eralized spiro nodes. Generalized spiro nodes are ring nodes at which a ring system can be split into separate ring systems without breaking rings. The identification of these nodes is very simple: for each node of connectivity four or greater, temporarily remove all bonds incident to it; if the ring system becomes two or more separate subsystems, the node is a generalized spiro node.

CONCLUSION

The enhanced SSSR search program described here works efficiently and with a low failure rate. This has made it possible to process the entire ring file in the CAS Registry system. Further improvement can be made by the implementation of the proposed new phase 3 in the SSSR search algorithm. With the new phase 3, the whole algorithm becomes mathematically rigorous and program failure can be eliminated.

ACKNOWLEDGMENT

We thank G. G. Vander Stouw for encouragement and constructive discussions and M. E. Olowinski for providing us with CAS Ring File processing data. We also thank the CAS Documentation Turnover Group for editorial support.

REFERENCES

(1) Patterson, A. M.; Capell, L. T.; Walker, D. F. *The Ring Index*, 2nd ed.; American Chemical Society: Washington, DC, 1960.
(2) *Ring Systems Handbook*; Chemical Abstracts Service, American Chemical Society: Washington, DC, 1988.
(3) Balaban, A. T.; Filip, P.; Balaban, T. S. *J. Comput. Chem.* **1985**, *6*, 316-329.
(4) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93-102.
(5) Corey, E. J.; Petersson, G. A. *J. Am. Chem. Soc.* **1972**, *94*, 460-465.
(6) Wipke, W. T.; Dyott, T. M. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140-147.
(7) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 78-82.
(8) Gasteiger, J.; Jochum, C. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43-48.
(9) Schmidt, B.; Fleischhauer, J. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 204-206.
(10) Zamora, A. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40-43.
(11) Mayeda, W. *Graph Theory*; Wiley-Interscience: New York, 1972.
(12) Plotkin, M. *J. Chem. Doc.* **1971**, *11*, 60-63.
(13) File 301 328-331 300 (*Chemname*, *Chemsis*, *Chemzero*), Dialog Bluesheet, revision April 1984, Dialog Information Retrieval Service, 1984.
(14) *Index Guide*, Appendix IV; Chemical Abstracts Service, American Chemical Society: Washington, DC, 1987.

Topological Structural Information in the CAS File: Statistical Occurrences of DARC Concentric Fragments. 1. Basic Carbon Substructures

GERARD CARRIER, ANNICK PANAYE, and JACQUES-ÉMILE DUBOIS*

Institut de Topologie et de Dynamique des Systèmes, associé au CNRS, Université de Paris VII, 75005 Paris, France

Received October 10, 1989

The Chemical Abstracts Service (CAS) Chemical File of some 7.5 million structures, structured in DARC topological space, was analyzed statistically for the frequency of occurrence of different topological substructures. These topological DARC fragments, known as FREL-Bs, are described concentrically around a single carbon atom with an environment limited to the first (A) and second (B) neighbor atoms and treated as a spanning graph tree. Occurrence searches were carried out for the 70 primitive carbon FRELs based on σ bonding. These represent potential acyclic or cyclic substructures with an ordered environment and are called FREL-AC or FREL-CY, respectively. Such ordered environments provide more than mere statistics of local and global descriptions; they lead to trend analyses. Certain broad correlations, linked to the idea of structural neighbors, can thus be observed. Analysis of the statistics associated with cyclic structures revealed some counterintuitive exceptions to the general rule that occurrence should be inversely related to complexity. As an example, the presence of various familiar chemical families such as steroids and terpenes can be detected within the context of the file and can be tagged by means of specific FREL associations.

INTRODUCTION

Many activities in chemical information, computer-assisted documentation (CADoc),[1-4] or computer-assisted design (CAD)[5-10] are based upon the exploitation of sets of substructures selected within a given context. Selection criteria vary with the application, but a search for the basic organizational parameters should help to formalize and elucidate original sets of substructures for new applications. Furthermore, an understanding of these parameters will permit informed improvement of the existing systems.

There is only one global statistical study that is frequently updated with regard to all known structural data.[11,12] This study deals with the familiar molecular fragments such as rings, chains, substituents, and functional groups. The significance of the statistical occurrence of such substructural fragments is complex. The selection of compounds into the database reflects human interest in the structures both for their properties and for the challenge they present to the synthetic chemist. It is because of this study that we can claim a global view of this large database and can use it indirectly to evaluate chemical knowledge. The database and its analysis provide data pertaining to the existence or nonexistence of compounds, and these can be used to draw conclusions as to the ease or lack of ease of access to structural families and even to specific structural entities. This is the basis of present-day chemical taxonomy, which can be exploited with the help of systematic nomenclature in an effort to understand familial relationships that may exist between various groups of chemical compounds.

In this paper, we outline another approach to this global view. This uses atoms, together with their bonds and local topologies, as structural primitives. The aim is a quantitative and qualitative analysis of the local and global structural environment of an atom.

A minimal structural fragment, such as an atom and its neighbors, constitutes the simplest basic fragment, but it is often an inadequate representation of the whole, particularly

0095-2338/90/1630-0110$02.50/0

in more complex molecules. Use of a larger fragment in the description leads to a FREL (fragment reduced to an environment that is limited), whose topology reflects some important structural elements such as atoms two and three bonds away from the central atom and not bonded to it, three-dimensional effects, and anomeric effects. Such second and third neighbors are identified in substructures designated SS_B or SS_C.

The choice of such a substructure with a variable, but well-defined and ordered, environment can facilitate the interpretation or the simulation of physicochemical properties of the parent structure. Substructures defined in this way often characterize a site, bonds, functions, and neighboring interactions. Various spectroscopies reflect certain types of substructural detail, and this can lead to the extraction from structure–spectra data of substructure–signal data pairs. There are also correlations between bulk physical properties, such as partition coefficient, and substructure. Correlations have been proposed[13,17] that link such physical properties to site contributions and topology, even to topography. The additive rules frequently used in studies of physicochemical behavior often focus on structural fragments that are selected for a variety of reasons.

In this study, we chose the set of carbon skeletal fragments ranging from the simplest (C) to the most highly substituted ($C–tBu_4$) and include the *70 basic saturated carbon skeletons* that are possible. To search for the occurrence of all of these fragments in the full database would be extremely difficult unless the file was organized in terms of these fragments, i.e., by topological concepts. The EURECAS file of the DARC system contains the database in just this organization and was used for substructural searches for fragments consisting of a central atom with ordered environments defined by the central atom's A and B neighbors.

With the EURECAS file, the object was to determine the frequency of occurrence of the basic skeletal moieties and so obtain new information pertaining to substructures, related substructures, and trends in this set of about 8 million chemical compounds. The analysis will not be linked to the chemical motifs such as rings and substituents that are commonly discussed. Topological data offer instead a different but complementary view of the database to that provided by Stobaugh,[11,12] although we share Stobaugh's reservations concerning the value of statistical results. Our analysis provides an insight that is valuable for the interpretation of structure–property relationships. From the taxonomical view, the interplay between a molecule's framework and its physicochemical properties is susceptible to topological analysis at the level of general carbon FRELs, which can be associated with the appropriate property data. It should be noted that a study of the environment shared by closely related substructures can lead to the recognition of some important relationships which can be detected by studying familially related substructures. (The substructure space can be ordered diversely and can even constitute a hyperstructure graph.) It is necessary to select the substructure and the neighborhood, but this choice, which is to some extent formal, must be made on the basis of the data included in chemical structure databases.

CONCENTRIC LABELED STRUCTURES

Many different types of substructures are used for documentation, database generation, correlation of physical properties, and structuring of chemical knowledge for artificial intelligence work, such as spectroscopy-assisted structure elucidation. Those defined concentrically about an atom or bond as origin are used frequently, and the concept of concentrically ordered atom environments is an important part of the DARC paradigm for the topological description of

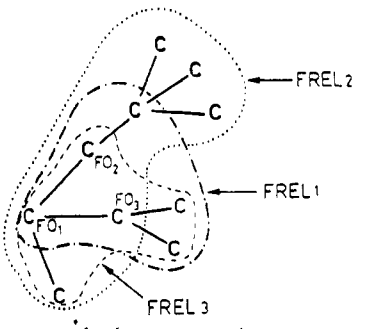
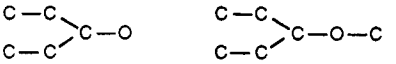


Figure 1. Structures and substructures. A structure or a substructure is considered an ordered concatenation of concentric substructures of fragments reduced to an environment that is limited (FREL). This structure is overlapped by seven FRELs (three of which are considered here) defined around the focus atom by two ranks of links and atoms. Each fragment is considered a topological environment around a site considered to be its focus.

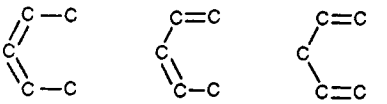
structures and substructures.[20,21] Concentric substructures are of particular interest in this work because it has been shown[19] that FRELs can be classified and organized with precision and by use of DARC, many varied families of FRELs can be created. We now propose to determine the criteria with which a set of common substructures can be used to derive and localize more finalized substructures.

The structure of any chemical compound can be derived by a more or less complex assembly of basic structural units, such as FRELs[1] (cf. Figure 1). Exhaustive evaluation of all the possible FRELs in a population of structures can lead to very large numbers, and it is therefore sometimes useful to select the most representative FRELs for further study. It is not easy, however, to evaluate the information content of a substructure; many commonplace intuitive concepts are found lacking when attempts are made to quantitate this. For example, it is often assumed that *simplicity* and *frequency* go together, but in the two substructures shown, that with one



42 100 68 000

less atom, i.e., the simpler one, occurs less frequently in the Chemical Abstracts Service (CAS) File.[22] Similarly, the frequency of occurrence of the isotopological substructure shown with the $CH_2(Et)_2$ motif depends upon the number of



1 200 108 400 13 200

bonds (0, 1, or 2) separating the two double bonds. The most commonly occurring of these three substructures is the one in which the double bonds are conjugated and separated by one single bond. This of course results from the resonance stabilization of such structures. These examples, however, point up the difficulties that can arise from the use of criteria which are too literal.

The frequency of occurrence of FRELs, measured in the EURECAS database by means of DARC,[22] allows their evaluation, and it then becomes possible, through study of relationships between different FRELs and their frequencies, to establish the criteria necessary to specify a homogeneous structural population. The analysis relies at each stage on both statistical data and chemical considerations. The population that is studied here is that of "carbon FRELs", an approach that corresponds to the natural exploratory efforts that would be undertaken by a chemist. This allows study of the carbon skeletons of both aliphatic and aromatic molecules. The

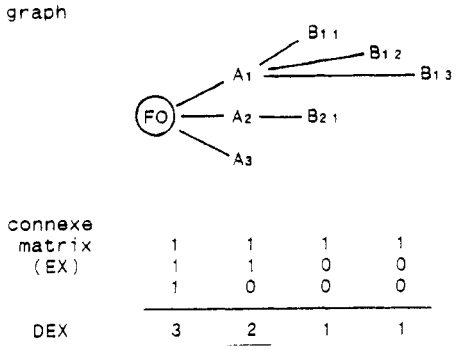


Figure 2. DARC encoding of the topology of a FREL; the existence descriptor (DEX). By application of an ordering function, each site A, B_{ij} is given a label of total ordering. The elements of the matrix connected to the graph are replaced by 1 or 0 depending upon whether or not they are occupied. Their summing according to a specified procedure provides a numerical expression of their topology. The graph ordering of an environment $A_1A_2A_3/B_{11}B_{12}B_{13}/B_{21}$ is different from its descriptor summation $A_1A_2A_3/B_{11}B_{21}/B_{12}/B_{13}$. The symbol ϵ is used as the generic term for environment, whereas C_i values are associated with ordered graphs and numeric descriptors of a given environment.

presence in a molecule of carbon is of course fundamental to organic chemistry, a tenet which is supported[18,19] by the percentage of carbon-containing compounds in the Cambridge Crystallographic Database (99%) and in the CAS File (94.47%). Use of carbon FRELs also permits easy subsequent approach to the notion of functionality, by substitution of heteroatoms for carbon.

REPRESENTATION OF SATURATED CARBON FRELS

FRELs are classified primarily by their focus and their depth. A FREL can focus on an atom (FREL-At) or on a bond (FREL-Li) and, in either case, can involve one row of neighboring atoms (FREL-A), two rows (FREL-B), or more. A FREL-At-B therefore is centered on one atom surrounded by two rows of atoms.[20,21]

(1) FREL Description. In this work FREL topology will be described by means of the descriptor of existence (DEX).[20,21] If a given topological site is occupied, an index is set to 1, otherwise, to 0. The DEX is simply the sum of these indices over the entire FREL. The topological sites are treated in strict order with the preference being given first to those A positions with the greatest number of B-position substituents, and the DEX that results is biunivocal for a depth of two rows of atoms, as can be seen from Figure 2. The DEX for FRELs centered on a particular atom ranges from (0000) to (4000) for FREL-A and to (4444) for FREL-B as can be seen from the 70 FRELs in Figure 3. A saturated carbon fragment will be defined by its focus and environment; a methyl group will be 1*(0000) and a *tert*-butyl group 1*(3000). All the foci in this study are carbons, and so the focus descriptor will be omitted. A methyl radical becomes 0000 and the *tert*-butyl is 3000.

A descriptor of an environment that is limited (DEL) accounts for not only the topology (described by the DEX) but also the bond or link chromatism (DLI) and that of the atom (DNA). Unsaturated bonds were not allowed in this first study: a reference topology was developed having single bonds only, and unsaturations could be introduced subsequently as perturbations of the basic topology.

(2) Occurrence of Saturated Carbon FRELS and Classification by Isomer Level. Searches in the EURECAS databases show that the frequency of occurrence of the 70 saturated carbon FRELs in Figure 3 is often very high, over 500 000 for the most common. To simplify the process, a search was first carried out in the MINICAS file, which is a 1% random sample of the EURECAS database. If more than 5000 retrievals were found here, an estimate of the number of retrievals in the EURECAS file was arrived at by extrapolation. Otherwise, the search was carried out in the larger file. The distribution frequencies were sorted by carbon number and subsorted by DARC ordering[21,22] to produce Figure 4 which shows the following: (1) The frequency of occurrence decreases as the carbon number increases. For the FREL-At-B with 12 or more atoms, the number of retrievals is always less than 1000, while for FRELs having 11 or fewer atoms, it is almost always on the order of several thousands or tens of thousands. (2) The frequency of occurrence falls as the topology grows more complex. This result could be expected and attributed to steric hindrance.

This global view, however, contains some more complex and sometimes contradictory local phenomena. If the 70 FRELs are broken down by isomer level, as in Figure 4, three major zones, 1, 2, and 3, along with a "root zone" can be defined according to the degree of branching at the focus. Examination of these zones reveals the following:

- Zone 1 contains FRELs that have a very high frequency of occurrence (>100 000) and which have a maximum of five carbons. These FRELs are not very discriminatory because they can be embedded in many structures. This zone includes all primary structures, some secondary structures, and a single tertiary structure.
- Zone 2 contains most of the secondary, tertiary, and quaternary FRELs. This is a result of the high variability of the environment. The variations show some continuity and often some important outliers.
- Zone 3 is characterized by the presence of FRELs with a very low frequency of occurrence (usually several hundred) that always have over 11 carbons. These are very selective FRELs.
- Zone R consists of the five root FRELs 0000, 1000, 2000, 3000, and 4000. These were removed from the analysis on an a priori basis because, having only one row A, they have no exterior environment.

It is in zone 2 that the principal features of the observed behavior can most easily be detected. In this subpopulation, six FRELs give more than 100 000 retrievals:

structure	DEX	occurrence
$CH_2(CHC_2)(CH_2C)$	2210	240 000
$CH_2(CC_3)(CH_2C)$	2211	178 000
$CH_2(CHC_2)_2$	2220	131 000
$CH(CHC_2)_2(CH_2C)$	3320	153 000
$CH(CC_3)(CHC_2)(CH_2C)$	3321	190 000
$C(CHC_2)_2(CH_2C)(CH_3)$	4320	121 500

The high carbon density for some of these does not appear to limit the occurrence rate.

Replacement of a hydrogen by a carbon, which undoubtedly increases the complexity of a FREL, does not necessarily lead to a lower rate of occurrence. As an example, three pairs of FREL-Bs with exceptional occurrence increases are

structural change	occurrence rate	adjunction
2111 → 2211	29 000 → 178 000	B_{21}
3310 → 3311	59 000 → 76 900	B_{13}
3222 → 3322	242 → 43 500	B_{33}

A logical assumption is that the number of carbon atoms in the FREL should play a major role in determining the FREL's discriminating power. If this assumption is correct, then the number of carbons will be a primary explanatory factor of the occurrence rate. The choice of 11 as the maximum number

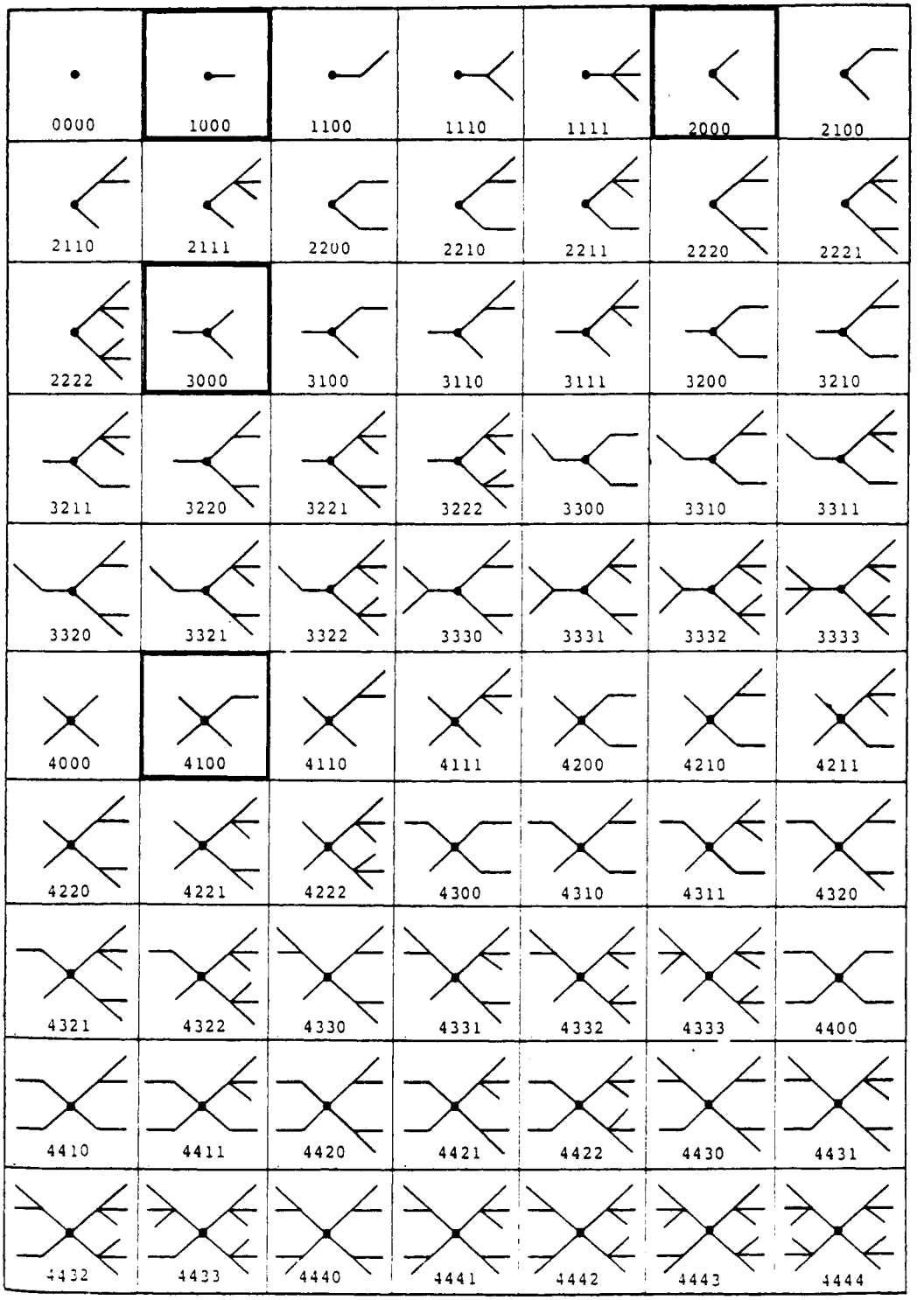


Figure 3. Set of the 70 ranked FREL-Bs centered on an atom with its environment labeling. The primary root (1000) leads to a set of 4 FRELs, the secondary root (2000) to 10 FRELs, the tertiary root (3000) to 20, and the quaternary root (4000) to 35. The DEX numbers describe the environment of the focus atom. The methane FREL 1 (0000) and the other FRELs are fully described with their focus and their environments, e.g., 1 (1000)...1(4444).

of carbons ensures that the population of the 70 saturated carbon FREL-At-B moieties is reduced to 48 significant FREL-At-B species, or the set of FRELs contained in zones 1 and 2. This reduced population will serve as a foundation for the study of the carbon FRELs whose properties we shall attempt to interpret. If, as a first approximation, we can say that occurrence decreases as internal steric hindrance increases, then this occurrence–topological complexity relationship must be refined further because it is not obeyed in all cases. FRELs cannot be defined merely by their topology; they must be envisaged in terms of FREL filiations and their partial or total embedment within rings. These two aspects of the problem are examined in the next section.

FILIATION PATHWAYS AND OCCURRENCES

Having studied the various FRELs separately, it is now possible to examine the relationships which associate them when a carbon atom is added to, or removed from, a FREL. In this way, the influence on the FREL of the environment of its focus can be determined. (The FRELs concerned are included in a genealogical graph, and their relationships with one another are called *filiations*. The fundamental filiations depend upon small topochromatic changes. These are studied

on the level of the matrices and their location within the filiation graph. The location of the changes in a FREL become an essential aspect of the transfer of information within this special space.) In Figure 5, each FREL can be viewed successively as (a) a loose FREL, (b) a FREL with a loose focus, or (c) a FREL with a loose environment. Thus, for a FREL-B, substitutions are allowed, not only in row B (the frontier) but also according to the following rules. For (a), substituents are allowed on the focus and on the internal environment (row A); for (b), they are allowed on the focus; and for (c), on row A. A maximum number of free sites is implied for each of the row B atoms. Thus, any FREL can be associated with the corresponding son FRELs by "site adjunction" and with the corresponding father FRELs by "site ablation". Father FRELs are substructures of the son FRELs. The terms ablation, and adjunction, and frontier are defined in the Glossary.

For example, ablation of a single site of FREL-At (3321) can lead to any of three father FRELs, all of which have a loose environment—1E. A single adjunction of the same FREL can lead to three son FRELs. There are only two son FRELs if this FREL-At is viewed as having a loose environment and only one if it is seen as having a loose focus (Figure 6). For *n* allowed adjunctions and ablations, this FREL gives, respectively, 19 son FRELs and 27 father FRELs, all of which are loose. It is of interest to determine whether these "filiation relationships" have a well-defined influence on the occurrence rate of the related FRELs.

In the case of the saturated carbon FRELs, there are four possible adjunctions that would allow a FREL to be converted to its son. These are termed "filiation pathways" and are shown in Figure 7 and summarized here.

Adjunction on the focus

→ adding a row A (pathway 1)

Adjunction on row B

→ adding a row B

passage Me → Et (pathway 2)

passage Et → iPr (pathway 3)

passage iPr → tBu (pathway 4)

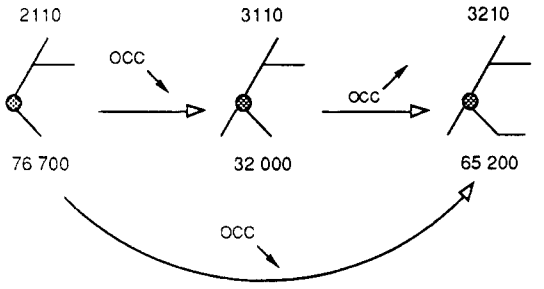
From the 48 FRELs that were selected with 11 or fewer carbon atoms, it is possible to generate 98 filiations by using one or other of the 4 possible filiation pathways. The influence of each of the pathways has been analyzed in terms of the occurrence variations between father and son FRELs. The four pathways have some interesting properties, described below.

(1) The FO–H → FO–Me Filiation. There are 27 pairs of FRELs in which the son differs from the father by an adjunct on the focus. In 24 of these cases, the transition FO–H → FO–Me leads to an occurrence decrease, and in only three cases is there an occurrence increase (cf. Figure 8a). This anomaly seems to be of minor importance; only a few 32XX structures are affected by it.

(2) The FO–Me → FO–Et Filiation. There are 24 pairs of FRELs that result from this filiation. For 17 of them, there is an occurrence increase, while for the 7 shown in Figure 8b the occurrence decreases. The increase is easily understood as a consequence of the greater possibilities for substitution offered by an ethyl group compared to the methyl group. A partial explanation for the decrease may lie in the increased steric crowding that accompanies an increase in topological complexity. The enhancement effect produced by the new substitution possibilities would be countered by the steric effects.

The notion of complexity cannot be applied, however, to all the pairs, because it would not allow an explanation for the increase in occurrence as filiations 4211 → 4311 or 4220 → 4320 take place, while filiations 4200 → 4300 and 4210 → 4310 are accompanied by decreases. These anomalies must be explained in some other way, and it will be shown subsequently that some FRELs contain information concerning possible cyclic structures. This information is not considered in this particular statistical study.

Filiations that involve an adjunct either on the focus (FO–H → FO–Me) or on row A (FO–Me → FO–Et) have opposite effects on occurrence rates. For most such pairs of FRELs, the father FREL has an occurrence rate that is greater or less than that of the son FREL. To evaluate the roles of these two types of transition, it is useful to consider second generation relationships (father–grandson). Such a relationship can be discerned if one considers occurrence variations on the level of successive pairs with the first substitution on the focus and the second on row A. It is observed that the corresponding occurrence rate is consistently decreased. Adjunction on the focus augments the complexity of even very slightly hindered FRELs and has more effect than adjunction on row A, which increases the substitution possibilities.



(3) The FO–Et → FO–iPr Filiation. In these 24 filiations, the trend to decreased occurrences is less clearly marked than in the preceding transitions. Only 14 of the 24 show such a decrease. For less hindered pairs in which filiation produces, for the first time, an isopropyl group, there is relatively little increase in occurrence. The same substitution applied to FRELs that already possess an isopropyl group (either independently or as part of a *tert*-butyl residue) leads to a remarkable increase in occurrence. Only the pairs 4410 → 4420 and 4211 → 4221, which show slight increases (2840 to 4470 and 1560 to 8900, respectively) fail to be so affected. Again, it can be seen that if increased substitution possibilities lead to increased occurrence, the increased occurrence must be partly independent of topological complexity. In this type of transition, the FRELs in which steric crowding is greatest occur most frequently in the file.

(4) The FO–iPr → FO–tBu Filiation. This filiation is marked by a general tendency to lower occurrence frequencies. There are only 3 exceptions to this among the 20 pairs.

To summarize, it seems to be generally the case that frequency of occurrence decreases whenever local connectivity increases. The single exception to this is the FO–Me → FO–Et filiation where transition from an unsubstitutable row A to a substituted row B leads to a sudden increase in the number of available free sites and, concomitantly, of substitution possibilities. This increase in occurrence reflects, however, the strict conditions imposed upon the query, as can be seen from the fact that the direct filiation FO–H → FO–Et leads to a decrease in occurrence frequency—a trend countered by only 5 of the 24 cases (cf. Figure 8c).

Nevertheless, a certain number of the FREL pairs represent exceptions to the rule of "increasing complexity → decreasing occurrence". In Figure 9 are grouped all the son-FREL exceptions whose occurrence is greater than that of the father (the FO–Me → FO–Et filiation is excluded). It should be noted that the FRELs in Figure 9 may be involved in different types of filiation such as those leading to or from cyclic structures.

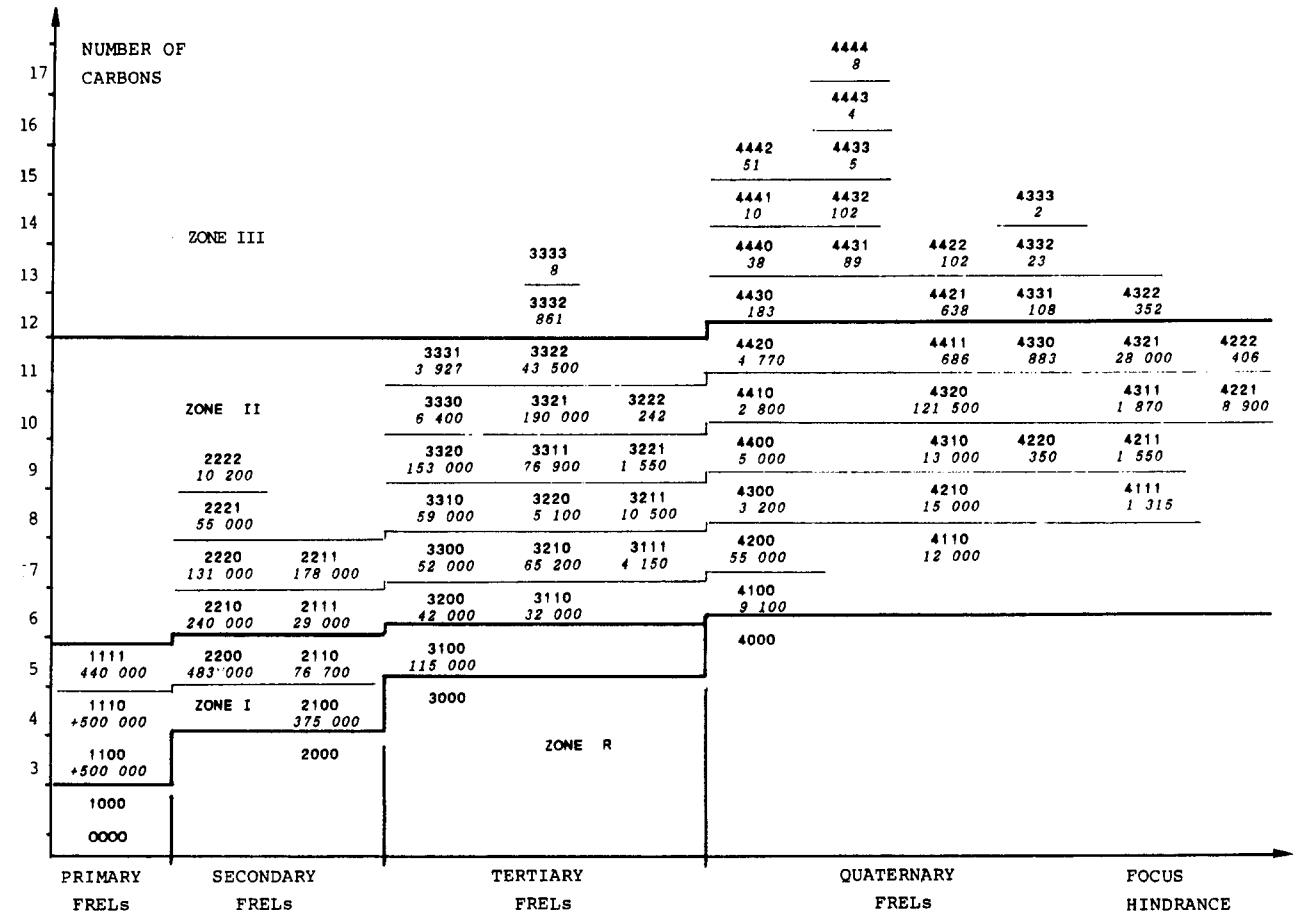


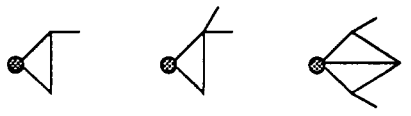
Figure 4. Carbon environment table based on FREL occurrences in the chemical corpus. First axis: Number of carbons (FO + E = 1 + $\sum A_i B_{ij}$) expressed in the DEX compressed matrix. Second axis: Complex ordering based on group radial defined by the connectivity of FRELs (P, S, T, Q) and secondary ordering of the B_{i1}, B_{i2}, B_{i3} environment of C in $-C\sum_{B_{i1}}\sum_{B_{i2}}\sum_{B_{i3}}$ in the group. All the saturated groups up to and including C–tBu are included in this FREL table.

In all, 15 of the FRELs considered here are distinguished by their failure to obey the above rule. Among these, two groups can be identified. A group of seven FRELs show rather small variations in occurrence frequency and are involved in a single abnormal pair. The remaining eight show very great variation in occurrence frequency, and all can be involved in several aberrant pairs, particularly if double transitions are considered.

FREL OCCURRENCE AND CYCLIC NATURE OF BONDS

The presence of rings in a structural formula depends only upon bond linkage, i.e., upon topology. Differentiation between cyclic and noncyclic bonds will be considered first before turning to the questions of bond multiplicity and atom nature.

(1) FRELS and Ring Closure. The presence of rings multiplies greatly the number of possibilities. There are, for example, 180 different existence FRELs centered on a depth B atom and with a focus atom belonging to one or two three-membered rings. To study these, it is necessary to



consider ring-closing bonds within the FREL-B. The combinatorial increase far exceeds the interest of the new FRELs. In many cases, FO, A, and B are atoms belonging to cycles whose closures are outside the FREL spanning tree. These cases are treated more elaborately in that the FRELs are considered to be embedded in cyclic structures, and this must be considered in their description and the resulting DEX and DLI values.

(2) Acyclic and Cyclic FRELS (FREL-AC and FREL-CY). (2.1) Defining the Different FREL Categories. Basic carbon FRELS are fundamentally topological and spanning graphs, and all other supplementary information about them *colors* this basic fact. The process of moving to a more colored graph, one including stereo or ring closure data, for example, involves the addition of some chromatic data. Ring closure of an open graph is precisely this sort of enhancement. The term cyclic FREL is not rigorous but is very useful because it effectively expresses the fact that FREL of the spanning graph contains some cyclic information, and it is compatible with the chemist's intuition. So far in this paper, the possibility that the bonds in FRELs might have a cyclic nature has been ignored. It is possible to specify whether or not a bond is involved in a ring, and thus two major categories of FREL are created: FREL-AC, all bonds are acyclic; FREL-CY, at least one cycle involves some of the FREL bonds. The second category can be subdivided into FREL-cyc, where all bonds are cyclic, and FREL-cyc/acyc, composed of an arrangement of cyclic and acyclic bonds forming a chemically feasible substructure. FREL-CX corresponds to the group of all these colored FRELs. The character of the bonds is open; they may be cyclic or acyclic, and so all possible combinations will be taken into account.

(2.2) Occurrence Significance. If a study of the global FREL populations is to be carried out by using the FREL-CX occurrence frequencies, allowing for the possibility that rings may be present, the occurrence rate of each FREL-AC and FREL-CY must be identified.

The logical addition or union is identical with the frequency

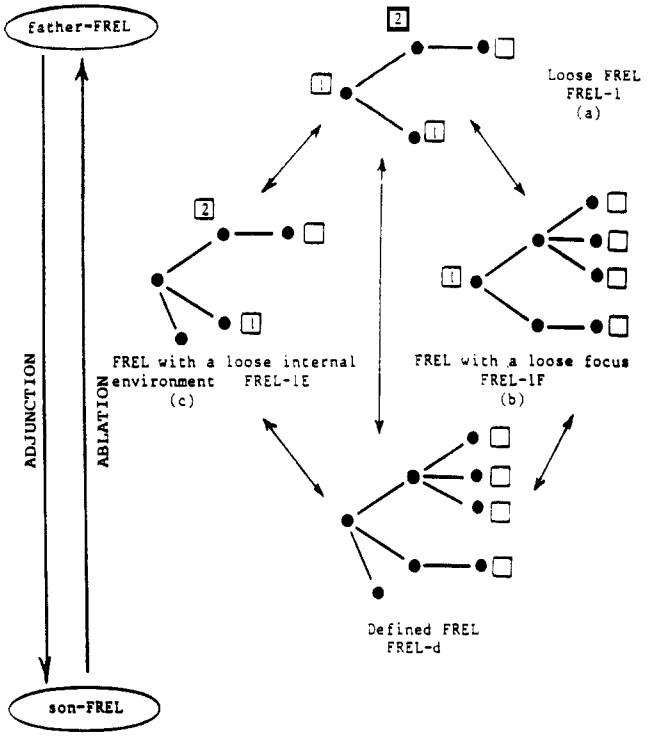
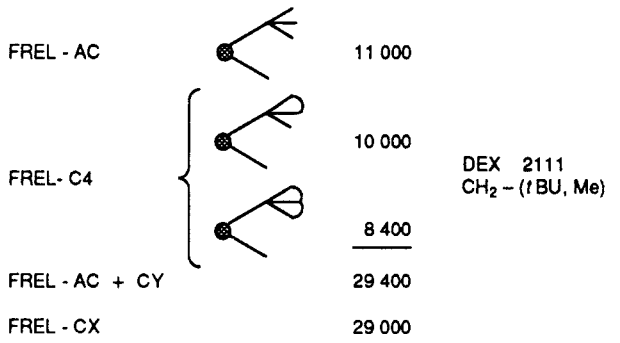


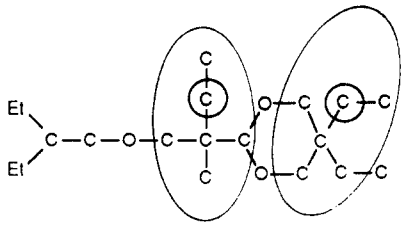
Figure 5. Localization of allowed substitutions in the various FREL categories: on the focus and the internal environment (FREL-1) on the focus (FREL-1F); on the internal environment (FREL-1E). Substitutions are implicit on the environment frontier. Filiation (FREL father → FREL son) by site adjunction or ablation. The free site index (□) may contain an internal digit which indicates the number of allowed substitutions. The absence of a digit indicates that all possible substitutions are allowed in conjunction with the connectivity of the node.

of occurrence of the FREL-CX but the arithmetic sum can, in certain cases, be higher. For example, in the case of the FREL 2111 [C–(tBu, Me)], the occurrence breakdown is



This difference stems from the fact that certain structures can contain two FRELs with the same DEX. In the present case, 400 structures have this characteristic; an example is

RN: 23546-62-5



molecular formula: $C_{18}H_{36}O_2$

Name: *m*-dioxane, 5-ethyl-2-[1-{(2-ethylbutoxy)methyl}-1-methylpropyl]-5-methyl- This phenomenon is of limited scope because the variation in occurrence between the

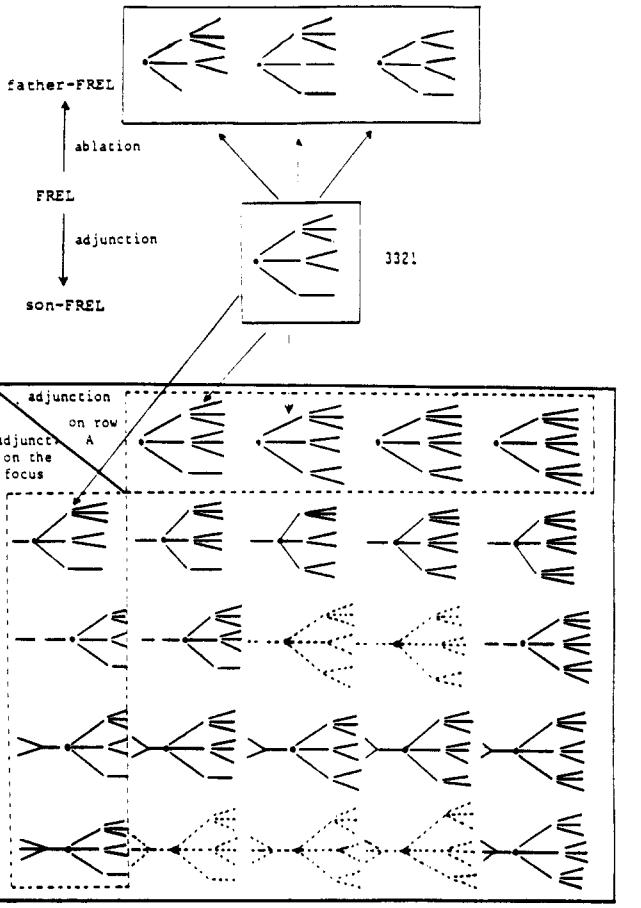
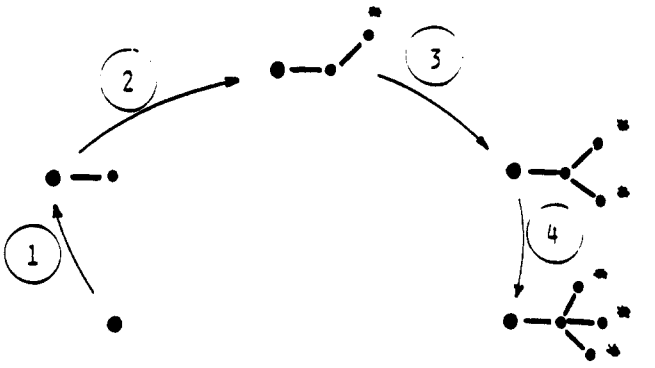


Figure 6. Filiation of the FREL-At (3321) (–CtBuiPrEt) to FREL-At (4444). Three sons are generated with one adjunction and 19 with *n* adjunctions. Five redundant sons (drawn with a dotted line) are indicated to localize their neighborhood relationships.



PATHWAYS	1	2	3	4
Cases of decreasing occurrence (number)	24	7	14	17
Cases of increasing occurrence (number)	3	20	10	3
Prevalent direction	88% ↓	74% ↑	58% ↘	85% ↓

Figure 7. Filiation pathways and their action upon occurrence. Adding a node to the graph induces filiations. If the group admitting the adjunction involves a Me–Et transition, the occurrence level of the FREL son increases. In other cases, the occurrence level decreases.

FREL-CX and the arithmetic sum of the occurrences of FREL-AC and -CY is estimated to be only a few percent, and

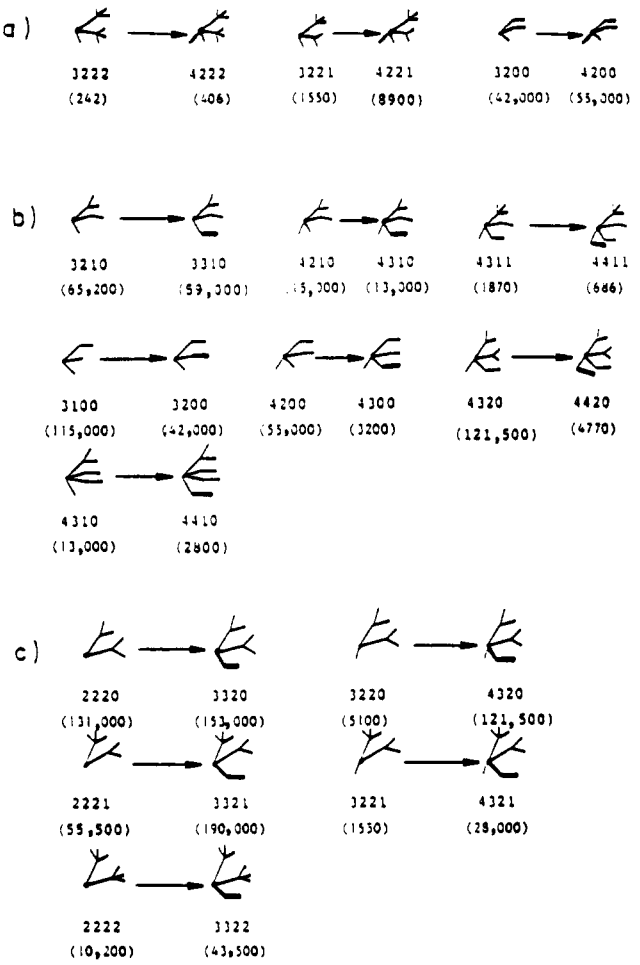


Figure 8. Occurrence dependence on simple methyl adjunction to the focus and first neighbor (a, b); combination of the two adjunctions or ethyl introduction on focus (c).

consequently it is assumed that the occurrence frequency of the FREL-CX is a sufficiently representative measure of the total number of FRELs (AC and CY).

(3) Cyclization and Occurrence in FREL Table/Zone 2. The cyclic character of bonds in the FRELs in zone 2 (Figure 4) is of interest because this zone includes all topologies with 6–11 carbons and also happens to include all the FREL anomalies that have been discussed here. FRELs of zone 1 with less than six carbons were not studied, even though their occurrence frequencies are high, because they show a generally poor selectivity.

(3.1) Acyclic FRELs (FREL-ACs) Simplified Coherence Table. It is straightforward to study the occurrence frequencies of FREL-ACs because for each DEX there is a single corresponding FREL containing acyclic bonds, and a table of isomeric and homologous FRELs with the corresponding occurrence frequencies can be assembled:

↑		3310	3322		4420	4411	4330	4321	4222
		7	9		7	10	3	9	9
homologs		3330	3321	3222	4410	4320		4311	4221
		50	1500	9	40	50		20	10
2222		3320	3311	3221	4400	4310	4220	4211	
2500		1500	750	155	250	100	15	30	
2221		3310	3220	3211	4300	4210		4111	
1650		2360	460	315	320	75		350	
2220	2211	3300	3210	3111	4200	4110			
4000	8900	2080	5900	2600	5500	2640			
2210	2111	3200	3110			4100			
64800	11000	17200	13000			9200			
								isomers →	

[Regular occurrence decrease of acyclic fragments in the CAS File for zone 2 (6–11 atoms) is seen. Spectacular decreases in vertical columns from bottom to top and some irregularities in decreases in the horizontal rows from left to right (e.g., n = 6 and 3310, 3321) are seen.]

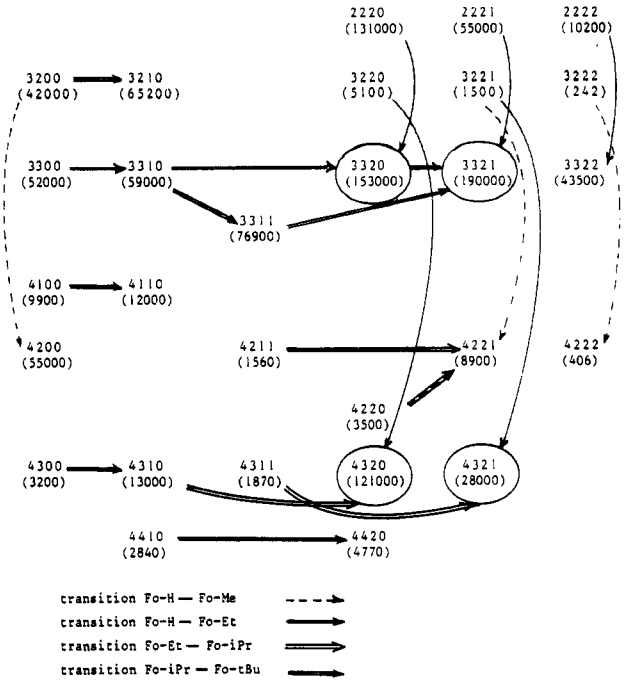


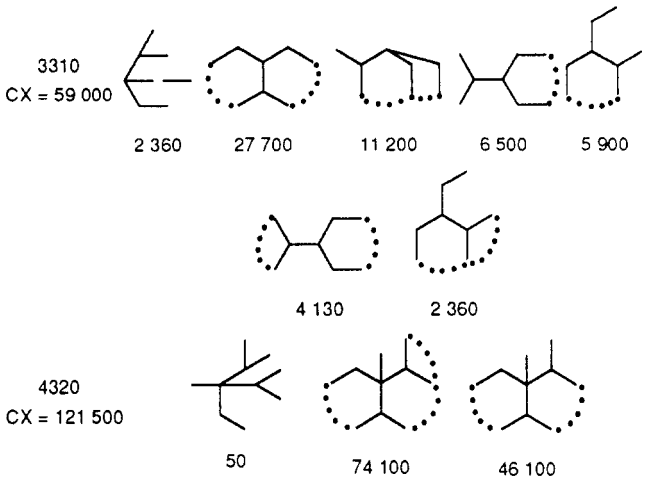
Figure 9. Genealogical environment of FRELs; normal decreases in occurrence and odd situations. The father–son relationships marked by arrows correspond to occurrence decreases. Four examples of surprising increases are shown (in circles). FRELs 3320, 3321, 4320, and 4321 show maximal occurrence in comparison to their predecessors. Moreover, FREL 3320 ($CEtiPr_2$) leads to a surprising increase.

The weak occurrence level of these FRELs is immediately apparent. In general, FREL-CXs have a very weak population of FREL-ACs, usually on the order of a few percent and sometimes as low as 1%. The occurrence frequencies noted for these FREL-ACs are very orderly, and they clearly decrease when the steric crowding increases in the FREL focus (Sy < Ty < Qt). The same is true when the number of carbons in the FREL increases. For this category of FREL, topology is indeed the parameter that takes account of occurrence. These occurrence/topology correlations can be summarized as follows: *The occurrence of FREL-ACs tends to decrease when their local complexity increases in the FREL table.* With one exception (3321), this rule is observed in the vertical columns of the FREL table shown, and it is also fairly well obeyed in the horizontal row, there being just 4 inversions (2211, 3210, 4200, and 3321) of the 41 possibilities. The anomalies detected in the FREL-CX (e.g., with FRELs 3320, 3321, and 4321) are not observed here and may be due to the *cyclization potential of certain FRELs* and some archetypal rings of certain chemical families (alkaloids, steroids, etc.) rather than the contribution from acyclic FRELs.

(3.2) Cyclic FRELs (FREL-CYs). (a) Breakdown of Cyclic and Acyclic FREL Occurrences. For each DEX of zone 2 of the diagram shown in Figure 4, the occurrence frequencies of the various FREL-CYs have been investigated. This takes into account the cyclic character of bonds. Only the more frequently occurring FRELs for each DEX were considered.

In most cases, transition from a FREL-AC to a FREL-cyc/acyc with only one cyclization leads to an increase in occurrence frequency. It should be recalled that FREL-CYs are characterized by the presence of the cyc shape and of the set of cyc/acyc shapes. The relative importance of these shapes can be determined by calculation of their *representativity index* (the percentage of the occurrence of a CY shape compared to that of the corresponding FREL-CX). These indices vary from 1% to 87%, and in view of this wide variation, it is clear that a FREL-CY could, in some cases, have maximum rep-

resentation with a minimum of different shapes or, conversely, require a large number of shapes to be almost entirely defined (cf. the contribution for the 3310 and 4320 FRELs).



For FREL 4320, 2 of 11 chemically feasible FREL shapes cover 99% of its representation, while, as expected for the 3310 FREL, a potential precursor of the 4320 FREL (less two carbons), 6–7 FRELs are needed to ensure the same level of representation.

(b) Evaluating the Influence of a Cyclic Bond on the Representation of a FREL. All the cyc/acyc and cyc FRELs of a single DEX whose sole difference was the cyclic nature of just one single bond were grouped together. It was possible to study 45 FREL pairs in this intermediate zone. When going through a transition that involves change of an acyclic to a cyclic bond, 32 pairs show a decrease in frequency occurrence. This decrease is very large in nine cases and is small or insignificant in the remainder, because of the weak representivity of these FRELs. A significant increase characterizes 13 FREL pairs. Transformation of the first acyclic single bond leads to an occurrence increase (Figure 10). Transformation of the second bond, in contrast, leads to an occurrence decrease, and it comes as no surprise to discover that FRELs containing four cycles with a common carbon are rare.

(c) Filiation of cyc–acyc FRELs (cyc/acyc). Analysis of the filiation pathways starting from the principal cyclic residues, considered roots, and adding only acyclic bonds (Figure 11) leads to the logically satisfactory result that occurrence frequency decreases as the number of acyclic single bonds increases.

Certain substitution sites are more favorable than others, and this is particularly true of ring junctions. In each of the five families studied, it was observed that certain son FRELs are remarkable because their occurrence is noticeably higher than that of their isomers and even, in some cases, of the father FREL. The five sets of substructures shown in Figure 11 reveal a generally monotonal distribution of FREL occurrences with only a few exceptional occurrence values (2–4 for 10–15 FREL-Bs). This is significant for further identification of hard-core knowledge embedded in structural data. Such results prove the value of combining a formal approach with experimental data. They also help limit the set of possible candidates for various chemical computer-aided design programs in more realistic proportions.

Thus far, this statistical evaluation reveals the importance of the topological factor, i.e., the framework of the molecules. Although we are dealing with cyclic compounds, it is difficult to compare these occurrences with those of rings in the CAS File as calculated by Stobaugh,[11] since the FRELs here correspond to families of compounds with rings of very diverse sizes. On the other hand, these data on ring occurrences are interesting to compare with our data to the extent that they represent a source of complementary structural information and stem from different taxonomies.

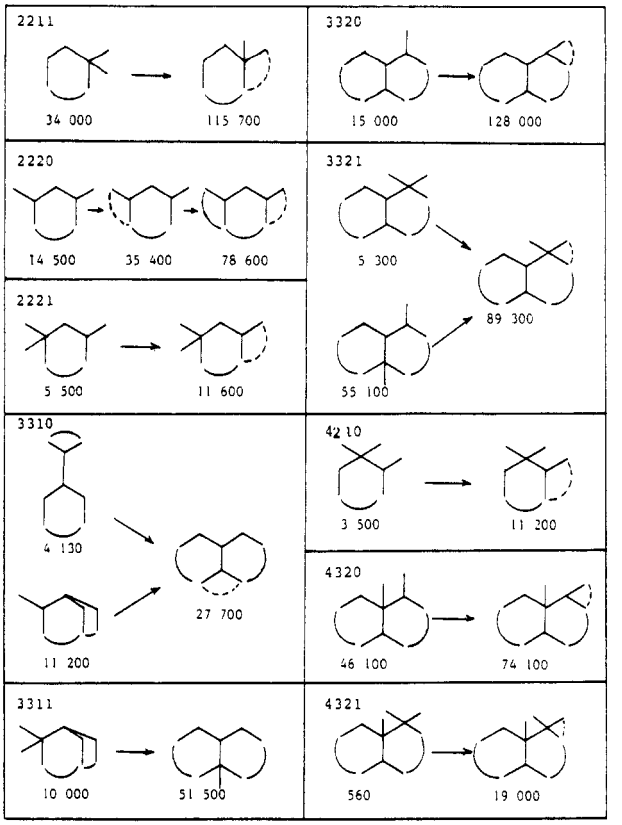


Figure 10. Thirteen FREL pairs characterized by an important occurrence increase when going from an acyclic to a cyclic bond. These are exceptions to the more frequent opposite tendencies.

(d) FREL–Structure Relationships. Structures with frequently occurring cyclic FRELs are often substructures of steroids. The steroid ring system, which consists of three fused six-membered rings, together with a single five-membered ring, containing mostly single bonds, contains six cyclic FRELs with a high frequency of occurrence. These FRELs, shown in Figure 12a, are 2210, 2211, 3321, 4320, 3311, and 3320. Of these six FRELs, four are characteristic of the steroid ring system, and a large majority of the corresponding structures are steroids (71% for 4320 and 3311, 65% for 3321, and 60% for 3320).

The steroid skeleton with possible substitution at C_3 or C_{17} (steroid numbering) has some 36 000 members in the CAS File. If a pseudolinear chain is specified at C_{17}, as in cholesterol, and free sites are established at all the ring atoms, the population decreases to about 12 000 compounds. This group leads to four other FRELs, shown in Figure 12b, 4320, 2210, 3210, and 3321. The 3321 occurrence frequency accounts for 76% of the 12 000 and 4320 for 72%. In other words, a more specific steroid family with a particular side chain has been identified. With this chain constraint, some steroids are eliminated at the retrieval step.

Three other FRELs are not characteristic of the steroids, but can be involved in terpenes, as shown in the pentacyclic triterpene oleanane (Figure 12c). Since double bonds are known to occur within the steroid ring system, it will be subsequently shown, by an occurrence search, that the steroid family is much larger. Thus, we can estimate that the steroid family must have at least 100 000 members. Three other FRELs are not characteristic of the steroids, but can be involved in terpenes, as shown in the pentacyclic triterpene oleanane (Figure 12c).

(3.3) Overview and List of Most Important FRELs. Figure 13 shows the 12 most frequently occurring FRELs which lead

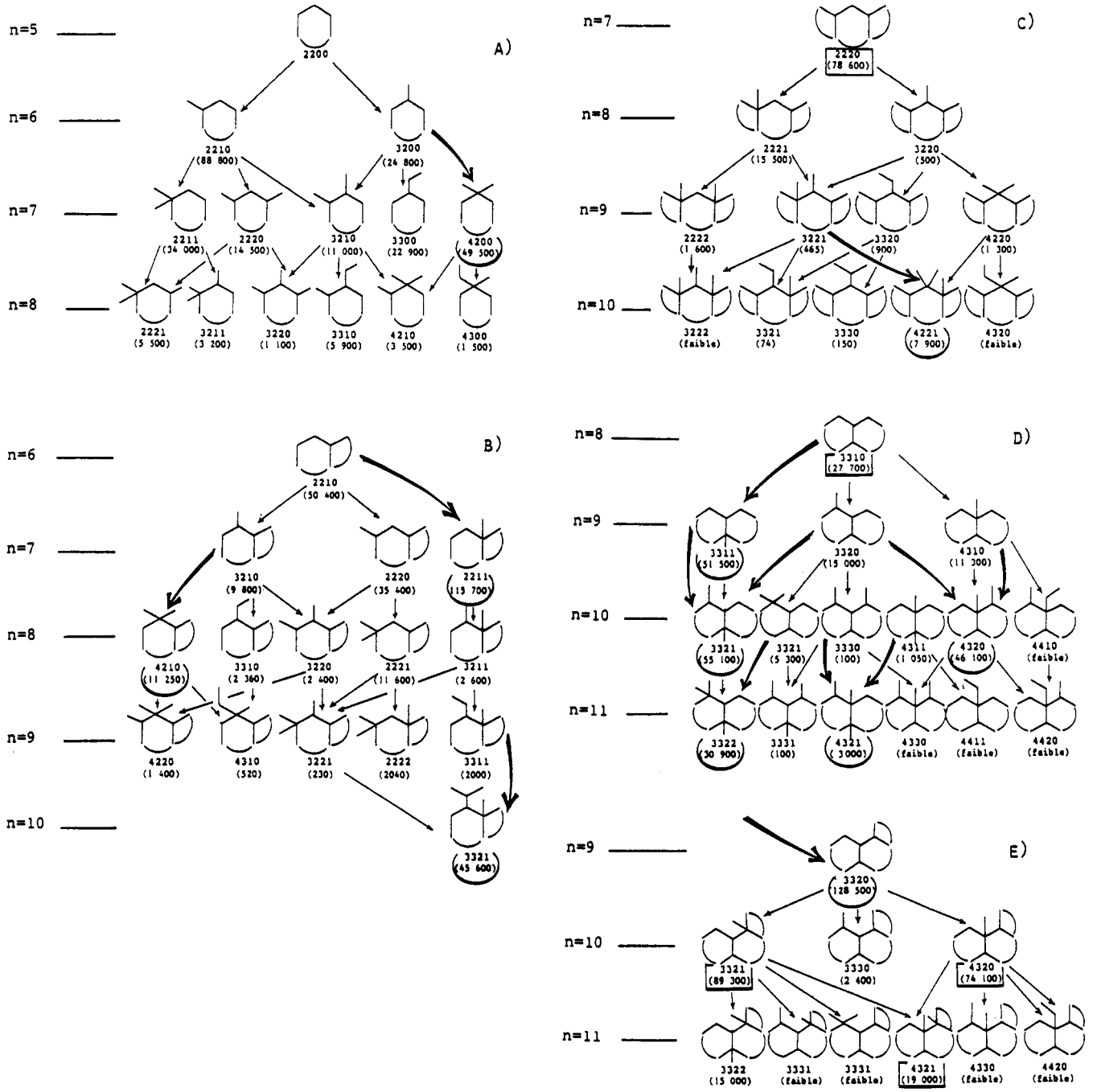


Figure 11. Presentation of significant FRELs (issued from simple rings) and their occurrence values. The half-circled FRELs are those with the most surprising occurrences among their isomers (*n* levels). Dark arrows between FREL pairs show important variations of occurrences. Half-squares draw attention to compounds with complex filiation relationships (relatively high occurrence).

to 20 FRELs possessing cyclic information. The high occurrence frequencies of these FRELs as compared to those of the equivalent FREL-CXs (cf. Figure 4) show that they play an important part in their representativity. They are the main cause of the various anomalies observed in the occurrence/topology correlations of the FRELs in the carbon FREL table (Figure 3).

The most valuable units for the analysis of the chemical database are the FREL-ACs or well-defined graphs going from 1 to 19 atoms, organized according to a concentric description two atoms deep or, at most, a transversal chain of five atoms.

Occurrences of FREL-ACs are few when compared to those of FREL-CYs involved in rings, but their evaluation, within a few percent of the AC + CY total, can increase sharply when the FREL-ACs become complex and correspond, for example, to the highly substituted graphs 2111 [CH_2(tBu, Me), 38%], 2210 [CH_2-(Et, iPr), 27%], or 2220 [CH_2-(iPr)$_2$, 24%]. The convenient rule, "The frequency of occurrence of an acyclic FREL decreases as its complexity increases", is quite satisfactory. When it fails, it does so because of an incontinuity in the tendencies seen in the FREL table. These deviations correspond to FREL-ACs involved in cyclic compounds. A logical interpretation of the contribution of ring occurrences found by a FREL-AC is difficult. Another practical rule, however, summarizes the observed tendencies quite well: "FREL-CY occurrence frequencies tend to decrease as their degree of cyclization increases".

Remarkable deviations from this tentative rule are associated with important chemical populations. Several FREL-CYs do not obey the rule, in that they have more cyclic contributions and very high occurrence frequencies that stand in contrast to those of their nearest FREL neighbors. Thus, high occurrence frequencies are associated with a high degree of complex cyclization, in the terpene and steroid families, for

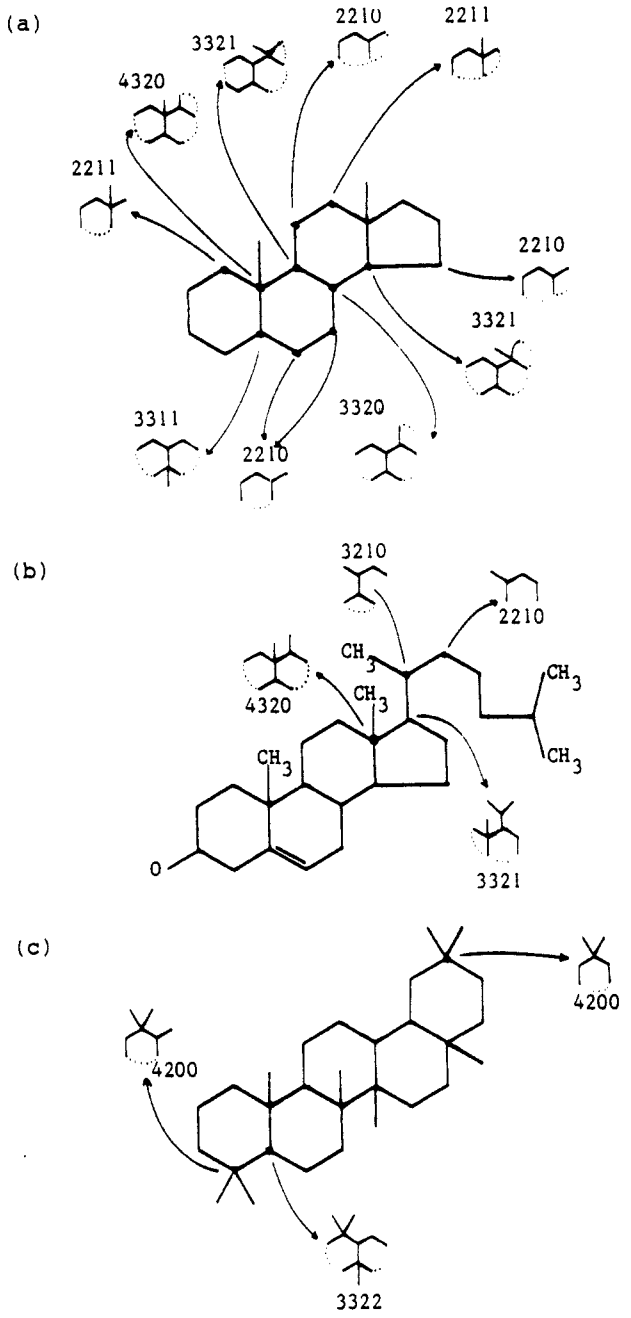


Figure 12. Display of some remarkable saturated carbon FRELs issued from some natural product families such as steroids, (a) androstane and (b) cholesterol, or polyterpenes such as (c) oleanane.

example. Some FRELs are characteristic of the steroid ring and the steroid family. Other exceptions apply to particular families of natural products, even though the FRELs in question have no chromatism.

At this point in the study of topological fragments, the set of 70 basic FRELs reflecting the organization of the carbon skeleton limited to B turns out to be the most interesting. The CAS File is a summary of a sort of the history of chemistry, and from this file it is possible to extract a *coherence table*, which associates the intuitive and quantitative ideas of FREL complexity and their occurrences surprisingly well. Our most refined analysis, limited here to primary, secondary, and tertiary carbon FRELs, confirms the chemist's natural perceptions quite well, though not without some surprises. This should eventually permit applications linked to sets of ordered FRELs within the context of expert systems and of correlation methods, particularly when the parent hydrocarbon method is applicable.

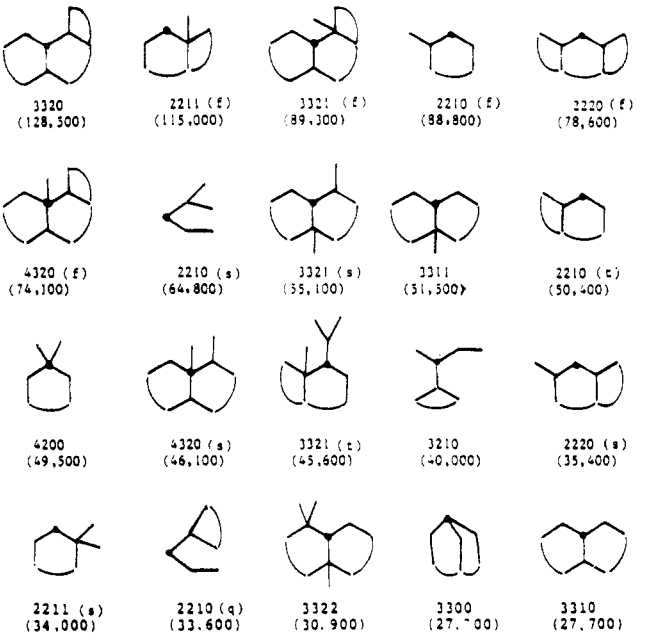


Figure 13. List of the 12 most frequent saturated carbon FRELs bearing more than 5 atoms, with or without cyclization potential. These FRELs are ranked according to decreasing occurrence. Identical FRELs with a different cyclic environment are identified and ranked within their own class (f for first, s for second, t for third, and q for fourth).

CONCLUSION

It has been shown earlier how, starting with these basic FRELs, a set of mono- and polyfunctional isotopological colored FRELs can be derived for use in various search strategies for the correlation of structurally related properties (physical increments weighted by their environments). Knowledge of FREL occurrence frequencies is already in use for the building and management of databases, assisting in areas such as coverage choice, substructure hierarchy, and query bit screens.

Even in behavior analyses or in computer-assisted synthesis, these carbon FRELs can be used to guide the synthesis of a target product by searches for FRELs that can be associated with synthetic intermediates or with comparable neighbors. It should also be noted that several hundred normal and colored FRELs of maximum B depth and their associated properties will constitute a solid base of physicochemical interpretation tools. These will facilitate the evolution of new physicochemical relationships in, for example, spectroscopy (effect of the environment upon an oscillator such as the carbonyl group) and even justify a search for new force fields (MMi simulation method). The FREL concept has already led to very complex developments, but has not previously been elucidated in a coherent computerized set. Several FREL sets such as EURECAS with 14000 FRELs and infra-FRELs and CAD systems such as DARC-EPIOS and DARC-SIRS[20] use FRELs whose choice is dictated by the application at hand.

These topochromatic searches, approached through carbon FRELs and extended to colored FRELs, should, on the one hand, confirm the hope that it should be possible to extract the *dormant chemical knowledge* embedded in large chemistry databases such as the CAS File. On the other hand, such searches could provide a holistic understanding of the chemical behavior of molecules. Research currently is being conducted in areas where topological correlations have already furnished important solutions such as the *gem*-6 effect and some distortion in steric effect explanations, but creation of sets of typical environments will allow modeling and simulation of global properties using complex local information.

GLOSSARY[24]

FREL	fragment reduced to an environment that is limited. A substructure determined concentrically around a focus.
FREL-AC	FREL with all acyclic bonds.
FREL-CX	FREL whose bonds can be either cyclic or acyclic.
FREL-CY	FREL with at least two bonds embedded in a cycle.
frontier row and controlled growth of a FREL	A FREL is composed of concentric layers of atoms distributed around the focus atom. The *depth*, i.e., the number of rows, is application-oriented. The last outer row is called the *frontier row*. The information content of a site, node, or bond may be total or partial. When partial, it defines a loose environment. In Figure 5, FRELS can be loose on two levels, the focus or the environment (any row); thus, according to the location, an *adjunction* will lead to a "following" or "son" FREL.
chromatic data	Information dealing with the nature of specific bonds and atoms and their arrangements (i.e., stereochemistry).
color	The progressive addition of structural information, called chromatic data, to the ground-supporting graph or framework of a molecule. This describes the generation process.
ablation and adjunction operators or operations	The DARC system handles the complex system of structures and reactions considered chromatic graphs and does so through a coherent set of operators. These are both necessary and useful for *graph management* of structures. Two formal operators, elementary *ablations* (ab) and *adjunctions* (ad), have been proposed as contributing to a formal classification of fundamental chemical reactions. The naming of these two operators must be independent of the common names used in chemistry, such as reaction, addition, elimination, substitution, or rearrangement. Ab and Ad operators manage both topological and chromatic changes. All chemical transformations can be described in terms of a combination of structural ab and ad operators. Ab and Ad can take place on structures and substructures—in this case FRELs.
hyperstructures (HS)	An HS can group different structures or different substructures (here, an HS of FREL). The mathematical handling of HS is not developed in this paper. The concept is used for enumeration.
DEL	descriptor of a structural environment that is limited. The atoms and bonds of this environment are ordered and labeled. This descriptor is divided into three successive steps that lead to three partial descriptors: DEX, DLI, and DNA.
DEX	descriptor of existence of nodes without identification of atoms (hydrogens are omitted).
DLI	descriptor of links (other than σ), nature "n" and location "l".
DNA	descriptor of the nature of atoms (n, l).
loose focus	A FREL in which at least one focus atom presents a nonzero residual valency.
FREL descriptors	The DEL/DEX/DLI/DNA procedures are used to develop FREL descriptors. In this paper, FREL-Bs are limited to their DEX description.
nature of descriptors	All *descriptors* cover two rows A_i and B_{ij} in succession and have a matrix form leading to a string of characters.
loose environment	A FREL in which at least one internal environment atom presents a nonzero residual valency.
biunivocal descriptor	A compound is described by a single expression corresponding to one and only one compound.
topochromatic ordering	Ordering sites according to priority rules based on topology and chromatism.
topological actions	Actions that are based only on topological priorities.
FREL table (FT)	A carbon environment table based upon FREL occurrences in a chemical database. [All the saturated alkyl groups up to and including $C–C(Me)_3$ are included in this FREL table.]

REFERENCES AND NOTES

(1) (a) 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 2, p 33 (ISBN 012 0760509). (b) Dubois, J.-E. DARC System in Chemistry. In *Computer Representation and Manipulation of Chemical Information*; Wipke, W. T., Heller, S. R., Feldman, R. J., Hyde, E., Eds.; Wiley: New York, 1973; p 239 (ISBN 0471 95595 7). (c) Attias, R. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102.
(2) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93.
(3) Crowe, J. E.; Lynch, M. F.; Town, W. G. *J. Chem. Soc. C* **1970**, 990.
(4) Adamson, G. W.; Lynch, M. F.; Town, W. G. *J. Chem. Soc. C* **1971**, 3702.
(5) Schwenzer, G. M.; Mitchell, T. M. In *Computer-Assisted Structure Elucidation*; Smith, D. H., Ed.; ACS Symposium Series 54; American Chemical Society: Washington, DC, 1977.
(6) Gray, N. A. *Computer-Assisted Structure Elucidation*; Wiley: New York, 1986; p 243.
(7) Dubois, J.-E.; Sobel, Y.; In Zupan, J. *Computer Supported Spectroscopic Databases*; Wiley: New York, 1986; p 26.
(8) Franke, R. *Theoretical Drug Design Methods*; Elsevier, New York, 1984; p 274.
(9) Pensak, D. A.; Corey, E. J. In *Computer-Assisted Organic Synthesis*; Wipke, T. W., Howe, J. W., Eds.; ACS Symposium Series 61; American Chemical Society: Washington, DC, 1977; p 1.
(10) Warr, W. A., Ed. *Chemical Structure*; Springer-Verlag: Berlin, 1988.
(11) Stobaugh, R. E. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 76.
(12) Stobaugh, R. E. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 180.
(13) Dubois, J.-E.; Carabedian, M.; Dagane, I. *Anal. Chim. Acta* **1984**, *158*, 217.
(14) Bawden, D. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 14, and references cited therein.
(15) Bremser, W.; Franke, B.; Wagner, H. *Chemical Shift Ranges in Carbon-13 NMR Spectroscopy*; Verlag Chemie, Weinheim, 1982; p 859.
(16) Chretien, J. R.; Szymoniak, J.; Dubois, J.-E.; Poirier, M. F.; Garreau, M.; Deniker, P. *Eur. J. Med. Chem.* **1985**, *20*, 315.
(17) Panaye, A.; Doucet, J. P.; Peguet, P.; Dubois, J.-E. Proceedings of the 10th International CODATA Conference *CODATA Bull.* **1986**, *64*, 32.
(18) Passlack, M.; Bremser, W. In ref 7; p 92.
(19) Dubois, J.-E.; Panaye, A.; Attias, R. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 74.
(20) Dubois, J.-E.; Viellard, H. *Bull. Soc. Chim. Fr.* **1968**, *3*, 900.
(21) Dubois, J.-E.; Viellard, H. *Bull. Soc. Chim. Fr.* **1968**, *3*, 905.
(22) (a) Dubois, J.-E.; Mathieu, G.; Peguet, P.; Panaye, A.; Doucet, J. P. *J. Chem. Inf. Comput. Sci.* **1990** (in press). (b) Carabedian, M.; Dagane, I.; Dubois, J.-E. *Anal. Chem.* **1988**, *60*, 2186.
(23) EURECAS File corresponding to the CAS Registry Structure File translated into the DARC language. Estimation of occurrences was obtained from the CAS database which, when coded into the DARC EURECAS database, contained 7.5 million structures. When the number of responses was in excess of 5000, the search was carried out in the MINICAS file, a 1% random sample of the EURECAS database.
(24) Further information is presented in the Glossary published in: Dubois, J.-E.; Panaye, A.; Attias, R. DARC System: Notion of Defined and Generic Substructures. Filiation and Coding of SS Classes. *J. Chem. Inf. Comput. Sci.*, **1987**, *27*, 744. See also: Dubois, J.-E. Conceptual Description of System Transformations and Reactions. In *Proceedings of the 8th International CODATA Conference*; (1982). North-Holland Publishing: Amsterdam, 1982; pp 155–166 (ISBN: 0 444 86668 X).