DARC System: Notions of Defined and Generic Substructures. Filiation and Coding of FREL Substructure (SS) Classes

JACQUES-EMILE DUBOIS,* ANNICK PANAYE, and ROGER ATTIAS

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

Received April 4, 1986

A set of structural fragments defined as specific substructures and generic substructures (SS) is presented by means of such notions as residual valency, free site, and chromatic fuzziness. Different classes of *fragments reduced to an environment that is limited* (FRELs) are described. Loose FRELs and infra FRELs are basic proposals, leading to a wide variety of associated SS. The information contained in these SS is distributed among the topological and chromatic sites within a concentric organization. These FRELs are presented as derivatives of structures or of other SS by *controlled trimming* relative to the molecular skeleton (topology) or to the nature of its sites (chromatism). As opposed to *trimming*, one can also obtain FRELs by a controlled, step by step, *generative DARC construction* of the SS. This type of construction creates filiations with generic substructures as starting points. Filiations between different FRELs can be very numerous. These relationships issuing from ordered generation laws are needed in various computer-aided-design strategies. For information processing, SS handling is facilitated by using alphanumerical descriptors or synonyms, based on letters. A good correspondence, at least for most frequent substructure types, is proposed for textual names and descriptors.

INTRODUCTION

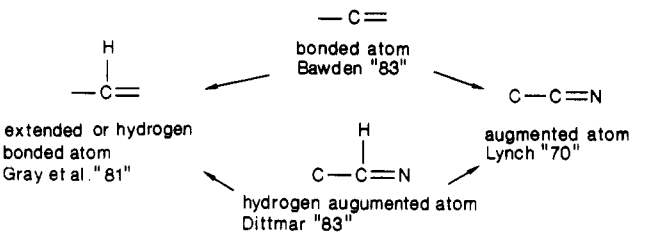
The taxonomy of molecular chemistry is based on the use of substructures (SS) to define large sets with common features or properties: families, types, etc. Usual substructures in structural chemistry regroup chemical functions, substituent groups, and rings. Hence, a principle of substructure analogy or similarity is used to distinguish groups with defined physical characteristics.

These substructures constitute the structural primitives of chemical taxonomical organization.[1] Furthermore, they have inspired a good deal of work whose aim is to describe, in a modular way, the construction of a chemical compound[2,3] to account for its properties by associating these substructures through the use of various formulas that range from strict additivity to complex interactions.[4]

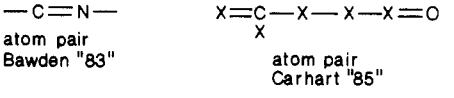
The notion of substructure remains nonetheless complex and difficult to define since it covers very diverse aspects: joined or disjointed SS, completely defined or fuzzy SS, two- or three-dimensional modeling.[5-7] A substructure can be considered in the structural space alone or in its association with certain physical characteristics such as pK, spectroscopic data, partition coefficients, or pharmacodynamic activity.[8] At that point, certain properties associated with local functions of a molecule are expressed.

An overall taxonomy, bearing on the whole set of substructures, would be extremely useful, since SS are essential factors in the basic software of both documentation and artificial intelligence. However, the diverse horizons of chemistry imply varied divisions of structural reality. Thus, the objectives of an SS relating to spectroscopic property[9] are quite different from those of SS or synthons used in computer-aided synthesis.[10]

The pressure of growing needs gave rise to various proposed terms: atom cluster, augmented atom, atom ganglia AA, hydrogen-augmented atom, bond sequences, octuplet, and so on.[8,11] Such a series of names with little or no logical relationship among them can scarcely prove adaptable to new cases, and the temptation is thus to use even more new and nonsystematic terms. Thus, an extended atom[12] where the number of hydrogens borne by the central atom and the nature of the related bonds are indicated can be considered a hydrogenated bonded atom. This is a compromise between the idea of a hydrogenated augmented atom (specifying the number of hydrogens)[11] and that of a bonded atom (merely indicating bond multiplicity).[8]



Furthermore, very similar and simple names can take on more than one meaning. Thus, atom pair can designate either two atoms linked by a bond and the external links[8] or the description of the bond (except for hydrogens) around these two atoms and the path that links them.[13] In the first case, one is dealing with a lone motif of two directly linked atoms, whereas the second case concerns an aggregation of motifs where the two atoms are connected only through a multiatom chain.



It is possible, by reflecting on the definition of substructure characteristics, to identify SS classes and to define them for easy use in chemical informatics? The aim of such research is to achieve a formal taxonomical framework within which to regroup, associate, and systematically rank substructure notions (from high to low details).

We shall first treat the characteristics and the representation of defined substructures. Then we shall move to generic SS, using the method of *information ablation* starting from defined structures and substructures. Filiations between various substructure classes are obtained by a reverse method, that of progressive generation of a complex substructure from a simpler substructure by *information adjunction*. Finally we shall tackle the problem of diversely codifying these substructure classes on different levels of complexity. This paper is limited to SS centered on atoms or bonds in chains or

0095-2338/87/1627-0074$01.50/0

Table I. Usual Types of DARC Bonds

	indifferent	acyclic	cyclic
indifferent	···	·:·	*:*
simple	—	·	*
double	=	:	*
triple	≡	⋮	*
aromatic	···		***
delocalized or tautomeric	···	·:·	*:*

spanning trees. Rings will be discussed later in a subsequent work.

CONCENTRIC DEFINED SUBSTRUCTURES OF DEFINED STRUCTURES: DEFINED FRELS

A substructure results from a systematic construction for which we dispose of basic elements (atom, bond) and assembly laws (valency). A simple splitting law distinguishes SS within a structure according to whether their centers are located either on atoms or on bonds.

We deal only with SS stemming from defined structures, i.e., structures where bonds and atoms are explicitly shown. These structures can be represented by a graph, connected or not, including a set of atoms (Mendeleef's table) and the usual bonds (Table I).

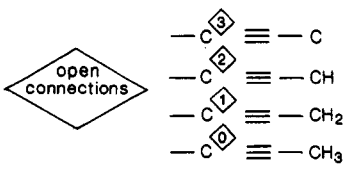
We shall not consider "Markush" structures that represent a population of compounds rather than a compound alone.[14]

Concepts of Residual Valency and Free Site. A SS is obtained by trimming a structure.[15] This trimming can be *topological*, suppressing one or more bonds, one or more atoms and related bonds, or *chromatic*, suppressing the multiplicity of bonds or the nature of atoms.

The topological indeterminations are handled through the notions of free site and residual valency.[15,16]

If all the atoms in structural formulas were expressed, the representation of the substructure would clearly localize the trimming. However, hydrogens are most often transparent, and their presence is reestablished by taking into account the usual atom valencies. Their absence must, therefore, be mentioned explicitly by an additional indication on the graph and the sites affected by a *residual valency*. For each heavy atom (carbon or heteroelement), we indicate the maximum number of links (MNL) that can be created on this site with the help of residual valency indexes. This lack of an occupied valency in the substructure is shown by a localized lozenge on this site. The index values of local residual valency are inscribed in the lozenge (Figure 1). It signifies that possible adjunctions on the site are controlled by this number and that the maximum connectivity of the site corresponds to the adjunction of either hydrogen atoms or monovalent X groups.

In the following example, the residual valency can be satisfied by either a double valency or a single one, i.e., the carbon site is either the hybridized sp^2 or sp^3 carbon site. This complex notion of residual valency regroups sites with different degrees of oxidation.



The allowed operations of adjunction or of "substitution" are managed by two principles mindful of the fact that an added hydrogen is not seen as a substitution: residual valency index (RVI), the difference between the usual valency and the occupied valency of a molecular site; free site index (FSI), the allowed number of substitutions.

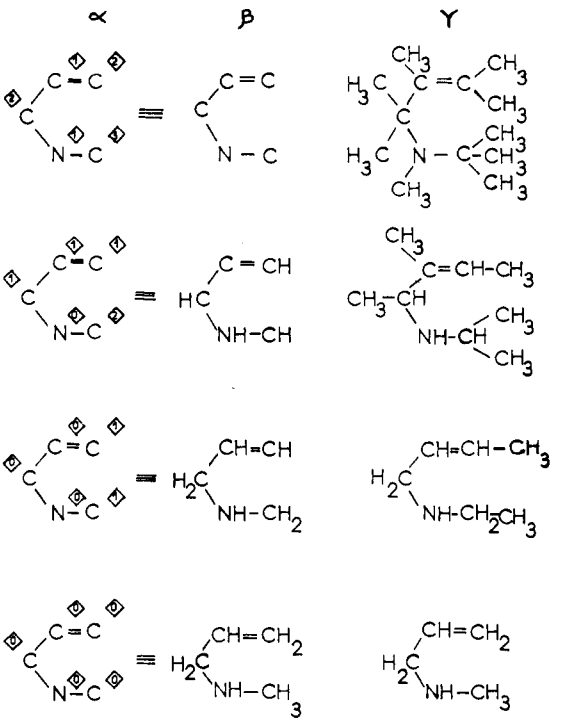
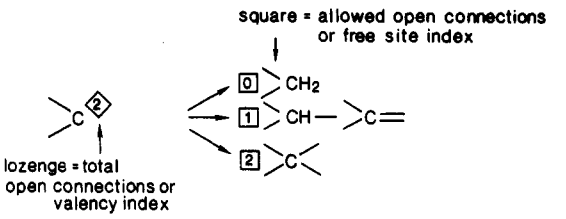


Figure 1. Substructure representation. From bottom to top the α substructures or FRELs become more general. The β formulas correspond to the allowed hydrogenated substructure. The γ molecules are the simplest and correspond to α and β FRELs. For α substructures, the nonsaturation of heavy carbon atoms or heteroelements whose usual valencies are incomplete is indicated by residual valencies and their values.

This FSI number is shown in a square or rectangle as index of the free site. In the previous example, the number of subclasses can be limited by introducing the FSI value. Thus the two indices RVI and FSI lead to three types of products for this labeled substructure. In terms of substitution, products coming from a divalent site can be (0) nonsubstituted (hydrogens), (1) monosubstituted by a monovalent radical (since they are monohydrogenated, monofunctionalized by a radical or a divalent atom), and finally (2) disubstituted.



The residual valency introduces chemical reality into a formal combinatorics of operators applicable to a graph. The notion of free site specifies the number of adjunctions allowed on the site. With this last notion, we determine the subfamilies allowed on the level of defined structures. A free site allows one to combine substructures that differ according to the number of their hydrogens.

The notions of free site and residual valency are linked, and the number of the former cannot be greater than the number of the latter. Table II groups the operations allowed for an atom of valency four. In certain cases, one can simplify by taking only the square indicator of the free sites.

Concentric Substructure. The chemist perceives very diverse substructures. They reflect the chemical notions of skeleton, chemical functions, cycles and acyclic parts, etc. These classical substructures enable us to manage the relationships of a large number of compounds for problem solving in be-

Table II. Allowed Operations of Substitution Adjunction on Residual-Valency Atoms[a]

Loose Valency \ Free site	1	2	3	Unspecified
1	• —	—	—	• —
2	• — •	• — •	—	• — •
3	• — •	• — •	• — •	• — •

[a] Residual valencies correspond to the difference between (usual) valency and occupied valency. Free sites define the number of allowed substitutions. The number of free valencies is thus equal to or greater than that of free sites.

havior analogy. They stem from structural similarities noted in studying syntheses, reaction mechanisms, and chemical properties. Oddly enough, the contributions of substructures stemming from the documentation field have often been developed in more general ways. To this day, these numerous substructures are far from constituting coherent sets.

In the DARC system, structural information is expressed through the description of the topological organization of sites in molecules.[17]

We have sought a local view of SS capable of being inserted in a more global, more holistic view of molecules.[18,19] Each site is then defined by its relations with its neighbors, and together they define a SS. Usually, an active substructure is centered on an active site captured with its close and relevant environment.

Seizing a structure in the DARC system[19-21] depends on the ELCO concept (environment that is limited, concentric, and ordered) of a site (atom or bond). Within the substructure framework, this concentricity of a site's environment is materialized by the FREL idea. Indeed this ELCO formulation is at the basis of the DARC coding of either a molecule (structure S) or a fragment (SS). The progressive introduction of sites to construct either a structure or a substructure leads to a set of S or SS affiliated among themselves so that their filiations are best represented by a conversion graph called HS (hyperstructure). An HS graph summarizes all the filiations existing between the structures located at its nodes.

Thus the description of S or SS is similar and the trilogy of "molecule, fragment and family" is handled with a strict synchronous ordering leading to the derived *trilogy S, SS, and HS*. To implement a substructure search system,[16] definitions of fragments of varied and increasing specificity leading to FRELs allow, in turn, definitions of filiation rules based on their ordered structural parameters. These ordering rules are totally original as compared to the fragments defined by the augmented-atom notion.[6] Since these first contributions to structural descriptions, many concentric substructures, more often centered on atoms than on bonds, have been studied both in documentation and in design.[8,22-24]

The FREL idea, part of the DARC system from the first,[17] asserted itself gradually and led to FREL classes adaptable

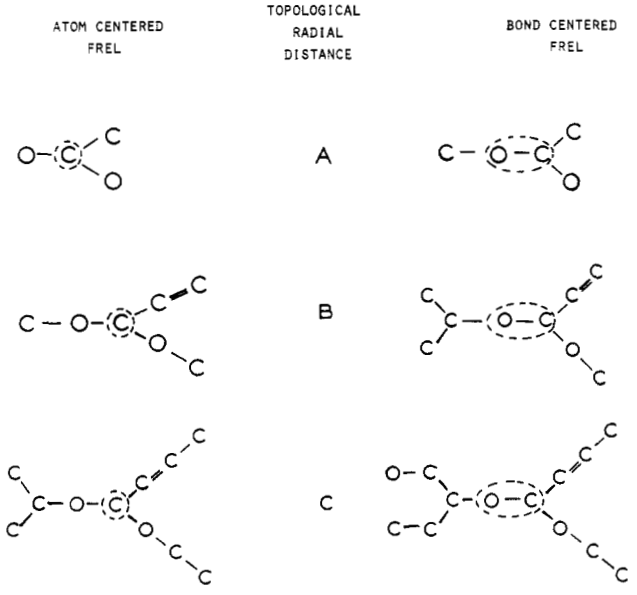


Figure 2. Concentric substructures: FRELs = FRagment defined by an Environment that is Limited. A letter A, B, C, ..., designates the topological radial distance of the environment around the focus: atom or bond.

to treat very diverse situations. Since this is an important element of the structural language needed in data-base-management systems[16,25-27] and in expert systems,[28,29] it is essential, at this stage, to provide some definitions for the use of these notions. At the same time, the FREL idea must remain flexible enough to evolve.

Definition. A fragment reduced to an environment that is limited or FREL is a substructure determined concentrically around a focus.

This focus is most often an atom or yet a bond (with its two atoms). If one considers that the focus is condensed, one can also envisage other, more complex focuses, such as rings. A FREL thus includes two parts: a focus (FO) and an environment (E). The latter is limited by a frontier where one finds the structural information farthest from the FO. That part of the environment situated between the FO and the frontier is called the internal environment.

We designate the FREL's range by a letter, A, B, C, or D, expressing the topological distance between the FREL's frontier and its FO (Figure 2).

When the chromatism of all sites is perfectly defined as to the multiplicity of bonds and the nature of atoms, the FREL is called a defined FREL or FREL-d. In ordinary cases, however, we use the general term FREL, often without the index d.

For frontier atoms, a residual valency is implicit. The residual valency index of these atoms depends on their connectivity and on their anterior bonds.

If the frontier is formed only of monovalent atoms such as F, Cl, Br or of divalent atoms doubly linked, =O or =S, no adjunction can take place later. The extension power of a FREL depends, therefore, on the number of peripheral atoms capable of hydrogen adjunction or of group X substituents after the allowed residual valencies and the reality of usual valencies are considered.

GENERIC FRELS

In large files of several million compounds, chemical diversity leads to great numbers of such substructures. Thus, from the millions of compounds of the Chemical Abstracts Service file handled by EURECAS,[30] algorithmic research has shown us that B FRELs centered on three connected atoms contain seven to eight FREL atoms and that there are an average of

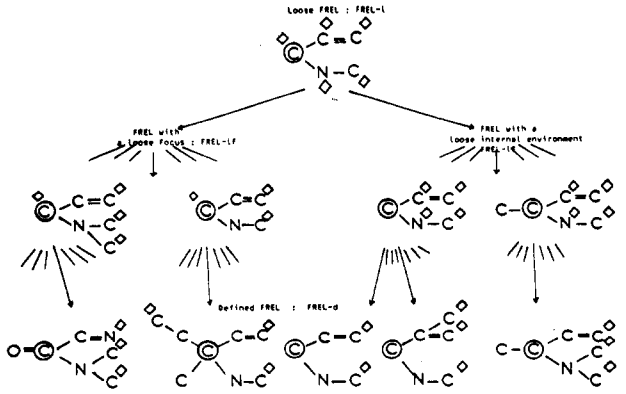
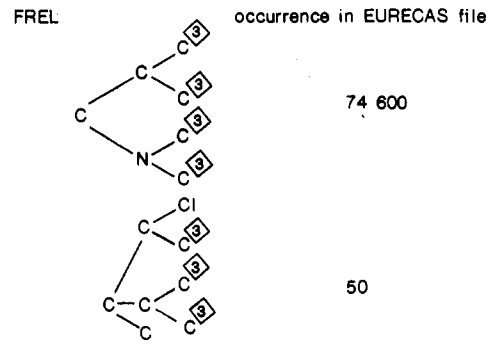


Figure 3. FREL and topological generic features. Each atom with residual valencies can be substituted. In order to simplify the figure, residual valency indexes are not specified; their values are maximum. By definition, final FREL-d have some external residual valencies.

five such FRELs per compound. These millions of constituting B FRELs, associated with various redundancies, correspond to approximately 500 000 original B FRELs. This file of B FRELs and their occurrence constitutes the core of organic chemistry.



These original FRELs can be used as screens, but for structure-file management, one must call on superscreens or major access keys. To this end we have used generic FRELs. The FRELs are gradually deprived of certain information associated with their topology and/or chromatism. The FRELs are all the more general when they result from successive bleachings of the defined FRELs. The ideal solution would be to establish the generic FREL file by algorithmic laws applied to the chemical compound file so as to achieve numerous filiations among all the substructures going from the definition of an SSd to the most stripped SSgs. This problem of generic FRELs is tackled below.

Topological Generic Features: Loose FRELs. In order to recognize that a group -sButyl is present in a triptyl radical, one must allow substitution possibilities within the FREL. To this end, residual valencies must be introduced not only on the frontier but also on the focus and/or on the internal environment.

The following definitions are essential in this methodology: FREL with a loose focus or FREL-lF, FREL where at least one focus atom presents a nonzero residual valency; FREL with a loose environment or FREL-lE, FREL where at least one internal environment atom presents a nonzero residual valency; loose FREL or FREL-l, FREL where at least one focus atom and one internal environment atom present a nonzero residual valency.

The augmented atoms (where the number of hydrogens borne by the central atom is not specified) are loose-focus FRELs. Sequences associated with atoms and bonds or defined by connectivity belong to the class of loose FRELs.[11]

Figure 3 regroups several examples of these loose FRELs. By considering the values of sites and residual valencies, one

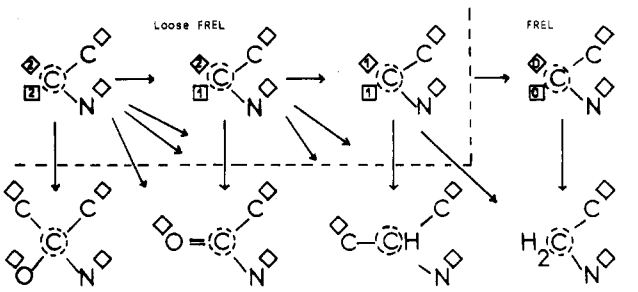


Figure 4. Loose FREL, residual valency, and free site. ◇: maximum residual valency. Indicating the values of residual valencies and free sites makes it possible to manage regroupings of defined FRELs (FREL-d) by loose FRELs (FREL-l).

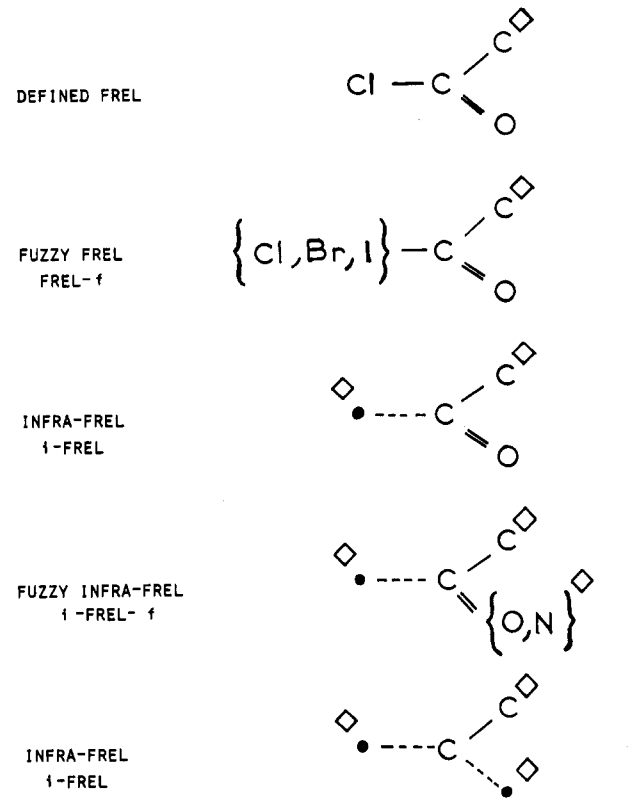


Figure 5. Chromatic generic features, infra FRELs. When chromatic information regarding an atom and/or a bond is suppressed, defined FRELs lead to more generic FRELs of the same topology, infra FRELs (i-FRELs).

can separate into subgroups those FRELs associated with a single heavy graph (Figure 4). These notions are coherent with our terminology oriented to reaction modeling and substructure search procedures.[15,16]

Chromatic Generic Features: Infra FRELs and Fuzzy FRELs. Thus far, real SS are obtained by information ablation starting from structures whose information entirely concerns an atom and its related bond or bonds. Generically richer SS can be obtained by finer trimming of FRELs than that described heretofore by operating only on an atom associated with its bond.

In the DARC system, site information is decomposed into three successive levels. The first is achromatic and only defines topology; the second and third concern the nature of the bond and of the atom. In other words, a site is gradually colored until it is totally defined. Here we proceed in reverse order, suppressing step by step an atom and its associated bond. Thus generic FRELs are created by gradually bleaching the sites (Figure 5). Consequently, sites can be classified through their chromatism: monochromatic site, site (atom or bond) where

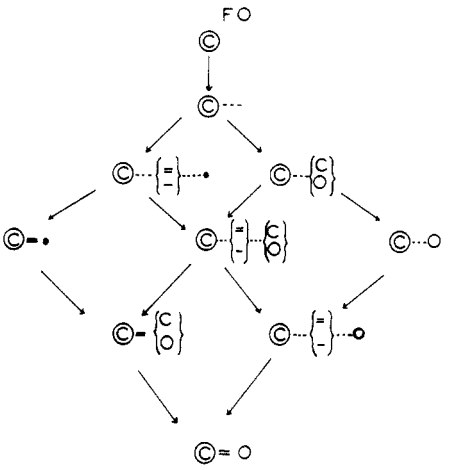
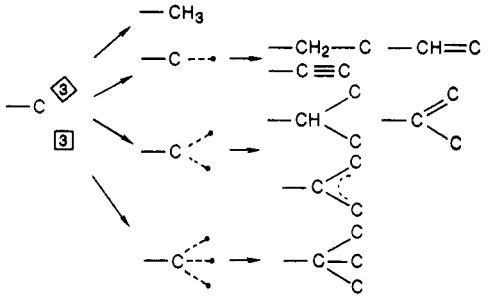


Figure 6. Site generation with intervention of fuzzy steps for atom and bond. The simple generation of a carbonyl can lead to complex filiations when one considers that the bond multiplicity is extracted from a simple/double (alcohol–ketone rapprochement) list and the nature of the atom from a carbon/oxygen (olefin–ketone rapprochement) list. Thus FRELs related to a carbonyl differ according to their generation pathways.

a single chromatic value is allowed; achromatic site, site whose existence alone is indicated.

Further definitions must be derived: defined FREL (or FREL-d), all the FREL sites are monochromatic; infra FREL (or i-FREL), FREL with at least one achromatic site while the others are monochromatic; undefined infra FREL (or i-FREL-u), FREL where all sites are achromatic.

While loose valency and free site notions manage substructure adjunction possibilities, achromatic sites can be used to impose a given number of adjunctions.



To carry out certain regroupings or to extract sets of analogous structures, it is sometimes useful to allow several chromatic values on certain sites: multiple, double, or triple bond, halogen atom, etc. This chromatic fuzziness is an intermediate solution between sites with only one chromatism and sites with no chromatism (In fact, the entire chromatic table is allowed. Multiplicity of bonds is indicated in Table I, and nature of atoms in the Mendeleef table.) Site discoloration is not total then, but partial. By introducing the notion of a multichromatic site, we can extend the number of FREL classes (Figures 5 and 6): multichromatic site, site where a defined list of chromatic values is allowed; fuzzy FREL (or FREL-f), FREL where at least one site is multichromatic; fuzzy infra-FREL (or i-FREL-f), FREL where at least one site is achromatic and one is multichromatic.

TAXONOMY, GENERATION, AND FILIATION OF GENERIC SUBSTRUCTURES

The different FREL classes proposed above have been defined by successive ablations starting either from a structure (defined FREL) or from a substructure (generic FREL). *This procedure corresponds to the chemist's perception, from the real to the generic.* The opposite intellectual construction leads to different information. To go from the generic to the real by a series of successive adjunctions creates various *generation pathways*. A series of successive conversion graphs increasingly rich in information can be derived with various generative rules.

DARC generation considers the structure under study as a target and reconstructs it step by step starting from a focus. The atoms and bonds comprising the environment of this focus are added concentrically around the focus. The environment generation or construction is thus composed of a series of elementary topochromatic adjunction operations concerning each introduction of an atom/bond couple.

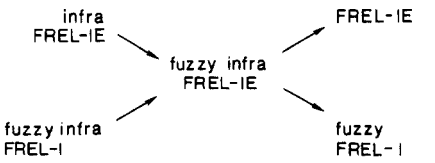
The formal and ordered association thus established between structure and substructures constitutes a powerful organizational tool. Progressive construction must be carried out with a certain precision. Before analyzing a generative tool, one must specify the value of its basic operations and the sequence of operations to implement these.[19]

Site Generation. A basic topochromatic adjunction involves adding a topological site or chromatic link (bond) or node (atom) information. To introduce or "build" an atom and the bond preceding it (when a focus is used for generation), three basic topochromatic adjunctions are needed: a *topological adjunction* expressing existence E of the node and the link (for connected SS, node and link are inseparable) and *two chromatic adjunctions*, one for the chromatism of link L, the other for the chromatism of node N. The order in which these two chromatisms are introduced leads to two different intermediate substructures: L>N, that is, the multiplicity of the bond is specified before the nature of the atom; and N>L, that is, the nature of the atom precedes the bond multiplicity.

If a single intermediate stage of fuzzy chromatism adjunction is introduced, it leads to a five-step generation. An atom and its bond will then cover nine different substructures (Figure 6).

Environment Generation. When the environment around the generation focus comprises n atoms (and bonds), $3n$ basic topochromatic adjunctions are required to build the corresponding graph. Since the building steps of n atom–bond couples can follow each other or be interlarded with each other, the number of potential intermediate substructures increases rapidly. This number becomes even greater with the intervention of several atom layers around the focus.

FREL Hierarchization. Numerous filiations can be established, not only among substructures but also among their classes. The diversity of these filiations is reminiscent of that obtained by hierarchizing the notions of skeleton, cyclic–acyclic, aromatic–alicyclic, and saturated–nonsaturated.[31] Each FREL class can be seen as deriving from FRELs with broader chromatic or topological generic features and leading to FRELs with more limited generic features.



According to the criteria involved, various organizations can be established. In Figures 7 and 8, we have shown such organizations, giving preference respectively to the presence or absence of the free site and the precision of the chromatism. In these representations, our starting points are the defined substructures as they are extracted from the structures. This corresponds to a chemist's perception of reality. However, in ordered generation of SS, the starting points of hierarchizations are the most generic substructures: the undefined loose infra FRELs.

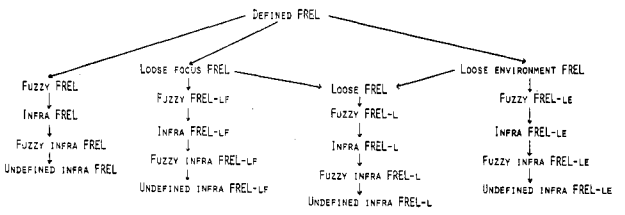


Figure 7. FREL class hierarchization with topological generic features taken as first criterion. FRELs are regrouped in four branches according to the eventual free-site localization. Then, in each branch, the chromatic generic features grow from the defined to the undefined, passing through fuzzy stages.

These filiation exercises point out the diverse aspects of the "potential correspondence" between defined FRELs and structures. In artificial intelligence, their existence allows for more and more numerous recognition strategies of property–structure association. For certain FREL–property associations, some important extensions or simulations are connected to the fuzzy FREL notion.

CODED DESCRIPTION OF FREL CLASSES

The generation state can be symbolized alphanumerically. The precision required for differentiating among the classes of FREL and infra FREL demands coded expressions with several figures. However, in ordinary speech, short synonyms are needed for the most often used classes.

Generation Degree Yields Three-Figure Alphanumerical Description. To characterize FREL classes, we specify the FREL range and then qualify each layer of the FREL environment. Each successive concentric layer is expressed by a progressive level descriptor where three figures specify topology, bond chromatism, and atom chromatism. One must indeed show whether the focus or the internal environment includes residual valencies. In addition, chromatic information relative to atoms and bonds is separated.

For the topology, numbers one or two characterize layer saturation: (1) the layer is incomplete (this means that the preceding layer has free sites); and (2) the layer is complete. A loose-focus FREL will thus have an A layer whose topology will be symbolized by one. The topology of all layers of a defined FREL will be symbolized by two.

Chromatism is shown by seven values from zero to six corresponding to combinations of one, two, or three chromatic values: (0) all sites are achromatic; (1) both achromatic and multichromatic sites are present; (2) all sites are multichromatic; (3) achromatic, multichromatic, and monochromatic sites are present together; (4) both achromatic and monochromatic sites are present; (5) both multichromatic and monochromatic sites are present; and (6) all sites are monochromatic.

For each existing layer, the descriptor ranges from 100 to 266, for A FRELs from the undefined loose infra FREL to the defined FREL. Defined FRELs correspond to the succession of the descriptor 266.

This numerical descriptor can grasp a very large number of situations: 98 theoretically for the A FREL, some of which are shown in Figure 9. For example, bond sequences are designated by a descriptor with the value (120) or (160), appearing once or several times when one specifies the cyclic or acyclic nature of the bonds as well as their multiplicity.

Not all generic FRELs are of equal importance. In the final part of this work, we shall pay special attention to homogeneous sequences of topochromatic adjunctions.

Homogeneous Generation and Descriptor Synonymity. A homogeneous generation deals with topochromatic adjunction E, L, or N which can only be carried out on a row n if all the topochromatic adjunctions of the type considered have been carried out for $(n - 1)$ rows.

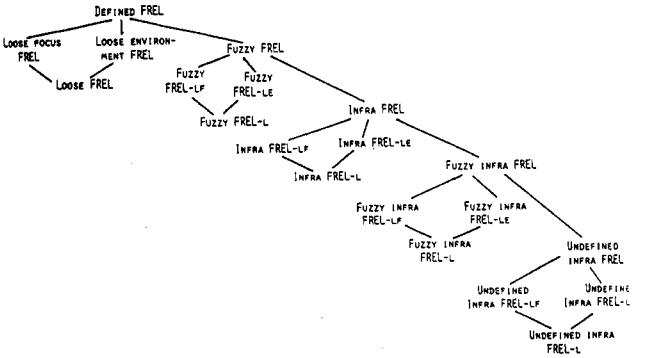


Figure 8. FREL class hierarchization with chromatic generic features taken as first criterion. Each step of chromatic generic features is broken down into three states of loose FRELs according to the free site localization on the focus and/or the internal environment.

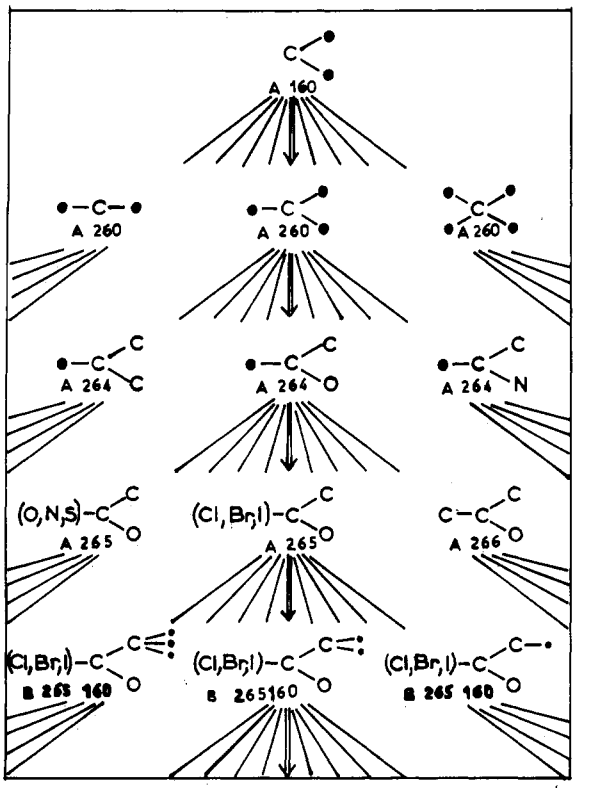
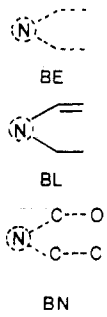


Figure 9. FREL class description by a three figure alphanumerical code. Three numbers describe the extent of saturation, e.g., bond chromatic data are given the second number (0–6) for a saturated site.

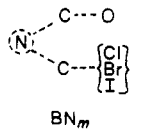
Generation concentricity can have different strategies. The sequence of three topochromatic operators can be carried out successively by site, by layer, or by environment. The resulting types of generation obviously lead to different SS (Figure 10) in the conversion graph or hyperstructure.

For each environment layer A, B, C, ..., the SS specifies the degree of progress of topology generation, of bond chromatism, and of atom chromatism.

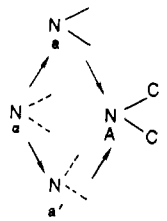
Environment Generation: Two-Letter Code. Among all the homogeneous concentric generation FRELs, some correspond to the imposition of a single type of chromatism on the whole graph. A shorter coding can be used: a first letter specifies the concentric layer depth of the FREL, a second specifies the existence alone as E or either the link chromatism L or the node chromatism N.



Multiple chromatism (m) is shown with an m index:



Layer Generation: One-Letter Code. FRELs whose last row only shows chromatic indetermination are very often used, particularly in spectroscopy.[29] In agreement with FREL range designation, a single capital letter designates a FREL whose bond and atom chromatisms are specified. A Greek letter shows whether simple existence is indicated, and a small letter, either with or without a prime, designates a FREL whose bond or atom chromatisms are respectively absent.



The nomenclatures proposed are precise, general, and, for those with one or two letters, very concise. In the latter, the symbolism chosen is undoubtedly simple to use and extend, as is evidenced by a comparison of the most commonly employed terms with the single-letter DARC designation (Table III).

CONCLUSION

The substructure notion is essential in artificial intelligence. Indeed, in inference procedures of expert systems, one uses ordinary substructures, often born of structural similarities suggested by the existence of common properties. In reality, these substructures have often proved interesting for different uses. Nonetheless, the sets of substructures comprising the accepted primitives in molecular chemistry are not coherent. Their development, in different fields of chemistry, has been too strictly handled with classical documentation tools. We feel that a more conceptual approach to substructures might be central to SS taxonomy as it influences different structured fields of computer-aided design in chemistry.

Clearly, the concept "to be a substructure of", developed in the DARC taxonomy of SS, can be important for chemical reasoning. Developing expert systems with recognized SS means being able to multiply previously used analogies and to extend their range of application. Searching for similitudes or measuring distances between situations by using the structural tool means creating a real SS phenomenology. In this way, one has at one's disposal precise research tools for structural similarity (defined structures and fuzzy substructures) with which to generate inference situations. Therein candidate substructures can be confirmed by the relations they create.

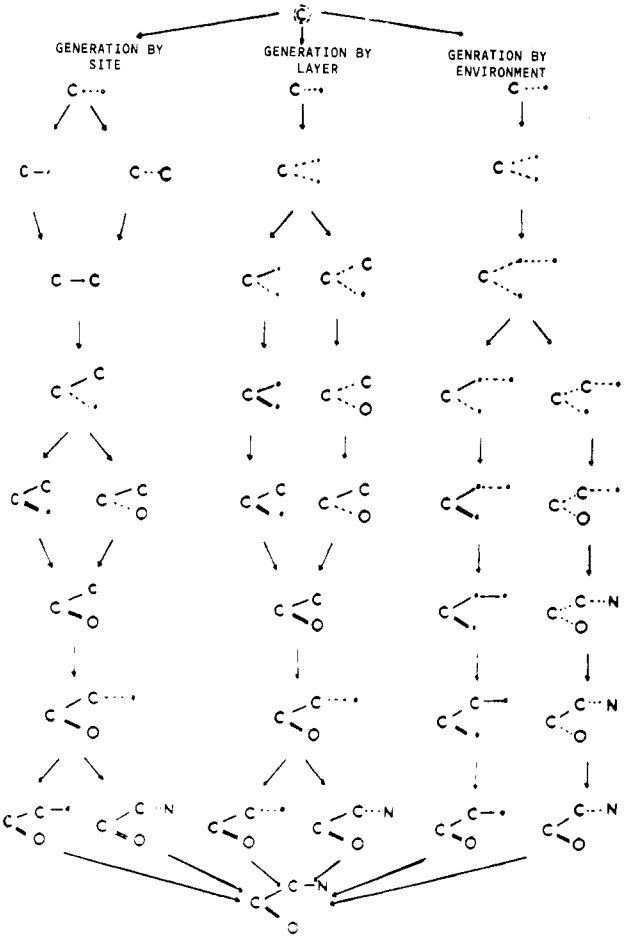


Figure 10. Site generation, layer generation, and environment generation are compared. In this example, we have chosen to generate the carbon of layer A before the oxygen. For each of the three kinds of generation, the left-hand pathway indicates bond coloring before that of atoms; in contrast, the right-hand pathway shows atom coloring before that of bonds. With these three kinds of generation and the reversal of priorities in atom or bond coloring, most of the usual concentric substructures are revealed. Some appear on several generation pathways.

Intuiting a substructure that transcends a given field (e.g., the hypothesis of the isoprene unit SS as a component of the structure of terpenes) involves conceptual views. Though such intuitive choices of SS in artificial intelligence are usually *problem* or *solution oriented*, we propose here that their implementation be expressed within a logical and general system. Thus, such associations or correlations linking properties and candidate substructures could be tackled, more and more, by computerized inference strategies.

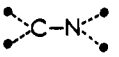
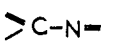
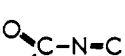
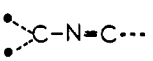
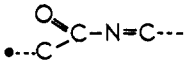
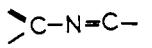
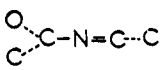
ACKNOWLEDGMENT

We express our sincere gratitude to the referees for their kind and constructive suggestions.

DEFINITIONS

DARC: Description Acquisition Retrieval and Conception
structure (S): structural formula whose sites can be labeled by concentric ordering rules
substructure (SS): structural fragment organized and labeled as a structure with pending bonds; obtention either by controlled trimming of SS or S or by progressive generation
hyperstructure (HS): conversion graph embedding a population of S or SS located on its nodes; it is usually a network of filiations

Table III. Compared Nomenclatures of Concentric Substructures: Often-Employed Terms Are Compared with the Single-Letter DARC Designations

N	FO	C-N
Simple atom		Simple pair
	∝	
connected atom		augmented pair
	a	
bonded atom		bonded pair
	a′	
atom cluster		
	A	
Augmented Atom (AA)		octuplet
	β	
Augmented AA		augmented octuplet
	b	
ganglia AA		
	b′	
atom ganglia AA		
	↓B	

generation principle: algorithmic organization rules that generate SS and S as structural entities and locate them in the HS conversion graph

generation steps: an S graph can expand by elementary adjunction and contract by ablation of structural data

site: node or bond in S or SS; it can be created by generation

focus (FO): structural starting site for constructing S or SS; FO can be a site or a condensed site

environment (E): structural part around a focus; it is organized in modula and layers

ELCO: Environment that is Limited, Concentric, and Ordered

FREL: Fragment Reduced to an Environment that is Limited; a substructure determined concentrically around a focus

topological data: information dealing with the molecular framework

chromatism: nature of bonds and atoms in addition to their simple topological location

chromatic data: information dealing with the nature of specific bonds and atoms

achromatic site: site without any chromatic information

monochromatic and multichromatic sites: sites provided respectively with one or more items of chromatic information or values concerning their generation and/or description

topochromatic sites: sites for which all the topological and some or all of the chromatic information is provided

color, coloring: the progressive addition of structural information called chromatic data on the ground-supporting graph or framework of a molecule; this describes the generative process

bleaching: the reverse of coloring, describing the trimming process

generic data: structural framework including sites (bonds or atoms) lacking their full share of chromatic information

residual valency index (RVI): difference between usual valency and occupied valency of a molecular site

free site index (FSI): allowed number of substitutions

loose FREL (FREL-l): FREL where at least one focus atom and one internal environment atom present a nonzero residual valency; derived terms include FREL with a loose focus (FREL-lF) and FREL with a loose environment (FREL-lE)

defined FREL (FREL-d): all the FREL sites are monochromatic

infra FREL (i-FREL): FREL with at least one achromatic site while the others are monochromatic

undefined infra FREL (i-FREL-u): FREL where all sites are achromatic

fuzzy FREL (FREL-f): FREL where at least one site is multichromatic; derived term, fuzzy infra FREL (i-FREL-f)

REFERENCES AND NOTES

(1) International Union of Pure and Applied Chemistry. *Nomenclature of Organic Chemistry*; Pergamon: Oxford, U.K.; 1979; Sections A–F, H.

(2) Hyde, E.; Matthews, F. W.; Thomson, L. D.; Wiswesser, W. J. "Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds". *J. Chem. Doc.* **1967**, *7*, 200–204.

(3) Thomson, L. D.; Hyde, E.; Matthews, F. W. "Organic Search and Display Using a Connectivity Matrix Derived from Wiswesser Notation". *J. Chem. Doc.* **1967**, *7*, 204–209.

(4) Leffler, E. J.; Grunwald, E. *Rates and Equilibria of Organic Reactions*; Wiley: New York, 1963.

(5) Lefkowitz, D. "Substructure Search in the MCC System". *J. Chem. Doc.* **1968**, *8*, 166–173.

(6) Crowe, J. E.; Lynch, M. F.; Town, W. G. "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part 1. Noncyclic Fragments". *J. Chem. Soc. C* **1970**, 990–996.

(7) Ash, J. E.; Hyde, E. *Chemical Information Systems*; Wiley: New York, 1974.

(8) Bawden, D. "Computerized Chemical Structure-Handling Techniques in Structure–Activity Studies and Molecular Property Prediction". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 14–22.

(9) Smith, D. H. *Computer-Assisted Structure Elucidation*; ACS Symposium Series 54; American Chemical Society: Washington, DC, 1977.

(10) Wipke, W. T.; Howe, J. *Computer-Assisted Organic Synthesis*; ACS Symposium Series 61; American Chemical Society: Washington, DC, 1977.

(11) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. "The CAS ONLINE Search System. 1. General System Design and Selection, Generation, and Use of Search Screens". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.

(12) Gray, N. A. B.; Nourse, J. G.; Crandell, C. W.; Smith, D. H.; Djerassi, C. "Stereochemical Substructure Codes for ^{13}C Spectral Analysis". *Org. Magn. Reson.* **1981**, *15*, 375–389.

(13) Carhart, R. E.; Smith, D. H.; Venataraghavan, R. "Atom Pairs as Molecular Features in Structure–Activity Studies: Definition and Applications". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.

(14) Welford, S. M.; Lynch, M. F.; Barnard, J. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 5. Algorithmic Generation of Fragment Descriptors for Generic Structure Screening". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 57–66.

(15) Dubois, J. E.; Panaye, A.; Picchiottino, R.; Sicouri, G. "Système DARC: Structure de L'invariant d'une réaction". *C. R. Seances Acad. Sci., Ser. 2* **1982**, *295*, 1081–1086.

(16) Attias, R. "Darc Substructure Search System: A New Approach to Chemical Information". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102–108.

(17) Dubois, J. E.; Laurent, D.; Viellard, H. "Système de Documentation et d'Automatisation des Recherches de Corrélation (DARC). Principes Généraux". *C. R. Seances Acad. Sci., Ser. C* **1966**, *263*, 764–767.

(18) Dubois, J. E. "DARC System in Chemistry" in *Computer Representation and Manipulation of Chemical Information*; Wipke, W. T., Heller, S. R., Feldmann, R. J., Hyde, E., Eds.; Wiley: New York, 1974; Chapter 10, pp 239–264.

(19) Dubois, J. E. "Ordered Chromatic Graph and Limited Environment Concept" in *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic: London, 1976; Chapter 11, pp 333–370.

(20) Dubois, J. E.; Viellard, H. "Système DARC VII. Théorie de Génération Description I. Principes Généraux". *Bull. Soc. Chim. Fr.* **1968**, *3*, 900–904.

(21) Dubois, J. E. "Prévisions d'activité pharmacodynamique à l'aide du système DARC". *Man. Comput.* **1972**, 309–330.

(22) Adamson, G. W.; Lynch, M. F.; Town, W. G. "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part II. Atom-Centered Fragments". *J. Chem. Soc.*, (C), **1971**, 3702–3706.

(23) Adamson, G. W.; Creasy, S. E.; Lynch, M. F. "Analysis of Structural Characteristics of Chemical Compounds in the Common Data Base". *J. Chem. Doc.* **1973**, *13*, 158–162.

(24) Adamson, G. W.; Bush, J. A.; McLure, A. H. W.; Lynch, M. F. "An Evaluation of a Substructure Search Screen System Based on Bond-Centered Fragments". *J. Chem. Doc.* **1974**, *14*, 44–48.

(25) Dubois, J. E.; Laurent, D.; Fatome, M.; Andrieu, L.; Pignalosa, C.

"Définition et exploitation d'une banque de données orientée vers Les propriétés radioprotectrices de composés chimiques". *Automatisme* **1975**, 320-328.
(26) Dubois, J. E.; Bonnet, J. C.; Goldwaser, D.; Attias, R. "The DARC System: A Chemical Information System Based on the Topological Encoding of Chemical Compounds". *Proceedings of Eurim II*; Batten, W. E., Ed.; 1976; pp 135-144.
(27) Dubois, J. E.; Bonnet, J. C. "The DARC Pluridata System: ^{13}C-NMR DATA Bank". *Anal. Chim. Acta* **1979**, *112*, 245-252.
(28) Chretien, J. R.; Szymoniak, J.; Dubois, J. E.; Poirier, M. F.; Garreau, M.; Deniker, P. *Eur. J. Med. Chem.-Chim. Ther.* **1985**, *20*, 315-325.
(29) Dubois, J. E.; Carabedian, M.; Dagane, I. "Computer Aided Elucidation of Structures by Carbon-13 NMR. The DARC-EPIOS Method: Characterization of Ordered Substructures by correlating the Chemical Shifts of Their Bonded Carbon Atoms". *Anal. Chim. Acta* **1984**, *158*, 217-233.
(30) EURECAS is the commercial name of the CAS file handled by DARC structure management system since 1979 (CNIC and Telesystem marketing).
(31) Gordon, J. E. "Chemical Inference. 2. Formalization of the Language of Organic Chemistry: Generic Systematic Nomenclature". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 81-92.

Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors

RAMASWAMY NILAKANTAN,* NORMAN BAUMAN,* J. SCOTT DIXON,† and R. VENKATARAGHAVAN

Lederle Laboratories, Pearl River, New York 10965

Received July 22, 1986

A new molecular descriptor, the *topological torsion* (TT), is described for use in statistical SAR studies. The TT consists of four consecutively bonded non-hydrogen atoms along with the number of non-hydrogen branches. This descriptor is essentially the topological analogue of the basic conformational element, the torsion angle. A comparative study of this descriptor and the *atom-pair* descriptor (Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73) using the trend vector and similarity probe methods is presented. These methods are described in detail in Carhart et al. The atom-pair and TT descriptors capture and magnify distinct aspects of molecular topology; judicious use of both of these descriptors could significantly enhance the hit rate in routine screening programs in the pharmaceutical industry.

INTRODUCTION

Several different molecular descriptors have been used in structure–activity (SAR) studies. The interest of the medicinal chemist is in capturing molecular features responsible for pharmacological activity. Although a specific three-dimensional arrangement of atoms may be necessary for activity, the features essential for activity are often found in the topological description of the molecule. Indeed, the possible three-dimensional conformations and pharmacological activity are implicit in the topological description if only we knew how to extract them.

A search of the literature shows essentially two types of molecular descriptors. The first is "holistic" in the sense that the descriptor is *one number*, usually representing some important physical property of the molecule as a whole. Some examples of such descriptors are the estimated 1-octanol–water partition coefficient[1] and the shape index of Kier.[2] The molecular shape indices of Hopfinger[3-5] can also be included in this category. These descriptors are either measured or calculated algorithmically and have been used with considerable success in SAR studies. A drawback, however, is that distinct pieces of information about the molecule, such as the constituent atom types, the bond types, and the pairwise connections, are not preserved. This tends to restrict their application to series of molecules that have a high degree of structural similarity.

The second category of descriptors consists of several distinct pieces of information strung together and may include such things as atomic species, bond types, and connectivity of pairs of atoms. Examples in the literature include the AA (augmented atom) and gAA (ganglia-augmented atom) descriptors of Hodes and co-workers,[6-8] the interactively selectable descriptor set used by Varkony and co-workers[9] for examining structural similarities among compounds, and the triplet ganglia fragments (a unit of three connected non-hydrogen atoms together with the terminal bonds) used in structure–activity studies by Tinker.[6,10] Another example of descriptors of this category are the linear subfragment descriptors of Klopman,[11] which are all the linear subfragments of a molecule that contain from 3 to 12 heavy atoms.

In our laboratory, a simple descriptor of the latter type, called the *atom pair*, has been used successfully in structure–activity studies.[12] The atom pair is defined as a substructure composed of two non-hydrogen atoms and an interatomic separation measured in bonds along the shortest path connecting the two atoms. The description includes the number of heavy-atom connections and the number of π electron pairs on each atom. The atom-pair descriptor corresponds to a clearly identifiable structural feature and is sufficiently easy to calculate to allow handling of large numbers of structures. The atom-pair descriptor can capture possible long-range correlations between atoms in active molecules. In this paper, we propose a new short-range descriptor that is intended not to replace the atom pair but to complement its predictive power. Like the atom pair, it will be found to correspond to clearly identifiable molecular features and to be easy to calculate. We term this the *topological torsion* (*TT*) descriptor.

DEFINITION

We define a topological torsion as a linear sequence of four consecutively bonded non-hydrogen atoms, each described by its atomic type, the number of non-hydrogen branches attached to it, and its number of π electron pairs. For compactness in coding, the number of branches excludes those that go to make the torsion itself. Thus, for the two end atoms of the torsion

* Address correspondence to these authors.
† Present address: Smith Kline and French Laboratories, L-940, Philadelphia, PA 19101.

0095-2338/87/1627-0082$01.50/0