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

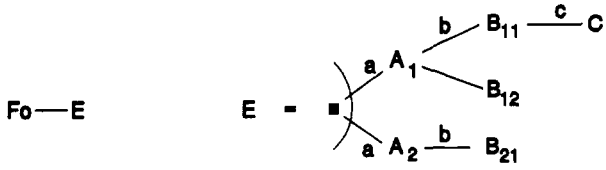


Figure 1.

cessive concentric environments of a focus.

ORDERED CHROMATIC GRAPHS—A REMINDER

The concept of chromatic information was introduced as a part of the DARC topological system in 1966–1968 and was revisited in 1973.[2] At the same time, Mowshowitz[3] proposed an information index for graphs which he called the chromatic information content. More recent developments concerning the extension of Polya's theorem are also based upon the introduction of different vertices, handled as colors, on a basic or reference compound, usually considered as a noncolored or black supporting graph. Each research group has tended to develop some convention[4] concerning chromatism or coloring. [In our system, as in this paper, the expressions **color** and **coloring** are used to specify any new information (**chromatic information**) that is added to a virtual spanning tree. In the field of **topological chirality**, the reference may be different. Coloring implies some new information in a comparison scenario. For instance, the Möbius ladders of the graph are considered to be **"uncolored"** if they have the same constitution (CH_2 and O chains) and if the vertices at which they are joined to the Möbius framework (usually N) are identical. Colored ladders correspond to different constitutional features (lengths and nature), distributed differently and thereby giving other signatures to the ladders. In both situations, for DARC coding, their coloration would be linked to the presence of heteroatoms in their substructures (ladders).] The DARC chromatic graph, its order, and the physical organization of the $ELCO_i$ environment of a focus have been described elsewhere.[4] Here, we shall summarize only some basic aspects in order to facilitate the understanding of the chromatic GBB proposals.

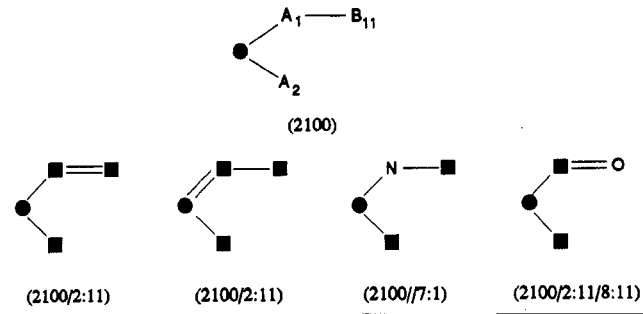
CHROMATIC GRAPH

A chromatic graph in which X is the set of vertices and U the set of edges linking these nodes has the vertices (X) and the edges (U) differentiated symbolically by the use of one or more colors.

A chemical structure is expressed as a chromatic graph whose vertices are atoms and edges bonds just as for a syntactic graph. The coloring refers to concepts and the relationships that unite them.

A chromatic graph may be written as $G_x(X,U,\chi_x,\chi_u)$. The graph $G(X,U)$ is termed a topological graph associated with the chromatic graph $G_x(X,U,\chi_x,\chi_u)$. In chemistry, chromatism and topology are interdependent. The coloring of a vertex or node depends upon the nature of the atom. The values of this χ_u coloring are derived from the $\tau\chi$ list of elements in the Mendeleev Table. The coloring thus imposes a maximum degree of connectivity on the vertices (valences), and it also determines certain types of coloring for the environments of the first neighboring row, A. Lists of primary and secondary chromatic information are used to determine the priorities used in labeling or ordering the graph. Priority decisions can be made by considering topological data or chromatic data. In some cases both aspects are used, and in such cases, the decision is influenced by a topochromatic set of data. One example of this is the Cahn–Ingold–Prelog rule in stereochemistry, where the primary ordering of the first neighbors is topological, but the final decision is chromatic in the sequence Cl > Me > t-Bu > H. A more topological priority rule would be t-Bu > Cl > Me > H, since the t-Bu group is the most highly substituted, i.e., it has the highest connectivity number.

Table I. Description of Chromatic FRELs

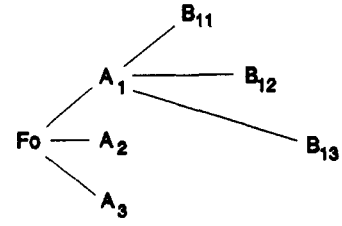


PROGRESSIVE DESCRIPTION OF A COMPOUND

The basic topological graph $G(X,U)$ is first organized as an ordered graph, either by a major or reference order or by finalized (application oriented) orders. This order is imposed on the graph, which consists of a focus and an environment. Both of these are labeled with the same topochromatic ordering rules used in the DARC/DUPEL ordering algorithm.[5] A comparison with the well-known Morgan algorithm[6] further clarifies the role of chromatism. Where the DARC/DUPEL algorithms use topology and chromatism, the Morgan algorithm gives precedence to the topological or uncolored graph of a molecule so that the focus is often determined by purely topological considerations. DARC operates with topochromatic functions that ascribe particular importance to chromatic information. A noteworthy feature of this system is the emphasis it places in the information management steps. In the DARC system, most of the information associated with a topochromatic table (Mendeleev Table and bond-type list) is divided into two main categories: primary chromatism for the most frequently occurring features and a secondary chromatism list for the numerous but less frequently occurring items of chemical information,[7] such as isotopes, charges, and so on. [This paper[7] is one of a series of 10 papers published between 1968 and 1976 describing the DARC code in documentation. Six other papers, published between 1970 and 1980, deal with the representation of stereochemical features and population correlation problems, and their description aspects are dealt with in a parallel series of papers (1968–1980). All of these articles were published in French journals: A complete bibliography is available from J.-E.D. upon request.] Ring information is handled as chromatic information, as are unsaturation and geometric and conformational information.

REFERENCE FRELS AND CHROMATIC FRELS

Some 70 primitive and basic carbon FREL-At-B (FRELs based on B depths, as shown below)



are available to build the framework of a topological description.[8] In Table I, the molecules, represented by a circle for the focus and a square for the vertices of the environment organized in first and second neighbor rings A and B, are described by progressive input of topological and chromatic information. The topological descriptor points only to the

Table II. GBB Representations and Their Ordered Numerical Vectors

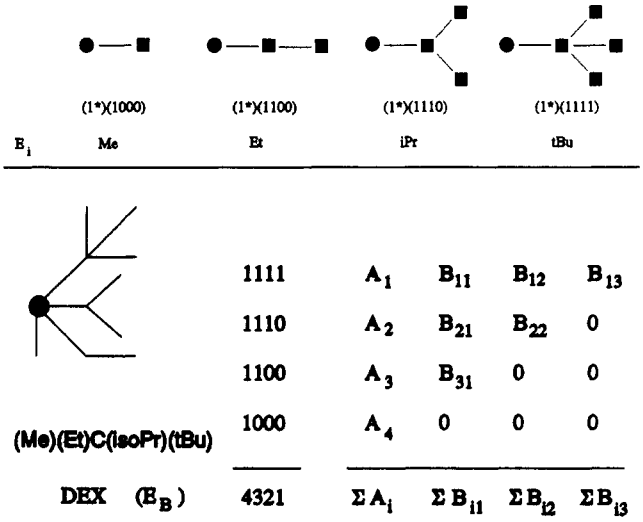


Figure 2. DEX of a FREL-At-B with four branches borne by the focus.

existence of the vertices (atoms, excluding H), and this is expressed by the descriptor of existence, DEX. In Table I, butane is considered as methyl 2-propane (focus on carbon 2); the environment has two A_1 and A_2 atoms and one B_{11} atom, and its description then is simply 2100, where there are zero B_{22} and B_{33} vertices. This existence graph provides the basic description of colored fragments while retaining the same topological organization modified by the introduction of chromatic features at either the bond and/or the node level. The A_i and B_{ij} coordinates are used for further enhancements of chromatic information. In Table I, the unsaturation information is introduced for edges (the double bond is labeled with "2:" in positions 1 and 11). This chromatic information leads to a "descriptor of links" (DLI) bound to the DEX in a DEX/DLI relationship. The nature of the vertices (DNA) is added in color form and localized by the DEX coordinates. The full Descriptor of this Environment Limited in B is written DEL=DEX/DLI/DNA. In Table I, nitrogen (designated with its atomic number 7) is introduced in the third example, DEL=DEX/DNA and localized on the A_i vertex, and in the last example, a bivalent oxygen leads to a DEL=DEX/DLI/DNA descriptor locating the oxygen as "8:11" in the B_{11} position and its double bond "2:11" marked by its terminal B_{11} edge location.

GRAPH BASIC BRANCHES (GBB)

We have retained as primitive graphs of the FREL-At-B, the four graphs of Table II, which are considered to be the basic branches. With this fundamental set of GBB we can describe numerically all the branches or groups allowed in the FREL-B space.

DARC topological descriptors are the simplest vectors that can be used in building matrices of important FRELs with several branches described in DARC code.

The four GBB branches or primitive graphs shown in Table II, which also correspond to the simplest FRELs, (methyl, ethyl, isopropyl, and tertiary-butyl) constitute a set of primitives that can be used to derive the descriptors of more complex carbon FRELs. Complex FRELs can be decomposed in the GBB vectors. The vectors in the matrix (Figure 2) are chosen in descending order (1111 > 1110 > 1100 > 1000).

In statistical studies of DARC topological structure files and in organizing database management systems (DBMS), it is most useful to identify the branches with 1–4 atoms—and more beyond the B_{ii} limits—that are neighbors of the foci. In this way it is possible to be very specific about the role of local information.

Table III. Examples of Topological and Topochromatic GBBs

(τ)GBB/Indices	1000	1100	1110	1111
^{X}GBB/Indices	1	2	3	4
a or b	1	2,1 2,2	3,1 3,2	
a or b	1	2,1 2,2		
Z A or B	Z 1	Z Z 2,1 2,2	Z Z 3,1 3,2	Z Z 4,1 4,2

As the FREL groups are thus identified in terms of classical fragments, comparisons with some fragment-based systems can be made more easily. It should be noted that embedding the GBBs in the FRELs permits a good evaluation of the near environment. This is shown later in connection with GBB occurrences.

Such an important increase in the information potential of more or less substituted aliphatic groups is only one aspect of the general problem of structure evaluation. In fact, interesting properties and reactivities are largely controlled by chromatic information. These considerations have led us to **extend the GBB concept in order to create basic sets of topochromatic (or colored) GBB.**

CHROMATIC GRAPH BASIC BRANCHES (*X* GBB)

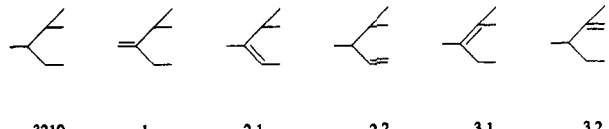
In generalizing the concept of GBB to that of *colored* (or *topochromatic*) GBBs, we obtain important specifications concerning the coloring of the FREL-At-A and FREL-At-B with regard to the identity of atoms *A* and *B*. All this information, localized on the level of the *GBB, can be used to describe the molecules and their component FRELs*. A table of the functionalized graphs (GBB) can be established, and some of its elements are shown in Table III. To these GBB can be added some simple or complex chromatisms associated with the common chemical functional groups, i.e., –OMe, —C═O, —C≡N, –X, –CO–OMe, and —C≡CH, since their existence vectors correspond to the GBB. Table III shows that the number of primitive chromatic GBBs used is very low indeed. The table contains only five olefinic and three acetylenic bonds and seven atom vertices colored with various *Z* valencies. These sets are small enough to provide excellent search keys.

For conjugated systems, the description is slightly more elaborate and will not be developed here. It should be emphasized that in such cases the information is distributed between two vectors and includes the focus. The chromatism mapped onto a $G(X,U)$ graph corresponds to the presence of chemical functional groups, and we call such fragments "functionalized FRELs".

FREL DESCRIPTION THROUGH THE GBB

Several functionalized FRELs can have the same DEX. In other words, they correspond to a common carbon FREL-At-B where the chemical function introduced can be located in the various positions of the FREL. In FREL 3210, for example, the five positions of the double bond correspond to an introduction into each of the five branches, as shown in Table IV.

Table IV. FREL (C=C) Occurrences of Isomeric Environments in the 1988 CAS File



3210	1	2,1	2,2	3,1	3,2
Occ: 62,500	7,700	12,000	6,200	15,100	5,800

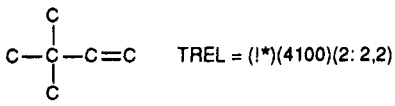
This is the only FREL-At-B that can have the five GBB with a double bond; localization of the double bond is limited by the connectivity constraints on the neighboring atoms, F_o, A_i, and A_iB_{ij}.

The occurrences in Table IV correspond to those of the B-FREL (C=C) in the CAS 1987–1988 File (approximately 8×10^6 compounds). It should be noted that the total occurrence of the 3210 (C=C) FRELs amounts to 46 800—a much lower value than that of the corresponding 3210 carbon At-B-FREL. The descriptors of these five B-FRELs conform to DARC writing (Table I), but a generic descriptor mentioning GBB categories can also be considered.

The category descriptors are useful when measuring search statistics in files in order to extract information for each of the olefin isomeric FRELs having the same topological DEX (e.g., Table III). Thus, it is possible to investigate the influence of the neighboring FREL sites containing a specific (C=C) GBB. The concept of the population descriptors (TREL = TREX + TRLi) was proposed in the **Population Correlation Theory** of the DARC system for isofocal compounds.[9] Here we propose to generalize the description of the FREL populations by introducing the various locations of their sites of chromatism in their basic GBB category indices shown in Table III. As an example, using the data in Table III:

$$\begin{aligned}\text{TREL} &= \text{TREX} + \text{TRLi} \\ &= (1^*)(3210)(2{:}|1/2,1;2,2/3,1;3,2|)\end{aligned}$$

The chromatic information 2: is found in the 1 GBB, the 2,1 and 2,2 GBB as well as in the 3,2 GBB. Usually, the number of categories is less than the five theoretically possible, as in the example given below:



TREL = (!*)(4100)(2: 2,2)

where the FREL (C=C) and the TREL (C=C) are identical.

STATISTICAL STUDIES OF LOCAL ENVIRONMENT INFLUENCE

Because of the DARC concentric description, the FREL organization (here mainly FRELs-At-B) of a structural database permits the elucidation of local situations with the help of quantitative expressions of local environments (E_i). Statistics can bear upon occurrences of E_i environments, or FRELs, in the file. FREL concentric fragments or carbon substructures have been investigated as sources of statistical surveys of large structural files and have been extended to ethylenic FRELs.[10] These results, combined with graph basic branches, open the way to more elaborate correlations. Thus, the effect of a local structure limited to an E_B environment on a GBB makes it possible to define occurrence categories such as Occ(GBB) = f(FREL-B) = f(DEL). The occurrences related to categories concerning the GBB are very variable. Table V shows the evolution of the (1100)(2:|2,1|) ● = ■-■* branch, taken from the CAS File (1987–1988) according to the FREL-B environments containing it. For some environments (DEX: 1100, 2210, 3311, 3200), significant occurrences are noted, while for other values, such as 3320 and 3322, the frequency of occurrence is much lower. Interpretation of these data requires a much more detailed analysis of the structure of the environments and, in particular, of their acyclic (FREL/AC) and cyclic (FREL/CY) natures. The FREL/CY effectively groups together the cyc-acyc FRELs and the cyc-FRELs. In the 2210 FREL (C=C), the acyclic contribution is limited to one-tenth of the sum of the occurrences of the four isomer categories shown in Table VI.

Table V. ● = ■-■* Branch Sensitivity to Local FREL-B Environments: Its Occurrences Throughout the CAS File Organized with DARC Topological FREL-At-B[a]

DEX	Occ	DEX	Occ
1100	141 700	3211	4 300
2100	22 000	3300	9 300
2200	134 200	3310	11 500
2210	112 800	3311	53 500
2211	39 300	3320	5 900
3100	38 600	3321	14 800
3200	42 900	3322	3 600
3210	12 000		

Table VI. Occurrences of the Various FRELs (C=C) Contributions in the CX Occurrence of Aliphatic E_i (AC) with Cyclic (CY) Isomeric E_i[a]

Occ:	240,000	11,300	41,700	39,500	20,000
	(1*)(2210)		(1*)(2210)(2:(2,1))		

[a] Accumulated occurrences = 112 800, which should be compared to the (2210) carbon FREL global occurrence (240 000).

A GBB/E_B correlation search is based on information associated with the category occurrence distribution. We proposed some guidelines[2] by means of statistical studies of the FREL-At-B, noting in particular that *occurrence in the reference carbon FRELs tends to decrease as internal steric hindrance increases for acyclic compounds and also when cyclization becomes greater.* In this paper, when introducing a double bond into a DEX, we generally observe that FREL (C=C) occurrence decreases as compared with the corresponding carbon FREL occurrence. However, exhaustive analysis of the chemical corpus reveals some exceptions. The combined use of these tools, the FREL-B and GBB, allows for essential local topological analysis and description.

CONCLUSION

The concept of a fundamental set of graph basic branches is complementary to that of the concentric ELCO and of FREL. It allows a more detailed analysis of molecular neighbor problems. It makes it possible to cross-check GBB information with that of augmented atom (AA) substructures and of their associated descriptors (HA and TW). As compared with these substructures, the GBB corresponds to a higher organization by its ordered concentricity linked to the DARC description. Its focus, which is the root of the branches, is eccentric, whereas in AA substructures, the focus is central. Its usefulness has been demonstrated both in analyzing and managing files and in computer-aided structural systems that use pattern or shape recognition methodologies. Moreover, this concept of the graph basic branch description leads to powerful new tools for knowledge extraction strategies to be implemented on very large molecular files. This dual view of structures, concentric site layout, and branch organization of the structural vicinity, can be extremely valuable for properties depending upon shape and local distances. These combined tools should also prove very useful in the creation of knowledge autoproduction generators from heterogeneous sources with different structural descriptors.

REFERENCES AND NOTES

(1) Dubois, J.-E.; Viellard, H. Théorie de Génération-Description: I. Principes Généraux. *Bull. Soc. Chim. Fr.* **1968**, 900, 905, 913.

(2) Dubois, J.-E. Ordered Chromatic Graph and Limited Environment Concept. In *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: New York, 1976; Chapter 11, pp 335–370.

(3) Mowshowitz, A. Entropy and Complexity of Graphs. 4. Entropy Measures and Graphical Structure. *Bull. Math. Biophys.* **1968**, *30*, 533–539.

(4) Walba, D. M. A. Topological Hierarchy of Molecular Chirality and Other Tidbits in Topological Stereochemistry. In *New Developments in Molecular Chirality*; Mezey, P., Ed.; Klumer Academic Publishers: Hingham, MA, 1991; pp 119–129.

(5) Dubois, J.-E.; Bonnet, J. C.; Laurent, D.; Viellard, H. Formalisation du Code DARC-DUPEL et Applications. *Bull. IRIA, Suppl. Bull. IRIA* **1971**, Note 2 - 1-16.

(6) Morgan, H. L. Generation of Unique Machine Descriptions for Chemical Structures, a Technique Developed at the Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–110.

(7) Dubois, J.-E.; Panaye, A.; Viellard, H. DARC-Théorie de Génération Description: V. Description d'Entités Nécessitant un Chromatisme Secondaire: Ions, Radicaux, Isotopes, Sels, Hydrates. *Bull. Soc. Chim. Fr.* **1973**.

(8) (a) Dubois, J.-E.; Panaye, A.; Attias, R. DARC SYSTEM: Notions of Defined and Generic Substructures. Filiation and Coding of FREL Substructural (SS) Classes. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 74–82. (b) Dubois, J.-E.; Hennequin, F.; Bossu, M. Utilisation du Système DARC à des Fins Documentaires. *Bull. Soc. Chim. Fr.* **1969**, 3615–3623.

(9) (a) Dubois, J.-E.; Laurent, D. Système DARC. XI. Théorie de Population-Corrélation II. Description d'une Population Chimique. *Bull. Soc. Chim. Fr.* **1969**, 2449–2455. (b) Dubois, J.-E.; Hennequin, F.; Boussu, M. Utlisation du Système DARC à des fins Documentaires: Methodes de Préparation des Cétones Aliphatiques Saturées. *Bull. Soc. Chim. Fr.* **1969**, 3615–3623.

(10) Carrier, G.; Panaye, A.; Dubois, J.-E. Topological Structural Information in the CAS File: Statistical Occurrences of DARC Concentric Fragments. 1. Basic Carbon Substructures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 110–121.