Banques de Données du Système DARC-C13/NMR-Crystallodata-Mass Spectra-Graphic Input and Output. *Entropie* **1977**, *78*, 53-55.
(18) Bremser, W. Expectation Ranges of ^{13}C NMR Chemical Shifts. *Magn. Reson. Chem.* **1985**, *23*, 271-275.
(19) Carabédian, M.; Dagane, I.; Dubois, J.-E. Elucidation by Progressive Intersection of Ordered Substructures from Carbon-13 Nuclear Magnetic Resonance. *Anal. Chem.* **1988**, *60*, 2186-2192.
(20) Small, G. W.; McIntyre, M. K. Structure Elucidation Methodology for Disaccharides Based on Carbon-13 Nuclear Magnetic Resonance Spectrum Simulation. *Anal. Chem.* **1989**, *61*, 666-674.
(21) Christie, B. D.; Munk, M. E. Structure Generation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87-93.
(22) Bremser, W. Structure Elucidation and Artificial Intelligence. *Angew. Chem., Int. Ed. Engl.* **1988**, *27*, 247-260.
(23) Attias, R.; Dubois, J.-E. Substructure Systems: Concepts and Classifications. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 2-7.
(24) Carabédian, M.; Dubois, J.-E. A Combined Model of Multi-Resonance Subspectra/Substructure and DARC Topological Structure Representation. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 564-574.

A Combined Model of Multi-Resonance Subspectra/Substructure and DARC Topological Structure Representation. Local and Global Knowledge in the ^{13}C NMR DARC Database

MICHEL CARABÉDIAN and JACQUES-ÉMILE DUBOIS*

Institut de Topologie et de Dynamique des Systèmes de l'Université Paris 7, associé au CNRS, URA 34, 1 rue Guy de la Brosse, 75005 Paris, France

Received March 27, 1991

The structural and spectral information in a ^{13}C NMR database can be represented by means of a model which relates substructural fragments to subspectral features for multiple resonances. The substructural part of this model contains a concise DARC description of the structural part with a partially generic $ELCO_b$ which is associated with **all** the spectral information pertaining to the focal atom (F_o) and its neighboring carbons (A_i). In the spectral information, the concentric environmental view is shifted from the focal atom to the neighbor positions. This leads to overlap in the views and redundancy in the information and a dissymmetrical physical perception which formally, is broader than the substructural view. New substructural/subspectral local and global knowledge functions of this model are managed with holographic techniques. Formalized local and global knowledge is described statistically by juxtaposition of the $\delta^{13}C_{F_o} \times \delta^{13}C_{A_i}$ correlation plane supporting the 3D occurrence distributions. Use of the inferential ability of these planes is facilitated by a table which correlates the repartitioning of the σ- and π-bonds in F_o-A_i atom pairs.

INTRODUCTION

In most structure elucidation systems, ^{13}C NMR data are assigned a classical one-to-one relationship with specific carbon atoms in the structure. With this single resonance subspectra/substructure model, the central carbon of each substructural fragment is considered exclusively, and its $\delta^{13}C$ signal, usually described as falling within a band in the overall scale of ^{13}C chemical shifts, is the only datum that is used to identify the substructure in question.

A more penetrating analysis of the relationships between structural environments and spectral properties of such atoms is possible.[1] This approach uses substructural fragments defined by DARC, in which a structure is treated as a focal atom (F_o) and its environment. The *e*nvironment is organized so as to be *l*imited, *c*oncentric, and *o*rdered, hence the acronym ELCO. Use of these substructural fragments has shown that a single resonance–single substructure model cannot account for the complex relationships between structure and spectra.[2]

The multi-resonance spectra/substructure model that is described here has advantages in the analysis of these relationships. A number of different ELCOs have been studied[2] and of these, the specific ELCO that seems to be most appropriate here is the $ELCO_b$. This is the classical atom-centered fragment which consists of a central carbon (the focus atom, C_{F_o}), its immediate neighbors, and the bonds attaching it to its immediate neighbors. The next neighbors, atoms β to the focus atom, are not explicitly defined, and this imprecision affords the $ELCO_b$ with a generic quality which permits the grouping together of numerous diverse environments for the focus atom. The data which define these environments are all derived from the ^{13}C NMR DARC PLURIDATA III ($^{13}CDP_{III}$) Database.[3] This database contains 15867 structures and allows the generation of environments having statistically significant populations.

Extended Perception of Environment. The shielding of a ^{13}C atom is determined in a complex manner by different factors. Localized site contributions are considered to be approximately additive[4] and any atoms in positions α, β, γ, and δ relative to the central or focus atom may affect the chemical shift. The δ position corresponds to the D shell in the concentric description used by DARC, but the most significant contributions usually come from the γ positions—the DARC C shell. There are few correlation functions which permit the interpretation of experimental spectra in terms of these interactions, and so purely empirical models must be used.

In the classical single-resonance model, the chemical shift of the focal carbon is used in isolation but in the multi-resonance model, the focal carbon chemical shift is considered together with the chemical shifts of as many as four carbons directly bonded to it. If these neighboring carbons are in turn regarded as focus atoms, then a much deeper perspective on the actual external environment of the $ELCO_b$ is revealed. The local spectral information is enriched by the data redundancy in the $ELCO_b$ and by the broader perception of its external environment. This multi-resonance model leads to more complex statistics than those proposed for use with the single-resonance model. The $H(E,\delta)$ holographic curves used in

0095-2338/91/1631-0564$02.50/0

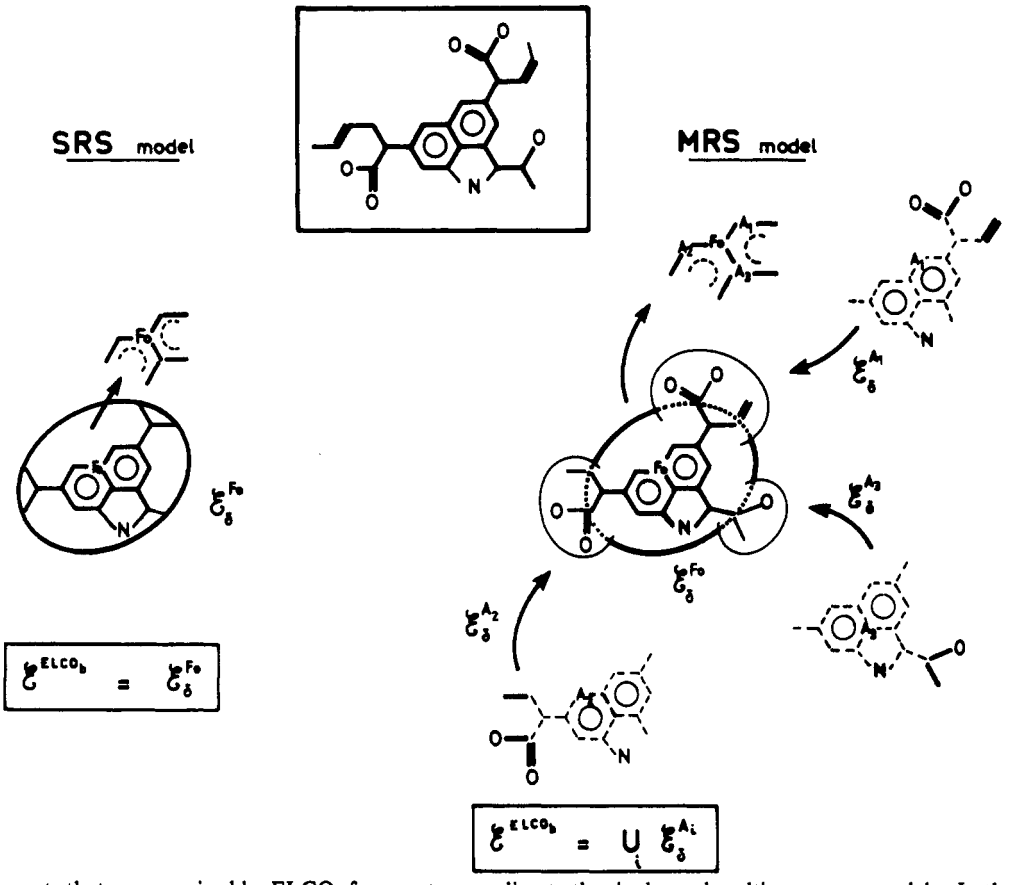

Figure 1. Environments that are perceived by $ELCO_b$ fragments according to the single- and multi-resonance models. In the multi-resonance model, the overall contribution of the environment is described by the union of the local environments.

the single-resonance model[2] correspond to the unidimensional variation in the chemical shift. Multi-resonance analysis, in contrast, uses a multidimensional framework and allows consideration of occurrences linked, not to a single chemical shift but to pairs of chemical shifts observed for atom pairs consisting of a focus atom (C_{F_o}) and any one of its immediate neighbors (C_{A_i}). The multidimensional properties of the $ELCO_b$ substructure and the associated statistics are discussed in the next section. The interactions measured in ^{13}C NMR spectra on the formal and semiempirical levels are considered, and various holographs derived from the $^{13}CDP_{III}$ database are described.

Perception of Environment: The Multi-Resonance Model. Modeling of the environment of a ^{13}C nucleus using an $ELCO_b$ module requires a significant compression of the structural information determining its chemical shift. The $ELCO_b$ is defined as the focal atom, its first neighbors, and the bonds joining them. The active environment outside the $ELCO_b$ is treated as a global perturbation which combines the different contributions of atoms β, γ, or δ to C_{F_o}. In the single-resonance model, all spectral information is associated with the focal carbon.

In the multi-resonance model, the physical responses of all the $ELCO_b$ carbons to this external global perturbation are considered, and so this perturbation is evaluated more precisely and its propagation is considered more broadly. This permits a better perception of the environmental effect on the $ELCO_b$ and the resulting spectral behavior and compensates for the partial neglect of structural information concerning the truly active environment of the $ELCO_b$. The spectral knowledge that is derived reflects the behavior of the C_{F_o} and the C_{A_i} atoms and permits more efficient handling of the global perturbation of the $ELCO_b$.

A structural/spectral representation of the $ELCO_b$, based upon the notion of "ordered couples" (i.e., C_{F_o} and C_{A_i}) offers two major advantages to the multi-resonance model:

(1) The physical effect of the external environment can be perceived from overlapping viewpoints by exploiting the dissymmetry of the C_{F_o} and C_{A_i} observation sites.

(2) Any perception can be oriented along axes defined by these atom pairs (C_{F_o} and C_{A_i}) or by specially defined atom pairs, and some local contributions of the environment can thereby be identified.

These two advantages of the multi-resonance model allow richer and better organization of the relationships between the $ELCO_b$ topology and the spectral data. The spectral and structural components have complementary properties of value to the multi-resonance model. On the one hand, the $ELCO_b$ is both concise and generic, while the spectral data permits directional organization and differential perception of the effects of nearest neighbors. These ideas can be perceived more effectively from the schematic representation shown in Figure 1.

An atom in position n of an $ELCO_b$ relative to the focus atom is, by definition in the $n-1$ position of one of the focus atom's nearest (C_{A_i}) neighbors and it will have an effect upon both the focus atom and the neighbor. Considering the various atom pairs (C_{F_o} and C_{A_i}) in an $ELCO_b$, the central environment of the focus atom is partially overlapped by the peripheral environments, each centered on a C_{A_i} atom. Over the whole molecule, the $ELCO_b$ which is implicitly described by the multi-resonance model can be extended as far as atoms in the ϵ-position relative to the focus atom and the overall environment corresponds formally to the union of these local environments.

In practice, these local environments can themselves be deduced from the various $ELCO_b$ that are centered on the different C_{A_i} atoms. The depth of the structural information that is implicitly considered by the multi-resonance model is

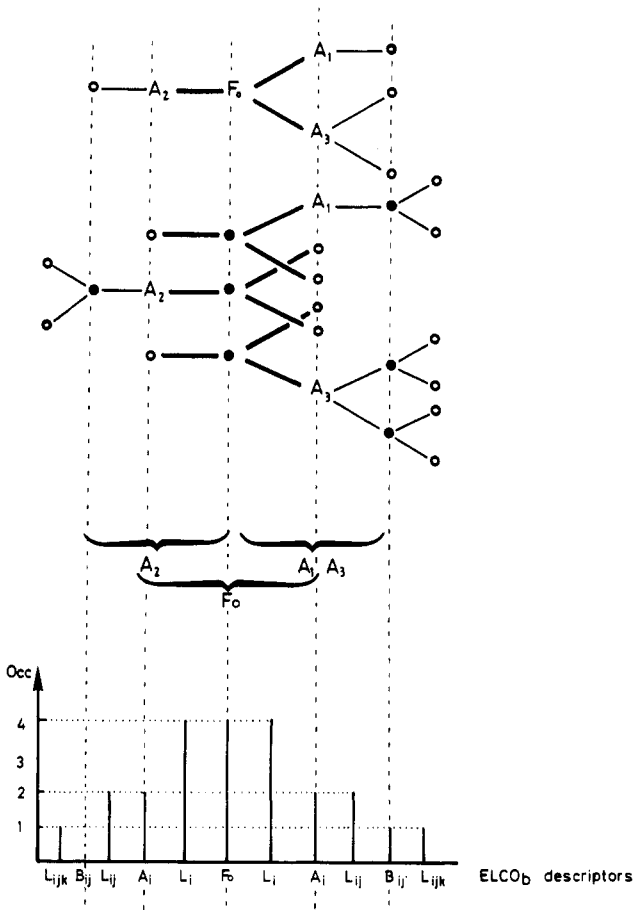

Figure 2. Overlapping of $ELCO_b$ centered on the C_{F_o} carbon atom and its C_{A_i} neighbors. The degree of overlap of structural information depends on the C_{F_o} connectivity and diminishes as one moves from the focus toward the B_{ij} positions.

depicted in Figure 1. The primary advantage of this model stems from the dissymmetry of the observation sites, C_{F_o} and C_{A_i}, which puts the local contributions to the $ELCO_b$ into perspective and provides orientation to their perception. The overlapping of the structural information described by the $ELCO_b$ centered on the C_{F_o} and C_{A_i} atoms is shown in Figure 2 where it can be seen that the focus atom and its bonds (L_i) are all that is common to all the $ELCO_b$ that are considered. They therefore correspond to maximum overlap. In general, for an $ELCO_b$ with nC_{A_i} neighbors, this information is described $n + 1$ times for all the $ELCO_b$ *centered upon the focus atom* and *its* C_{A_i} *neighbors*. Each of these C_{A_i} atoms, with its bonds, is described twice in this union, and its specific contribution, i.e., the B_{ij} atoms and their L_{ijk} bonds, is described once. Thus each carbon plays multiple roles in determining the environments of the different atoms in the structure. It is the observation site for its own environment, and it also participates in defining the environment of the other atoms, most significantly, those of its neighboring carbons. This dual role of focus and neighbor (F_o and A_i) is elucidated and organized according to the F_o–A_i directions by the multi-resonance model. In the structure elucidation strategy proposed for EPIOS,[5] these directions represent the axes used in the progressive expansion of candidate structures. Here they support the representation of the structure/spectra correlational knowledge.

Order and Knowledge Organization. The creation of the various forms of knowledge from the $^{13}CDP_{III}$ databank is based on structure/spectra pairs, and consequently, the order imposed upon this graph, that is the $ELCO_b$, is important in ensuring the coherence of any extension to these patterns. The overall organization of a knowledge base depends on the ordering of the ELCO topology,[6] which in turn follows from the concentric generation, in DARC, of the ELCO sites. This indexing fixes the coordinates of all the sites, including those occupied by the carbons whose shifts are being considered. The influence of the ordering is illustrated here by a morphological analysis of the holographs produced on the various $\delta^{13}C_{F_o} \times \delta^{13}C_{A_i}$ correlation planes defined by the multi-resonance model. The ordering functions used by the DARC system are usually objective-oriented and are based upon priority rules which use topological indices and chromatic information that is linked to atoms, bonds, and other structural features.[7]

Correlation Plane and Decision Surface. In characterizing an $ELCO_b$, the multi-resonance model uses the behavior of its focus C_{F_o}, and its neighbors, C_{A_i}. The EPIOS knowledge

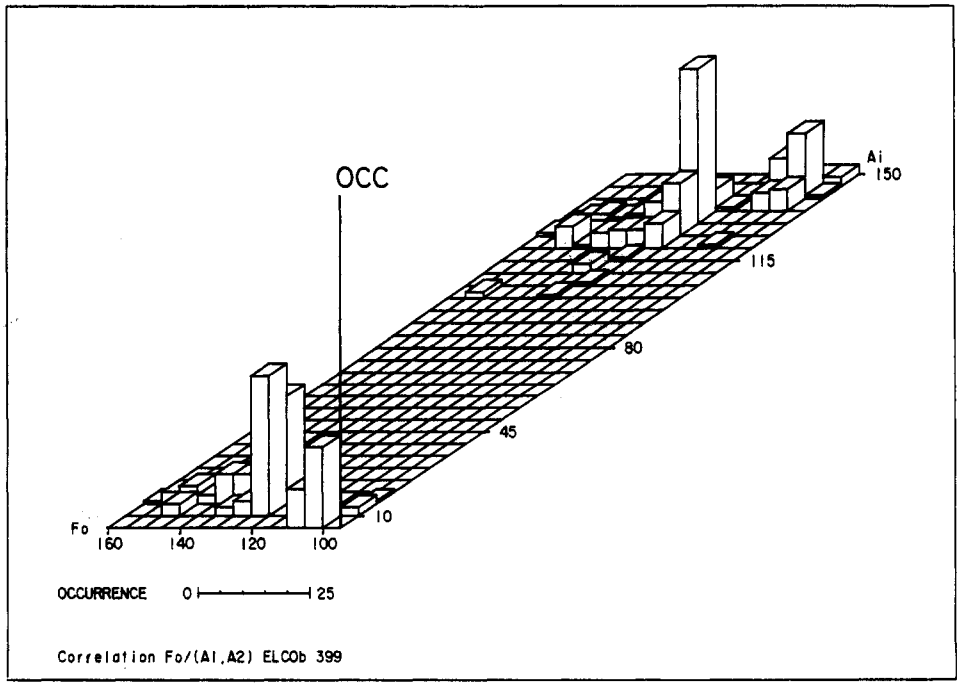

Figure 3. Holograph of the $ELCO_b$ C_{A_1}-C_{F_o}=C_{A_2}-. The distribution shape is determined by the values of the [$\delta^{13}C_{F_o}$, $\delta^{13}C_{A_i}$] pairs. Its profile is determined by the occurrence (OCC) of these pairs in the database spectra.

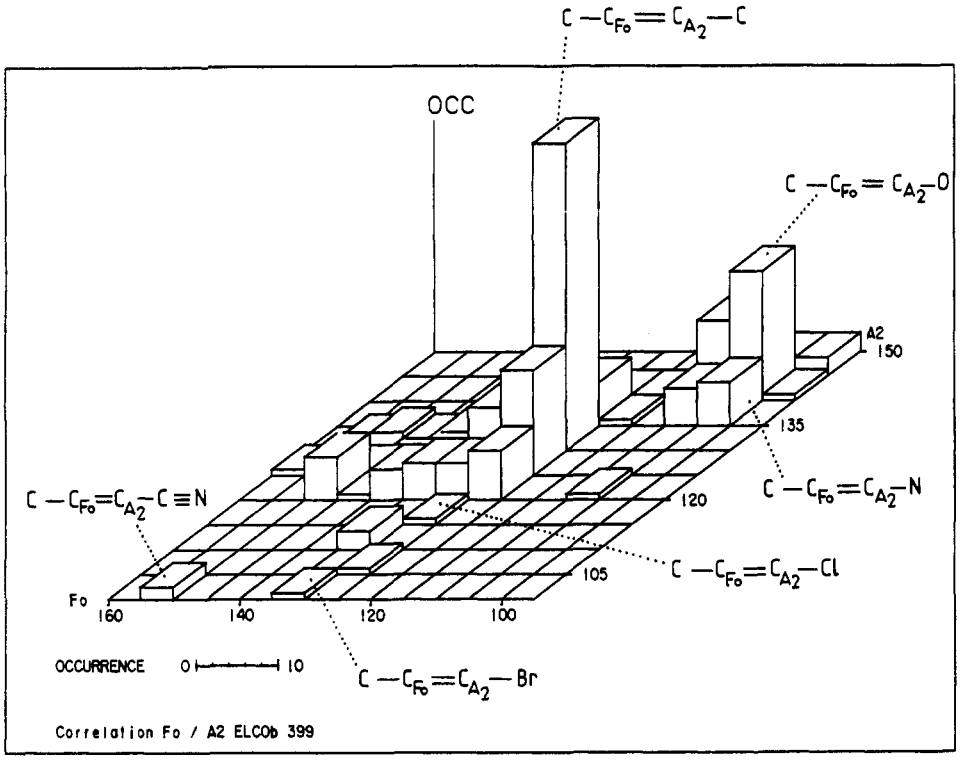

Figure 4. Holograph of the $ELCO_b$ C_{A_i}-C_{F_o}=C_{A_2}-. Component part of the $\delta_{F_o} \times \delta_{A_2}$ correlation plan of the holograph of Figure 3.

base[6] relates each $ELCO_b$ to a part of the $\delta^{13}C_{F_o} \times \delta^{13}C_{A_i}$ correlation plan shown in Figure 3 and not merely to an interval on the $\delta^{13}C$ chemical shift scale.

The structural information contained in an ELCO consists of the carbon number, the connectivity, and the bond types and can be used to define a **spectral map** whose contours express the NMR responses of the various carbon atoms to the environmental effects they experience. These environments are those that are found in the 15 867 structures associated with the $^{13}CDP_{III}$ databank which was used to create the EPIOS knowledge base. The spectral map, which is made up of the occurrence frequency of different $[\delta^{13}C_{F_o}, \delta^{13}C_{A_i}]$ pairs observed for an $ELCO_b$, offers a detailed topological representation of its behavior and generally enables one to identify the subpopulations that correspond to the different C_{A_i} carbons. In the C_{A_1}-C_{F_o} = C_{A_2}- case illustrated in Figure 3, the C_{A_1} and C_{A_2} carbons, with quite different chemical shifts, form separate clusters. The C_{A_1} carbons, which are sp^3-hybridized, all give signals in the lower part of the three-dimensional plot, while the C_{A_2} (sp^2) carbons all appear in the upper part of the diagram. When the C_{A_i} carbons in an $ELCO_b$ are almost identical, the peaks due to their chemical shifts can be merged.

The topological description of the ELCO is such that it allows isolation of the chemical shift of each (C_{F_o}, C_{A_i}) atom pair and differentiation of its four basic components (i = 1, 2, 3, and 4) with the corresponding carbon pairs. The relationships between structural and spectral information is guided by the order in the ELCO, and the integration of this order into the description of the structural/spectral relationships supports their organization and their use. A plot of the occurrence frequency of chemical shifts for the atom pair C_{F_o}, C_{A_2} of the $ELCO_b$ considered in this example is shown as a $\delta_{F_o} \times \delta_{A_2}$ plot in Figure 4.

The highest occurrence frequency corresponds to fragments containing a carbon in position β of the focus atom, i.e., α to C_{A_2}. This is the most common type of olefinic carbon in the database. A second distinct peak results from structures in which an oxygen is β to the focus atom. The structural origins of the various regions in this spectral map are summarized in Table I.

Table I. Variation Ranges of the C_{F_o}, C_{A_2} Pairs of $ELCO_b$ in Figure 3

C - C_{F_o} = C_{A_2} -	δC_{Fo}	δC_{A_2}
C - C_{F_o} = C_{A_2} - C	115–160	100–145
C - C_{F_o} = C_{A_2} - O	95–110	135–150
C - C_{F_o} = C_{A_2} - N	95–110	120–145
C - C_{F_o} = C_{A_2} - P	140–150	120–145
C - C_{F_o} = C_{A_2} - Cl	125–130	115–120
C - C_{F_o} = CA_2 - Br	125–135	100–110

On such maps related to $ELCO_b$, there are more or less well-defined areas that correspond to specific elements of the $ELCO_b$, which identify the nature of the atoms β to the focus atom. The origin of the effects perceived by the carbons of an $ELCO_b$ cannot always be expressed as clearly by recourse to increased specificity.

To characterize an $ELCO_b$, it is necessary to consider all the influences on each of the carbons. In spectral interpretation, indexed spectral maps such as those shown in Figures 3 and 4 function as decision surfaces that allow the selection of $ELCO_b$ which are compatible with the spectrum. The only $ELCO_b$ that are retained are those for which every carbon can be assigned to at least one chemical shift of the query spectrum. This has the advantage that it provides information on the possible connections among the carbons of the structure that is sought.[8] In this way, EPIOS can extract from a spectrum information that is only retrieved with difficulty using the single-resonance model.

Use of Order in Representation of Multi-Resonance Relations. The content of the current EPIOS knowledge base is represented in Figures 5–8. Each $\delta_{F_o} \times \delta_{A_i}$ correlation plane shows the contour of the superimposed maps derived from the 8587 $ELCO_b$ that are in the database. Each chemical shift pair (δ_{F_o}, δ_{A_i}) is associated with the number (Occ) of $ELCO_b$ whose C_{F_o} and C_{A_i} carbons are compatible with these values, i.e.:

$$\delta_{F_o} \in \Phi\delta^{13}C_{F_o} \text{ and } \delta_{A_i} \in \Phi\delta^{13}C_{A_i}$$

The topological order implicit in the ELCO description can

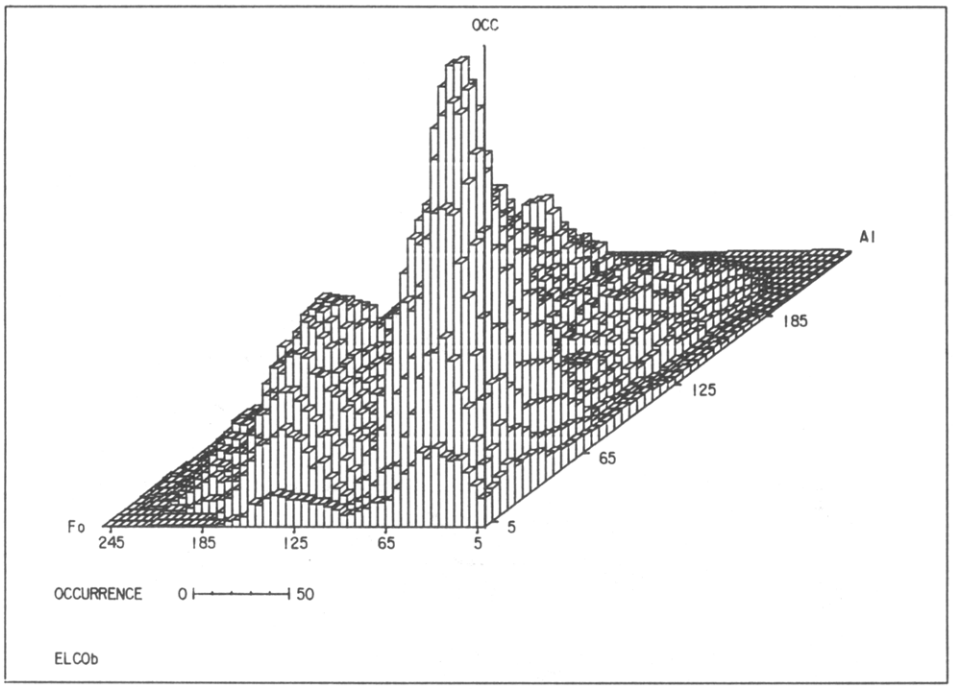

Figure 5. Projection on the correlation plane of the Φ_{F_o}, Φ_{A_1} atoms. Each δ_{F_o}, δ_{A_1} pair is associated with a number of $ELCO_b$ compatible with these values: $\delta_{F_o} \in \Phi_{F_o}$, $\delta_{A_1} \in \Phi_{A_1}$.

Table II. Constitution of $ELCO_b$[a]

$ELCO_b$	no. of $ELCO_b$	A_1	A_2	A_3	A_4	N_C	%
primary	58	C				20	34.5
secondary	788	C				631	80.1
			C			395	50.1
		C	C			307	38.9
tertiary	3223	C				2752	85.4
			C			2164	67.1
				C		1624	50.4
		C	C	C		1026	31.8
quaternary	1128	C				1113	98.7
			C			1076	95.4
				C		892	79.1
		C	C	C	C	387	34.3
						387	34.3

[a] N_C is the number of $ELCO_b$ with carbon in the A_i position.

be used to model the observed multi-resonance relations as determined from the reference spectra. It also determines the morphology of the distributions that are found in the four correlation planes.

Morphology of Multi-Resonance Spectral Shift Distribution. Entry of a multi-resonance relationship on a $\delta_{F_o} \times \delta_{A_i}$ plane is made when there is a carbon atom pair on these positions (F_o and A_i) of the $ELCO_b$ concerned. The order in which A_i positions are defined about the C_{F_o} therefore determines the $ELCO_b$ subpopulations that are considered in each correlation plane. As a result, it plays a direct role in the morphology of the distributions obtained. Table II shows the number of primary, secondary, tertiary, and quaternary carbons, which are centers of $ELCO_b$, present in these distributions. In all, 5197 of the 8587 available $ELCO_b$ are considered. Eliminated from this representation were the 2788 $ELCO_b$ that were centered on a heteroatom and 602 which were insufficiently characterized. Each $ELCO_b$ with N (C_{F_o}, C_{A_i}) pairs will appear in N correlation planes. As an example, pairs from four types of $ELCO_b$ (primary to quaternary) are found in the $\delta_{F_o} \times \delta_{A_1}$ plane because they all can possess a carbon in the A_1 position. This is therefore the largest population. The $\delta_{F_o} \times \delta_{A_4}$ plane, on the other hand, contains only 387 quaternary $ELCO_b$, each of which has a carbon in the A_4 position.

The number of $ELCO_b$ types in the subpopulations of each plane are shown in Table III. It can be seen that the tertiary $ELCO_b$, which are present in large numbers (3223, Table II) dominate the overall morphology except in the $\delta_{F_o} \times \delta_{A_4}$ plane, where they cannot be present.

Statistics of the composition of the different $ELCO_b$ are given in Table IV. This table shows the proportion of σ- or π-bonds linking C_{F_o} and C_{A_i}. (A π-bond in this context refers to any multiple bond.) These σ- and π-bonds are evidence of sp^3 and sp^2 hybridization, respectively, of the carbons and are useful in the interpretation of the distributions that are obtained. Two major peaks can be seen on the δ_{F_o} and δ_{A_i} axes on either side of the 90-ppm point, and these correspond to these two carbon types. The lowfield peak is absent from the $\delta_{F_o} \times \delta_{A_4}$ plane because sp^2-hybridized carbons cannot have a C_{A_4} neighbor.

The different holographs shown in Figures 5–8 represent a general and detailed view of the set of multi-resonance spectral relationships defined by the database and show graphically the complexity of the environmental effects experienced by the $ELCO_b$ carbons. The degree of hybridization of C_{F_o} and C_{A_i} is, to a first approximation, an adequate criterion for the interpretation of these holographs and will be used in the next section.

Table III. Number of $ELCO_b$ with Different Focus Carbons in the Four $\delta_{F_o} \times \delta_{A_{i3}}$ Planes

plane	no. of $ELCO_b$	nature of the focus atom C_{F_o}			
		primary	secondary	tertiary	quaternary
$\delta_{F_o} \times \delta_{A_1}$	4516	20 (0.4%)	631 (14.1)	2752 (60.9)	1113 (24.6)
$\delta_{F_o} \times \delta_{A_2}$	3635	395 (10.9)	2164 (59.5)	1076 (29.6)	
$\delta_{F_o} \times \delta_{A_3}$	2516	1624 (64.6)	892 (35.4)		
$\delta_{F_o} \times \delta_{A_4}$	387				

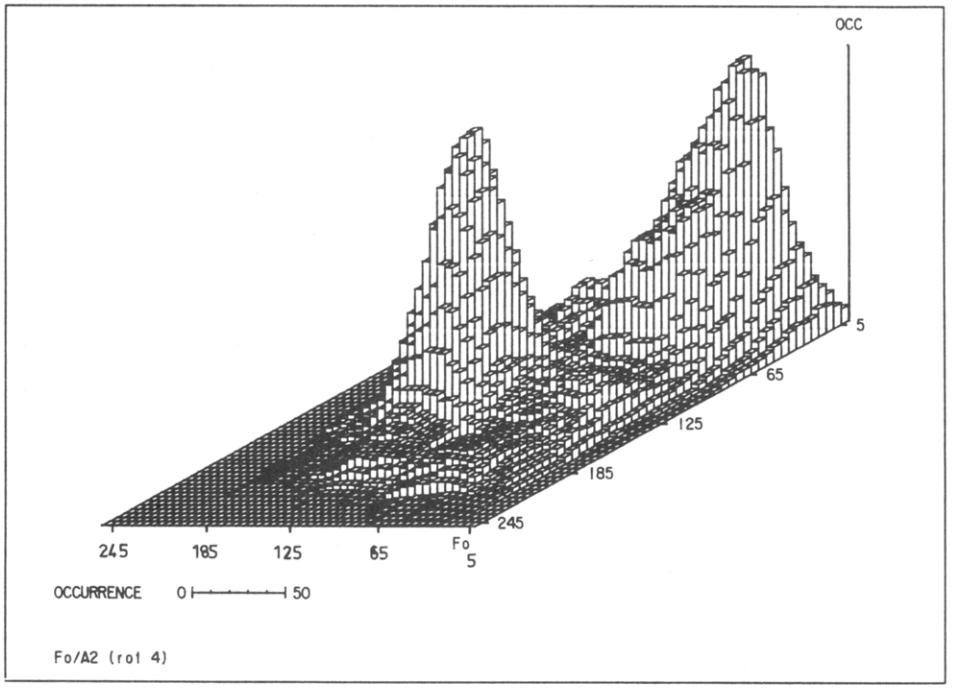

Figure 6. Projection on the correlation plane of the Φ_{F_o}, Φ_{A_2} atoms. Each δ_{F_o}, δ_{A_2} pair is associated with a number of $ELCO_b$ compatible with these values: $\delta_{F_o} \in \Phi_{F_o}$, $\delta_{A_2} \in \Phi_{A_2}$.

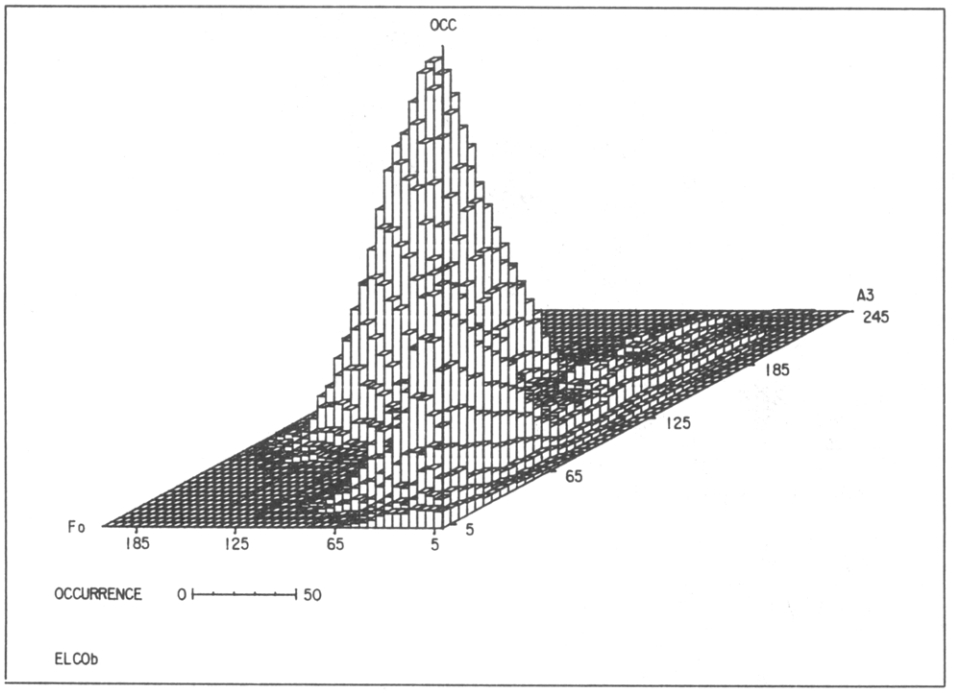

Figure 7. Projection on the correlation plane of the Φ_{F_o}, Φ_{A_3} atoms. Each δ_{F_o}, δ_{A_3} pair is associated with a number of $ELCO_b$ compatible with these values: $\delta_{F_o} \in \Phi_{F_o}$, $\delta_{A_3} \in \Phi_{A_3}$.

The σ/π Assignment Table of F_o–A_i Pairs. Labeling the component carbons of $ELCO_b$ as F_o^σ, F_o^π, A^σ, or A^π allows the definition of an assignment table of F_o–A_i atom pairs. This table can be used as shown in Figure 9 to characterize the four major zones defined in each correlation plane by the dividing line at $\delta^{13}C = 90$ ppm. The 90-ppm value is not an absolute limit, strictly separating sp^3 carbons on the one hand from sp^2 and sp carbons on the other. Rather, it permits location on the different planes of the peak occurrence frequencies and has adequate statistical validity to support the description of the observed distributions.

These have been simplified by means of the localization tables shown in the four panels of Figure 10, which show the most frequently occurring chemical shifts pairs ($\delta^{13}C_{F_o}$, $\delta^{13}C_{A_i}$) in each of the four quadrants. These pairs, generally situated at the center of the peaks formed in the zones, provide a view of the evolving morphology of the distributions according to the (C_{F_o}, C_{A_i}) pair under consideration. The order adopted in determining the rank of these pairs stems from the $ELCO_b$ description, and its influence is seen in the morphological evolution that it propagates.

(*a*) *The $\delta_{F_o} \times \delta_{A_1}$ Plane.* The $\delta_{F_o} \times \delta_{A_1}$ plane shown in Figure 5 contains the largest $ELCO_b$ subpopulation, with 4516 members (Table III) distributed into four clearly defined zones. All combinations of sp or sp^2 (π) or sp^3 (σ) carbons are among the atom pairs defined in this $ELCO_b$. A schematic summary of the information in this plane is shown in Figure 10. The most frequently occurring fragment, representing 47% of the

Table IV. Description of the $ELCO_b$ Present on the Correlation Planes

$\delta_{F_o} \times \delta_{A_1}$	L_1 σ%	$\sum L_i$ σ%	n_{1i} 0	1	2	3	$\sum L_{1i}$ σ%
P	65.0	65.0	15.0	30.0	40.0	15.0	71.0
S	94.1	82.3	8.1	30.7	47.3	13.9	68.2
T	99.4	79.9	13.9	33.7	43.7	8.7	69.4
Q	100.0		39.6	32.4	23.7	4.3	77.6
$\delta_{F_o} \times \delta_{A_2}$	L_2 σ%	$\sum L_i$ σ%	n_{2i} 0	1	2	3	$\sum L_{2i}$ σ%
S	43.8	68.6	10.2	28.3	44.3	17.2	78.8
T	82.0	78.9	1.7	27.7	53.6	17.0	71.6
Q	100.0		40.7	32.9	22.6	3.8	77.4
$\delta_{F_o} \times \delta_{A_3}$	L_3 σ%	$\sum L_i$ σ%	n_{3i} 0	1	2	3	$\sum L_{3i}$ σ%
T	32.4	71.2	4.2	28.3	55.2	12.9	77.2
Q	100.0		3.5	26.2	47.4	22.9	84.3
$\delta_{F_o} \times \delta_{A_4}$	L_4 σ%	σ%	n_{4i} 0	1	2	3	$\sum L_{4i}$ σ%
Q	100.0		3.5	26.2	47.4	22.9	84.3

[a] L_i is the bond between the C_{F_o} focus and its C_{A_i} neighbor; $\sum L_i$ is the set of L_i bonds; n_{ij} is the number of B_{ij} neighbors borne by C_{A_i} $\sum L_{ij}$ is the set of bonds between C_{A_i} and its B_{ij} neighbors; σ% is the proportion of the bonds that are σ-bonds.

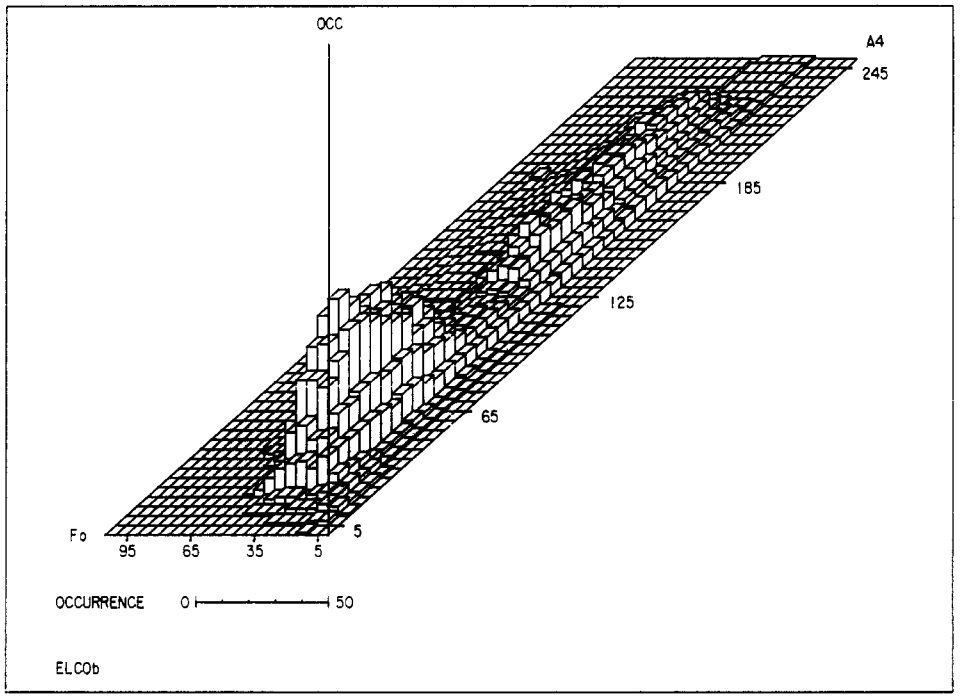

Figure 8. Projection on the correlation plane of the Φ_{F_o}, Φ_{A_4} atoms. Each δ_{F_o}, δ_{A_4} pair is associated with a number of $ELCO_b$ compatible with these values: $\delta_{F_o} \in \Phi_{F_o}$, $\delta_{A_4} \in \Phi_{A_4}$.

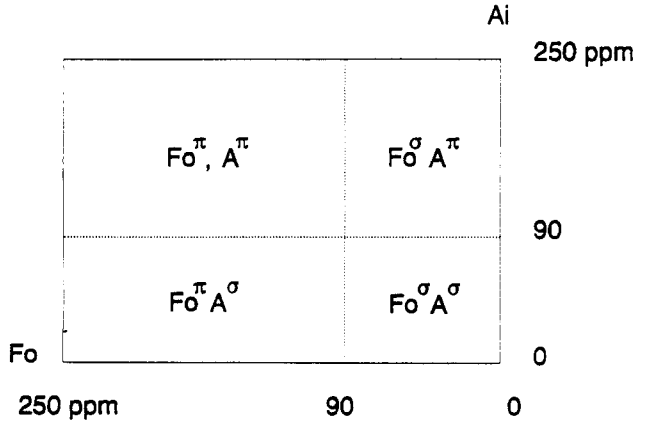

Figure 9. Assignment table F_o–A_i. The four different occurrence zones of the σ/π character of the C_{F_o}–C_{A_i} carbon atom pairs.

total population, is that in which two sp^3 carbons are joined together. At this peak, 309 $ELCO_b$ have $\delta^{13}C_{F_o}$ = 40 ppm and $\delta^{13}C_{A_i}$ = 25 ppm. The large proportion of σ-bonds at the focus atoms, noted in Table IV as $\sum L_i$, and the assignment of the A_1 position to carbons that also have only σ-bonds are the reasons for this. The number of σ-bonds increases with the $ELCO_b$ connectivity patterns shown in Table IV, going from 65% for the primary carbons to 94.1% for the secondary carbons and 99.4% for the tertiary carbons.

The next most important zone in the $\delta_{F_o} \times \delta_{A_1}$ plane is that pertaining to the (F_o^{π}, A_1^{σ}) pairs, which are mostly associated with secondary and tertiary $ELCO_b$ centered on sp^2 carbon atoms. The proportion of A_1 carbon atoms (68.4%) involved in these two zones F_o^{σ}, A_1^{σ}, and F_o^{π}, A_1^{π}, as a result of their spectral behavior, agrees with that observed in the structural description of the $ELCO_b$ in which 68.2–77.6% of the L_{ij} bonds at C_{A_1} carbon atoms are σ-bonds (see Table IV).

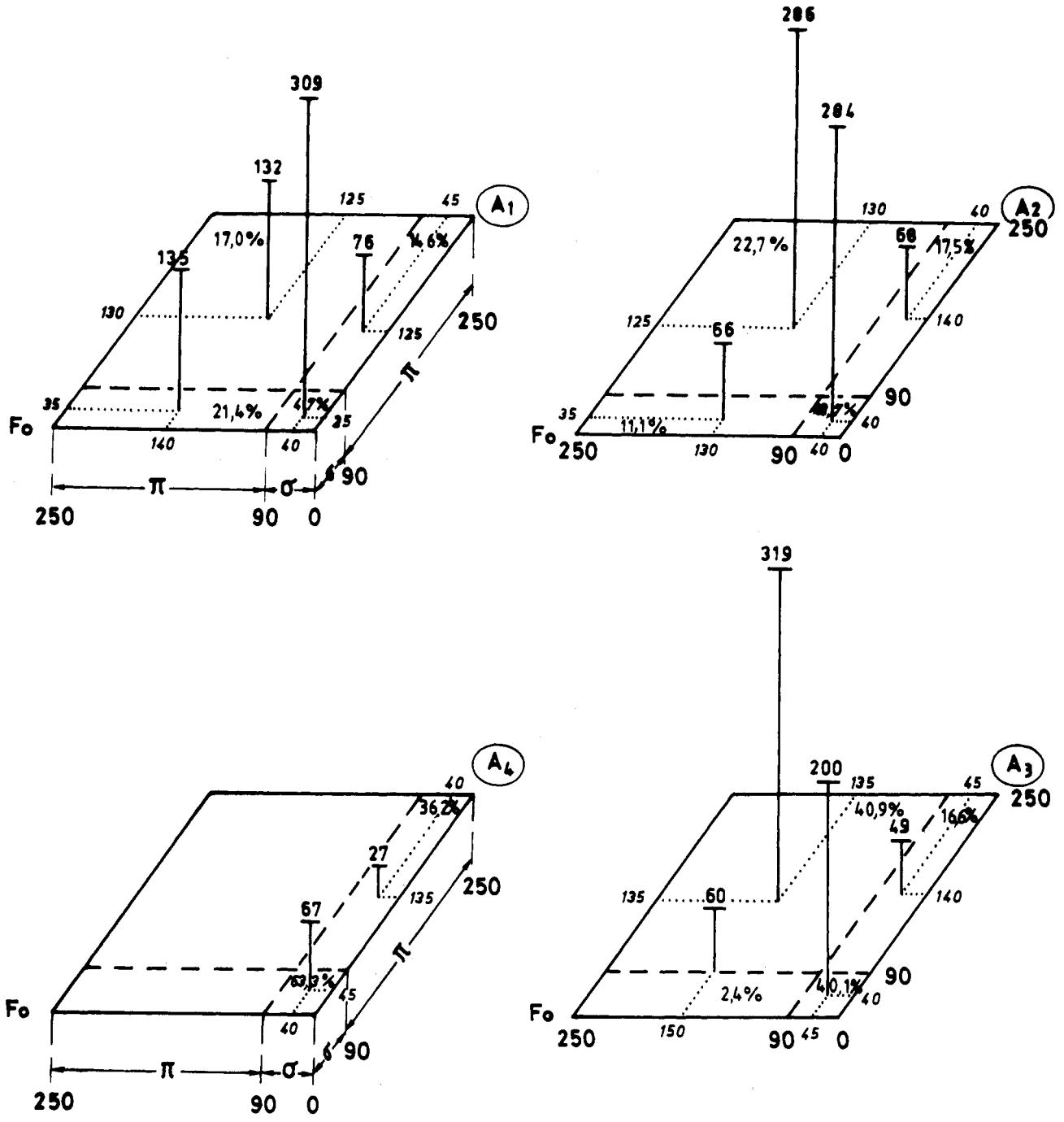

Figure 10. Diagrammatic description of the $H(\Phi_{F_o}, \Phi_{A_i}, ELCO_b)$ holographs (Figures 5–8) from the F_o–A_i assignment table (Figure 9). Each σ/π zone of the correlation planes is characterized by its $(\delta_{F_o}, \delta_{A_i})$ peak of maximum occurrence.

The remaining 31.6% of the $ELCO_b$ is found in the F_o^{σ}, A_1^{π} and F_o^{π}, A_1^{π} zones, which involve an A_1 carbon carrying a π-bond. The maximal occurrence in $(F_o^{\sigma}, A_1^{\pi})$ is low (76 $ELCO_b$ for $\delta^{13}C_{F_o} = 45$ ppm, $\delta^{13}C_{A_1} = 135$ ppm) because this A_1 position is usually occupied by an sp^3 carbon. However, many $ELCO_b$ reflecting aromatic rings or conjugated systems (cf. Figure 11) are represented in the (F_o^{π}, A_1^{π}) zone where they form a homogeneous peak (height 132) comparable to the $(F_o^{\pi}, A_1^{\sigma})$ peak of height 135.

(*b*) *The $\delta_{F_o} \times \delta_{A_2}$ Plane.* In the $\delta_{F_o} \times \delta_{A_2}$ plane, which is shown in Figure 6 and summarized in Figure 10, the distribution is characterized by a diagonal symmetry. The $(F_o^{\sigma}, A_2^{\sigma})$ and (F_o^{π}, A_2^{π}) zones contain 71.4% of the $ELCO_b$ in the two major peaks which have equivalent maximal occurrences (284 and 286, respectively). The remaining 28.6% of the $ELCO_b$ is divided between two very low peaks in the mutually complementary $(F_o^{\pi}, A_2^{\sigma})$ and $(F_o^{\sigma}, A_2^{\pi})$ zones. This results mainly from the suppression of the peak in the $(F_o^{\pi}, A_2^{\sigma})$ zone relative to the peak in the (F_o^{π}, A_2^{π}) as compared to their homologues in the preceding plane. It is an expression of the reduced number of L_2^{σ} bonds joining C_{F_o} atoms to their C_{A_2} neighbors. This goes from 94.1% to 43.8% for secondary $ELCO_b$ and from 99.4% to 82.0% for tertiary $ELCO_b$.

(*c*) *The $\delta_{F_o} \times \delta_{A_3}$ Plane.* In the $\delta_{F_o} \times \delta_{A_3}$ plane, which is shown in Figure 7, the $(\delta^{13}C_{F_o}, \delta^{13}C_{A_3})$ distribution is simply a further development of the evolving distribution that was seen in the $(\delta^{13}C_{F_o}, \delta^{13}C_{A_1})$ and $(\delta^{13}C_{F_o}, \delta^{13}C_{A_2})$ planes. In this case, the suppressed $(F_o^{\pi}, A_3^{\sigma})$ frequency occurrence leads to a shift of this zone's maximum occurrence to the 90-ppm limit. This therefore provides a justification for this approximate limit and for the simplified (π, σ) notation. It is impossible, for reasons of local connectivity, to associate an F_o^{π} focus atom with a carbon A_3^{σ} neighbor; the existence of an A_3 position in an $ELCO_b$ means it must be ternary or quaternary. Only

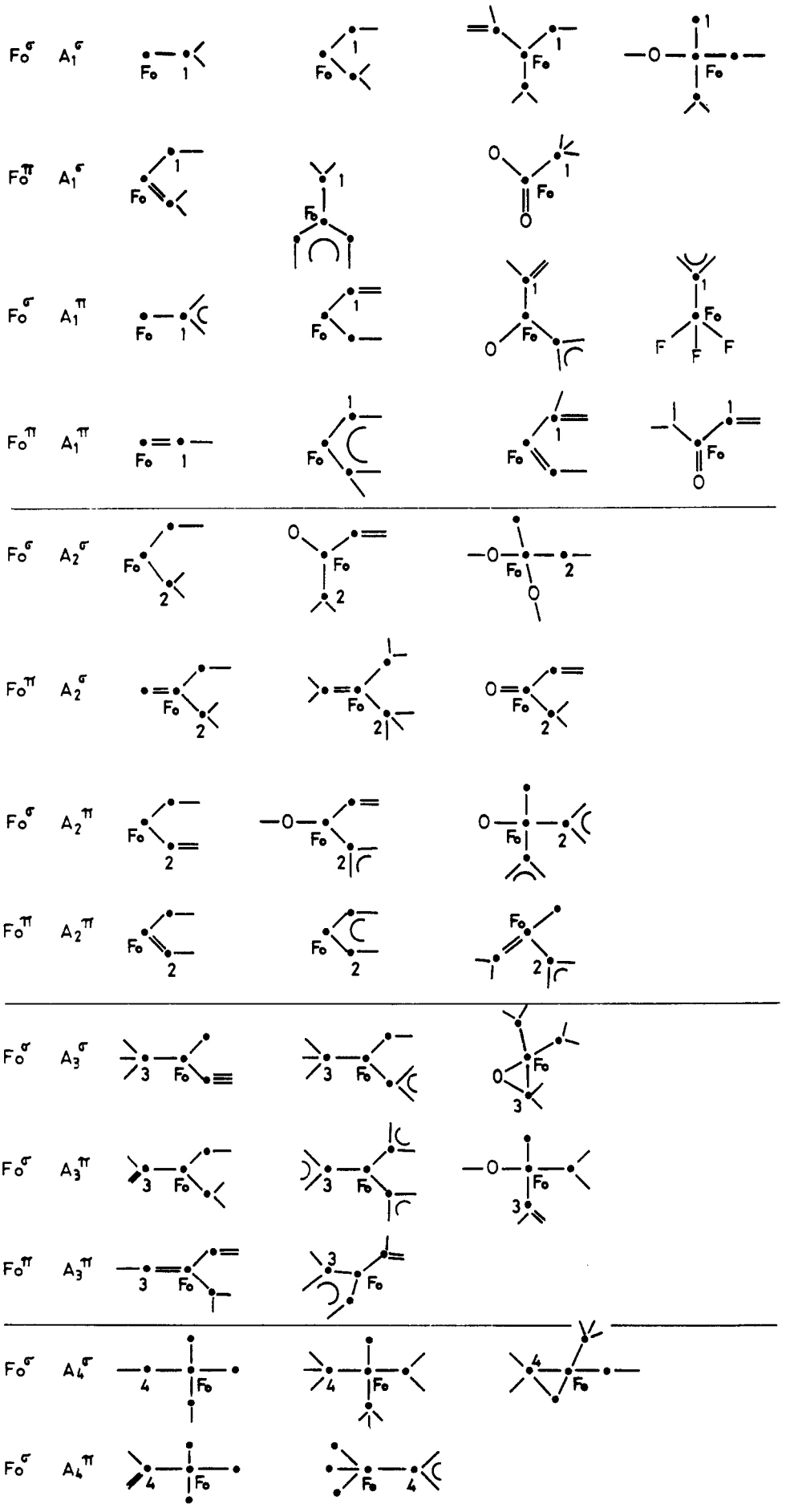

Figure 11. Examples of $ELCO_b$ primitives in the different $H(\Phi_{F_0}, \Phi_{A_l}, ELCO_b)$ holograph zones.

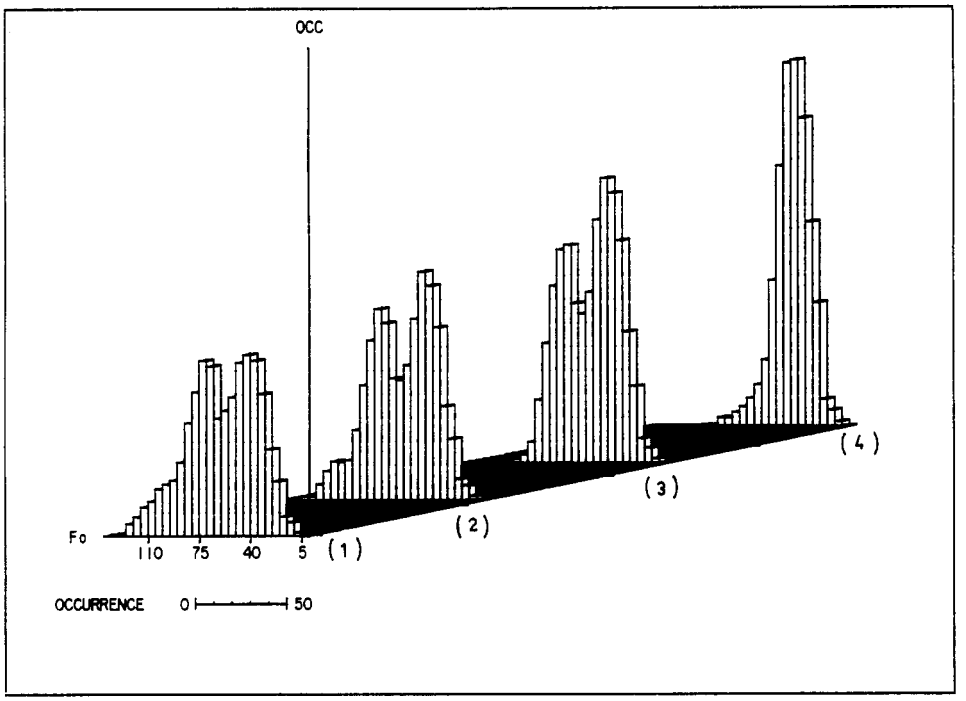

Figure 12. $H(\Phi_{F_0}, ELCO_b)$ components of the different $H(\Phi_{F_0}, \Phi_{A_i}, ELCO_b)$ for the quaternary $ELCO_b$. The absence of a heteroatom in position A_4 of these $ELCO_b$ causes the disappearance of the secondary peak ($\delta_{F_0} \approx 70$ ppm) observed in the first three distributions.

ternary $ELCO_b$ may have a π-bond at their focal atoms, and such bonds may exist only where there is a $C_{A_3}{}^{\pi}$ neighbor. Thus the 2.4% of the $ELCO_b$ that are in this zone are actually ($F_0{}^{\pi}$, $A_3{}^{\pi}$) zone elements for which the chemical shifts of carbon A_3 fall below the arbitrary 90-ppm limit. Such $ELCO_b$ typically have a heteroatom α to the π-bond, i.e., in the A_1 or A_2 position and by means of conjugation, this heteroatom causes the signals of the $A_3{}^{\pi}$ carbons at the β-position to move to higher fields (cf. Figure 11). In this $\delta_{F_0} \times \delta_{A_3}$ plane, the reduced spread of the $\delta^{13}C_{F_0}$ chemical shifts may also be noted. This results from the absence of carbonyl and thiocarbonyl groups which, on the previous planes, were associated with very low field shifts, more than 200 ppm from the internal reference.

(*d*) *The $\delta_{F_0} \times \delta_{A_4}$ Plane.* In the $\delta_{F_0} \times \delta_{A_4}$ plane (Figure 8), the zones corresponding to an $F_0{}^{\pi}$ focus must be empty because such foci cannot be present in quaternary $ELCO_b$. The frequent appearance of heteroatoms in the A_4 position causes the elimination of 741 (65.7%) of the quaternary heteroatoms of this distribution (Table III). These heteroatoms are however, implicitly considered in the different correlation planes where their influence on the $ELCO_b$ foci has been seen to impinge upon the distributions of the $\delta^{13}C_{F_0}$ values.

All the $\delta^{13}C_{F_0}$ distributions for quaternary $ELCO_b$ are grouped together in Figure 12. The secondary peak, centered at 70 ppm and observed in the $F_0{}^{\sigma}$ zones ($\delta^{13}C_{F_0} \leq 90$ ppm) of the first three distributions, is due to the presence of heteroatoms. Since the 741 quaternary $ELCO_b$ having a heteroatom at position A_4 were eliminated from the correlation, this peak may not be present in the fourth distribution (cf. Figure 12).

CONCLUSION

Representation of basic knowledge is indispensable to its use. Examples of this truism include understanding the rules governing the influence of structural environment or using semitheoretical relationships derived from database information. This knowledge is understood here to mean the association of structural and spectral information. In the case of ^{13}C NMR spectra, the $\delta^{13}C$ chemical shift is a very sensitive indicator of structural changes in the environment of the nucleus in question. Structural/spectral knowledge is approached via structural and spectral approximations by a study, in a large file, of the effect on chemical shift of fragments defined as Environments that are Limited Concentric, and Ordered (ELCO).

The chemical shifts of carbon atoms in $ELCO_b$ can be represented in terms of the order imposed by the $ELCO_b$ topology description in models which express multi-resonance relations. This order controls the overall organization of the EPIOS knowledge base by arranging the characteristic behavior of $ELCO_b$ on four correlation planes, $\delta_{F_0} \times \delta_{A_i}$, which function as decision surfaces governing the $ELCO_b$ selection during either structural elucidation or spectral simulation.

In this paper, the practical consequences of the order introduced into the $ELCO_b$ on the representation of the $\delta^{13}C$ shifts in the multi-resonance model have been analyzed. The adequacy of this concentric order of structural environmental effects is supported by the coherence of the relations that emerge between structural and spectral information and the ability to extract self-consistent categories or families of chemical shifts. These relations are shown to have discriminatory power and thus selectivity in spectral interpretation in the EPIOS structural elucidation system is ensured.

For any $ELCO_b$, multi-resonance spectral representation is richer than the single-resonance alternative because it allows exhaustive probing of the environment of the focal atom. The dissymmetry of the sites of the $ELCO_b$, the focal atom C_{F_0}, and the neighbors C_{A_i} provide a representation of this environment. This perspective provides the multi-resonance spectral model with its enhanced global and local perception. The spectral maps proposed here facilitate the perception of remote structural environment and the identification, in structural terms, of the maps' components based upon their positions in the maps. This will be discussed in a future paper.

ACKNOWLEDGMENT

We are indebted to B. Dubois for her patience and helpful contributions in editing this paper. Our very sincere thanks to Bill Milne for his kind advice on the final presentation of our papers.

REFERENCES AND NOTES

(1) Dubois, J.-E.; Carabédian, M. Modeling of Alkyl Environment Effects on the ^{13}C Chemical Shift. *Org. Magn. Reson.* **1980**, *14*, 264-271.
(2) Dubois, J.-E.; Carabédian, M. Single-Resonance Subspectra/Substructure (SRS) Investigations on the ^{13}C DARC Databank. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 557-564.
(3) (a) Dubois, J.-E.; Bonnet, J. C. The DARC Pluridata System: The BCNMR Bank. *Anal. Chim. Acta* **1979**, *112*, 245-252. (b) Dubois, J.-E. Nouvelles de Pluridata: Banques de Données du Système DARC-C13/NMR-Crystallodata-Mass Spectra-Graphic Input and Output. *Entropie* **1977**, *78*, 53-55.
(4) Stothers, J. B. *Carbon-13 NMR Spectroscopy*; Academic Press: New York, 1972.
(5) Dubois, J.-E.; Carabédian, M.; Ancian, B. Elucidation Structurale Automatique per RMN du Carbone 13: Méthode DARC-EPIOS. I. Recherche d'une Relation Discriminante Structure-Déplacement Chimique. II. Description de l'Élucidation Progressive par l'Intersection Ordonnée de Sous Structures. *C. R. Acad. Sci., Ser. C* **1980**, *290*, 369-372, 383-386.
(6) Dubois, J.-E.; Carabédian, M.; Dagane, I. Computer-Aided Elucidation of Structures by ^{13}C NMR. The DARC-EPIOS Methods: Characterizing Ordered Structures by Correlating the Chemical Shifts of Their Bonded Carbon Atoms. *Anal. Chim. Acta* **1984**, *158*, 217-233.
(7) Dubois, J.-E.; Panaye, A.; Attias, R. DARC SYSTEM: Notions of Defined and Generic Substructures. Filiations and Coding of FREL Substructures (SS) Classes. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 74-82.
(8) Carabédian, M.; Dagane, I.; Dubois, J.-E. Elucidation by Progressive Intersection of Ordered Substructures from Carbon 13 Nuclear Magnetic Resonance. *Anal. Chem.* **1988**, *60*, 2186-2192.

DARC Topological Descriptors for Pattern Recognition in Molecular Database Management Systems and Design

J.-E. DUBOIS,* G. CARRIER, and A. PANAYE

Institut de Topologie et de Dynamique des Systèmes de l'Université Paris 7, associé au CNRS, URA 34, 1 rue Guy de la Brosse, 75005 Paris, France

Received July 17, 1991

The DARC environment module called FRELs (Fragment Reduced to an Environment that is Limited) are concentric ordered graphs. We show that FRELs can be vectorized or built with four Graph Basic Branches (GBB). Colored branches can be derived by the introduction onto these GBBs of chromatic edges or vertices. Statistics of GBB occurrences within a set of carbon-centered topological FRELs are excellent tools for the investigation of environmental features in large structural databases. Occurrences of chromatic information, such as the statistics of "double-bond GBBs" within an isomeric set of olefinic FREL-Bs, are also easily carried out with the GBB concepts. A very small set of topological and chromatic GBBs provides excellent category descriptors and allows for more detailed analysis of molecular neighbor problems. They also facilitate cross-checking of GBB information within augmented atom substructures.

INTRODUCTION

Similarity tools are useful in many fields such as information systems and computer-aided chemical systems where structural knowledge is essential. These tools are based on the use of substructures, chosen so as to be optimal for certain pattern-recognition procedures. The substructures are defined on different levels of complexity involving, in the first place, their local atomic connectivities or their geometrical shapes, real or standard. The information is usually expressed by physical descriptors such as σ or E_s which are related to the structural primitive. The most popular similarity tools are associated with fragments of molecular entities, with indices forming similarity scales, and with pairing molecular groups or atoms with differential physical values.

DARC TOPOLOGICAL REPRESENTATION

An essential part of the structural information in chemistry depends upon the topological arrangement of the atoms in molecules—the connectivities and the nature of the atoms and bonds. Topology deals well with the aspect of connected sets of nodes and disconnected graphs but not with that of *distance* between vertices (atoms) and edges (bonds). Fortunately, molecules are not only conceptual topological objects but also geometrical objects whose definition depends upon geometric bond lengths and valence angles. In fact, real distances are present, implicitly or transparently, in topological descriptions, and topological descriptors are successful in various applications because they summarize the structures of the chemical species without any direct explicit identification. Various forms of topological representation have been considered, including tabulation of atoms and bond descriptors, connectivity tables, and matrices, with or without ordering of the molecular graphs, which are often considered as spanning trees and graph loop creations by ring closure steps. They are usually used for molecular description since the local information in a structural framework can be considered to be influenced by its "locality", i.e., its environment. The topological distances and the standard distances between vertices and edges of a "frozen graph" correspond to what researchers call path evaluation or distance.

In the DARC system, we have proposed a topological description of the *vicinity features*, and we term this the structural environment E of a focus atom F_o. This model, furthermore, is enriched by the imposition of order on the graph sites which are organized by locating the A_i and B_{ij} neighboring atoms[1] on concentric circles or spheres around F_o. The "distances" from the focus to the first A_i atoms are called a, and those between the A_i and the B_{ij} atoms are named b. They correspond to the standard bonds between these atoms. Information concerning a, b, c and A_i, B_{ij}, C_{ijk} reveals the inherent molecular arrangements of these sites, as can be seen from Figure 1. Any substructure within a structure can be considered to have an atom focus or a bond focus. The description of the substructure is called a **FREL** (Fragment Reduced to an Environment that is Limited). We have described various families of FRELs elsewhere.[2]

It is important to establish a quantitative description of such FRELs, and this may be done by the connex matrix associated with the **colored graph information** of the molecule or substructure. To date, we have produced topological descriptions by concatenation of the matrices corresponding to the suc-

0095-2338/91/1631-0574$02.50/0