Handling Genericity in Chemical Structures Using the Markush Darc Software

Pierre Benichou* and Christine Klimczak

Questel.Orbit, Le Capitole, 55, Avenue des Champs Pierreux, 92029 Nanterre Cedex, France

Philippe Borne

INPI, 26 Bis, Rue de Saint-Petersbourg, 75800 Paris Cedex 08, France

Received June 18, 1996[⊗]

Markush Darc expresses the genericity which is present in patent chemical structures by the use of generic groups as well as infinite and closed set superatoms. New tools are offered to the user for performing more efficient searches: superatom translation attributes and variable positions of attachment may now be input in the structure query. Operation of these new functionalities is described within screen search—Bit Screen and FRELs screen—and the exact atom-by-atom match. The contribution of superatoms to the Bit Screen content, the reduction of FRELs subgraph, and the effects of superatom translation on the backtracking mechanism of atom-by-atom search are discussed. Associated optimizations of the whole system were necessary to earn maximal efficiency from these strategic changes in the software.

INTRODUCTION

Patents may be described in broad generic terms to prevent competitors from slightly modifying an invention without changing its main properties. In the chemistry field, generic structures are claimed in patent specifications. Such descriptions, which do not only cover the specific compound found in the invention, are disclosed in Markush structures. Furthermore, a given Markush structure may represent an infinite set of substances, which were not necessarily covered by the invention.[1] Computer storage and retrieval of Markush structures requires typical notations.[2,3]

The main mechanisms of variation are position variation, substituent variation, frequency variation, and homology variation.[4] The position variation is allowed to connect alternately a substituent to a list of neighbors. It will be further designed in the paper as a variable position of attachment. The substituent variation refers to a substituent with fixed connections, which is given alternative values. In most cases, this type of variation will be represented by a generic group. Position and substituent variations may be combined in a Markush structure. The frequency variation is represented by a multiplier coefficient, e.g., n, referring to a partial substructure. The homology variation may generate finite sets of chemical radicals, as halogens, or infinite sets of substructures such as “alkyls”. Groups defined exclusively in terms of properties are also included.[5] For instance, protecting or dying groups may be quoted as homology variation. One can easily imagine the complexity of Markush structures and of the softwares which have been developed for retrieving them. Matching generic terms against specific instances or specific instances against generic terms represented a major problem to be solved in the development of appropriate programs.

Markush Darc is first a chemical database management system for the creation of Markush Darc databases, which contain structures stored as chromatic graphs.[6] This extremely powerful software also permits, via a connection-table based substructure search system,[7] to retrieve both Markush and specific compounds found in patents.

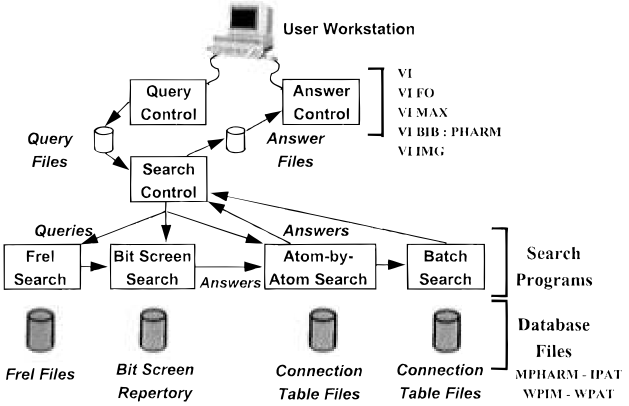


Figure 1. Markush Darc architecture.

Markush Darc belongs to the Darc family, the first public online structural search system, which evolved from the research activity of Professor Dubois’ team.[8] Markush Darc, which is referenced in a French patent,[9] was derived from Generic DARC, which permits a limited genericity to be expressed only in queries. Markush Darc was developed by Questel.Orbit, with the active collaboration and support of the French Patent and Trademarks Office, INPI, and Derwent Information Ltd.

The INPI Pharmsearch databases[10] focus on pharmaceutical patents and are composed of a bibliographic file (PHARM) and an associated structure index (MPHARM) containing Markush structures. On the other hand, Derwent produces the WPIM structure database (World Patents Index Markush) and WPIL the companion bibliographic database.[11,12] These databases cover chemical patents from the Pharmaceutical and Chemistry industries. Another Markush system, MARPAT, covers the chemical patent file from Chemical Abstracts Service (CAS).[13,14]

For a better understanding of the present paper, the general Markush Darc service architecture is presented in Figure 1. A user query may be entered through an easily learned formal language. Markush structures are stored in connection tables, but the databases are also described by FRELs files and a Bit Screen repertory, which are used as screens before the atom-by-atom search. Indeed, the final aim of the system

[⊗] Abstract published in *Advance ACS Abstracts,* January 1, 1997.

S0095-2338(96)00036-4 CCC: $14.00

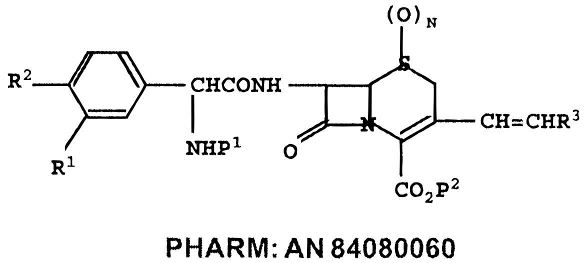
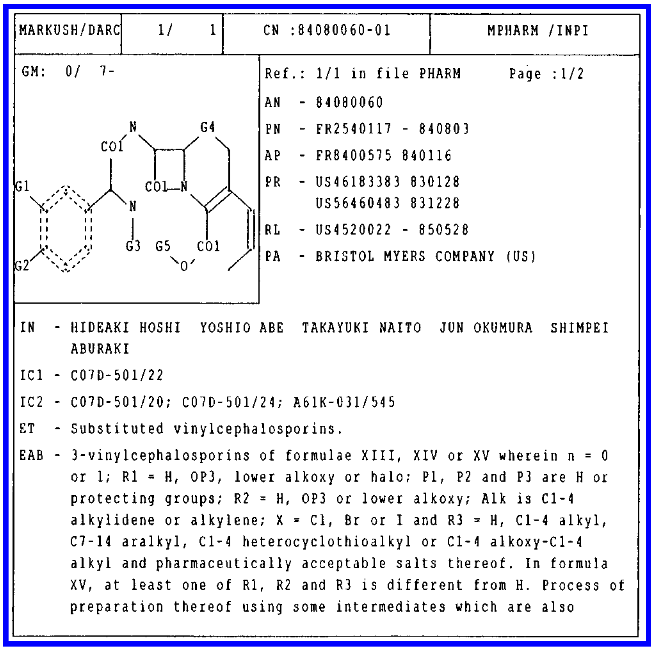


Figure 2. Example of chemical patent from database: R, generic groups; P, protecting groups.

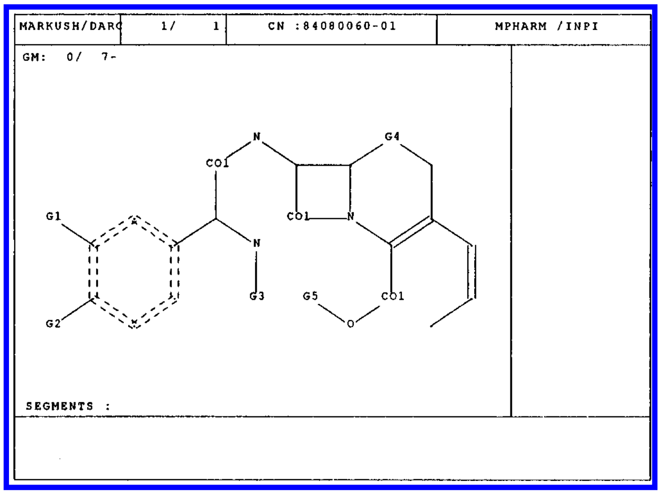


```
MARKUSH/DARC  1/    1   CN :84080060-01        MPHARM /INPI
GM:  0/  7-
Ref.: 1/1 in file PHARM      Page :1/2
AN  - 84080060
PN  - FR2540117 - 840803
AP  - FR8400575 840116
PR  - US46183383 830128
      US56460483 831228
RL  - US4520022 - 850528
PA  - BRISTOL MYERS COMPANY (US)
IN  - HIDEAKI HOSHI  YOSHIO ABE  TAKAYUKI NAITO  JUN OKUMURA  SHIMPEI
      ABURAKI
IC1 - C07D-501/22
IC2 - C07D-501/20; C07D-501/24; A61K-031/545
ET  - Substituted vinylcephalosporins.
EAB - 3-vinylcephalosporins of formulae XIII, XIV or XV wherein n = 0
      or 1; R1 = H, OP3, lower alkoxy or halo; P1, P2 and P3 are H or
      protecting groups; R2 = H, OP3 or lower alkoxy; Alk is C1-4
      alkylidene or alkylene; X = Cl, Br or I and R3 = H, C1-4 alkyl,
      C7-14 aralkyl, C1-4 heterocyclothioalkyl or C1-4 alkoxy-C1-4
      alkyl and pharmaceutically acceptable salts thereof. In formula
      XV, at least one of R1, R2 and R3 is different from H. Process of
      preparation thereof using some intermediates which are also
```

Figure 3. Relationship between patent text and patent structure. The screen was obtained by activating Markush Darc "VI BIB" command.

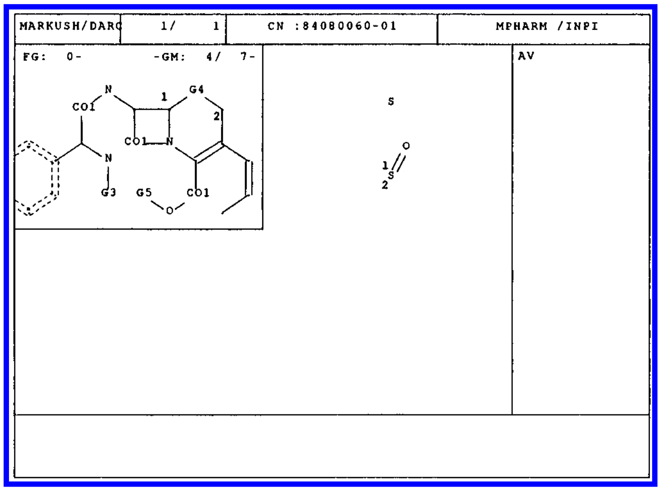
is to provide routines for matching the query substructure with the database. Since an exact search against the entire database would be computationally inconsistent with an online service, screening steps are necessary for reducing the number of candidates to be searched during the atom-by-atom step. The Bit Screen search (BS) completes the RE search or may be performed directly.

This paper will focus mainly on the improvements which have been made on the different search steps during the last four years. These improvements will be exemplified with a PHARM patent referring to the cephalosporin molecule (Figure 2). Variations in this patent are expressed as generic R groups, protecting P groups, or facultative position. Figure 3 represents the correspondence between the patent text, issued from the bibliographic file, PHARM, and the associated Markush structure. This screen may now be obtained directly in Markush Darc using the "VI BIB" command, newly available for the MPHARM database. Figure 4 shows the invariant part of the associated Markush structure. This invariant part contains the generic groups G1 and G2 corresponding in the patent to the generic groups R2 and R1 and the generic groups G3 and G5 corresponding in the patent to the protecting groups P1 and P2. As exhibited in Figure 5, a generic group is described by instances with attachment bonds to the father group, which is memorized in the upper-left window. This G4 generic group corresponds in the patent to the cyclic sulfur with a facultative acyclic oxygen neighbor.



```
MARKUSH/DARC  1/    1   CN :84080060-01        MPHARM /INPI
GM:  0/  7-
SEGMENTS :
```

Figure 4. Invariant part of markush structure: G, generic groups.



```
MARKUSH/DARC  1/    1   CN :84080060-01        MPHARM /INPI
FG:  0-      -GM:  4/  7-                                AV
```

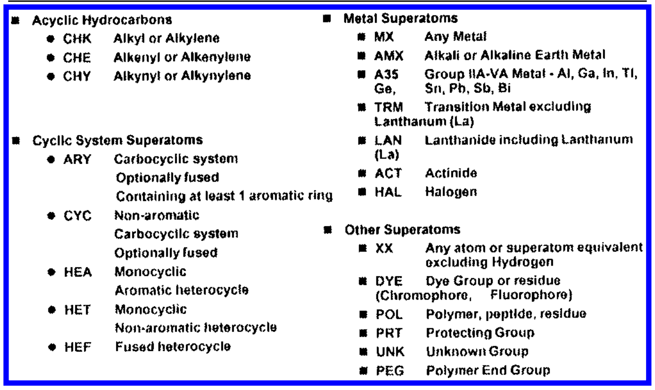
Figure 5. Displaying of generic group contained within a Markush structure. Numbers around sulfur atom refer to the corresponding attachment bond in the father group (see in the upper left window).

The matching of specific substructures against specific substructures, using the generic groups instantiation, has functioned well for several years in the Markush Darc software. Thus, the instantiation mechanism, which allows a three-depth level of embedment of generic groups containing up to 50 instances, will be outside the scope of this paper. More recent developments will be pointed out, including translation between specific and generic substructures and variable positions of attachment in query formulation. How some problems inherent to the implementation of previous features in screens and atom-by-atom searches have been solved will be explained. The decisive progress brought to Markush Darc functionalities by these recent improvements will be underlined. This structuring of the paper will make us often merge the discussion on user commands with other aspects. If necessary, a further paper will focus exclusively on the user's side.

DEFINITIONS OF SUPERATOMS

Markush Darc expresses generic terms as "Superatoms", entered by two or three characters code, either in databases or in queries structures. Questel.Orbit and our two partners, INPI and Derwent, have spent a great deal of time and effort defining superatoms in a nonambiguous and exhaustive way. Table 1 shows Markush Darc superatoms. Acyclic hydrocarbons and cyclic systems correspond to infinite sets of components. Metal and halogen superatoms generate only

Table 1. Markush Superatoms



■ Acyclic Hydrocarbons	
● CHK	Alkyl or Alkylene
● CHE	Alkenyl or Alkenylene
● CHY	Alkynyl or Alkynylene
■ Cyclic System Superatoms	
● ARY	Carbocyclic system Optionally fused Containing at least 1 aromatic ring
● CYC	Non-aromatic Carbocyclic system Optionally fused
● HEA	Monocyclic Aromatic heterocycle
● HET	Monocyclic Non-aromatic heterocycle
● HEF	Fused heterocycle
■ Metal Superatoms	
■ MX	Any Metal
■ AMX	Alkali or Alkaline Earth Metal
■ A35	Group IIA-VA Metal - Al, Ga, In, Tl, Ge, Sn, Pb, Sb, Bi
■ TRM	Transition Metal excluding Lanthanum (La)
■ LAN	Lanthanide including Lanthanum (La)
■ ACT	Actinide
■ HAL	Halogen
■ Other Superatoms	
■ XX	Any atom or superatom equivalent excluding Hydrogen
■ DYE	Dye Group or residue (Chromophore, Fluorophore)
■ POL	Polymer, peptide, residue
■ PRT	Protecting Group
■ UNK	Unknown Group
■ PEG	Polymer End Group

closed sets based on the Mendeleev classification. The remaining superatoms are characterized by a group property or by an extension of the vagueness as for unknown groups.

Exact definitions of aromaticity in rings and XX content must be provided since the major problem was to accord its definition. For carbon rings, in the INPI Pharmsearch database, the Hückel's $4n + 2\ \pi$ electrons rule is applied, but, in the Derwent WPIM database, aromaticity is limited to six-carbon rings alternating bonds, like benzene. The search softwares have been adapted to this different definition of aromaticity in carbocycles. Fortunately, such a decision is unique and is justified by the aromaticity concept which differs according to the definition. Any carbocyclic system which does not contain ring matching in the above definitions will be considered as CYC superatom. On the contrary, the presence of only one aromatic ring or more in a fused carbocyclic system will identify an ARY superatom. For monocyclic heterocycles, six-atom rings with six normalized bonds are aromatic. For the five-atom rings, the combined occurrences of cyclic normalized bonds, cyclic double bonds, and acyclic bonds with external oxygen or sulfur must equal or exceed 2. Aromatic heteromonocycles will be classed as HEA superatoms and nonaromatic heteromonocycles HET.

The XX superatom corresponds to any defined Mendeleev element or to any group which is equivalent to another superatom. For example, two acyclic connected nitrogens will not correspond to a XX superatom, but one acyclic nitrogen or a ring system will do. It must be mentioned that all the peptide shortcuts may be considered as superatoms and included within the XX definition.

The superatoms internal description have to be taken into account in two search steps: Bit Screen and atom-by-atom search.

SUPERATOMS INTERNAL DESCRIPTION AT THE BIT SCREEN STAGE

Like classical Bit Screen searches, the bits generated by the absence or presence of given features in the structure query are matched against the bits generated by each database structure or candidates already selected by mean of FRELs. If any bit generated from the structure query is not present in the candidates's Bit Screen, this candidate will be eliminated. In the Bit Screen search, determining the internal subgraph for a single superatom is not necessary. As summarized in Table 2, four main types of treatment are applied to structures during the database setup or to queries during the Bit Screen search: characterization by the element symbols, by the bonds and the functions, by the attributes, and by the ring systems. According to the considered features, the range of occurrences or only the presence or absence of this feature will be determined. In addition, the total number of atoms is stored in the Bit Screen for small molecules containing less than seven nodes. This information will be available if the small molecule limitation is required by the user. In database structures, the Bit Screen is generated via a down-top mechanism which examines generic groups from the deeper level of embedment to parent groups. It is obvious that when a superatom is encountered, its content must be elucidated, or at least hypothesized since particular atom natures, or bond natures, or chemical functions, or ring systems may be implicitly included within the superatom. For the small molecule limitation, superatoms may also contribute to increase the total number of atoms.

In a way to narrow the potential content of carbon chains and ring superatoms, attributes are displayed in databases structures and may also be entered in the structure query. These chain/ring (CR) attributes are searchable during the Bit Screen step and also in the final atom-by-atom match. Table 3 gives attribute definitions together with superatom natures. Carbon chain superatoms may be straight or branched or exhibit size variations, while ring superatoms may be monocyclic or fused, saturated, or unsaturated. For instance, in Figure 6, the CHK superatom possesses less than six carbons since it is referenced by a "LO" attribute through the attached number. Attributes may be implicit according to the superatom definition: they are not effective for HEA, which is defined as unsaturated and monocyclic.

In Markush databases, the use of parameters listed in Table 4 is more efficient than attributes for narrowing superatom definition. These parameters are calculated from textual notes displayed with the instances of generic groups. Small amounts of not searchable free text may also be contained in these textual notes. An expert system has been developed for finding and resolving inconsistencies between superatom attributes and textual note parameters during the database setup. The values issued from parameters or attributes are restricted to an interval where predefined mathematical rules are observed. An example of parameters list is displayed in Figure 7 for a structure from WPIM database. The textual note at the bottom of the screen refers to the HET superatom numbered 3. In this case, if the CR command had been activated, the chain ring attributes should also be displayed,

Table 2. Types of Treatment during Bit Screen Search[a]

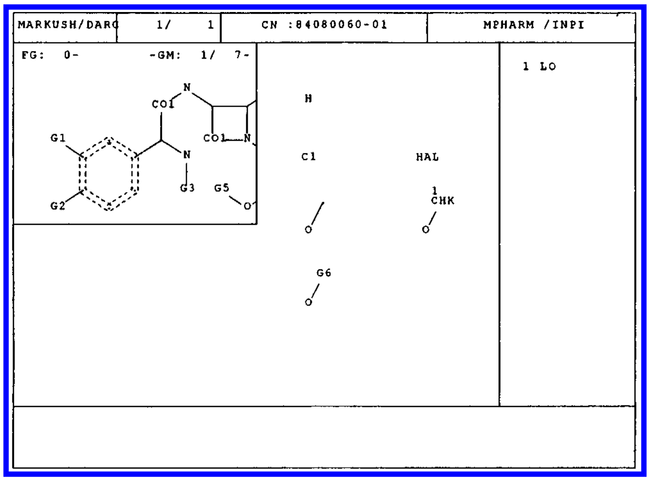
elements symbols	bonds functions	attributes	rings systems	small molecules limitation
range	bond types	charge	range	used only if other specifications required
presence or absence for	nature of atoms	valence	presence or absence for rings systems	
Mendeleïev atoms	two bonds functions	stereo	type	
metals	C environment	mass	internal description	
peptides		polymer		
		peptide		

[a] The fifth type—small molecule limitation—is used only on user requirement.

Table 3. Chains/Rings (CR) Superatom Attributes[a]

superatoms	attributes		
CHK (alkyl, alkylene	STR (straight)	BRA (branched)	
	LO (low)	MID (middle)	HI (high)
CHE (alkenyl, alkenylene)	STR (straight)	BRA (branched)	
	LO (low)	MID (middle)	HI (high)
CHY (alkynyl, alkynylene)	STR (straight)	BRA (branched)	
	LO (low)	MID (middle)	HI (high)
ARY (aryl)	MON (monocyclic)	FU (fused)	
CYC (cycloaliphatic)	MON (monocyclic)	FU (fused)	
	SAT (saturated)	UNS (unsaturated)	
HEF (fused heterocycle)	SAT (saturated)	UNS (unsaturated)	
HET (nonaromatic monocyclic heterocycle)	SAT (saturated)	UNS (unsaturated)	

[a] Size attributes definitions: LO, 1−6 carbons; MID, 7−10 carbons; HI, >10 carbons.


MARKUSH/DARC 1/ 1 CN :84080060-01 MPHARM /INPI

FG: 0- -GM: 1/ 7-

1 LO

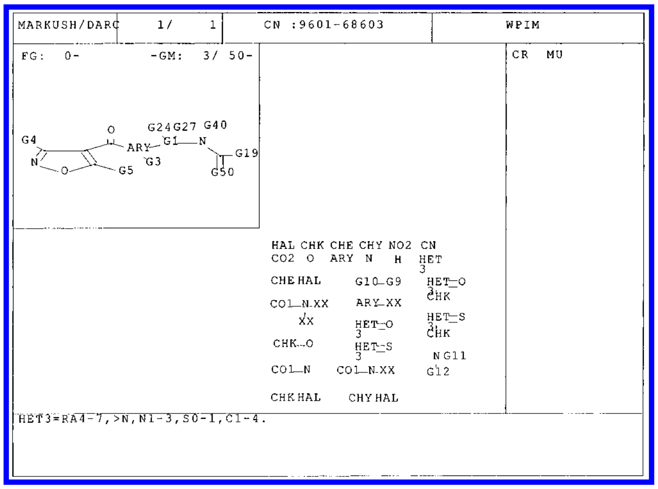
Figure 6. Superatom chain/ring attribute. Display of the generic group G1 in the Markush structure associated with the cephalosporin patent. The LO value refers to the CHK superatom labeled with the corresponding number.

Table 4. Superatom Parameters[a]

	>	≫	E	Y	BX	NR	RA	C	Het	X
CHK	X	X						X		
CHE	X	X	X					X		
CHY	X	X	X	X				X		
CYC	X	X	X	X		X	X	X		
ARY	X					X		X		
HEA	X	X					X	X	X	X
HEF	X	X	X	X	X	X	X	X	X	X
HET	X	X	X	X	X		X	X	X	X
other	X	X	X	X	X	X	X	X	X	X

[a] >, type of internal atom attached to parent structure for one attachment bond; ≫, type of internal atom attached to parent structure for two attachment bonds; E, number of double bonds; Y, number of triple bonds; BX, number of any other bonds; NR, number of rings (fused system); RA, number of ring atoms; C, number of carbon atoms; Het, number of specified heteroatoms; X, number of any other heteroatoms.

referring to the same or to other superatoms. This structure exemplifies the complexity of the checking algorithms which have been developed for deducing all the features implied by the presence of superatoms. First, the nature of the heteroatoms must be analyzed. Thus, all possible atom natures are deduced from the ranges indicated for heteroatoms and for the total atom count. Features of ring descriptions are also examined. The more complex tests concern the possible presence of bonds, characterized by their nature and their atoms ends, and of chemical functions which are composed of two consecutive determined bonds. Indeed, not only the internal description of the superatom must be envisaged but also the bonds of junction with their external environment. In the case of Figure 7, it is important to mention that the attachment node of the generic group G3, which contains the HET superatom in one of its instances, is a ARY node in the father group and that the superatom attachment atom is a nitrogen. Building up the possible junction bonds and functions will be deduced from these informations. Figure 8 shows a specific compound from the WPIM database, which presents an interesting possibility for a potential internal description of carbon chain superatoms. Indeed, the combination of conjugated double bonds and of

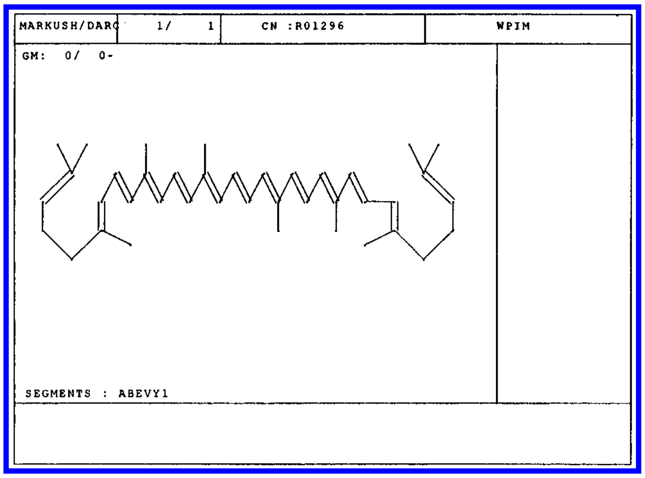

MARKUSH/DARC 1/ 1 CN :9601-68603 WPIM

FG: 0- -GM: 3/ 50-

CR MU

HET3=RA4-7,>N,N1-3,S0-1,C1-4.

Figure 7. Text notes parameters for generic superatoms. Display of the generic group G3 of a structure which contains 50 groups. The HET superatom labeled with the number 3 is described by the corresponding parameter list: RA4−7, number of ring atoms between 4 and 7; >N, the ring attachment is through a nitrogen atom; N1−3, number of nitrogen atoms between 1 and 3; S0−1, C1−4, equivalent intervals for sulfurs and carbons.


MARKUSH/DARC 1/ 1 CN :R01296 WPIM

GM: 0/ 0-

SEGMENTS : ABEVY1

Figure 8. Specific structure exhibiting particular carbon chain features.

tertiary carbon nodes leads to a particular carbon environment, which belongs to the chemical functions taken into account in the Bit Screen. This type of information is decisive for establishing the Bit Screen of carbon chain superatoms, due to the absence of discriminating atom natures or bonds in such superatoms. However, it must be mentioned that the Bit Screen corresponding to the superatoms in the structure query is less extended than in the databases's structures since the query Bit Screen only describes the not facultative features contained in the query. Any facultative feature will be eliminated in the query, while it will be cumulated with other features in the database.

In conclusion, the way of treating superatoms in the Bit Screen has been based on the maximal potential explosion of features induced by these superatoms in the databases and on the minimal extent in the structure query. In the other screening search, a completely different strategy has been chosen.

GRAPH REDUCTION INTO SUPERATOM MOIETIES DURING FRELS GENERATION

Screening steps are usually based on fragment generation in topological or non topological structural systems.[15−18] Markush Darc uses the unique process of FRELs developed by Dubois et al.,[19,20] and Attias et al.[21] FRELs are locally limited fragments centered around a focus atom, branching out at two levels. To be chosen as a focus, a graph node must possess at least two specifically defined neighbors. The first level is represented by the focus neighbors, which are designated as the A positions. The second level includes the neighbors of the A positions, designated as the B positions. The FRELs code defines the nature of the focus, the number of A positions, the nature of bonds and atoms for the A and B positions. Variations in the query such as generic groups induce the generation of a list instead of a specific A or B position. The positions around the focus are sorted by the codes for atoms and bonds. One completely defined FREL in the database is converted into a fuzzy FREL by permuting the defined positions. This permutation allows the retrieval of the equivalent query FRELs issued from the choice of one instance of a list corresponding to a variable position.

The list of FRELs is not fixed in the database since new FRELs are generated when entering new structures in the database. This inverted FRELs file allows the retrieval of a list of candidates for each FREL value. The structure query is broken down into a set of fragments which match FRELs in the database. The lists of candidates associated with the FRELs values generated for a single focus node are merged via an union operation. The intersection of the lists of candidates generated around different query foci is thus operated to obtain the definitive list of RE answers.

Before implementing superatom translation, a statistical study was performed to determine the impact on FRELs generation. The envisaged implementation was first to convert all superatom positions into their specific content using chemical grammars.[22] This operation would have corresponded to a narrow translation of the generic superatoms. Simulations show that this solution would have led to an explosion of FRELs files, in terms of instances and size. Such data were inconsistent with a correct database management and with a normal CPU-consuming FRELs search. The only solution which could be retained, regarding FRELs principles, was to reduce graphs. Graph reduction consists in collapsing nodes together to arrive at a simpler graph. All the nodes belonging to a developed substructure corresponding to the definition of a chain or a ring superatom are collapsed together and replaced by a unique node which value corresponds to the implicit chain or ring superatom. This operation is similar to the broad translation of a developed substructure into a superatom. Total graph reductions have been described[23,24] and result in the replacement of the whole graph by a simpler one with the atoms of the initial graph systematically collapsed when possible.

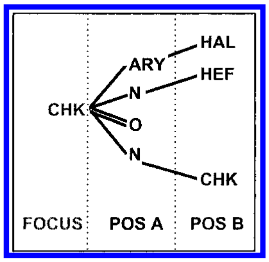


Figure 9. Reduced FREL. Like a classical FREL, a reduced FREL is composed of one focus, A positions which correspond to the focus' neighbors, and B positions, which correspond to the A positions' neighbors. The mechanism for generating this reduced FREL is explained in the text: CHK, saturated carbon chain superatom; ARY, aromatic ring superatom; HEF, fused heterocyclic system; HAL, halogen atom.

The solution of the total graph reduction was nevertheless rejected, and only partial reductions allowing the generation of "reduced FRELs" were performed. This partial graph reduction is illustrated by comparing Figures 4 and 9. The reduced FREL in Figure 9 is centered around the chain superatom focus CHK, which corresponds in the invariant part of the cephalosporin structure displayed in Figure 4 to the carbon issued from an acyclic CO1, collapsed with its carbon neighbor. This carbon neighbor is directly connected to an aromatic ring in Figure 4. All the ring nodes are then collapsed into an ARY superatom, which is represented in Figure 9 as an A position of the reduced FREL. The nodes of the chain substructure and of the ring substructures which are directly connected in Figure 4 will be further designated as the junction points of both systems. In Figure 4, the cyclic system corresponding to ARY in Figure 9 is twice substitued by the generic groups G1 and G2. In Figure 9, one of the possible combinations of the instances of both groups is illustrated. The hydrogen instance of G1, which may be seen in Figure 6, has been selected and is thus not memorized in the reduced FREL in Figure 9. The only neighbor of the ARY A position will be the HAL B position issued from the other group. The second A position in Figure 9 corresponds in Figure 4 to the nitrogen atom which is connected to the four nodes-ring belonging to the second cyclic system of the invariant part of the cephalosporin structure. All the nodes of this cyclic system are thus collapsed into the HEF B position in Figure 9. The third A position of the reduced FREL is the oxygen atom linked by a double bond to the carbon chain in CO1 in Figure 4. The last A position in Figure 9 is a nitrogen atom. In Figure 4, this atom is connected to the generic group G3. The CHK B position in Figure 9 corresponds to one of the instances of this generic group. Thus, the resulting subgraph still includes two layers around the focus, but the number of nodes implicated in this reduced FREL exceeds more than two layers in the initial structure.

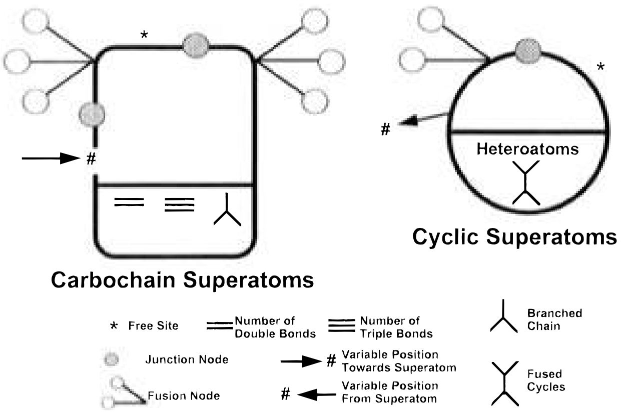


Figure 10. Superatom moieties during graph reduction. All symbols and legends are explicited in the text.

A reduced FREL will be composed of three types of nodes: nodes collapsing in a cyclic superatom, nodes collapsing in a chain superatom, and nodes which do not belong to any superatom. However, the aim of the operation will be not only to reduce the whole structure but also to generate all the possible reduced FRELs of the structure. For that reason, it is better to speak of a partial reduction instead of a total reduction.

The initial algorithm of FRELs generation has been deeply modified and examines all the nodes to collapse through a neighbor-to-neighbor forward process starting from the initial FREL focus. This step results in creating objects whose properties are summarized in Figure 10. Resulting fragments are first identified as cyclic or acyclic. Acyclic fragments will correspond to carbochain superatoms in Figure 10. They are characterized by the carbon count and the possible unsaturated bonds and branching points, parameters which will be exploited for assigning a nature—CHK, CHE, CHY—and an attribute value—LO, MID, HI or BRA, STR, as described in Table 3—to the resulting superatom. Cyclic fragments will correspond to cyclic superatoms in Figure 10. They are characterized by the possible presence of heteroatoms and rings fusion. The nature and attributes of the resulting superatom will also be deduced from the above parameters. As already settled, the junction nodes are connected to a node outside of the superatom. These junctions nodes are represented by empty circles in Figure 10. During subsequent FRELs generation, the former unique FREL focus will be replaced by a set of “real foci” which corresponds to the set of junction nodes. The A positions will be searched as external neighbors of these junction nodes. The internal neighbors, which belong to the same superatom that the junction nodes, will be eliminated from the set of A positions. In the structure query, free sites may also be assigned to the superatom, as indicated in Figure 10. Connections other than those entered in the query will thus be allowed by free sites. This could result in the subsequent change of the nature or valency of the superatom. For instance, two free sites on a monocyclic saturated monosubstituted structure may switch it into a potential polycyclic unsaturated polysubstituted structure.

The fragments generated in the above step may correspond to nonachieved superatoms. Indeed, progress in the graph is interrupted when encountering adjacent generic groups in a chain section or a single generic group within a ring. The reason for interrupting the growth of the superatom in such cases will be explained later. The nodes on the superatoms moieties where the progress has been stopped are designated as fusion nodes, and each fusion node possesses a set of authorized fusion addresses, which correspond to neighbors’ addresses in the connectivity table. These fusion addresses are represented as grey circles in Figure 10. Generating only superatoms moieties is also caused by stopping progress at the level of a variable point of attachment. The direction of the variable positions of attachment (from the moiety or toward the moiety) is also represented in Figure 10.

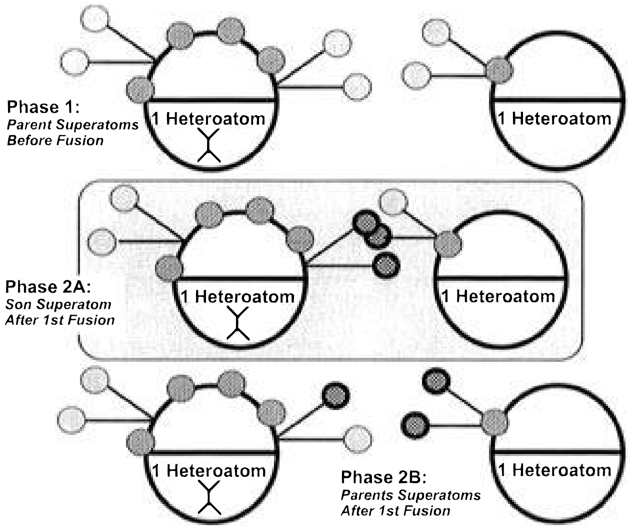


Figure 11. Fusion of Superatom Moieties during Graph Reduction: empty circle, junction node; grey circle, authorized address of a fusion node; black circle, forbidden address of a fusion node. The other symbols are referenced in the legend of Figure 10. The mechanisms described by this figure are explained in the text. The left part of each phase is quoted as G0 moiety in the text, and the right part as G4 moiety.

The next chapter explains how the superatoms moieties are fused, and why entire superatoms were not directly generated.

ASSEMBLING SUPERATOMS IN REDUCED SUBGRAPHS DURING FRELS GENERATION

For a better understanding, the following explanations will be focused on the cephalosporin structure given in Figures 4 and 5. The invariant part of this structure contains a fused heterocyclic system with an internal generic group G4 (Figure 4). In addition, this ring system has four substituents, carried on four different junction nodes. This fused heterocyclic system gives in phase 1 of Figure 11 two superatom moieties since the progress has been stopped from part to part of the generic G4 group. The left moiety—or left parent superatom—in phase 1 of Figure 11 contains all the ring atoms, excepted the generic group G4. The four junction nodes, the nitrogen heteroatom, and the ring fusions, which are observed in the structure in Figure 4, are thus represented in this parent superatom. In the following text, this moiety will be designated as the G0 moiety, because all the collapsed nodes are issued from the G0 group of the structure. The right parent superatom in phase 1 of Figure 11 corresponds to the second instance of the G4 group in Figure 5, i.e., the cyclic sulfur node double-bonded with an external oxygen atom. Consequently, this parent superatom contains one junction node. This moiety will be further designated as the G4 moiety since it is contained within the G4 generic group. Another moiety is generated with the other instance of G4, i.e., the sulfur atom alone, but this will not be illustrated for needs of simplification.

The number of fusion nodes and fusion addresses on both moities in phase 1 of Figure 11 will now be explained. The two fusion nodes of the G0 moiety correspond to the two ring nodes from part to part of the generic group G4 in Figure 4, since the progress in the graph has been stopped by the detection of a generic group included within the cyclic system. Each of these fusion nodes may be connected to any of the two instances of the G4 group in Figure 5. Thus, each fusion node possesses two fusion addresses, which correspond to the addresses of the G4 instances in the table of connectivity. Before any fusion, all the fusion addresses are authorized and represented as grey circles in Figure 11. In the G4 moiety, the junction node is also the fusion node. In Figure 11, the fusion addresses of the right moiety are then directly attached to the junction node. These fusion addresses correspond to the two bonds of attachment of this G4 instance, which are numbered 1 and 2 in the Figure 5.

The fusion step will consist in attaching the complementary fragments using the fusion addresses and will result in phases 2A and 2B of Figure 11. Phase 2A represents the son superatom resulting from the fusion of the moeities of phase 1, while phase 2B represents the state of the parent superatoms after this fusion. It is important to notice that this fusion will change the distribution of the authorized and forbidden addresses in the initial moieties. Such a mechanism is more complex than a simple molecular puzzle. It could be designated as “informational lego”.[25] In the son superatom, the addresses of the G0 and G4 moieties which have been used for the fusion are no longer available. This is materialized in phase 2A of Figure 11 by the two black circles which are linked together. But an extra connection with the second instance of the G4 moiety is forbidden to avoid an abnormal fused ring containing two occurrences of the G4 moiety. Then, the second fusion address in the G0 part of the son superatom is also black-colored in phase 2A of Figure 11. In phase 2B of Figure 11, the forbidden addresses on each moiety are calculated in such a way that redundant fusions become impossible. On the contrary, the fusion of the G0 moiety with the other G4 moiety issued from the other instance of this generic group remains authorized. Then, all valid combinations will be operated. One may imagine the number of combinations and controls to be performed in more complex examples.

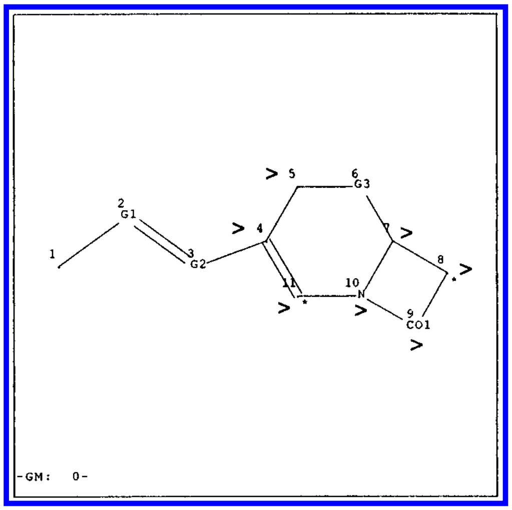


Figure 12. Structure query entered for a possible match with the database structure in Figure 4: G1 and G2, generic groups containing 50 instances representing CHK and any carbon chain in the interval C1−C49; G3, generic group similar to G4 group in Figure 5; >, broad translation; *, free site.

Fragmenting superatoms between adjacent generic groups in chain segments or when entering a cyclic group avoids the generation of too high amount of identical superatoms. The following discussion will be focused on the example in Figure 12. This structure query is derived from the cephalosporin database structure displayed in Figure 4. The similar fused heterocyclic system is entered in this query. The query G3 generic group in Figure 12 is also similar to the database structure G4 generic group in Figure 4. In addition, all ring substituents were replaced by free sites, except for the unsaturated carbon chain, which was constituted of three carbon nodes with a double bond in the right lower area of the Figure 4. In the latter, each carbon node implicated in the double bond is replaced by generic groups G1 and G2 in Figure 12. These groups contain 50 instances: CHK, C1, C2, ..., C49. Entering the query in such a way corresponds to the user’s intention to look at any carbon chain with one double bond located at variable positions in this chain. Figure 13 shows the variation of the number of generated superatoms starting from the end node numbered 1. With all methods, instantiating the first layer will create 50 chain superatoms. These 50 chain superatoms will all contain two carbon nodes and will result from the combination of the query node numbered 1 with the first node of each of the 50 instances of the G1 generic group in Figure 12. With our optimized method which interrupts the progress of the chain between two adjacent generic groups, there is no more superatom created until the 50th layer of progress. Between the first and the 50th layer of progress, the first 50 superatoms moieties which were created by the first layer are completed. Since the 50 instances of G1 group are of growing size, the shorter moieties are first completed but are set in standby until all the moieties issued from the same group instantiation are completed. Then, the 50th layer of progress in the graph will be reached with 50 superatoms moieties resulting from the combination of the carbon node numbered 1 with each instance of the G1 group. In each moiety, the last collapsed node will represent a fusion node, since the progress has been stopped between the consecutive generic groups, G1 and G2. At the 50th layer of progress in the graph, 50 new superatoms moieties are created, as indicated in Figure 13. These moieties are obtained by instantiating the G2 generic group. This optimized method will then generate 100 superatoms moities before starting the process of fusions. The fusion of the first complementary moieties create the first complete superatom numbered 101. Further fusions of other moieties create identical superatoms regarding the reduced subgraph: indeed, all these carbon chain superatoms have one double bond and are connected to the same external neighbor. They are thus redundant and are eliminated. Elimination steps were previously operated after each new moiety generation, but in this case all the moieties were distinct, because their fusion nodes were issued from different instances, and no one was thus eliminated. After the last fusion, the total number of generated superatoms is still 101, since all the complete superatoms were identical to the product of the first fusion and immediately eliminated. The ultimate phase consists in eliminating the 100 initial parent superatoms. This elimination is triggered off by the fact that all their fusion addresses have become forbidden. The total operation resulted in the creation of one complete superatom with the transitory existence of 100 moieties. This is not the case of a nonoptimized method which would combine the instances of G1 and G2 groups, leading to 2500

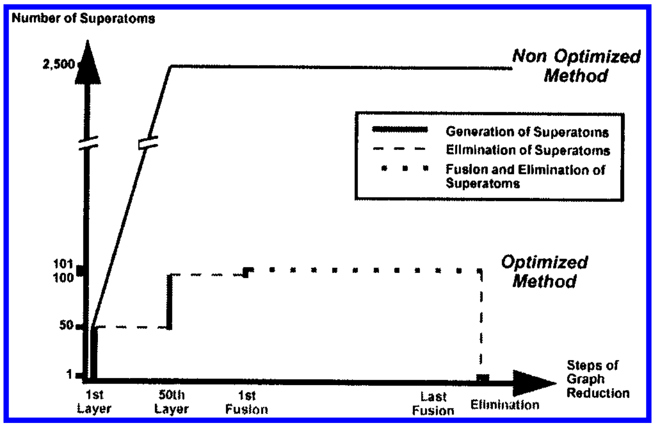


Figure 13. Generation of superatoms during the graph reduction. The number of generated superatoms is plotted against the different steps of the reduction of the graph in Figure 12. The events on the horizontal axis are as follows: the layers of progress along the graph (from 1 to 100) which produce 100 superatoms moities by the optimized method, followed by the steps of fusion of these superatoms moieties, to end, the elimination of the remaining nonfused superatoms moieties. During the process of the optimized method, the generation of superatoms is represented by vertical bars. The fusion and elimination of superatoms is operated between the first and the last fusion step. The remaining phases eliminate the redundant superatoms, but, as explained in the text, no one elimination is possible in this particular case during the layers of progress along the graph. See further explanation in the text.

intermediary superatoms. These superatoms appear identical regarding the needs for reduced graphs, but for eliminating the duplicates, it is necessary to wait for their completion, since next layers could create differences. Indeed, one may conclude that two superatoms are identical only when they are both completed. In Figure 13, the maximal number of intermediary generated superatoms is then 2500 against 101 with our optimized method.

Without the above optimization and considering that common situations may be much more complicated that the query proposed in Figure 12, thousands and thousands of intermediate superatoms, possibly identical once completed, would be created. This feature would cause strong problems of physical limits and CPU needed for looking at all the possible fusions. Despite these optimizations, some database structures may generate excessive amounts of superatoms during the reduction of graphs. These too broad and complicated structures are called nasties and cause problems in any step of the databases setup or during the screening or the atom-by-atom search. The consequence of nasties are also negative for the meaning of chemical patents and concern patentees, patent examiners, and databases producers.[26,27]

SUPERATOMS TRANSLATION ATTRIBUTES AND VARIABLE POSITIONS OF ATTACHMENT DURING FRELS GENERATION

In the above part, the graph was reduced into chain and cyclic superatoms without regarding if the translation of these superatoms was required. The user may easily require superatom translation by using the translation attribute TRA. Four values may be assigned to this attribute: narrow (NT), broad (BT), any (ANY), equal (EQ). The EQ value is set by default. In the graphic visualization, the NT attribute is symbolized by the "<" sign, and BT by ">", as in Figure 12. The Narrow translation (NT) attribute, when applied to a superatom, will retrieve the superatom itself plus specifics belonging to the family. The narrowest level refers to defined structures or Mendeleev atoms. The Broad Translation (BT) attribute is mostly used on specific atoms to retrieve the specific atom and the corresponding superatoms. A native hierarchy of superatoms within the translation process was established, starting from the specific level, going toward a broader level which is reached with chain, rings, and closed set superatoms. However, in the last category, MX is broader than any other metal superatom. The XX superatom represents the broadest level. At the moment, the flag values are not implemented in a way to distinguish between the different levels, but this functionality is in progress. Then, BT value allows a match with broader groups, and NT value allows a match with narrower groups.

This native hierarchy is slightly modified by the combination of free sites and translation flags. Indeed, ARY with two free sites and BT attribute matches HEF. In the same conditions, CYC matches ARY and HEF, while HEA and HET match HEF. With one free site and BT attribute, CHK matches CHE and CHY, and CHE matches CHY. It could be said that the presence of free sites makes HEF broader than ARY, which is not the case when not considering free sites.

The ANY value represents the simultaneous presence of NT, EQ, and BT.

In the graph reduction operated during FRELs generation, the presence or absence of translation attributes is not considered. After the graph reduction, when definitive FRELs are generated, this value is taken into account for assigning the atom values to foci, A positions and B positions. Indeed, two codes have been created for describing the generic superatoms in FRELs. One value will be assigned to the explicit superatom and the other one to the implicit superatom obtained from the reduction of a corresponding developed substructure. In the database, the atom value for an explicit superatom differs from the value for the corresponding defined structure which may be reduced to this implicit superatom. In the query, both values are assigned to nodes affected by translation flags. The set of FRELs generated around the same focus will have to take into account these different values. All the lists of candidates corresponding to these FRELs generated around the same focus will be merged. By the same way, in case of free sites or undefined bond values, all the possible values will be applied to the corresponding position.

Variable positions of attachment are used when a substituent has several possible points of attachment. Figure 14 is derived from Figure 12 by varying the attachment of two acyclic substituents on the ring system. Variable positions are entered using a dummy atom symbolized by "#". The possible sites of attachment to which each dummy atom is attached to the ring are listed in the upper-right of the screen. A total number of 10 dummy atoms is authorized for a structure query. Each dummy atom may be attached to 50 positions. In the absence of the variable position of attachment function, users would have to enter separate structure queries and to perform boolean union on the respective answers sets. However, to avoid excessive combination of genericity, a variable position of attachment may not point to a generic group or a generic superatom with NT or ANY attribute. For similar reasons, it is not permitted for a generic superatom with NT or ANY attribute or for a group with a repeating unit attribute to be directly connected to a dummy atom.

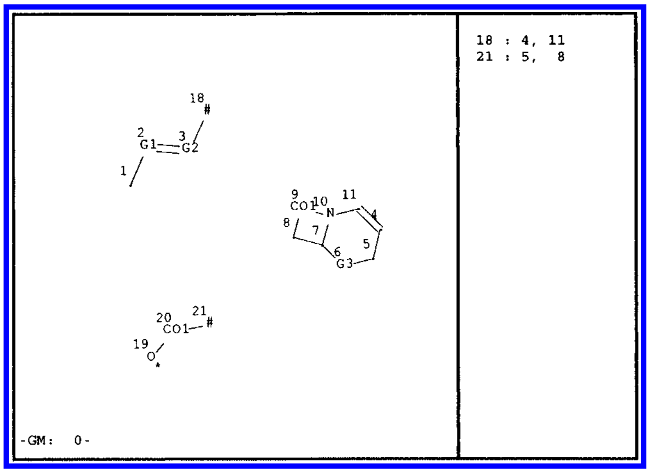


Figure 14. Structure query with variable positions of attachment. This structure has been derived from Figure 12: #, dummy atom used for handling variable positions of attachment; 18:4,11 list of possible attachment for the dummy atom numbered 18.

The presence of variable positions of attachment in the query graph naturally stops any progress during the graph reduction. Nevertheless, as mentioned in Figure 10, the presence and the direction of a variable position of attachment is included within the descriptors of the superatom moieties issued from the first step of graph reduction. During the fusion step, variable positions will be considered in a similar way that the fusion addresses. Each possibility of attachment issued from the variable position will be managed like the fusion addresses, with flags for authorizing or forbidding this attachment.

Since the FRELs system is specific to the Darc family, it was necessary to present some aspects of the implementation of the superatom translation and of the variable positions of attachment in the RE search. The database structures which have passed through the screening step will become candidates for the atom-by-atom search. How the superatom translation and the variable positions of attachment are managed in this definitive match is presented in the next chapter.

SUPERATOM TRANSLATION ATTRIBUTES AND VARIABLE POSITIONS OF ATTACHMENT DURING ATOM-BY-ATOM SEARCH

The atom-by-atom Markush Darc final search is based on backtracking, which is the generalized method for this exact search of generic chemical structures.[28] Figure 15 summarizes the main operations which may occur when running the atom-by-atom search on Markush structures.

Forward is guided by the numbering of the structure query. In this exemple, atoms 1 and 2 of the query match candidates nodes c1 and c2. This may correspond to a generic−generic nodes match or to a specific−specific nodes match. It is then hypothesized that the query atom 3 is a generic group G1. The main atom-by-atom routine calls an "instantiator" subroutine which manages the instances of the encountered generic groups. This management includes the choice of the instances, the elucidation of the obtained subgraph, and the memorization of all the operations on each generic group for a subsequent orientation of the backward process. In this case, the first instance of this query generic group G1 is selected. The first node of this first instance is numbered 20 in the general table of connectivity. This query node 20 is assigned to the candidate node c3. The forward goes on along this instance of G1 until the third node of this instance, numbered 22, is reached. Indeed, this query node again corresponds to a generic group G2 embedded in G1, and an instantiator subroutine is again required. Once again in this example, the first instance of this generic group is selected. The node numbered 30 of this instance is a two-depth attachment node, i.e., it is directly connected with a node contained within the grandfather group of G2. The system then comes back to the invariant graph after matching of this node with the candidate c5 node. It must be noticed that another type of instantiators is called if the generic group is encountered in the candidate instead than in the query structure. It is usual that query and candidate instantiators operate simultaneously during the atom-by-atom search. The memorization of all the operations by the instantiators is very important, since the backward process has to run back along the query and candidate structures following exactly the road opened by the forward process. If any assignment of a candidate node to a query node becomes impossible in a later stage, this backward process will be trigged off, and the marks memorized by the instantiators will allow the stopping of this backward process and the starting again of a forward process with new groups instances.

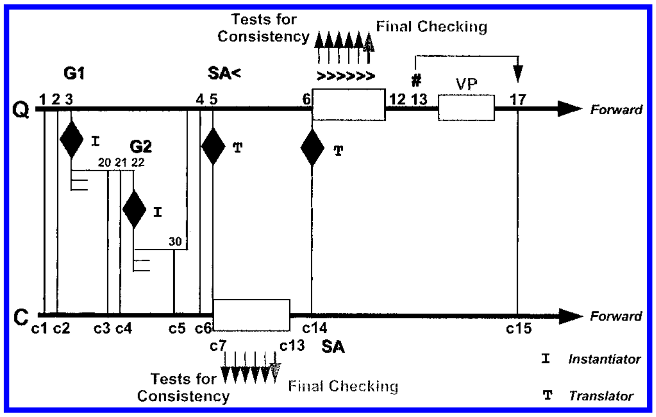


Figure 15. Superatom translation and variable position treatment in the atom-by-atom search. The upper axis represents the forward process along the invariant part of the query structure. The query nodes are numbered on this axis: Gi, generic group; SA<, superatom with narrow translation; >, broad translation; #, dummy atom used for handling variable positions of attachment (VP). After the use of an instantiator, the secondary axis represents the forward process along the generic groups of the query structure. The nodes are also numbered on these secondary axis. The lower axis represents the forward process along the candidate structure. This process is guided by the forward along the query graph. The candidate nodes are numbered on this axis and designated with the letter "c" before this number. The vertical connections between the query and the candidate axis represent the matches between query and candidate nodes. See further explanation in the text.

In Figure 15, it is then hypothesized that the query atom 5 is a generic superatom, labeled with narrow translation. A "translator" subroutine is required, and assigns six candidate nodes to this single query node. The six candidate nodes are collapsed into a defined structure corresponding to the query superatom. For instance, the query node 5 could be ARY with NT translation, and the six corresponding candidates atoms could be contained within a benzene ring. The translation operation requires additional tests of consistency for each new assignment of a candidate node and a final checking when the translation is over. For instance, the superatoms attributes are checked. In the previous example, it would not have been possible to assign an isolated benzene ring to the ARY superatom if the attribute FU−Fused−had been restricted to this superatom. The use

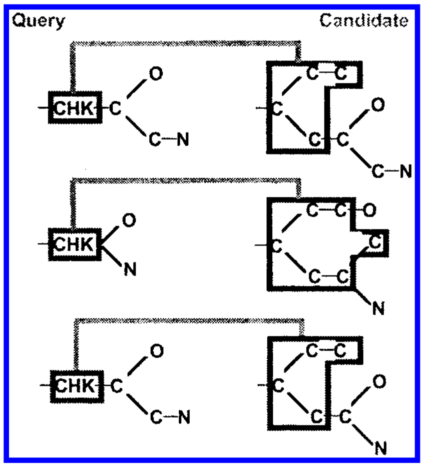


Figure 16. Problem of the Superatom environment for carbon chains. Comments in the text.

of the text notes annexed to the database structures, as in the Bit Screen search, is not yet available for the atom-by-atom search.

The next query node is then a defined atom with BT translation, and the translator module does not detect any inconsistency between this query node and the candidate node c14 which is a superatom. Once again, this superatom could be ARY. Then, this ARY superatom is assigned to six query nodes with BT translation, considering that all the tests are correct.

In conclusion, translation operations cause accidents in the normal graph forward and need to abandon the classical one-to-one match between query and candidate atoms. A subsequent backward process will free in the same time all the nodes collapsed in a translated superatom, either in the query or in the candidate.

The last hypothesis concerns the query node numbered 13, which is a dummy atom corresponding to a variable position of attachment. In this case, one possible attachment, the query node 17, matches the candidate node c15, but the system has jumped from the query node 13 to the node 17. Before any further progress along the forward axis, candidate nodes must be assigned to the "jumped" query nodes 14, 15, and 16. Thus, variable positions of attachment induce another type of accident in the normal forward process of the atom-by-atom search operation.

The above description of Figure 15 shows that the translation operation runs as far as additional nodes consistent with the superatom nature may be collapsed into the corresponding defined structure. This is an exact rule for ring superatoms, since cyclic superatoms may not be contained within another ring. However, a chain superatom may be contained within a chain structure and implicated in part-generic to part-generic matches. Figure 16 presents situations where excluding carbon nodes from the translated chain superatom makes sense. Indeed, the user's intention is quite different when entering each type of query substructure. The atom-by-atom search is thus able to restitute the carbon nodes which had been included within the chain superatom if these nodes are required as external neighbors.

JOINED OPTIMIZATIONS OF SEARCH STEPS

For improving and not decreasing the performances of the Markush Darc software with superatom translation and variable positions of attachment, joined optimizations were necessary. Powerful tools are now offered to users, but this may result in increasing vagueness in the query formulation. As a consequence, screening steps produce more answers which are treated in a more complex and CPU-consuming atom-by-atom search. It must be outlined that too excessive vagueness is sometimes expressed in users queries. Optimizations were also necessary for accompanying FRELs reduction, which could result in a very limited loss of specificity of the RE search. On the contrary, screening was made more discriminating after the following optimizations.

A new mechanism allowed to keep FRELs with less than two A positions after reduction and even no A position. The only condition is that the unique A position resulting from reduction is necessarily specific. This limitation already existed for FRELs with two A positions and was thus extended to FRELs with one A position.

Developments were performed for transforming a variable A position described in the query FRELs by a list of instances into a set of specific A positions. Instances of FRELs are generated around the focus changing the specific value of the A position. This is performed for the shorter list if this query focus is connected to many variable A positions.

In some cases, the instances of a generic group or the possible attachments of a variable position of attachment in the query give exactly the same result in the generated FREL. Such "false variabilities" are identified and replaced by a specific position.

The three above optimizations avoid the rejection of many FRELs foci, and allow the generating of more FRELs and the increasing of the discriminatory power of RE search.

New features were introduced in the Bit Screen description, particularly features susceptible to be contained within the collapsed structures corresponding to superatoms after the FRELs reduction. The aim was to describe in the Bit Screen the elements which could have been lost by the FRELs reduction. In that way, the internal description of chain superatoms was diversified.

To manage the excessive vagueness of some queries, the limit for FRELs and Bit Screen searches was increased from 5000 to 50 000 answers. Other internal limits linked to these screening searches were also strongly increased, for limiting the risks of overflow.

In the minority of cases, no FRELs are still generated. In that condition, automatic Bit Screen is running without any user action.

If the user does not want to spend online connection during the atom-by-atom search, or if excessive vagueness of the query requires a long time for this search, it is possible to perform Batch search offline. Since it was expected that this batch search will be used more and more, the CPU time allowed during batch was increased. The candidates which could not be analyzed during the CPU time authorized for the online atom-by-atom search are called RX candidates. The possible number of RX candidates for one atom-by-atom search was increased from 400 to 5000, which corresponds to the number of candidates for any atom-by-atom search. Then, the search is no longer interrupted by excessive amounts of RX candidates.

For INPI users, the total number of allowed batches was increased from 3 to 10.

CONCLUSION

The intensive work performed on the Markush Darc software considerably enhanced the functionalities of this online service in the field of chemical patents. The translation capability was required by users and searchers,[29,30] and it is now assumed that Markush Darc has ended a significant disadvantage. Superatoms able to be translated cover the main fields of the chemical area. In addition to this translation capabiblity, users may now enter searchable variable positions of attachment in the structure query. Combined with the already existing generic groups, free sites, and variable bonds nature, the above features offer powerful tools to the user. Deep modifications of the screening searches—Bit Screen and the unique FRELs screen—and of the atom-by-atom search were necessary for implementing these new functionalities. The overall service was optimized and is still in progress, in a way of increasing search limits or allowing batch search in the best conditions.

However, such powerful capabilities may be used for entering excessive vagueness in the structure query. This point is general to the use of topological search systems since chemists may encounter problems for expressing their vocabulary with the syntax proposed in these systems.[31] The user's intention must be defined in the query formulation.[4] In that way, additional tools will be offered to the user for better expressing the intention of the entered query. These tools are in progress and will allow the narrowing of too broad queries. Particular attention must be given to the meaning of free sites and to the required level of translation. Recent developments also participate in the integration of structural and bibliographic data. Indeed, one must always keep in mind that the aim of the Markush Darc software is to display chemical generic structures which are attached to patents documents. Following this aim, we have achieved in giving the user the possibility of finding in databases all the structural answers to the input query. We must now better help the user to define which structural answers he is really looking for.

ACKNOWLEDGMENT

We thank French Patent and Trademarks Office, INPI, for the permanent support and interest given to Markush Darc developments; Derwent Information Ltd for funding and participation to the functionalities described in this paper; the technical team in Questel.Orbit, with particular mention to Frédéric Lellouche and Dominique Renaud whose contribution to Markush Darc project has been essential for several years; Barbara Thanni for her helpful comments on an earlier version of this manuscript; and Safia Boulkalem for her major help in the material realization of this paper.

REFERENCES AND NOTES

(1) Kaback, S. M. What's in Patent? Information! But Can I Find it? *J. Chem. Inf. Comput. Sci*. **1984**, *24*, 159−163.

(2) Lynch, M. F. The Sheffield University Generic Chemical Structures Project. *World. Pat. Inf.* **1993**, *15*, 135−141.

(3) Lynch, M. F.; Barnard, J. M.; Welford, S. M. Generic Structures Storage and Retrieval. *J. Chem. Inf. Comput. Sci*. **1985**, *25*, 264−270.

(4) Dethlefsen, W.; Lynch, M. F.; Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 11. Theoretical Aspects of the Use of Structure Languages in a Retrieval System. *J. Chem. Inf. Comput. Sci*. **1991**, *31*, 233−253.

(5) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. The Sheffield University Generic Chemical Structures Project-A Review of Progress and of Outstanding Problems. In *Chemical Structures*; Warr, A. W., Ed.; Springer-Verlag: Berlin, 1988; Vol. 1, p 151.

(6) Tarjan, R. E. Graph algorithms in Chemical Computation. In *Algorithms for Chemical Computations*; Christoffersen, R. E., Ed.; ACS Symposium Series 46. American Chemical Society: Washington, 1977; p 1.

(7) Barnard, J. M. Problems of Substructure Search and their Solution. In *Chemical Structures*; Warr, A. W., Ed.; Springer-Verlag: Berlin, 1988; Vol. 1, p 113.

(8) Attias, R. DARC Substructure Search System: A New Approach to Chemical Information. *J. Chem. Inf. Comput. Sci*. **1983**, *23*, 102−108.

(9) Questel S. A. The search process Markush DARC. French Patent, 9, 004,134, 1990.

(10) O'Hara, M. P.; Pagis, C. The Pharmsearch Database. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 59−63.

(11) Shenton, K. E. Graphic Retrieval of Patent Information. In *Proceedings of the 9th International Online Information Meeting*; London, 1985; p 43.

(12) Shenton, K. E; Norton, P.; Ferns, E. A. Generic Searching of Patent Information. In *Chemical Structures*; Warr, A. W., Ed.; Springer-Verlag: Berlin, 1988; Vol. 1, p 169.

(13) Fisanick, W. Storage and Retrieval of Generic Chemical Structure Representations. U.S. Patent, 4, 642,762, Feb. 10, 1987.

(14) Fisanick, W. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 145−154.

(15) 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 Structures Screening. *J. Chem. Inf. Comput. Sci*. **1984**, *24*, 57−66.

(16) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 10. The Assignment and Logical Bubble-up of Rings Screens for Structurally Explicit Generics. *J. Chem. Inf. Comput. Sci*. **1989**, *29*, 215−224.

(17) Holliday, J. D.; Downs, G. M.; Gillet, V. J.; Lynch,M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 14. Fragment Generation from Generic Structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 453−462.

(18) Holliday, J. D.; Downs, G. M.; Gillet, V. J.; Lynch,M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 15. Generation of Topological Fragment Descriptors from Nontopological Representations of Generic Structure Components. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 369−377.

(19) Dubois, J. E.; Panaye, A.; Attias, R. DARC System: Notions of Defined and Generic Substructures. Filiation and Coding of FREL Substructure (SS) Classes. *J. Chem. Inf. Comput. Sci*. **1987**, *27*, 74−82.

(20) Dubois, J. E.; Mathieu, G.; Peguet, P.; Panaye, A.; Doucet, J. P. Simulation of Infrared Spectra: an Infrared Spectral Simulation Program (SIRS) which Uses DARC Topological Substructures. *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 290−302.

(21) Attias, R.; Dubois, J. E. Substructure Systems: Concepts and Classifications. *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 2−7.

(22) Welford, S. M.; Lynch, M. F.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammars and their Role in the Manipulation of Chemical Structures. *J. Chem. Inf. Comput. Sci*. **1981**, *21*, 161−168.

(23) Gillet, V. J.; Downs, G. M.; Ling, A. I.; Lynch, M. F.; Venkataram, P.; Wood, J. V.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 8. Reduced Chemical Graphs and Their Applications in Generic Chemical Structure Retrieval. *J. Chem. Inf. Comput. Sci*. **1987**, *27*, 126−137.

(24) Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Lynch, M. F.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 13. Reduced Graph Generation. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 260−270.

(25) Panaye, A.; Doucet, J. P.; Cayzergues, P.; Carrier, G.; Mathieu, G. L'Elucidation Structurale: des Données Spectrales à la Reconnaissance Moléculaire. L'Actualité Chimique. **1988**, 103−111.

(26) Sibley, J. F. Too Broad Generic Disclosures: A Problem for All. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 5−9.

(27) Milne, G. W. A. Very Broad Markush Claims; A Solution or a Problem? Proceedings of a Round-Table Discussion Held on August 29, 1990. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 9−30.

(28) Ray, L. C.; Kirsh, R. A. Finding Chemical Records by Digital Computers. *Science* **1957**, *126*, 814−819.

(29) Meurling, A. CAS ONLINE and DARC: A Comparison. Database **1990**, *13*, 54−63.

(30) Schmuff, N. R. A Comparison of the MARPAT and Markush DARC Software. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 53−59.

(31) Simmons, E. S. The grammar of Markush Structure Searching: Vocabulary vs Syntax. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 45−53.

CI9600364