Inferring Extended Virtual Knowledge from an EPIOS Conversion Graph of Overlapping Substructures

Michel Carabédian and Jacques-Emile 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 September 29, 1993[⊗]

In the EPIOS system, the elucidation strategy is guided by the progressive intersection of ordered substructures. Pairs combining substructural and spectral information are extracted from a suitable DARC-^{13}C-NMR databank. They constitute the basic units for the overlapping process and preserve it from a blind combinatorial matching used to build a molecule and/or its spectra. Enriched knowledge, needed by the system to define a larger solution space, is achieved by investigating the topological relations governing the substructural overlapping. A comparison of the experimental space and its virtual extension through these topological relations is presented.

INTRODUCTION

The combination of the computer and the ^{13}C-NMR spectrometer opened the way for great progress in structural analysis investigation technique. Limited at first to piloting pulse methods in NMR experiments, then to storing their results, today the computer leads to ever more sophisticated information drawn from the spectrometer. Its application also extends to *intelligent* tasks such as interpreting and exploiting this information.

Among these tasks, one of the most ambitious is surely the structural elucidation of unknown compounds from their spectrum. Two major obstacles characterize this area: the size of the search space resulting from the extreme diversity of chemical compounds; the absence of any real expertise and of a general theoretical model describing relations between the behavior of a ^{13}C nucleus (δ^{13}C) and its environment, capable of guiding the heuristic exploration of this space.

As an answer to these obstacles, greater computer power is not always a sufficient solution. It is of far greater importance to endow the system with appropriate knowledge, ensuring its own expertise over a broad range of applications for greater efficiency to begin with. This issue of knowledge is not always given the attention it deserves. In the systems inspired by the DENDRAL precursor, its role is secondary. This goes far to explain the persistent difficulties encountered. The evolution of these systems, marked by successive versions of isomer generators, all justified by their predecessors' limits, illustrates the inadequacy of their basic strategy, relying too much on these generators' performances. The knowledge used for isomer enumeration is generally elementary and involves a simple transcription of the idea of chemical shift. It characterizes reference substructures simply by the δ^{13}C behavior of their central carbon. In a former article, we analyzed the limitations and disadvantages of this type of *single resonance/substructure* knowledge.[1]

In contrast to these systems, from its inception the knowledge issue was situated at the very heart of the EPIOS (elucidation by progressive intersection of ordered substructures) system's development.[2] A preliminary analysis of the structure/δ(^{13}C) relationship resulted in the definition of the *multiresonance/substructure* (*MRS*) model, based on the idea of linked carbon atom couples in ELCO (environment that is limited, concentric, and ordered)[3] structural primitives. This notion of *linked atom couples*, lying at the origin of the EPIOS system, entirely determines both its organization and its function.[2]

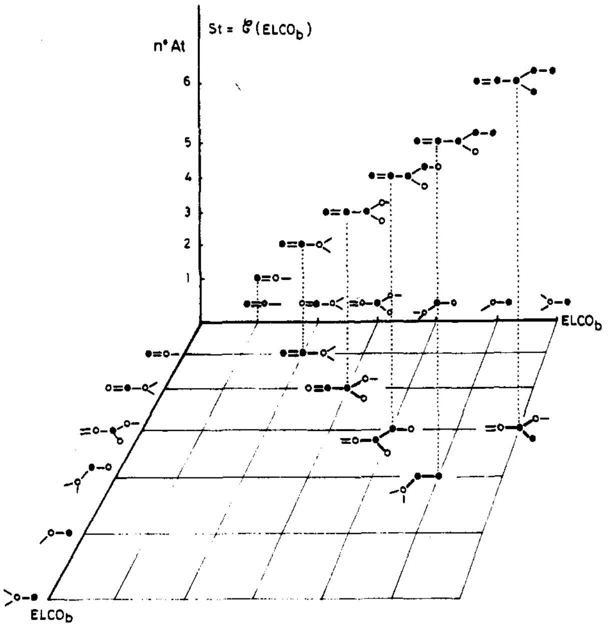


Figure 1. Diagram of the EPIOS principle of structure generation. Assembling the ELCO$_b$ primitives involves the progressive overlapping of all their focus/neighbor atom couples.

Figure 1 outlines the progressive assemblage of a structure, starting with its ELCO$_b$ primitives. The partial structure grows toward the complete structure by successively integrating the ELCO$_b$. Each ELCO$_b$ locally overlaps the structure formed by its predecessors and locally determines the integration conditions of its successors.

This method, proposed while most systems used only nonoverlapping substructures as constraints for their generator (CONGEN in DENDRAL,[4] ASSEMBLE in CASE,[5]), has today been adopted by the ACCESS[6] and COCOA[7] systems.

The conceptual operation, that ensures the growth of a structure by the progressive overlapping method, involves the search for *possible connections between ELCO$_b$ primitives*. If the knowledge required to guide the process is not included during the candidate structure generation step, starting with numerous alternative primitives, then the result will lead to

[⊗] Abstract published in *Advance ACS Abstracts*, March 15, 1994.

combinatoric difficulties comparable to those found by classical isomer generators.

To avoid these difficulties, we decided *to integrate the establishment of connections among $ELCO_b$ into the learning step of the EPIOS system.* These predefined connections are then a constitutive part of the knowledge acquired by the system which thus groups $ELCO_b$ primitives and certain relations among these primitives. However, merely limiting the combinatoric difficulties inherent in the elucidation is not the only advantage of this new knowledge. In this article, we show how these linked atom couples, formalized by the MRS model, enable us to enrich the *first level knowledge* directly extracted from our reference bank ^{13}C DP_{III} (16 000 structures, 195 000 $\delta^{13}C$) by a *virtual extension.* Moreover, the virtual knowledge resulting from extrapolation of previous knowledge exhaustively settles potential connections (i.e. those not observed in the reference population) of the 8587 available $ELCO_b$. This fresh knowledge determines an extension of the solution space covered by the system.

To the advantage of this extended solution space, we have added those obtained thanks to a physical organization of all the knowledge on the basis of $ELCO_b$ connections. This allow us to embed the EPIOS elucidation strategy in its own computerized architecture. We describe below this double exploitation of knowledge during the learning step.

FIRST LEVEL KNOWLEDGE

The choice of structural primitives, designed to support the knowledge of an elucidation system, largely conditions its performances. In EPIOS, the primitives chosen are the $ELCO_b$, composed of a central C_{Fo} carbon, its immediate C_{Ai} neighbors and their bonds with their outside environment (their atomic connectivity). This represents an optimal compromise to cover a broad solution space, thanks to their generic character, and a satisfactory characterization of their $\delta^{13}C$ behavior thanks to the MRS model.[3] More recently, Munk also adopted them as primitives of COCOA,[8] renaming them ACF (atom centered fragment).

Extracting these primitives from a reference structure population is the first step in modelizing the structural and spectral information contained in this population.

Figure 2 outlines the extraction process leading to the set of $ELCO_b$ constituting a reference structure. Each atom of this structure appears simultaneously in different $ELCO_b$ where it is either the focus or the neighbor of a focus. This information redundancy, associated with each atom, reveals the duality of this role in the molecule. It is an observation point for its own environment, participating simultaneously in defining the environment of its neighbor atoms. It is the *neighbor of its neighbors.* This redundancy is used in EPIOS to characterize the connections among the atoms of a structure. It is an essential element of information in an elucidation system whose aim is indeed to establish these connections in order to generate candidate structures.

However, the set of primitives obtained does not provide enough intrinsic structural knowledge to generate a structure. That requires combinatoric handling, seeking all assemblage possibilities for each primitive. This structural knowledge, uniquely composed of a set of primitives, is purely descriptive. It enables us to express the composition of a structure, in terms of primitives, but cannot predict their relative organization in that structure.

One way of surmounting this limitation consists in extracting, in parallel with the $ELCO_b$, their *mutual neighbor relations* in the reference structure. The knowledge thus acquired, grouping the primitives and their neighbor relations can then be called operational. It allows us to assemble a source structure (Figure 2) with no need for combinatoric handling. Since this knowledge stems directly from the reference population and expresses only information contained therein, we handle it as first level knowledge.

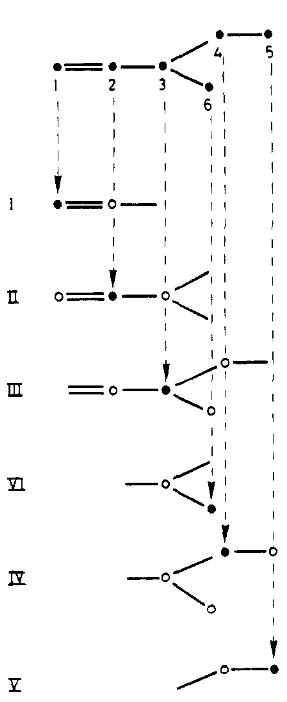


Figure 2. Extracting the $ELCO_b$ primitives from a reference structure. Each atom is simultaneously the focus of the $ELCO_b$ formed with its immediate neighbors and a neighbor of these latter.

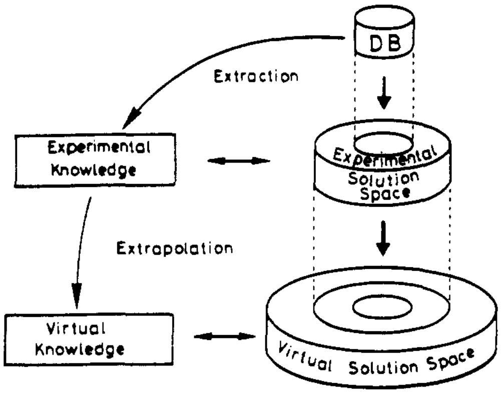


Figure 3. Generalizing knowledge and extending the solution space covered by a system.

The solution space covered by a system based on this first level knowledge is a limited generalization of the initial experimental population that it includes (Figure 3).

Such an extension of the solution space around the nucleus of experimental references used for learning is indispensable for elucidating new structures (i.e. absent from these references). This first level knowledge is limited to the true data of the bank. It concerns only the set of structures whose connections between $ELCO_b$ were observed at least once in the references. With the following example, we show both the extension and the limitations of the solution space associated with this first level knowledge. Thus, in Figure 4 a reference population is composed of three structures (A, B, C) from which eight $ELCO_b$ have been extracted (I–VIII). Neighbor relations, called Nr, observed for $ELCO_b$ II in structures A and B are Nr_1 (I–II–III) and Nr_2 (IV–II–V). They authorize the assemblage of new structures D and E,

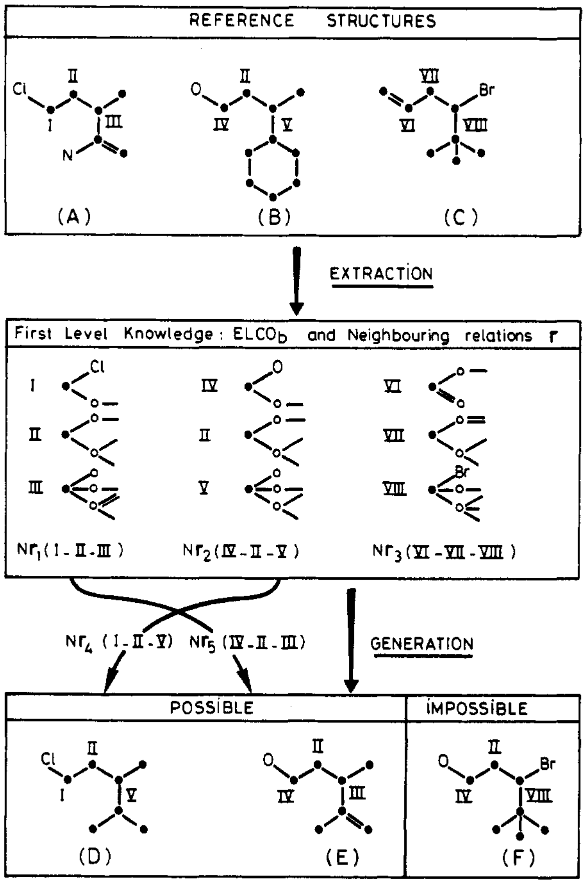


Figure 4. ELCO$_b$ I–VIII from reference structures A–C. Neighbor relations Nr$_{1-3}$ express the connections observed among these ELCO$_b$. The new relations Nr$_{4,5}$ are possible combinations formed from the Nr$_{1-3}$ relations. The ELCO$_b$ II connections in the new structures D and E are those observed in the reference structures A and B. The absence of an observed connection between the ELCO$_b$ II and VIII prohibits assembling the new structure F.

absent from the references. For structures D and E, the new relations Nr$_4$ (I–II–V) and Nr$_5$ (IV–II–III) are obtained by exchanging the neighbors (V and III) noted for ELCO$_b$ II. On the other hand, a target structure where ELCO$_b$ II is connected to an ELCO$_b$ other than I, III, IV, or V cannot be elucidated. This is the case for structure F, all of whose constituting ELCO$_b$ (IV, II, VIII), although known, cannot be connected because of the lack of an Nr (IV–II–VIII) relation.

This experimental knowledge, centered on the ELCO$_b$ and their neighbor relations, determines a so-called *first level solution space* (Figure 3) that permits certain predictions (here structures D and E) resulting from the overlapping method. The disadvantages of combinatoric permutation of all the ELCO$_b$ are avoided, but on the other hand, certain candidates (F) are inaccessible. *This first level solution space is then useful but not exhaustive.* To go beyond these limitations, we sought to increase the generic nature of the ELCO$_b$ neighbor relations in order to express the complete set of all the possible connections between available ELCO$_b$.

SECOND LEVEL VIRTUAL KNOWLEDGE

The *virtual* knowledge sought stems from extrapolation of the previous knowledge. It reveals, not observed facts, i.e. connections between two ELCO$_b$ in a reference structure, but facts that are potentially foreseeable during an elucidation, i.e. connections between two ELCO$_b$ in an unknown target. To define this virtual knowledge, we must, in an initial step, identify the conditions of a neighbor relation between two ELCO$_b$.

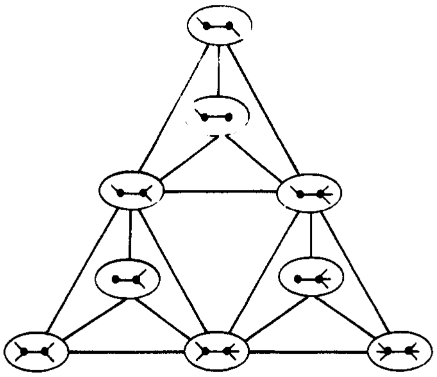
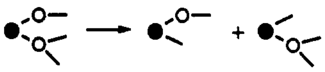


Figure 5. Graph nodes, being the elementary topological couples (TCo) of atoms of connectivity ≤4. The edges indicate the possible connections among these couples when they have an atom of same connectivity >1. The topology of any ELCO$_b$ can be decomposed on the basis of these nine elementary TCo (Figure 6).

Extracting the ELCO$_b$ from reference structures and assembling them to build target structures are reverse operations. They are respectively concerned with the observation and the establishment of an overlapping connection between two ELCO$_b$. This overlapping connection corresponds to the presence of a couple of linked atoms, simultaneously focus and neighbor in these two ELCO$_b$. The conditions of a neighbor relation between two ELCO$_b$ concern both their topology, i.e. atom connectivity, and their chromatism which specifies atom nature and bond type. These two components of chemical structures, topology and chromatism, inherited by the ELCO$_b$, are used for modeling knowledge. The generic nature of topology makes it possible to mask the diversity introduced by chromatism and thereby enables it to organize this diversity. In considering only their topology, the exhaustive set of possible ELCO$_b$ is reduced to 68 different topological ELCO$_b$. These topological ELCO$_b$ can be developped from a base of 9 generic topological couples, **TCo**, represented on a conversion graph (Figure 5).

In this conversion graph an edge joins two couples with at least one atom of identical connectivity (>1). The topology of an ELCO$_b$ with a focus of *n* connectivity is composed by overlapping, on this focus, *n* topological couples TCo with a node of identical connectivity:



(● is the focus, O a neighbor atom).

All ELCO$_b$ can thus be described in a four-dimensional topological space by its *n* components TCo*i* (1 ≤ *i* ≤ 4). In Figure 6a, we show the topology of the 10 ELCO$_b$ with secondary focuses and the 20 ELCO$_b$ with ternary focuses produced by the combination of their topological couples. The 3 primary ELCO$_b$ are part of the primitives. When we add the 35 ELCO$_b$ with quaternary focuses, we obtain the total of 68 possible different topologies for ELCO$_b$.

These elementary couples that also manage the ELCO$_b$ connections make it possible to localize their potential neighbors. Each ELCO$_b$ composed of *n* elementary couples is a potential neighbor of *n* sets of ELCO$_b$ containing one of these couples in its transposed form. These sets of potential ELCO$_b$ neighbors are seen in Figure 6a in the planes with one of these transposed couples as coordinate. Figure 6b shows the possible neighbors, according to the Fo/A2 couple, of the ELCO$_b$ – A1 – Fo – A2 < (–O–●–O<). These ternary focus

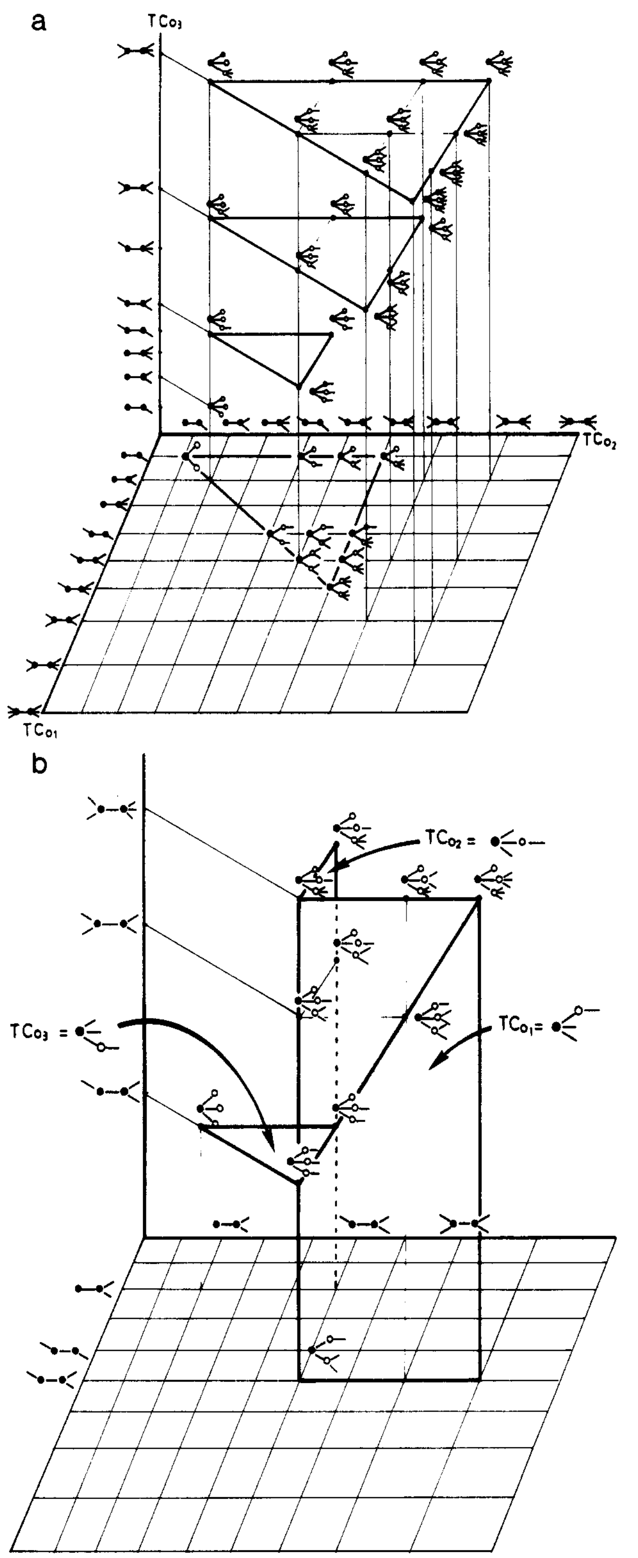


Figure 6. (a) Localizing the topology of $ELCO_b$ with primary, secondary, and tertiary focuses from their topological couple (TCo) components. (b) Localizing potential neighbors of $ELCO_b$ $-C_{A1}-C_{Fo}-C_{A2}<$ according to its Fo/A2 couple.

neighbors appear on the three planes of coordinate TCo_1═TCo_2═TCo_3 = >●–O–, transposed from the preceeding couple –●–O<.

The conversion graph of Figure 5 leads both to the topology of the $ELCO_b$ set and to their potential connections. It thus contains implicitly, in generic form, a global representation of the topology of the chemical structure set (composed of ≤4 connectivity atoms). It can be seen as the topological component, i.e. as the set of skeletons of chemical structures forming the complete solution space with which an elucidation system is confronted a priori.

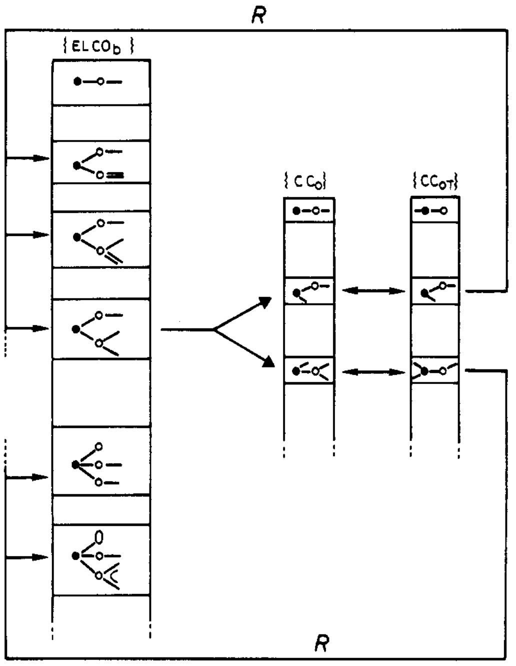


Figure 7. Transposed CCo_T forms of the elementary chromatic couple (CCo), composing an $ELCO_b$, indicating the set of its potential neighbors. The set 740 753 potential connections (679 304 of which are virtual) are established among the 8587 available $ELCO_b$.

In the EPIOS system, the 8587 available $ELCO_b$ are chromatic derivatives, specifying the atom nature and the bonds of the 68 topological primitives. They are composed of the combinations of 966 different generic chromatic couples, CCo, whose topology corresponds to the 9 generic topological couples TCo. During the elucidation of an unknown compound, their role is to ensure the adaptation of the complete solution space to the search for this particular target structure. This adaptation means reducing the space around the target structure from its spectrum onward. EPIOS thus deduces from its knowledge a specific generation graph including those $ELCO_b$ compatible with the query spectrum and all the possible connections of these $ELCO_b$.[2] This EPIOS elucidation strategy is inscribed in the physical organization of its structural knowledge. This organization and the enrichment it brings are described below.

MODELING AND EXTRAPOLATING STRUCTURAL KNOWLEDGE

First level knowledge, drawn from the reference bank, and virtual knowledge, extrapolated from the former, are loosely associated. Their physical organization is determined by their objective and stems from the idea of linked atom couples. Figure 7 provides an outline of this organization.

The set of 8587 $ELCO_b$ currently available is described on the basis of the 966 elementary chromatic couples CCo. These couples CCo of an $ELCO_b$ are the indices that designate, through their transposed form, CCo_T, the set of $ELCO_b$ satifying the topological and chromatic requirements of a potential connection. The extrapolation, carried out by an exhaustive search for the $ELCO_b$ containing the transposed couples, CCo and CCo_T, involves the elementary expansion (Figure 1) of each $ELCO_b$ toward the set of target structures to which it might belong. The established relations group those effectively observed in the reference structures as well as those extrapolated. Within this organization, each $ELCO_b$

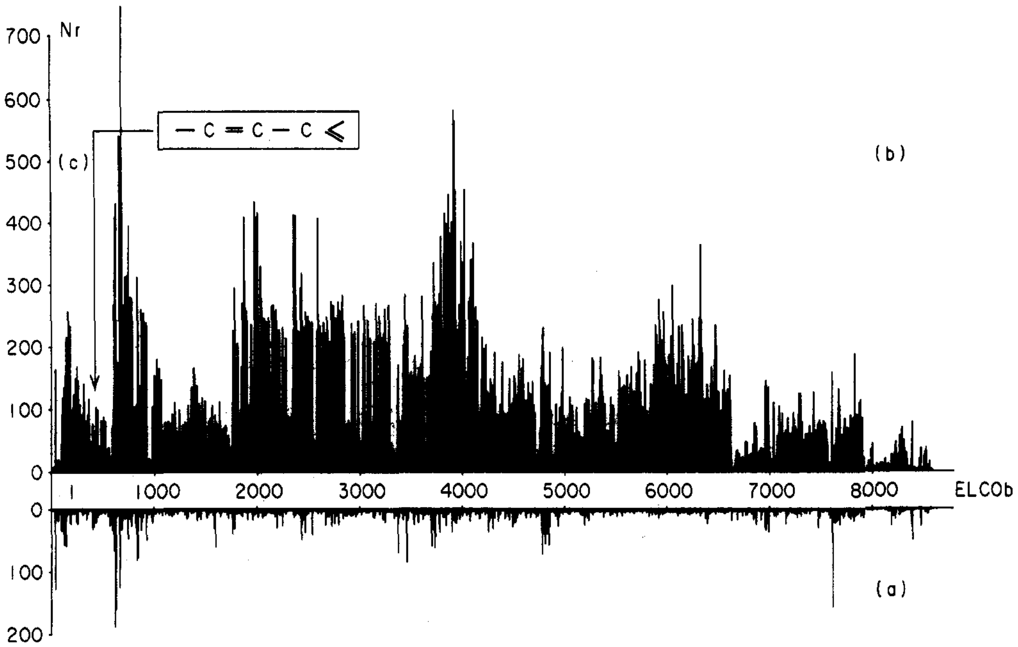


Figure 8. Distributions of the number of neighbor relations Nr (a) observed in the 16 000 reference structures and (b) potential for each ELCO$_b$. These distributions provide a comparative global image of first level knowledge (a) and its enrichment by virtual knowledge (b). (c) Virtual and observed neighbors for ELCO$_b$ $-C_{A2}{=}C_{Fo}{-}C_{A1}\leqslant$ are described in Table 1.

a priori *knows* all its potential neighbors. As an example, for ELCO$_b$

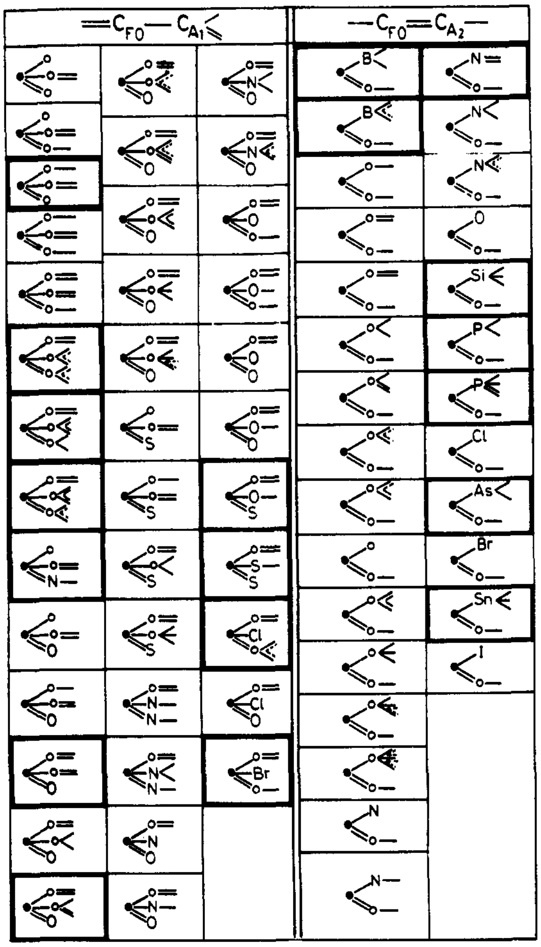
$$-C_{A2}{=}C_{Fo}{-}C_{A1}\leq$$

A set of 47 neighbor ELCO$_b$ are observed in 264 reference structures containing this motif: 26 ELCO$_b$ are linked according to the C_{Fo}/C_{A1} couple and 21 according to the C_{Fo}/C_{A2} couple. By an extrapolation search for virtual neighbors, 19 new virtual neighbor relations are established: respectively 11 and 8 supplementary ELCO$_b$ are associated with C_{Fo}/C_{A1} and C_{Fo}/C_{A2} couples. The set of these experimental and virtual neighbors is shown in Table 1. Here the number of relations linked to virtual knowledge is less than that of first level relations. In fact, statistically, the contribution of virtual relations is far greater than first level ones. The extent of the gain in virtual knowledge acquired by this generalization is very important. Indeed while 61 449 effective relations are observed, 679 304 virtual relations are established for the 8587 ELCO$_b$ studied. This enrichment (more than 90% of virtual relations) is seen in Figure 8. The two distributions shown allow for a visual global comparison between the number of observed (Figure 8a) and virtual (Figure 8b) relations associated with each ELCO$_b$.

Target structures containing fragment F of figure 4, inaccessible to first level knowledge, are now part of the solution space covered by virtual knowledge.

This extension of the solution space (Figure 3) ensures the elucidation of all structures composed of available ELCO$_b$. Each of these is a node of a graph detailing all possibilities of eventual expansion. This organization of structural knowledge thus a priori determines the set of environments that can be generated around each ELCO$_b$. This global knowledge graph, associating the 740 753 neighbor relations with the 8587 ELCO$_b$, explicitly provides all the structural information necessary to elucidate some unknown compounds. Interpreting its spectrum by EPIOS means selecting from this global graph a subgraph grouping only those elements, ELCO$_b$ and subspectra, compatible with the problem studied.

Table 1. Potential ELCO$_b$ Neighbors of $-C_{A2}{=}C_{Fo}{-}C_{A1}\leq$ According to Its Fo/A1 and Fo/A2 Couples[a]



[a] The virtual neighbors are outlined in boldfaced type.

EPIOS thus has, in its memory, a total display of its virtual solution space which it actualizes in order to solve a given problem.

CONCLUSION

Modeling knowledge required for structural elucidation and, more generally, for exploiting structure/δ^{13}C relations (simulation, spectrum assignment) often means obeying limitations dictated by the method chosen or by an algorithm developed concurrently. EPIOS, on the contrary, provides a unifying process involving the set of steps going from the acquisition of structural and spectral reference information to its exploitation in the form of elaborate knowledge. A single homogeneous method governs the extraction, the extrapolation, and the physical organization of this knowledge and leads, with no distortion, from the learning of the system to its way of functioning. Thus EPIOS relies on global knowledge outlining its application field and including not only structural primitives but also the exhaustive set of relations defined among these primitives.

This knowledge, drawn from 16 000 reference structures and characterized here by the 740 753 relations established among the 8587 available $ELCO_b$, is nonetheless insufficient to hope to cover the extreme diversity of chemical compounds. However, one of the advantages of the proposed modeling is to make it possible to identify the gaps and to fill them by an oriented experimental or statistical learning. This latter, based both on facts and on extrapolating from these facts, provides the conceptual and practical tools necessary to enrich the EPIOS system knowledge. In a forthcoming article, we shall show the use of these tools to exploit spectral information.

REFERENCES AND NOTES

(1) Dubois, J.-E.; Carabédian, M. Single-Resonance Subspectra/Subspectra (SRS) Investigations on the ^{13}C DARC Database. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 557–564.

(2) 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.

(3) Carabédian, M.; Dubois, J.-E. A Combined Model of Multiresonance Subspectra/Substructure and DARC Topological Structure Representation. Local and Global Knowledge in the ^{13}C NMR DARC Database. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 564–574.

(4) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference. 17. An Approach to Computer-Assisted Elucidation of Molecular Structure. *J. Am. Chem. Soc.* **1975**, *97*, 5755–5762.

(5) Shelley, C. A.; Hays, T. R.; Roman, R. V.; Munk, M. E. An Approach to Automated Partial Structure Expansion. *Anal. Chim. Acta* **1978**, *103*, 121–132.

(6) Bremser, W.; Fachinger, W. Multidimensional Spectroscopy. *Magn. Reson. Chem.* **1985**, *23*, 1056–1071.

(7) 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.

(8) Munk, M. E.; Lind, R. J.; Clay, M. E. Computer-Mediated Reduction of Spectral Properties to Molecular Structures. *Anal. Chim. Acta* **1986**, *184*, 1–19.