy FREL-B  $\rightarrow$  specific FREL-B. Each level, when rec-

**Table V.** Bicyclic Complex Focus (X = Free Site): FOXEs<sup>a</sup>

no.	substructure	no.	substructure
1		9	
2		10	
3		11	
4		12	
5		13	
6		14	
7		15	
8		16	

<sup>a</sup>Substructures or FOXEs here differ through location and number of substitution and/or chromatic bonding ( $\sigma$  or  $\pi$ ).

ognized, cancels out the previous level.

This process, which is repeated for all extracted FRELs, is followed by an occurrence calculation, so that each identified FREL is cited only once.

**Searching for Related Subspectra.** The spectral data related to FRELs and to FOXEs are simply extracted from the database.

**Module 3—Producing and Editing the Simulated Spectrum.** Editing of the spectra is one of the problems encountered when building spectral databanks. The display of a spectrum from a library requires either storage of the entire analogue curve or regeneration of the curve, using several significant parameters. The necessary transformations then lead to a coding of the different absorption bands and of their intensities,<sup>17,42</sup> with an eventual different resolution of the experimental bands.<sup>43,44</sup> Codification of spectra involves various simplifications, which imply a loss of information.<sup>1a,3a</sup> Gribov,<sup>5b</sup> by a simplified coding process, derives a spectrum which is quite different from the real spectrum. Zupan,<sup>1c,42,45</sup> on the other hand, using a multiple spectroscopic approach, obtains a reconstituted IR spectrum which is similar to the experimental spectrum. This process is undoubtedly assisted by the use of two supplementary parameters—the half-height width and the band shape, classified as either Gaussian or Lorentzian.

In this current version of SIRS, the simulated spectrum is derived by simple concatenation of identified bands. Some examples are shown in Figures 4 and 5.

Each bond, identified by the indices of the atoms involved in the molecular graph, is cited in turn. When one of the terminal atoms is a hydrogen, its index is assigned as zero

because hydrogen atoms are omitted from the input formula. Next, the number of the identified FREL and its level are indicated. A FREL number of zero corresponds to zero identification or to a bond handled as a complex focus. Finally, the FOXE is displayed.

The IR data related to the FRELs and/or FOXEs are then retrieved sequentially.

The simulated spectrum that is obtained can be displayed in several ways, as shown in Figure 4. In the first display, the simulated band is represented by a peak situated in the middle of the frequency ranges that is indicated, and with a length equal to the band intensity given in the database. The advantage of this display is that the limits within which the experimental band must be found are clearly shown. A different display, which more closely resembles the actual spectra, is obtained if the band is represented as a Lorentzian curve (or eventually a Gaussian curve for OH and NH bands).

#### EVALUATING SIMULATED SPECTRA: NOTIONS OF SPECTRA SIMILARITY AND RESEMBLANCE

The degree of correspondence between a simulated spectrum and the experimental spectrum is an interesting question whose answer depends upon the context. In a teaching environment, the appearance of the band is important to guide reasoning, but when using a structure elucidation system there will be more concern over the degree of correspondence between the two. These perspectives have been considered in the development of software for qualitative and quantitative comparisons.

To test the validity of the SIMULIR software, compounds were chosen to represent a broad range of the functional groups normally encountered in organic chemistry. The simulated spectra that were obtained were compared to standard spectra from databases. Most of the standard spectra were taken from the ALDRICH collection, which has some 10 000 IR spectra recorded between 4000 and 400  $\text{cm}^{-1}$ . A direct preliminary comparison was carried out with 100 compounds.

**Missing Information.** As a result of the simulation process, as described above, it is probable that the simulated spectra will be incomplete; some peaks, particularly in the fingerprint region, will be missing. There are other reasons for the incompleteness of simulated spectra:

- *Simulated spectra are abbreviated* when compared to experimental spectra. When simulating  $^1\text{H}$  or  $^{13}\text{C}$  NMR spectra, it is possible, except in cases of magnetic equivalence, to produce spectra in which there is a one-to-one correspondence between a specific atom and the corresponding chemical shift. Infrared spectra, by way of contrast, contain uncharacterized skeletal vibrations which have to be ignored. Moreover, the higher harmonics and combination bands are not generally recorded in correlation tables because of their low intensity. For these reasons, simulated spectra usually have fewer peaks than the experimental spectra.
- *The effect of molecular symmetry on transition moments is ignored* by the current version of SIRS. For instance, in the trans-disubstituted alkenes, the  $\text{C}=\text{C}$  stretching vibration disappears when the substituents are identical and it is weakened when the substituents are merely analogous. Thus, the simulated spectrum can suggest absent or weak signals in the experimental spectrum.
- *The physical condition of the compound* has a considerable effect upon the absorption intensity.
- *The knowledge stored in the SIRS database* is based upon the usual rules of interpretation, used

Table VI. Monocyclic Complex Focus (X = Free Site): FOXEs

no.	substructure	no.	substructure	no.	substructure	no.	substructure
1		11		21		31	
2		12		22		32	
3		13		23		33	
4		14		24		34	
5		15		25		35	
6		16		26		36	
7		17		27		37	
8		18		28			
9		19		29			
10		20		30			

in secondary collections of data. It is not exhaustive, and certain fragments, with their related signals, will be missing. The completeness of the database is most reliable for groups that occur commonly, and these are groups that have been encountered frequently in synthetic organic chemistry.

#### Similarity Indices of Experimental and Proposed Spectra.

Aside from some cases involving symmetry, the comparisons or matchings that were carried out show that the simulated bands can nearly always be identified in the experimental spectrum. The intention of the program was to ensure that these matches succeeded by basing them upon important absorption bands. The quality of the simulation can be quantitated with various types of criteria. Different similarity (dissimilarity) indices have been proposed. Those based on calculations of the Euclidean distance between two spectra<sup>44</sup> were eliminated because of the generally large range possible for each simulated frequency. Zupan's empirical formula, which takes band coincidence into account, seems better suited to this task. Two indices were therefore established:  $F_1$  (absolute) and  $F_2$  (relative). These differ according to whether or not they consider the number ( $m$ ) of simulated bands and are as follow:

$$F_1 = \sum_{i=1}^m A_i (B_i + C_i + D_i)$$

$$F_2 = \frac{1}{m} \sum_{i=1}^m A_i (B_i + C_i + D_i)$$

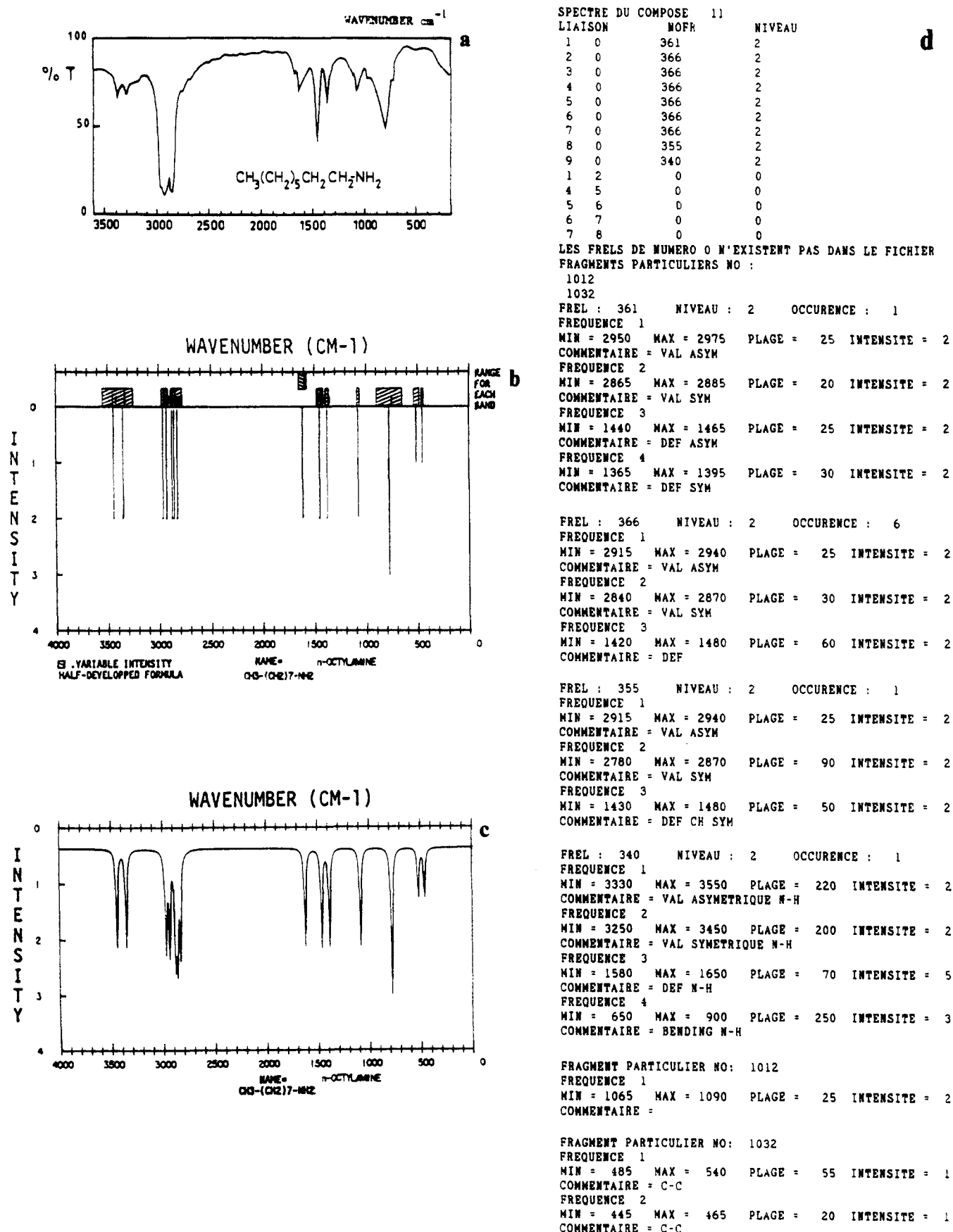
where  $F_2 = F_1/m$  and  $A_i$  is equal to 1 if the simulated band coincides with the experimental band, otherwise, 0. This coincidence is affected by three factors:

- $B_i$  represents the difference in the intensities of the two bands. The resemblance is better if the experimental and simulated bands have similar intensities.
- $C_i$  is linked to the breadth of the frequency range for the simulated band. The fit is better when the allowed range is narrow.
- $D_i$  a composite factor of shape, set to 0.2 if the simulated band is an N-H or an O-H absorption. This avoids the incorrect matching of these bands with the C-H vibrational bands in the same spectra region.

The simulated and experimental spectra are compared by double matching of their bands. First, for each simulated band, the best match is sought in the experimental spectrum. Next, the same search is carried out starting with the experimental bands. Such double exploration is necessary because one band in one of the spectra may be associated with several bands in the other spectrum as a result of the range allowed for each frequency. When this procedure was applied to the test population, the percentage of simulated bands correctly recognized in the experimental spectra varied between 70% and 100% with an average of 91%. On the other hand, the percentage bands in the experimental spectra that were recognized correctly varied from 20% to 90% with an average of 55%. The final matching derived from this double comparison carried out on the test population lead to values for  $F_1$  between 2 and 15 with an average of 8.1 and for  $F_2$  between 0.40 and 0.80 with an average of 0.58.

When these numerical results were compared with visual evaluation by spectroscopists, it was found that the minimum quality necessary in simulated spectra required 70% of the assigned simulated bands and a 0.47 average for  $F_2$ . It has





**Figure 4.** Simulation of the infrared spectrum of *n*-octylamine: (a) experimental spectrum, (b) simulated spectrum with peaks in the center of the authorized frequency ranges, (c) with absorption bands represented as Lorentzian curves, (d) listing of the recognized structural fragments with the associated frequencies. Similarity indices:  $F_1 = 9.9$ ;  $F_2 = 0.66$ .

been proposed<sup>2c,36a</sup> that the resemblance between an actual spectrum and one contained in a database is acceptable when there is 80% or higher coincidence between their lines. To date,

similarity indices have not proved to be sufficiently accurate for the evaluation of the quality simulated spectra, and qualitative appraisal of relevance by experts is often more

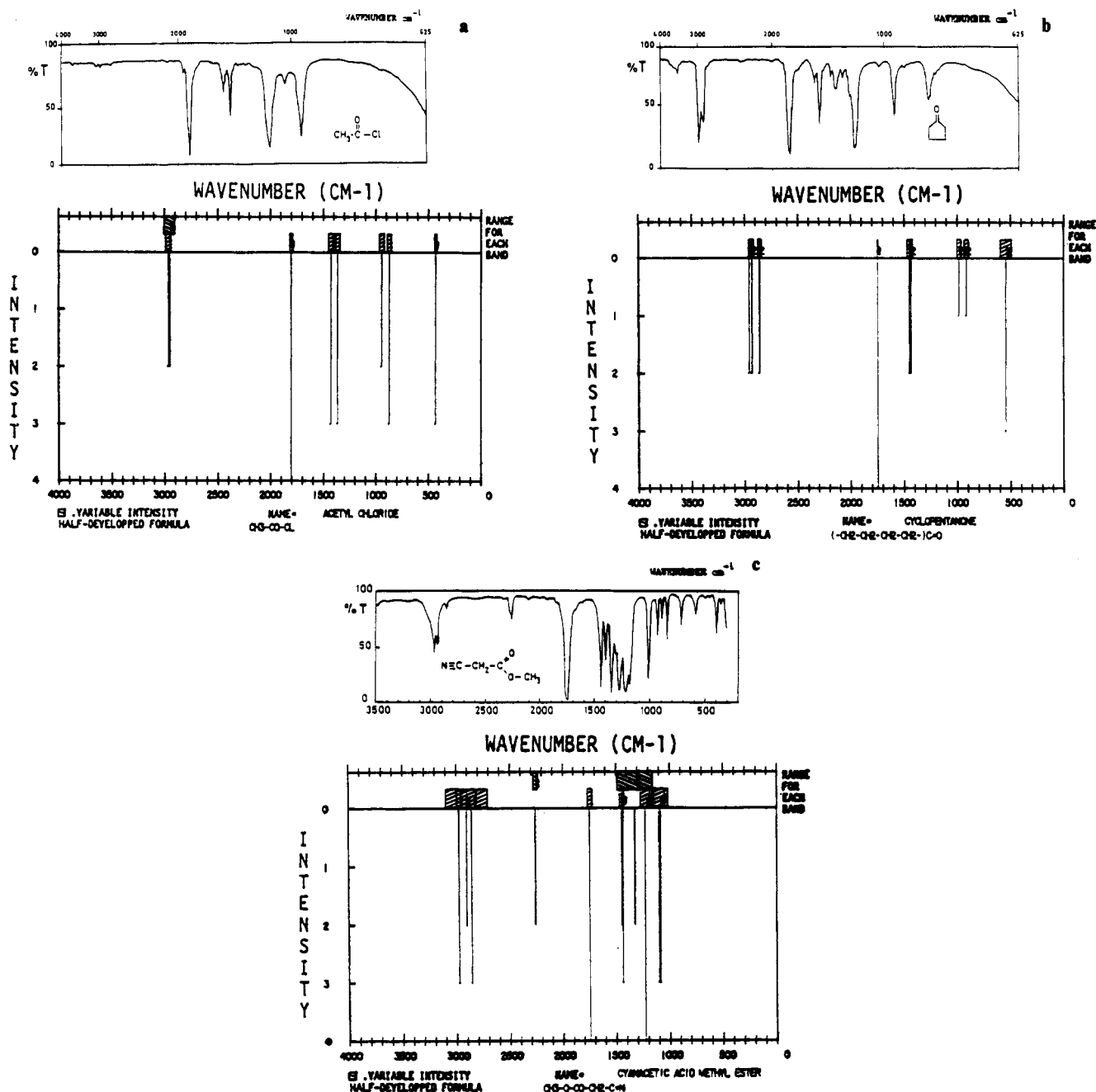


Figure 5. Examples of simulated infrared spectra: (a) acetyl chloride,  $F_1 = 5.4$ ,  $F_2 = 0.77$ ; (b) cyclopentanone,  $F_1 = 7.11$ ,  $F_2 = 0.78$ ; (c) methyl cyanoacetate,  $F_1 = 6.5$ ,  $F_2 = 0.54$ .

appropriate. Efforts to achieve better quantitative pattern recognition of spectra and to avoid some of the discrepancies that accompany similarity index methods are under consideration as a part of the SIRS development effort. It is hoped that, with the help of computerized simulation methods such as those employed by SIRS, it will prove to be possible to gain a better appreciation of the relevance of the generalization rules guiding the translation of the secondary information in the charts correlating the characteristic frequencies when compared to the diversity of the primary information (the published spectra). This in turn should help to improve the learning processes required to generate new primitive couples.

#### SIRS IMPLEMENTATION

In this section, the general construction of the database and the design of the IR spectral simulation software are described.

**Building the Database.** The IR database requires both structural information and the related spectral data (see Figure

6). An absolute number (NOFR) is used for each specific FREL or FOXE in the database. This guarantees the link between the structural description and the spectral data.

The structure input program can accept both graphic and alphanumeric information. These are registered as connection tables which are then ordered. The ordering program proceeds through successive layers of atoms around the focus, and the automatic verification procedures ensure the uniqueness of registered substructures and the coherence between the announced FREL level and the breadth of the proposed chromatism. By structural truncation, the father FREL is built and its identifier is sought. This can be optionally displayed to assist in validation.

Each FREL level is registered in its own file. To facilitate the extraction and comparison procedures, certain direct access files were selected for output. When simulating the IR spectrum, direct access to a given FREL is necessary and is accomplished by means of an index. It is also necessary to collect together all the son FRELs from one father FREL in

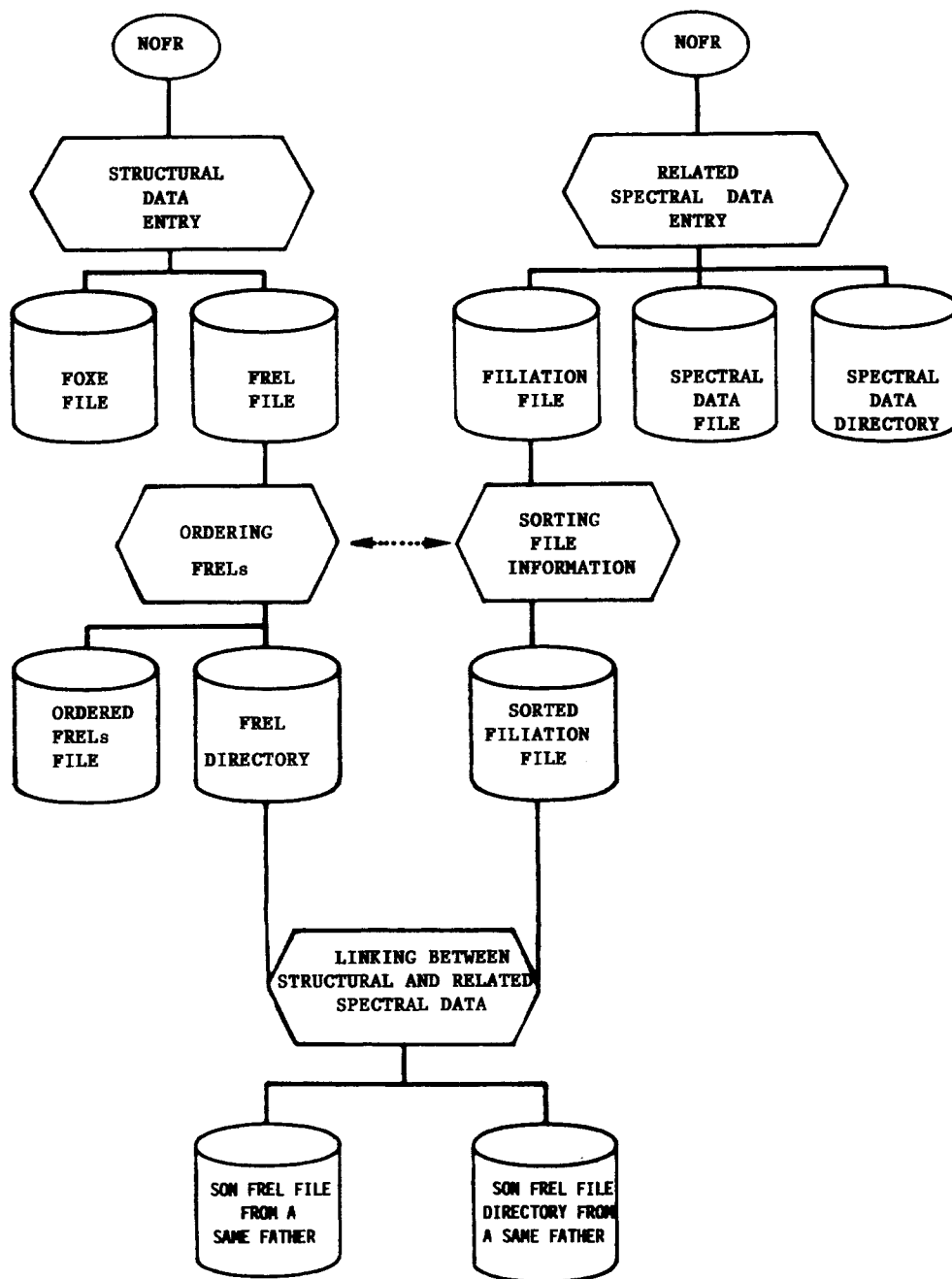


Figure 6. Building scheme of the database.

order to determine their number and position in the file. These data are grouped in an indexed sequential file.

Spectral information corresponding to records of greatly varying length are also grouped in an indexed sequential file. Validation tests are carried out online, and when appropriate, the registration process is terminated. Programs which handle deferred modification proceed very simply by erasing and rewriting, the NOFR identifier of the substructure in question ensuring the links between the spectral and structural files.

The database is maintained on a VAX 11/780. The structural descriptor files occupy some 700 blocks of 512 octets and those of related data some 160 blocks.

**Simulating IR Spectra.** The structural formula of the compound is acquired interactively by the same programs that were used to enter the structures that are in the database. Hybridization of atoms and cyclic character of bonds are deduced automatically and need not be specified. The structural formula can also be obtained by reading its connection table from a previously created registry file.

The name and structure can be entered with a view to final

output. For this purpose, they are stored in the FSD file and ultimately printed in a window that is appended to the simulated spectrum.


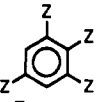

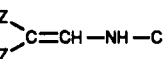

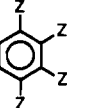
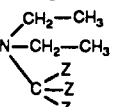
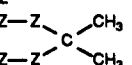

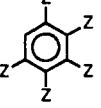

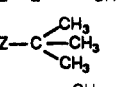

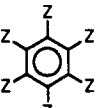

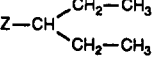

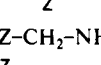
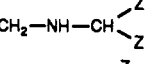
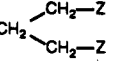
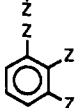
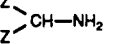
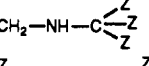
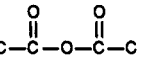
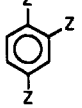
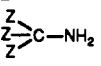
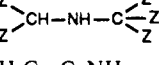
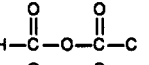

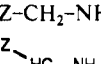
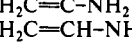
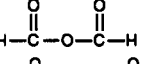

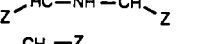
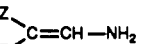
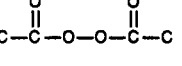
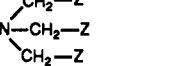

Displays are developed on a Tektronix graphics terminal and a Benson plotter. A list of identified substructures together with details of the related frequencies (range, intensity, and assignment) is available to the user.

The working version of SIRS-02 has 41 programs and subroutines, written entirely in FORTRAN IV, and occupies about 120 blocks. A version for personal computers is under development.

## CONCLUSION

The IR spectral simulation system DARC-SIRS, based upon associations between characteristic structural fragments and spectral lines, permits reconstruction of the IR spectrum of a molecule by means of recognition of characteristic fragments described topologically within the framework of the DARC

**Table VII.** Benzenic and Acyclic Complex Focus (Z Has Three Free Sites)

no.	substructure	no.	substructure	no.	substructure	no.	substructure
1		10		20		30	
2		11		21		31	
3		12		22		32	
4		13		23		33	
5		14		24		34	
6		15		25		35	
7		16		26		36	
8		17		27		37	
9		18		28		38	
		19		29			

system in the form of FREL bonds or complex foci, or FOXE. These descriptions, which are complementary, provide great flexibility for the isolation and recognition of structural fragments.

Starting from a database of nearly 700 structural fragments (primitives) organized by filiation, DARC-SIRS simulates the IR spectra of molecules containing the most frequently occurring elements (H, C, N, O, halogens) and displays the resulting spectra on a graphics terminal. When the desired substructure is absent from the database, it is possible however to estimate the associated data, albeit with less precision, by referring to the frequencies associated with analogous substructures. So as a result of the precision of the structural descriptors, particularly the topological DARC substructures of varying depth and precision, it is possible to generate the IR spectrum of any structure composed of the cited elements.

The quality of the simulated spectra can be evaluated quantitatively by the similarity indices between experimental and simulated spectra.

For computer-aided spectral simulation the current version, DARC-SIRS-02, performs well despite some restrictions that are currently being reexamined. More thought will be given to stereochemistry, and programs will be developed to account for unusual bonding, such as hydrogen-bonding in chelated species. This software will be extended gradually to cover other elements that are important in biologically interesting molecules or polymers (P, S, Si, etc). Finally, the proposed SIRS software, which is very open in its conception, encompasses the advantages of exploiting a database of primitive couples and some expert systems. It has been conceived for both computer-aided spectral interpretation and integration into a large, comprehensive computer-aided structure elucidation

system, as one of several complementary subsystems.

## REFERENCES

- (1) (a) Penca, M.; Zupan, J.; Hadži, D. *Anal. Chim. Acta* **1977**, *95*, 3. (b) Zupan, J. *Anal. Chim. Acta* **1978**, *103*, 273. (c) Zupan, J.; Penca, M.; Razinger, M.; Barlic, B. *Anal. Chim. Acta* **1980**, *122*, 103.
- (2) (a) Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. V.; Duffield, A. M.; Djerassi, C. *J. Am. Chem. Soc.* **1969**, *91*, 2973. (b) Duffield, A. M.; Robertson, A. V.; Djerassi, C.; Buchanan, B. G.; Sutherland, G. L.; Feigenbaum, E. A.; Lederberg, J. *J. Am. Chem. Soc.* **1969**, *91*, 2977. (c) Buchanan, B. G.; Feigenbaum, E. A. *Artif. Intell.* **1972**, *11*, 5. (d) Buchanan, B. G.; Smith, D. H.; White, W. C.; Gritter, R. J.; Feigenbaum, E. A.; Lederberg, J.; Djerassi, C. *J. Am. Chem. Soc.* **1976**, *98*, 6168.
- (3) (a) Shelley, C.; Munk, M. *Anal. Chem.* **1978**, *50*, 1522. (b) Shelley, C.; Munk, M. *Anal. Chim. Acta* **1981**, *113*, 507. (c) Trulson, M.; Munk, M. *Anal. Chem.* **1983**, *55*, 2137.
- (4) (a) Miyashita, Y.; Ochiai, S.; Sasaki, S. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 228. (b) Sasaki, S.; Abe, H.; Hirota, Y.; Ishida, Y.; Kudo, Y.; Ochiai, S.; Saito, K.; Yamasaki, T. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 211. (c) Sasaki, S.; Fujiwara, I.; Abe, H.; Yamasaki, T. *Anal. Chim. Acta* **1980**, *122*, 87. (d) Abe, H.; Yamasaki, T.; Fujiwara, I.; Sasaki, S. *Anal. Chim. Acta* **1981**, *133*, 499.
- (5) (a) Gribov, L. A.; Elyashberg, M. E.; Serov, V. *Anal. Chim. Acta* **1977**, *95*, 75. (b) Elyashberg, M. E.; Gribov, L. A. *Crit. Rev. Anal. Chem.* **1979**, *8*, 111. (c) Gribov, L. A.; Elyashberg, M. E.; Koldashov, V. N.; Pletjnov, V. *Anal. Chim. Acta* **1983**, *148*, 159.
- (6) Zippel, M.; Mowitz, J.; Köhler, I.; Opperkuch, J. *Anal. Chim. Acta* **1982**, *140*, 123.
- (7) (a) Woodruff, H. B.; Smith, G. M. *Anal. Chem.* **1980**, *52*, 2321. (b) Woodruff, H. B.; Smith, G. M. *Anal. Chim. Acta* **1981**, *133*, 545. (c) Tomellini, S. A.; Saperstein, D. D.; Stevenson, J. M.; Smith, G. M.; Woodruff, H. M. *Anal. Chem.* **1981**, *53*, 367. (d) Smith, G. M.; Woodruff, H. B. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 33. (e) Tomellini, S. A.; Stevenson, J. M.; Woodruff, H. B. *Anal. Chem.* **1984**, *56*, 67.
- (8) Tanabe, K.; Tamura, T.; Hiraishi, J.; Saeki, S. *Anal. Chim. Acta* **1979**, *112*, 211.
- (9) (a) Visser, T.; Van Der Maas, J. H. *Anal. Chim. Acta* **1980**, *122*, 357. (b) Visser, T.; Van Der Maas, J. H. *Anal. Chim. Acta* **1980**, *122*, 363.

- (c) Visser, T.; Van Der Maas, J. H. *Anal. Chim. Acta* **1981**, *133*, 451.
- (10) Debska, B.; Duliban, J.; Guzowska, B.; Hippe, Z. *Anal. Chim. Acta* **1981**, *133*, 303.
- (11) (a) Szalontai, G.; Simon, Z.; Csappo, Z.; Farkas, M.; Pleifer, G. *Anal. Chim. Acta* **1981**, *133*, 31. (b) Farkas, M.; Markos, J.; Szepesvary, P.; Bartha, I.; Szalontai, G.; Simon, Z. *Anal. Chim. Acta* **1981**, *133*, 19.
- (12) Passlack, M.; Bremser, W. In *Computer-Supported Spectroscopic Databases*; Zupan, J., Ed.; Ellis Horwood: Chichester, 1986, p 92.
- (13) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. *J. Am. Chem. Soc.* **1975**, *97*, 5755.
- (14) Carhart, R. E.; Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Djerassi, C. *J. Org. Chem.* **1981**, *46*, 1708.
- (15) Carabedian, M.; Dagane, I.; Dubois, J. E. *Anal. Chem.* **1988**, *60*, 2186.
- (16) Cabrol, D.; Rabine, J. P.; Rouillard, M.; Ricard, D.; Forrest, T. P. (University of Nice and Dalhousie University of Halifax) EXP'AIR PROGRAM v.2, 1989.
- (17) Schrader, B.; Bougeard, D.; Niggeman, W. Computer Evaluation of IR and Raman Spectra. *Comput. Methods Chem. (Proc. Int. Symp.)* **1977**, *80*, 37 (44BBAM).
- (18) Elyashberg, M. E.; Gribov, L. A. *J. Mol. Comput. Sci.* **1981**, *21*, 48.
- (19) Klopman, G.; McGonigal, M. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 48.
- (20) GAUSSIAN-80: AB-INITIO MO PROGRAM QCPE 446; Quantum Chemistry Program Exchange: Bloomington, IN, 1980.
- (21) Small, G. W. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 232.
- (22) Panaye, A.; Doucet, J. P.; Peguet, P.; Dubois, J. E. Proceedings of the 10th International CODATA Conference. *CODATA Bull.* **1986**, *64*, 32.
- (23) Peguet, P. Thèse de Docteur Ingénieur, Université Paris 7, 1985.
- (24) Socrates, G. *Infrared Characteristic Group Frequencies*; Wiley-Interscience Publications: New York, 1980.
- (25) Nakanishi, K. *Infrared Absorption Spectroscopy*; Practical Holden-Day Inc.: San Francisco, 1962.
- (26) Bellamy, L. J. *The Infrared Spectra of Complex Molecules*, 2nd ed.; Methuen: London, 1958.
- (27) Colthup, N. B.; Daly, L. H.; Wiberley, S. E. *Introduction to IR, Raman Spectroscopy*, Academic Press: New York, 1964.
- (28) Bellamy, L. J. *Advances in Infrared Group Frequencies*; Methuen: London, 1968.
- (29) Kirmman, A.; Janot, M. M.; Ourisson, G. *Structures et Propriétés Moléculaires: VII Fonctions Monovalentes. VIII Fonctions Divalentes. IX Fonctions Trivalentes. Monographies de Chimie Organique*; Masson et Cie: Paris, 1970.
- (30) (a) Dubois, J. E.; Viellard, H. *Bull. Soc. Chim. France* **1968**, 900. (b) Dubois, J. E.; Viellard, H. *Bull. Soc. Chim. France* **1968**, 905. (c) Dubois, J. E.; Viellard, H. *Bull. Soc. Chim. France* **1968**, 913.
- (31) Dubois, J. E. In *The Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: New York, 1976; p 330.
- (32) Dubois, J. E.; Laurent, D.; Viellard, H. *C. R. Acad. Sci., Ser. C* **1966**, *236C*, 764, 1245.
- (33) Dubois, J. E.; Laurent, D.; Viellard, H. *C. R. Acad. Sci., Ser. C* **1967**, *264C*, 348.
- (34) Attias, R. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102.
- (35) Dubois, J. E.; Carabédian, M.; Ancian, B. *C. R. Acad. Sci. Ser. C* **1980**, *290C*, 369, 383.
- (36) (a) Dubois, J. E.; Carabédian, M.; Dagane, I. *Anal. Chim. Acta* **1984**, *158*, 217. (b) Carabédian, M.; Dagane, I.; Dubois, J. E. *Systèmes Experts et leur Applications. INA* **1985**, *1*, 401.
- (37) Lioutas, A. Thèse de Doctorat, Université Paris 7, 1986.
- (38) Laurent, D.; Aranda, A. *J. Phys. Chim. Biol.* **1973**, *70*, 1068.
- (39) Dubois, J. E.; Mercier, C.; Panaye, A. *Acta Pharm. Jugosl.* **1986**, *36*, 135.
- (40) Dubois, J. E.; Panaye, A.; Attias, R. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 74.
- (41) Panaye, A. ASC: Adaptation Structurale Conventionnelle. Thèse de Doctorat, Université Paris 7, 1976.
- (42) Zupan, J.; Hadži, D.; Penca, M. *Comput. Chem.* **1976**, *1*, 71.
- (43) Delaney, M. F.; Warren, F. V.; Hallowell, J. R. *Anal. Chem.* **1983**, *55*, 1925.
- (44) Lowry, S. R.; Huppler, D. A. *Anal. Chem.* **1983**, *55*, 1288.
- (45) Razingier, M.; Zupan, J.; Penca, M.; Barlic, B. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 158.

## Computational Perception and Recognition of Digitized Molecular Structures<sup>†</sup>

M. LEONOR CONTRERAS,\* CARLOS ALLENDES, L. TOMAS ALVAREZ, and ROBERTO ROZAS

University of Santiago de Chile, Department of Chemistry, Casilla 5659, Santiago 2, Chile

Received March 7, 1990

Molecular structures containing both common and special alphanumeric characters are efficiently recognized by a program written in C. The program was designed to process type- and hand-printed structures. A scanner digitizes the corresponding images. Treatment of the binary information obtained in this way includes molecular graph perception and character recognition. Known and new image processing methods for molecular graph perception and an intelligent pattern-recognition principle for character processing were used. A graphic interface allows one to display and manipulate the recognized molecular images. Applications of the software to different areas such as molecular design, automatic input of structures to databases like ARIUSA, and others are also presented.

### INTRODUCTION

Representation of molecular structures allows one to describe and study sophisticated molecules such as vitamins, alkaloids, antibiotics, pheromones, organometallic complexes, etc., all of which may contain a 2-D stereochemical representation (dot and wedge convention) and delocalized bonds as in donor-acceptor complexes. Thus, the natural way of knowledge communication and management of information in chemistry is done using these structures. This is true in databases,<sup>1,2</sup> in CAMD,<sup>3</sup> in structure-activity relationships,<sup>4</sup> in synthesis design,<sup>5</sup> etc.

The structures themselves consist of two basic components: (a) a graph<sup>6</sup> or skeleton of the structure and (b) common and special alphanumeric characters (symbols, parenthesis, charges). A program that works with molecular structures must handle both components. That is what most interfaces

do for manual input of structures to computer systems.<sup>1-6</sup> The internal representation of that information through a connectivity table is known as recognition of the molecular structure by the system. This recognition of chemical structures is necessary for selective retrieval of information.<sup>1,2</sup> However the input of the structures, especially when they have more than 20 atoms and stereochemical specifications, is a time-consuming process normally requiring specialized people.<sup>1</sup>

In this paper we present a system which supports capture, perception, and recognition of type- and hand-printed molecular structures. In addition, as a part of this system, a graphic interface for the display and manipulation of the recognized structures is also presented.

### DESCRIPTION OF THE SYSTEM

The process basically consists of four steps: (a) scanning of molecular structures, (b) graph recognition, (c) character recognition, and (d) display.

<sup>†</sup> Presented in part at the XII Workshop of Systems Engineering, Santiago, Chile, July 1989.