DERWENT

WORLD PATENTS INDEX

Markush DARC User Manual

Edition 1
ISBN: 978 1 905935 14 7

THOMSON™

Published by Thomson Scientific

Visit the Thomson Scientific web site at http://scientific.thomson.com

Edition 1 published December1993

ISBN: 978 1 905935 14 7 (Edition 1)

Table of contents

Section 1: Introduction

Section 2: Connection to Markush DARC

Section 3: Constructing the query

Section 4: Structure representation conventions

Section 5: Searching

Section 6: Answer display and interpretation of the results

Section 7: Answer handling and Cross-file searching

Appendix 1: Markush DARC superatoms, shortcuts and attributes

Appendix 2: Markush DARC and Questel Plus command summary

Appendix 3: WPIM database

Appendix 4: MPHARM database

Appendix 5: Search examples

Section 1
Introduction

I. The Markush concept

I.1. Markush structures

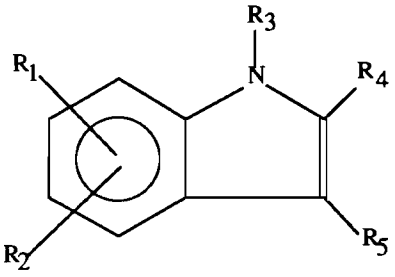
Markush structures are commonly found in Chemistry patents.
Markush structures are chemical structure diagrams which are used by patent applicants to express families of chemical compounds sharing some structural features.
These families of compounds may comprise large numbers of compounds and a single Markush structure may be used to express all of them.

Markush structures are diagrams describing atoms and bonds.

They have the following characteristics:

- *fixed parts* expressing the structural features common to all of the compounds of the family consisting of defined atoms linked by defined bonds

- *variable parts* attached to the fixed parts (substituents) or included in the fixed parts and usually represented by symbols such as R1, R2 etc...



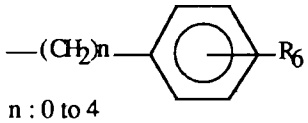
R1 : H, Cl, amino, nitro or hydroxy

R2 : as R1

R3 : H, methyl or ethyl

R4 : H, alkyl, alkanoyl, alkoxycarbonyl or cycloalkyl

R5 : H, alkyl, alkoxy, optionally substituted heterocycle or



n : 0 to 4

R6 : H, nitro, halogen or alkyl

Example of a Markush structure

Each variable part (here R1 to R6) may be defined by:

- a set of specific atoms (example in R1: Cl for chlorine)
- groups of atoms (example in R1: NO2 for nitro group)
- generic terms (example in R4: alkyl to cover a saturated carbon chain of variable length)
- drawings which may in turn comprise fixed and variable parts (this is the case here for the last value of R5, which contains R6: we say that the R6 substituent is *nested* in the R5 substituent).

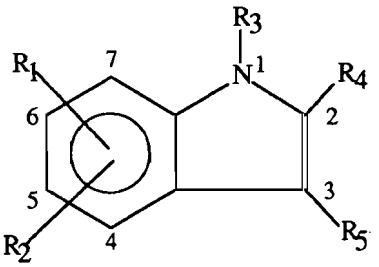
Some variable parts may be variably attached on the fixed parts: we say that there is an *undefined attachment* of the variable substituent on the fixed part: this is the case in the above example, where R1 and R2 may be attached to any position of the benzene ring.

Each member of the chemical family embraced by the Markush structure is a combination of the fixed part with the appropriate permutation of the variable parts.

In a patent, a Markush structure represents the broadest aspect of the invention which is claimed in the patent.
The term *generic* is commonly used to qualify this broadest aspect.

Some of the individual compounds which are covered by a Markush structure in a patent are specifically described in the patent.
They represent the *specific* aspects of the invention.

If we still consider the following example:



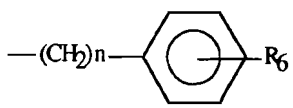
R1 : H, Cl, amino, nitro or hydroxy

R2 : as R1

R3 : H, methyl or ethyl

R4 : H, alkyl, alkanoyl, alkoxycarbonyl or cycloalkyl

R5 : H, alkyl, alkoxy, optionally substituted heterocycle or



n = 0 to 4

R6 : H, nitro, halogen or alkyl

Example of a Markush structure with ring atom numbering

The following list of compounds could be described specifically with specific values of the variables:

	R1	R2	R3	R4	R5	R6
Compound 1	4 Cl	H	H	CH3	benzyl	4 Cl
Compound 2	7 Cl	H	H	CH3	benzyl	4 Cl
Compound 3	4 Cl	7 Cl	H	CH3	benzyl	4 Cl
Compound 4	4 Cl	H	CH3	CH3	benzyl	H
Compound 5	H	H	H	cyclopropyl	H	-
Compound 6	6 OH	7 NH2	H	H	benzyl	3 NO2
Compound 7	H	H	H	cyclopropyl	benzyl	4 CH3
Compound 8	H	H	CH3	H	phenyl	H
Compound 9	6 OH	H	CH3	H	phenyl	H

The following possible situations may occur for the variable part values of a Markush structure:

- the variable part values are described by specific terms in the Markush structure: this is the case in our example for the values of R1

- the variable part value is described by a generic term in the Markush structure and there are corresponding specific terms from specific compounds which are specifically described in the patent specification: this is the case for the value alkyl in R4, which has one corresponding specific value CH3 from compounds 1 to 4

- the variable part value is described by a generic term in the Markush structure and there is no corresponding specific compound described in the patent specification: this is the case in R4 for alkanoyl and alkoxycarbonyl generic values, which have no corresponding specific values in the table of specific compounds.

I.2. Retrieval methods for chemical compounds

Various database storage and retrieval methods have been used for chemical compounds. The various methods can be classified in four basic types:

- textual systems
- fragmentation coding systems
- single compound graphic systems
- Markush (generic) graphic systems.

The textual or nomenclature type systems are the oldest.
The newest of these are Markush graphic systems.

The following table lists some of the characteristics of each type of chemical retrieval system.

	textual systems	**fragmentation coding systems**	**single compound graphic systems**	**Markush graphic systems**
databases	*specific compounds *generic compounds	*specific compounds *generic compounds	*specific compounds ONLY	*specific compounds *generic compounds
query formulation	*specific queries *generic queries	*specific queries *generic queries	*specific queries *generic queries	*specific queries *generic queries
retrieved answers	*low recall *low relevance	*high recall *low relevance	*low recall *high relevance	*high recall *high relevance
ease of learning and use	*difficult to learn *special training	*difficult to learn *special coding training	*easy to learn *easy to use	*easy to learn *easy to use
search and display characteristics	*complex strategy formulation *no structure display	*complex strategy formulation *no structure display	*language of the chemist	*language of the chemist *as in patents
generic to specific and specific to generic translation	no	no	no	yes

I.3. What is Markush DARC

Markush DARC is a database management system which allows the production and searching of databases containing Markush structures as they are found in chemistry patents.
Markush structures in Markush DARC databases are stored and searchable as they appear in patents: by their drawings, using the language of the chemist.
Storage and retrieval are graphical.

All of the features characterising Markush structures in patents are kept in Markush DARC databases:

- fixed and variable parts
- generic terms
- variable points of attachment
- specific and generic aspects of the patent.

Markush DARC offers much more complete retrieval of chemical information contained in patent documents than single compound database systems in which only a part of the family of compounds generated by a Markush structure is covered.

For searching a Markush DARC database, it is possible to use:

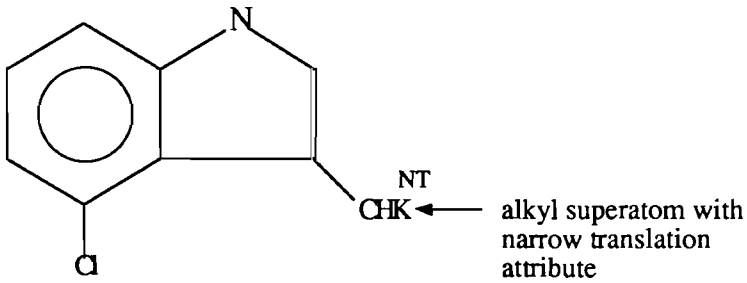
- specific queries where all the atoms and bonds are defined or
- Markush (generic) queries having fixed and variable parts, generic terms and variable points of attachment.

The storage and retrieval of chemical structures via graphical methods is built upon the concept that a chemical structure can be represented as a graph composed of points or nodes, with connections between certain nodes.
Each node usually represents an atom, with its identity recorded. A node can also represent a generic term. In Markush DARC, these generic terms are represented by pseudo atoms called *superatoms*.
Connections between the nodes represent bonds.
When a user conducts a search, the graph or connection table formed by the query is compared to the connection tables stored in the database.
Normally, each node and connection defined in the query must be exactly matched to the stored connection tables in order for a stored structure to be a valid answer.

In Markush DARC, the matching capability has been widened. It is possible to:

- use specific terms (i.e. specific atoms or groups of atoms) in the query to find corresponding generic terms (i.e. superatoms) in the relevant answers of the database: Markush DARC translates the specific terms into the corresponding generic terms. This is the broad translation capability.
- use generic terms (i.e. superatoms) in the query to find corresponding specific terms (i.e. specific atoms or groups of atoms) in the relevant answers of the database: Markush DARC translates the generic terms into the corresponding specific terms. This is the narrow translation capability.

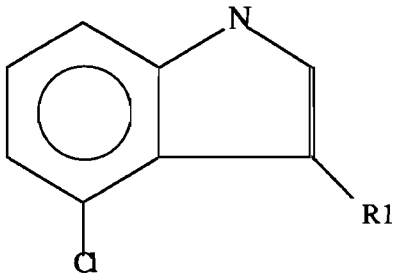
Let us consider the following query:



Example of a query with narrow translation on an alkyl superatom

The use of the narrow translation attribute on the CHK (alkyl) superatom allows the user to avoid specifying several specific carbon chain values because Markush DARC translates the generic term CHK into corresponding specific terms, e.g. methyl or isopropyl.

The following structure could be an answer to this query:



R1 is hydrogen or methyl

Example of a possible answer

The alkyl superatom of the query matches the methyl value of R1 in the answer.

All of the four possible ways for a match between a query and a database structure are allowed in Markush DARC, thanks to the translation capability:

Query structure	Database structure	Translation
specific	specific	equal
specific	generic	broad
generic	specific	narrow
generic	generic	equal

The *specific to generic translation* is particularly useful because, in a Markush structure found in a patent, generic terms of a group may be used without any corresponding specific term if there is no corresponding specific compound described or mentioned in the patent.

The *generic to specific translation* is also useful , because in a Markush structure specific terms may be used without any corresponding generic term if the generic aspect is not claimed for the corresponding value of the variable part.

It can also be useful to match exactly specific terms of a structural fragment in a query structure with specific terms of relevant answers in the database if the information that the structural fragment has been described specifically in the patent is of importance (equal translation).

In the same way, it may be important to know if a structural fragment has been claimed in a generic way, without any corresponding specifically described compound. An exact match of the generic term used in the query with generic terms of relevant answers in the database is required in this case (equal translation).

I.4. Overview of Markush DARC basics

I.4.1. G groups

In Markush structures, a wide variety of symbols are used to stand for variable parts. Conventionally, a capital R, qualified by a numerical suffix (R1, R2, etc...) is the most common way to represent variables. Other letters, such as A, X or Y, can also be used.

In Markush DARC, the variable parts are always represented by the letter G (for Group) followed by a numerical suffix in the range from 1 to 50.

Each variable group in a Markush will have two or more values.

Each of these values is a structural fragment which can itself be represented by a graph.

The parent graph, which corresponds to the fixed part of the Markush structure with the variable parts attached to it, is called group zero (G0). It contains variable groups (G1, G2, etc...).

Each variable group identified in group G0 is represented separately by a family of separate partial graphs which correspond to the values of the variable substituent.

These partial graphs may in turn contain variable groups. We say that these variable groups are *nested* within their parent group.
This process can be continued, in that the nested variables can have among their values other embedded variables. It is thus possible to have a number of levels of nesting.

I.4.2. Atoms and shortcuts

All atoms of the Mendeleev periodic table are possible in Markush DARC structures.

To facilitate and speed up the construction of queries, Markush DARC provides defined groups of atoms called shortcuts: example NO2 for nitro, Ph for phenyl, etc...
There are 21 shortcuts available in Markush DARC.

I.4.3. Superatoms

Superatoms are used in Markush DARC to represent *generic terms* used in patents to express a set of chemical structures having some shared aspect. The three basic types are:

- chemically defined closed set: these cover a fixed list of possible meanings, all of which have a common discrete chemical feature. An example is halogen (Br, Cl, F, I), covered in Markush DARC by the HAL superatom.

- chemically defined infinite set: these terms have quite a precise meaning to a chemist, but encompass an infinite variety of specific structures. Typical examples are alkyl (CHK superatom) and aryl (ARY superatom).

- set defined by property: a portion of a structure may be defined in terms of activity, such as chromophore or protecting group. Some of these have a corresponding superatom in Markush DARC (for example, the PRT superatom stands for protecting group).

There are 20 superatoms of these three types.

In addition, there are 30 superatoms corresponding to standard amino acids found in peptide chemistry. Each superatom corresponds to a defined amino acid.

Two additional superatoms are used in Markush DARC: one is used to cover any atom or group excluding hydrogen (XX superatom), while the other is used to cover an undefined group (UNK superatom).

I.4.4. X atom

In Markush DARC queries, it is possible to use a special symbol, the X atom, to specify any atom or superatom except hydrogen.

I.4.5. Bonds

Bonds are used as connections between nodes.
A node can be an atom, a superatom, a shortcut or a G group.

The possible values of bonds for query formulation are the following:

- single
- double
- triple
- normalised (aromatic or tautomeric)
- undefined
- Z bond to specify a list of possible values of bonds.
- peptide bond

Stereochemical bonds can be searched and displayed in the Markush DARC databases in graphic query input mode :

- up
- down
- racemic
- unknown.

For query formulation, there are 3 types of bonds which can be specified:

- acyclic
- cyclic
- undefined type: cyclic or acyclic.

I.4.6. Attributes

Attributes are labels used on atoms and superatoms to provide additional information.
There are four classes of attributes:

- atom attributes

    - charge
    - delocalized charge
    - abnormal mass
    - abnormal valency
    - deuterium
    - tritium

- superatom attributes

    - chain attributes, used to specify carbon chain length (low, medium or high) and branching (straight or branched)
    - ring attributes, used to distinguish monocyclic and fused rings and to specify the degree of saturation (saturated or unsaturated)
    - multiplier attributes, used in the databases on superatoms to specify the minimum and/or maximum degree of substitution which is possible for a given generic term
    - numbering attributes, used to label a particular superatom and link it to additional text information

- peptide attributes, which are of two types:

    - configuration attributes used on peptide superatoms to specify the configuration of an amino acid at its alpha-carbon: D, L or DL.

    - position attributes used on atoms, superatoms, groups and shortcuts which are substituents of a peptide superatom to specify the positions of substituents on the amino acid

- polymer attributes.

I.4.7. Translation attributes

Translation attributes are used in Markush DARC queries to specify the type of matching which is required between the query structure and possible answers in the database.

- broad translation to translate specific atoms of the query to corresponding superatoms in the database

- narrow translation to translate superatoms of the query to corresponding specific atoms or groups of atoms in the database

- equal translation to match specific to specific terms or generic to generic terms

- any translation to translate a generic or specific term to any term (superatom - including the XX superatom - and specific atoms)

For more details on translation attributes, please refer to *Section 3 - Constructing the query - : Attributes chapter.*

I.4.8. Text description

Additional information in the form of controlled or free text may be stored for each Markush DARC structure.

This is extra information about the entire structure or certain parts of it which cannot be handled graphically. It is not used for searching in the present version of Markush DARC.
It is displayed with the structure and is intended mainly as an aid in determining the relevance of an answer.

II. Markush databases

II.1. Structure databases and bibliographic files

The Markush DARC software supports two structure databases at present:

- MPHARM (Pharmsearch structures) produced by INPI (Institut National de la Propriété Industrielle)
- WPIM (WPI-Markush) produced by Derwent Publications Ltd.

There are also two corresponding training databases:

- IPAT: training file corresponding to MPHARM
- WPAT: training file corresponding to WPIM.

The companion bibliographic databases are:

- PHARM
- WPIL
- bibliographic training files:
    - ZPHARM
    - ZWPIL.

II.2. Database Contents

The MPHARM Markush DARC database contains structures (Markush structures and specific structures) found in the following pharmaceutical patents:

- Patents from the French Patent Office
- Patents from the European Patent Office
- Patents from the US Patent Office
- Patents from the British Patent Office (from 1992)
- Patents from the German Patent Office (from 1992)
- PCT patents (Patent Cooperation Treaty) (from 1993)

The WPIM Markush DARC database contains structures (Markush structures and specific structures) found in all the Chemical Patent Index (CPI) patents of Derwent in Sections:

- B (Farmdoc
- C (Agdoc)
- E (Chemdoc).

The training files have the following contents:

- IPAT: 1527 structure records from MPHARM
- WPAT: 1546 structure records from WPIM

Most of the Markush DARC system features are used in the same way from file to file; they are not database dependent.

Any features particular to a file are included in the relevant *database appendices (Appendices 2 and 3)* produced by INPI and Derwent Publications Ltd, which can be found at the end of this manual.

III. Help Desk Technical Assistance

Users should contact the Help Desk for any assistance they may require for searching Markush Databases.

The Help Desk is available daily (Monday through Friday) from 9.00 hours to 18.00 hours. The number to call in France is +33 (1) 46 14 51 00.

Technical Assistance is also available through Questel's Agents in Belgium, Germany, United Kingdom, Japan and USA. Addresses and telephone numbers of these Agents are given below:

In France:
Questel S.A.
Le Capitole
55, avenue des Champs Pierreux
92029 Nanterre Cedex - FRANCE
Customer Services: +33 1 46 14 56 60
Help Desk: +33 1 46 14 51 00
Switchboard: +33 1 46 14 55 55
Fax: +33 1 46 14 55 11

In the USA:
Questel Inc.
2300 Clarendon Boulevard
Suite 1111
Arlington, VA 22201 - USA
Telephone: (703) 527 7501
Toll-free number in North America: (800) 424 9600
Fax: (703) 527 7664

In Japan:
Maruzen Company Ltd.
Masis Center
P.O. Box 5335
Tokyo International 100-31 - JAPAN
Telephone: (03) 3 271 6068
Fax: (03) 3 271 6082

Kinokuniya Company Ltd.
ASK Information Retrieval Services Department
P.O. Box 55 Chitose - Tokyo 156 - JAPAN
Telephone: (03) 3 439 0123
Fax: (03) 3 439 1093

In Germany:
IuK Information Services GmbH
Merzhauser Str. 110
W-7800 Freiburg - GERMANY
Telephone: (0761) 459 0720
Fax: (0761) 40 9668

In Belgium:
CNDST
Bibliothèque Royale
Boulevard de L'Empereur 4
1000 Brussels - BELGIUM
Telephone: (02) 519 5656
Fax: (02) 519 5679

In the United Kingdom:
IS Information Service Ltd
Suite 5 - Hartley's Place
Church Lane, Wexham
Slough, Berks SL3 6LD
UNITED KINGDOM
Telephone: (0753) 512 513
Fax: (0753) 553 193

Section 2
Connection to
Markush DARC

I. Type of Terminal

You can search Markush DARC using a non-graphics or graphics terminal or a micro-computer equiped with a graphics emulation software.

The type of terminal you are using must be declared to the system the first time you connect to Markush DARC.

The terminal type is entered with the Options (OP) command which is entered at the ST command level. Prior to entering your terminal type you are requested to select the language in which you choose to work - either English or French. Also, you need to specify whether you require a cost estimation display of your search.

How you obtain the list of terminal types and software packages and select from it is shown below. The list of graphics terminals supported is displayed below. Also the types of emulation software packages. You define the terminal type by entering the number which is displayed to the left of the terminal or software name.

If you are working with a non-graphics terminal you may perform all the normal search operations on Markush DARC but you may not display structures on your terminal screen or printer.

Graphics emulation software packages are used with micro-computers - PC or Macintosh.

Example:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? OP
   FRANCAIS(1)/ENGLISH(2) ?  2
   DISPLAY COST OF THE SESSION : Y/N ? Y
   ** TERMINAL INFORMATION **
 0. NON-GRAPHICS TERMINAL              ----->   0
 1. TEKTRONIX 4010,4014                ----->   1
 2. TEKTRONIX 4020                     ----->   2
 3. TEKTRONIX 41XX,42XX                ----->   3
 4. VT 24X,330,34X  (TEKTRO MODE)      ----->   4
 5. VT 640 ( OR PCPLOT-IV EMULATOR )   ----->   5
 6. HEWLETT-PACKARD 2647A,2648A        ----->   6
 7. SECAPA                             ----->   7
          ** EMULATORS **
 8. VERSATERM-PRO (MACINTOSH)          ----->   8
 9. EMUTEK 5+,7+                       ----->   9
10. EMUTEK LEVEL 2                     ----->  10
11. INFOLOG                            ----->  11
12. ZSTEM 240                          ----->  12

   PLEASE ENTER YOUR TERMINAL CODE ------->  8
   TRANSMISSION RATE (BAUD) ?  2400
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

II. Connecting to Markush DARC

There are 3 ways of connecting to Markush DARC. These are listed below:

- Direct Connection
- Connection from Questel Plus
- Connection from Generic DARC

II.1. Direct Connection

Regardless of the type of terminal you are using the terminal must be equiped with a modem. To connect directly to the Questel computer and Markush DARC you may connect manually by dialing the number of the local telecommunications network and typing in the Network User Address and your Logon and Passsword.

Alternatively, if you have a communications software package you may automate your access procedure. The communications package must be configured for accessing the Questel computer.

The procedure which is followed either manually or automatically is as follows:

1. Connect to the local telecommunications network.

2. Enter the terminal identifier (this is optional depending on the network) and the network user identification (NUI). Neither of these is necessary when accessing the Questel Computer in France.

3. Enter the Network User Address (NUA) of the Questel Computer.

   In France the NUA is :**1061902007**
   Outside of France the NUA is: **208061902007**

4. Enter the letter Q

5. Enter your user number followed by ,MD, indicating that you wish to connect to Markush DARC.

   For example: **XXXXX:XX,MD**

6. Enter your password (4 character alphanumeric code). For security purposes your password does not appear on the screen. Your password should be kept strictly confidential to prevent "piracy". You may wish to change your password from time to time.

Example:

```
TRANSPAC 103023069
1061902007

COM


Nom du service / Service name : QUESTEL etc... ?
q


enter user number preceded by LOGON
04101:K5,MD
PASSWORD: ****
18 H 21   *   93.05.05


   April 29th,1993.

   MPHARM : Period covered
       US patents : 1985 week 28 to 1993 week 03
       EP and FR patents : 1985 week 36 to 1993 week 11
       DE and GB patents : 1993 week 02 to 1993 week 03
       PCT patents       : 1993 week 03
       BSM (Specific French Drug Patents) : Complete collection
       (certificates of addition pending).
```

On connecting to Markush any messages relating to the databases or new features are displayed, before you arrive at the ST command level.

```
****    LAST SELECTED DATA BASE :    WPIM      ****
****    BASE WPIM              -   30/04/93    ****
    187615 COMPOUNDS - LAST CN : 9226-74105

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

II.2. Connection from Questel Plus

If you have previously connected to Questel Plus and you wish then to connect to Markush DARC. For example, if you wish to cross-file search between a Markush bibliographic and a Markush DARC structure database. Or simply, if after having completed a bibliographic search on any database on Questel Plus you wish to move to Markush DARC to carry out another search. You need only enter the command ..ST MDARC.

Example:

```
  Search statement  1

?ASPART+/ALL

  ** SS 1: Results 3.428

  Search statement  2

?1 AND (92 OR 93)/PN

  ** SS 2: Results 735

  Search statement  3

?..JOIN TO WPIM VIA ASPART GEN

  Total number of terms extracted:  380

  Search statement  3

?..ST MDARC
```

When disconnecting from Questel Plus information relating to the cost of the online session is given. On connecting to Markush DARC messages pertaining to the databases or new features are displayed before going to the ST command level.

```
****    LAST SELECTED DATA BASE :    WPIM      ****
****    BASE WPIM                -   30/04/93   ****
    187615 COMPOUNDS - LAST CN : 9226-74105

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

II.3. Connection from Generic DARC

If you are connected to Generic DARC and you wish to pass to the Markush DARC system you need only enter the command MD (Markush DARC) at the ST command level.

Example:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,BL,BI,MD,INFO) ?  MD
```

Section 3
Constructing the query

I. Query input modes

Queries for Markush DARC can be constructed using either graphic or text input.

In both cases, the query is constructed of nodes (*atoms*) and connections (*bonds*).

The query can be further refined by permitting substitution (*free sites*), by restricting definitions (*attributes*) and by placing limits on acceptable answers (*other specifications*).

For query input the nodes (atoms) are defined using the following:

- element symbols
- shortcuts
- G groups (generic groups)
- superatoms

II. Query input basics

II.1. Element symbols

The element symbols are the conventional element symbols shown in the following table.
The normal valency of each element is listed under the atomic symbol in the table. To specify valencies other than the normal valencies, you may use the AV appropriate attribute.
This feature is discussed later in the manual.

H 1																	He 0
Li 1	Be 2											B 3	C 4	N 3	O 2	F 1	Ne 0
Na 1	Mg 2											Al 3	Si 4	P 3	S 2	Cl 1	Ar 0
K 1	Ca 2	Sc 3	Ti 4	V 5	Cr 3	Mn 2	Fe 3	Co 2	Ni 2	Cu 2	Zn 2	Ga 3	Ge 4	As 3	Se 2	Br 1	Kr 0
Rb 1	Sr 2	Y 3	Zr 4	Nb 5	Mo 6	Tc 7	Ru 3	Rh 3	Pd 2	Ag 1	Cd 2	In 3	Sn 4	Sb 3	Te 2	I 1	Xe 0
Cs 1	Ba 2	La 3															
			Ce 3	Pr 3	Nd 3	Pm 3	Sm 3	Eu 3	Gd 3	Tb 3	Dy 3	Ho 3	Er 3	Tm 3	Yb 3	Lu 3	
			Hf 4	Ta 5	W 6	Re 7	Os 4	Ir 4	Pt 4	Au 3	Hg 2	Tl 2	Pb 4	Bi 3	Po 2	At 1	Rn 0
Fr 1	Ra 2	Ac 3															
			Th 4	Pa 5	U 6	Np 5	Pu 4	Am 3	Cm 3	Bk 3	Cf 3	Es 3	Fm 3	Md 3	No 3	Lw 2	

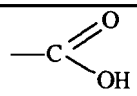
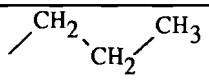
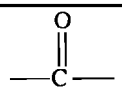
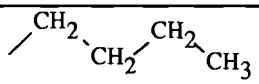
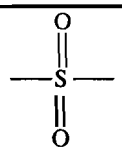
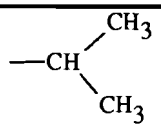
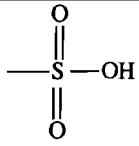
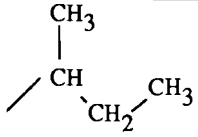
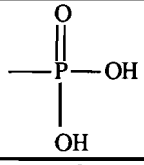
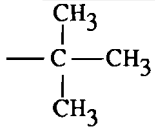
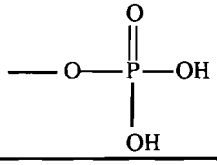
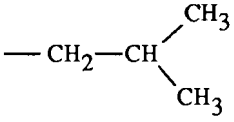
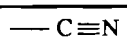
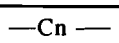
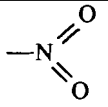
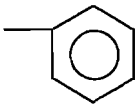
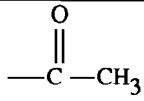
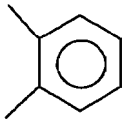
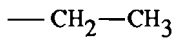
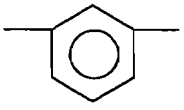
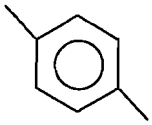
II.2. Special symbols

X is a symbol used to represent any atom or superatom except hydrogen. It is in effect a "wild card" for non-hydrogen atoms. This symbol is used only in queries. It is never used in stored structures of the databases.

Another symbol, the # symbol, is used to specify variable points of attachment of a node to another part of a structure.

II.3. Shortcuts

The shortcut symbols which are listed below may be used to represent the indicated groups. When a shortcut is used, the system automatically supplies the atoms, bonds, valencies and substitution indicated in the shortcut. Each shortcut is counted as the number of non-hydrogen atoms represented in the shortcut. The shortcut symbols are used in exactly the same fashion as element symbols.

--CO2		--nPr	
--CO1		--nBu	
--SO2--		--iPr	
--SO3		--sBu	
--PO3		--tBu	
--PO4		--iBu	
--CN		--Cn (CHA)	
--NO2		--Ph	
--Ace		--oBe--	
--Et		--mBe--	
		--pBe--	

Remarks:

- free sites on shortcuts are not allowed.

- the Cn shortcut, where n is a number, is a symbol for a chain of carbons with the number of carbon atoms indicated by the "n.". This shortcut can only be substituted on the terminal position.

For example,

C4 becomes -CH2-CH2-CH2-CH2-

None of the hydrogens on the carbons can be replaced with a substituent.

In graphic query mode, this shortcut appears in the shortcut menu as Cha (chain).

- the oBe, mBe and pBe shortcuts can only be substituted in the ortho, meta and para positions of the phenyl ring, respectively.

- for the divalent shortcuts, Cn, SO2, CO1, oBe, mBe, and pBe, the system will automatically supply a free site when these shortcuts are used as a terminal attachment in a query.

II.4. Superatoms

Markush DARC uses *superatoms* to represent chemical families which are not specifically defined. The superatoms describe the following types of chemical families:

- acyclic hydrocarbons
- cyclic systems
- metals
- others.

The symbols for the Markush DARC superatoms and the definition for each are shown in the following tables.

In the Markush DARC databases, each superatom represents a chemical family which has not been fully specified in the original patent document.
Superatoms may be used in Markush DARC queries to specify a chemical family with a single symbol.

The superatoms used in WPIM and MPHARM are discussed in more detail in the respective *Database Appendices* of this manual.

ACYCLIC SUPERATOMS		
CHK	**CHE**	**CHY**
alkyl, alkylene	alkenyl, alkenylene	alkynyl, alkynylene
CYCLIC SYSTEMS		
ARY	**CYC**	**HEA**
- in WPIM: monocyclic or fused carbocyclic system containing at least one benzene ring - in MPHARM: monocyclic or fused carbocyclic system containing an even number of C atoms and the same number of normalized bonds	cycloaliphatic monocyclic or fused, non-aromatic carbocycle optionally substituted by acyclic hydrocarbons	monocyclic heteroaryl, i.e. 5-membered heterocycle with 2 double bonds or 6-membered heterocycle with 3 double bonds alternating with 3 single bonds
HET	**HEF**	
non-aromatic monocyclic heterocycle	fused heterocycle	
ELEMENTS		
MX	**AMX**	**A35**
any metal	Alkali and Alkaline earth metals	Group III A-VA metals

TRM	LAN	ACT
Transition metals excluding Lanthanum	Lanthanides (including Lanthanum)	Actinides (including Actinium)
HAL		
halogen		

Other superatoms, which are used in the databases and are less important for searching, are listed in the following table:

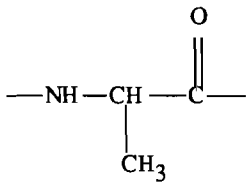
OTHER SUPERATOMS		
ACY	**DYE**	**POL**
acyl (i.e. residue left after removal of one or more OH groups from an acid)	chromophore or fluorescent group (including dye residue)	polymer, polypeptide residue
PEG	**PRT**	**XX**
polymer end group	protecting group	any atom or group excluding hydrogen
UNK		
undefined group		

In addition to these superatoms, Markush DARC provides thirty additional superatoms corresponding to the standard amino acids commonly found in Chemistry patents.

These peptide superatoms are listed in the following table:

ABU	alpha-aminobutyric acid	**LEU**	leucine
ALA	alanine	**LYS**	lysine
ARG	arginine	**MET**	methionine
ASN	asparagine	**NLE**	norleucine
ASP	aspartic acid	**NVA**	norvaline
ASU	alpha-aminosuberic acid	**ORN**	ornithine
CYS	cysteine	**PHE**	phenylalanine
GLN	glutamine	**PRO**	proline
GLP	pyroglutamic acid	**SAR**	sarcosine
GLU	glutamic acid	**SER**	serine
GLY	glycine	**STA**	statine
HCY	homocysteine	**THR**	threonine
HIS	histidine	**TRP**	tryptophane
HSE	homoserine	**TYR**	tyrosine
ILE	isoleucine	**VAL**	valine

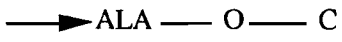
Each superatom corresponds to the amino acid substructure without its OH group i.e. to the amino acid when linked to another amino acid with an amide linkage.
Thus, free amino acids must be represented with the corresponding superatom substituted by a OH group.
For example, the ALA superatom represents the following substructure:



ALA (alanine) superatom

Free alanine as part of a peptide structure should be represented as follows:

Likewise, the methyl ester should be represented as follows:



or

N.B.: the use of either a single bond or a peptide oriented bond to link the oxygen atom to the ALA superatom depends on the database. The single bond is used in INPI databases and the oriented bond in Derwent databases.

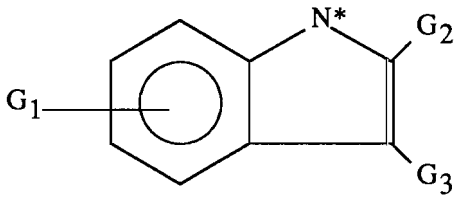
II.5. Variable groups

II.5.1. Definition

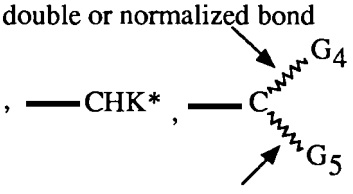
The position of a generic (variable) group can be indicated using a G group. The format for a G group is the letter "G" followed by a number. A single query can contain from one to twenty G groups (G1, G2, G3, ...G20). By definition, the main structure or substructure is called the parent group and is labelled G0 (G zero).

In a single query, a G group can be used only once; that is, G1 can only be used one time. If a query has two identical G groups, a different number must be used for the second G group.

Let us consider the following example:



Generic group G1 G1 : —H , —N , — NO_2



Generic group G2 G2 : —H , —CHK* , — C

Generic group G3 G3 : —H , — HET 2* , — CHK*

Generic group G4 G4 : —O , —N 2*

Generic group G5 G5 : — H , —N* , — CHK*

Markush DARC query

In this example, G0 is the parent group of the G1, G2 and G3 generic groups.
Group G2 is the parent group of groups G4 and G5.

The attachment points (AP) are the atoms of the generic group attached to the parent group.
In the above example, attachment bonds are shown in bold style. These link the attachment points to the parent group.

II.5.2. G group limits

Each G group can have up to 20 variations.

A G group may be nested in another G group.
For example, in the above query structure, groups G4 and G5 appear in group G2: we say that G4 and G5 are nested in G2.

Three levels of nesting are allowed in Markush DARC queries in addition to the main group G0.
In this example, group G0 is at level zero, G2 at level 1 and G4 and G5 at level 2: there are thus two levels of nesting. It would be possible to have for example a group G6 contained in G4 : in this case there would be three levels of nesting.

Connectivity is the number of attachments to a node. The connectivity is not related to the type or order of the attachment. The connectivity of a G Group may be zero to eight. A connectivity of zero for a G group means that the G group may appear as part of the query with no specified attachments to any other part of the query. A connectivity of six, for example, means that the G group is attached to six other atoms or groups in the query.

Markush structures which are stored in Markush DARC databases may contain up to 50 variable groups or occasionally 100 in some records, each containing up to 50 values. There may be up to 4 levels of nesting.

The following table gives a comparison between Markush DARC G groups in a structure query and Markush DARC G groups in database structures:

MARKUSH DARC G GROUPS		
	QUERY	DATABASE
Connectivity	zero to eight	zero to eight
Number of G groups	zero to 20	zero to 50
Number of values per G group	1 to 20	1 to 50
Nesting	3 levels	4 levels

II.6. Bonds

Bonds are used as connections between nodes.
A node can be an atom, a superatom, a shortcut or a G group.

The possible values of bonds for query formulation are:

- single
- double
- triple
- normalized bond (aromatic or tautomeric) : conventions used in Markush DARC for the normalization of aromatic and tautomeric bonds are explained in Section 4 - *Structure representation conventions* -
- Z bond to specify a list of possible values of bonds (any value mentioned above i.e. single, double, triple or normalized)
- peptide bond to represent amide linkages between two amino acids in a peptide structure: this bond is represented by an arrow and is oriented from the carbonyl group of the first amino acid to the amino group of the second amino acid. This bond is also used to link C-terminal amino acids to terminal substituents.

- undefined nature of bond

Stereochemical bonds can be searched (in query graphics mode only) and displayed in the Markush DARC databases:

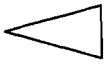
- up
- down
- racemic
- unknown.

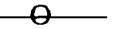
For query formulation, there are 3 types of bonds which can be specified:

- acyclic
- cyclic
- undefined type of bond: cyclic or acyclic.

Each bond value or type has its own symbol and own display in Markush DARC.

The following table shows the various bond values and types, with corresponding symbols and display.

MARKUSH DARC BONDS			
BOND VALUES	SI	SINGLE	
	DO	DOUBLE	
	TR	TRIPLE	
	NO	NORMALIZED	
	X	UNDEFINED	
	Z	LIST	
	PEP	PEPTIDE BOND	
STEREO CHEMICAL BONDS	UP	UP	
	DOWN	DOWN	
	RA	RACEMIC	
	UN	UNKNOWN	

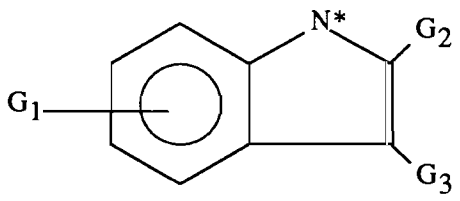
BOND TYPES	CY	CYCLIC	
	AC	ACYCLIC	
	CX	CYCLIC OR ACYCLIC	

Note on bond types:

- During query definition, the system will automatically assign the bond type to the bonds in the query, i.e. all bonds involved in a ring will be assigned the CY (cyclic) type, while all bonds not involved in a ring will be assigned the AC (acyclic) type.

- For a Z bond (list of bonds), the type applies to all alternatives in the Z list. All bonds in a Z list must have the same type.

If we consider the same query example as previously:



Generic group G1 G1 : —H , —N , — NO_2

Generic group G2 G2 : —H , —CHK* , — C (double or normalized bond; single or normalized bond; G_4, G_5)

Generic group G3 G3 : —H , — HET 2* , — CHK*

Generic group G4 G4 : —O , —N 2*

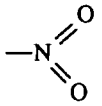
Generic group G5 G5 : —H , —N* , — CHK*

Markush DARC query

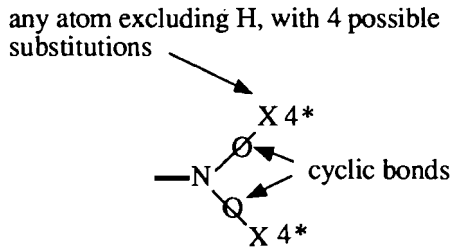
- The two Z bonds we have specified for the last value of G2 are acyclic according to the drawing and because we did not define a specific type. The system will assume these to be acyclic.

- In G4 we have specified a nitrogen atom with two possible substitutions (2*).

We will be able to retrieve primary or secondary amino groups, cyclic amino groups and e.g nitro groups.

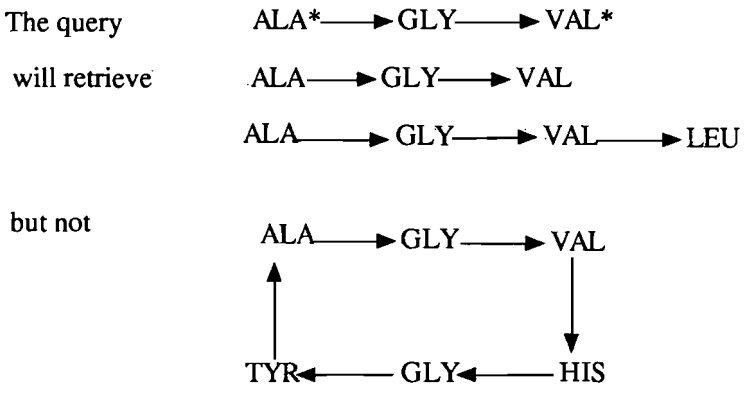


If we wish to retrieve only compounds having a cyclic amino group, we could draw the following query, instead of the nitrogen atom with two substitutions:



Cyclic amino group value

- For peptide searching, which will be explained in more detail in the *Specialised search capabilities* chapter later in this section, it is important to be careful when defining the bond type.
The following example illustrates the need for specifying the bond type in certain situations:



Peptide query

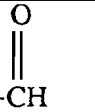
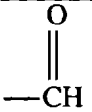
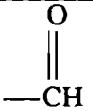
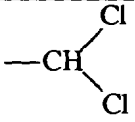
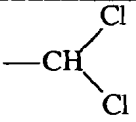
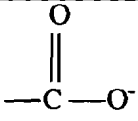
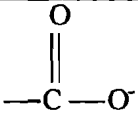
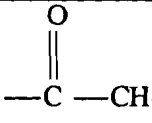
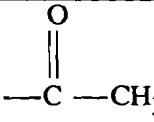
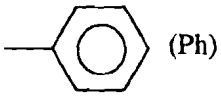
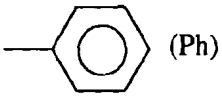
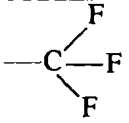
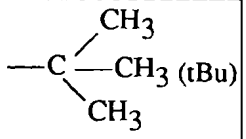
The query

will retrieve all these because the bond type is cyclic or acyclic.

II.7. Free sites

Free sites are used to specify the amount or degree of substitution which is acceptable on a given position.
If no free sites are specified, the system will complete the normal valency of the atom with hydrogens when executing the search.
The number of free sites specified indicates the maximum number of attachments which are acceptable at that position. Since free sites are a maximum, the retrieval includes all lower degrees of substitution. Free sites indicate the number of attachments; the type of bond is not considered. An atom attached by a double bond, for example, is considered a single attachment.

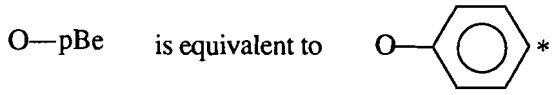
The following table shows what would be retrieved using various free site specifications on a carbon atom.

NO FREE SITE	1 FREE SITE	2 FREE SITES	3 FREE SITES
—CH_3	—CH_3	—CH_3	—CH_3
—CH_2^-	—CH_2^-	—CH_2^-	—CH_2^-
—$CH_2\bullet$	—$CH_2\bullet$	—$CH_2\bullet$	—$CH_2\bullet$
	—CH_2—Cl	—CH_2—Cl	—CH_2—Cl
			
	—C≡N	—C≡N	—C≡N
			
			
			
			
			
			

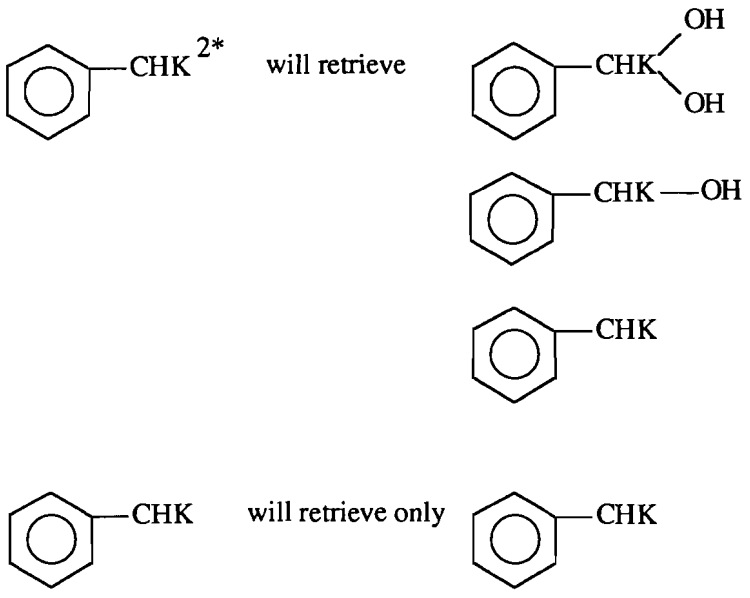
Free sites may be specified using the Free Site (FS) command or by designating the desired free sites as an attribute.

Free sites on divalent shortcuts: it is possible to use a divalent shortcut (CO1, SO2, Cn, oBe, mBe or pBe) with only one attachment, which is equivalent to a free site on the second point of attachment of the shortcut.

Example:



Markush DARC superatoms may also be substituted. Free sites can be used on a superatom to indicate the degree of substitution which is acceptable.
The free sites on superatoms work in the same way as free sites on atoms.
As shown below, a search for CHK^{2*} will retrieve both substituted and unsubstituted CHK superatoms. A search for a superatom without free sites will retrieve only the unsubstituted superatom.



Superatom free sites

II.8. Attributes

Attributes are labels used to provide additional specificity on atoms and superatoms . Searching without an attribute will retrieve answers irrespective of whether the attribute present in the database. If an attribute is specified in the query only answers containing tl attribute will be found.

II.8.1. Atom attributes

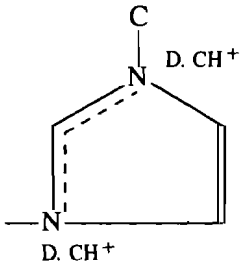
The following atom attributes are available in Markush DARC:

- charges

  Example: Cl^-

- delocalized charges

  Example:



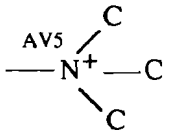
Example of delocalized charges

- abnormal mass (AM and the abnormal mass value)

  Example: Tc AM99

- abnormal valency (AV and the abnormal valency value)

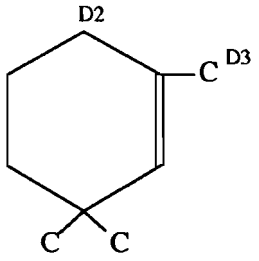
  Example:



The nitrogen atom in this example has an abnormal valency of 5 : 4 attachments and a positive charge, the normal valency for nitrogen being 3.

- deuterium (D, followed by the number of hydrogen atoms to be replaced by deuterium)

Example:



Example of deuterium attributes

- tritium (T, followed by the number of hydrogen atoms to be replaced by tritium)

II.8.2. Superatom attributes

The following atom attributes are available in Markush DARC:

- chain attributes, used to specify carbon chain length (low, middle or high) and branching (straight or branched)

- ring attributes, used to distinguish monocyclic and fused rings and to specify the degree of saturation (saturated or unsaturated)

- multiplier attributes, used in the databases on superatoms to specify the maximum degree of substitution which is possible for a given generic term. These attributes are not used in Markush DARC queries. This is additional information, which it is possible to display with the structure.

The chain and ring attributes for chain and ring superatoms are listed in the following table:

MARKUSH DARC CHAIN/RING (CR) SUPERATOM ATTRIBUTES			
SUPERATOMS	**CHAIN 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)
SUPERATOMS	**RING ATTRIBUTES**		
ARY (aryl)	MON (monocyclic)	FU (fused)	
CYC (cycloalkyl)	MON (monocyclic)	FU (fused)	
	SAT (saturated)	UNS (unsaturated)	
HEF (fused heterocycle)	SAT (saturated)	UNS (unsaturated)	
HET (non-aromatic monocyclic heterocycle)	SAT (saturated)	UNS (unsaturated)	
HEA (heteroaryl)	no attributes on HEA		

LO applied on a carbon chain means 1 to 6 carbon atoms in the chain.
MID means 7 to 10 carbon atoms.
HI means over 10 carbon atoms.

II.8.3. Peptide attributes

There are two types of peptide attributes:

- configuration attributes used on peptide superatoms to specify the configuration of an amino acid at its alpha-carbon: D, L or DL. It is possible to specify a single attribute or a list of possible alternatives

  Example: ALA^{D}
  $ALA^{D,L}$

- position attributes used on atoms, superatoms, groups and shortcuts involved in substitution on a peptide superatom to specify the positions of substituent on the amino acid: the list of amino acids with their positions is given in *Appendix 1* at the end of this manual. This attribute is numerical.

II.8.4. Attributes used on atoms and superatoms

II.8.4.1. Polymer attributes

The polymer attributes (PA) are used to specify the roles of atoms and superatoms in a polymer.

POLYMER ATTRIBUTE	SYMBOL
monomer/condensate	MC
cross-linker	XL
end group	EG
modifier/derivative	DE
grafting monomer	GM

II.8.4.2. Free sites

Free sites may also be considered as attributes. They may occur on atoms and superatoms.

II.8.4.3. Numbering attribute

Although the numbering attribute (NU) is used in Markush DARC databases, it cannot be used in Markush DARC queries for the moment.

This attribute is a numerical label that is attached to atoms or superatoms.
It is used in Markush DARC databases to make the link between labelled atoms or superatoms and additional textual information, in the form of text notes with corresponding numbers.
It is also used to indicate the atoms which form part of repeating units, to define the number of times such a group is repeated.

II.8.5. Summary table of Markush DARC searching attributes

The various attributes which can be used in Markush DARC for searching are summarised below:

MARKUSH DARC ATTRIBUTES		
ATOM ATTRIBUTES	charges	CH
	delocalized charges	CH (text input) DCH (graphics input)
	abnormal mass	AM
	abnormal valency	AV
	deuterium	D
	tritium	T
	free sites	FS
SUPERATOM ATTRIBUTES	chain/ring attributes	CR
	free sites	FS
	multiplier attribute (non-searchable, only displayable)	MU
PEPTIDE ATTRIBUTES	D, L or DL amino acid configuration (on peptide superatom)	D, L, DL
	position of substitution on the amino acid	SP

POLYMER ATTRIBUTES (on atoms or superatoms)	roles of atoms or superatoms in a polymer (on atoms, shortcuts or superatoms)	PA

II.8.6. Translation attributes

II.8.6.1. Types of translation attributes

Translation attributes are used in Markush DARC queries to specify the type of matching which is required between the structure query and possible answers in the database.

The different types of translation are:

- broad translation (BT) to translate specific atoms of the query to corresponding superatoms in the database. This type of translation can also be used to widen the meaning of a superatom. It allows the user to specify a superatom in the query and find this superatom or the XX superatom in the database.

- narrow translation (NT) to translate superatoms of the query to corresponding specific atoms or groups of atoms in the database.

- equal translation (EQ) to match specific to specific terms or generic to generic terms. Markush DARC automatically applies this type of translation by default.

- any translation (ANY) to translate a generic or specific term to any term (superatom - including the XX superatom - and specific atoms).

The following tables summarise the possible situations, depending on the type of query and the type of translation used:

Query (without free sites)	Type of translation	Answers
specific atom or group of atoms	EQUAL	specific atom or group of atoms
superatom	EQUAL	superatom
specific atom or group of atoms	BROAD	specific atom or group of atoms corresponding superatom other superatoms in the hierarchy XX superatom
superatom	BROAD	superatom other superatoms above it in the hierarchy XX superatom
superatom	NARROW	corresponding specific atom or group of atoms superatom other superatoms below it in the hierarchy
specific atom or group of atoms	ANY (equivalent to BROAD)	specific atom or group of atoms corresponding superatom other superatoms in hierarchy XX superatom

superatom	ANY (equivalent to BROAD + NARROW)	superatom other superatoms in the hierarchy specific atom or group of atoms corresponding to all superatoms in the relevant hierarchy XX superatom

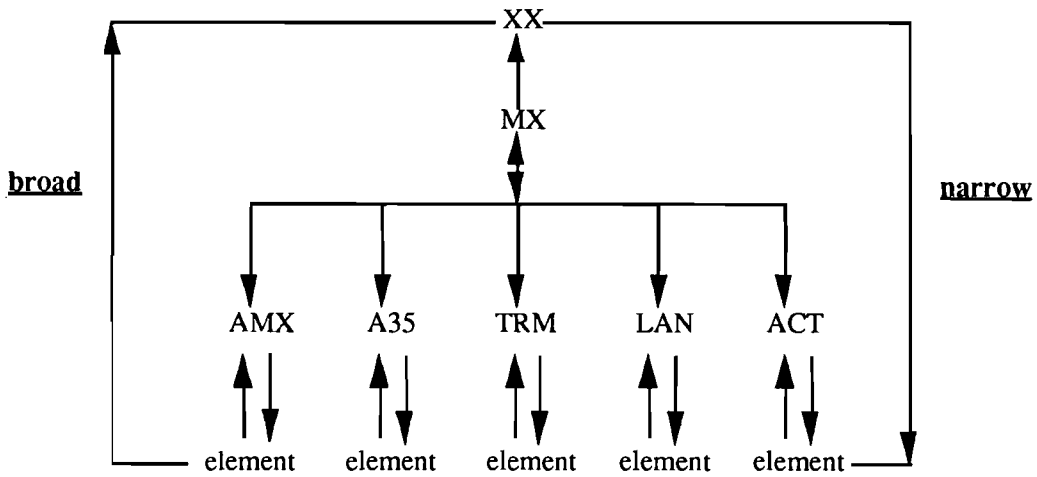
II.8.6.2. Hierarchy of superatoms within translation

Each class of superatoms is hierarchically organised in order to take into account the possible presence of free sites which may widen the meaning of a superatom if a broad translation attribute is used on the superatom.
For example, if we consider the class of ring superatoms. The presence of two free sites on the HEA (monocyclic heteroaromatic rings) superatom with BT (broad translation) will retrieve both the HEF (fused heterocyclic rings) and HEA (monocyclic heteroaryl) superatoms.

Likewise, the presence of free sites on a specific monocyclic heteroaromatic ring in addition to BT (broad translation) on every ring atom, e.g. a furan ring having one free site on the 2 and 3 positions, retrieves both HEA and HEF superatoms.

The hierarchy for each class of superatom is given on the following pages.

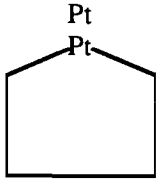
- Metals:
with 0 or 1 free site on metal or superatom:



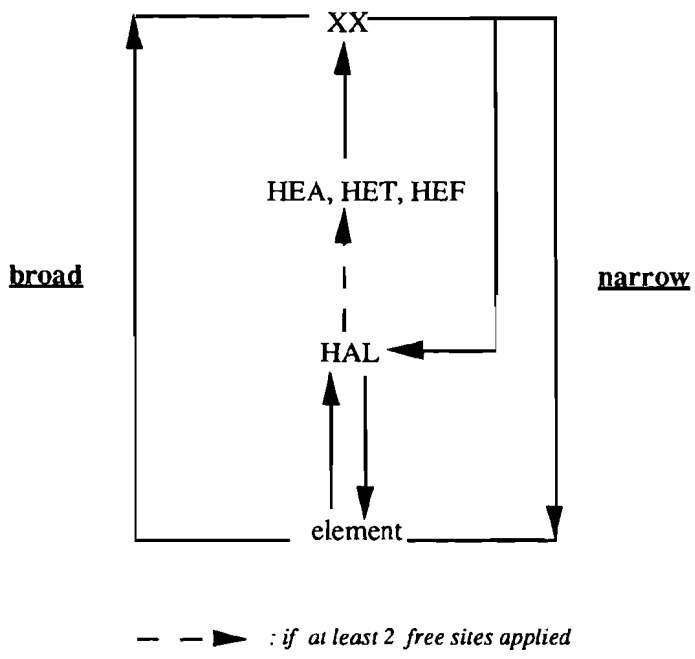
Hierarchy of metals and superatoms within translation

Note: In the present version of Markush DARC, the use of at least 2 free sites in conjunction with a broad translation attribute on a metal atom will retrieve the heterocycle superatoms HET, HEA and HEF in addition to the MX and XX superatoms.

Examples:

Query	Retrieval
Pt^{BT} (Pt with Broad Translation)	Pt TRM MX XX
$Pt^{BT\ 2*}$ (Pt with Broad Translation and 2 free sites)	Pt  TRM MX HET, HEA, HEF XX
TRM^{BT}	TRM MX XX
$TRM^{BT\ 2*}$	TRM MX HEA, HET, HEF XX
AMX^{NT}	AMX Li, Na, K, Rb, Cs, Be, Mg, Ca, Sr, Ba

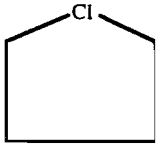
- Halogens:



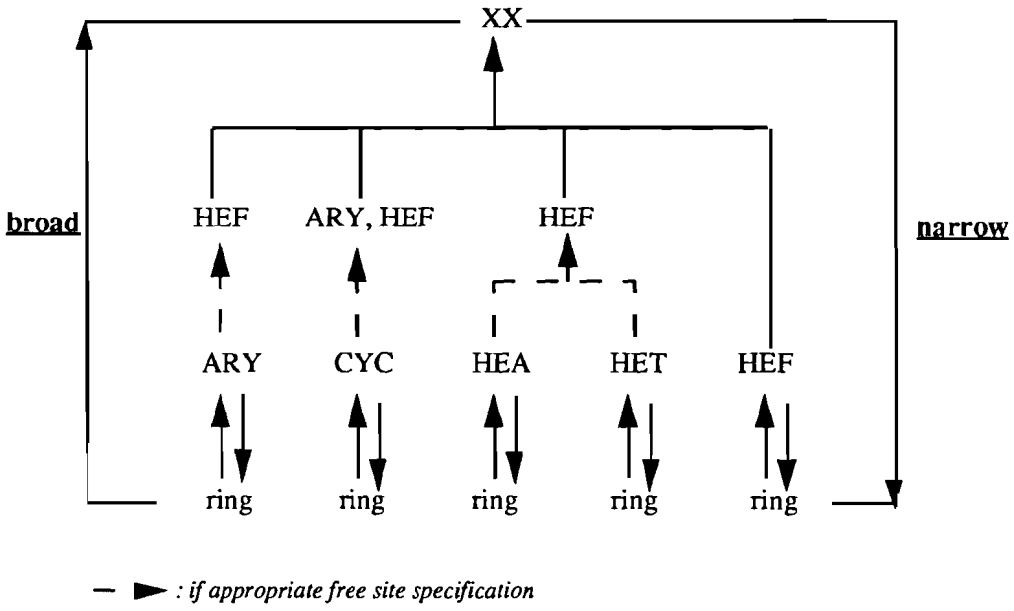
Hierarchy of halogens and superatoms within translation

If at least 2 free sites are applied to an halogen atom of a query in addition to Broad Translation, Markush DARC will retrieve the heterocycle superatoms HET, HEA and HEF in addition to the HAL (halogen) superatom.

Examples:

Query	Retrieval
HALBT	HAL, XX
HALBT 2* (HAL superatom with Broad Translation and 2 free sites)	HAL HEA, HET, HEF XX
ClBT 2* (Cl with Broad Translation and 2 free sites)	Cl C—— Cl  HAL C—— HAL HEA, HET, HEF XX

- Rings:

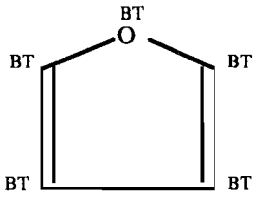
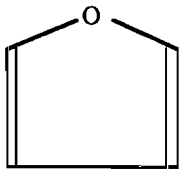
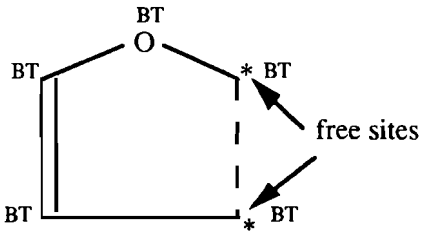
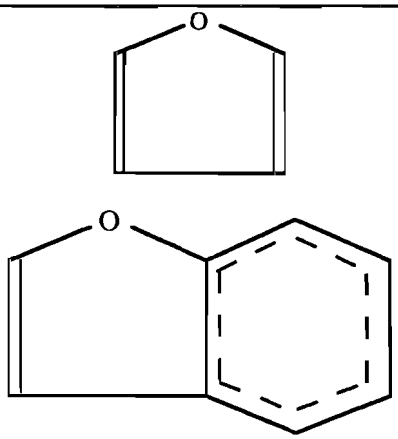


Hierarchy in ring superatoms within the translation process

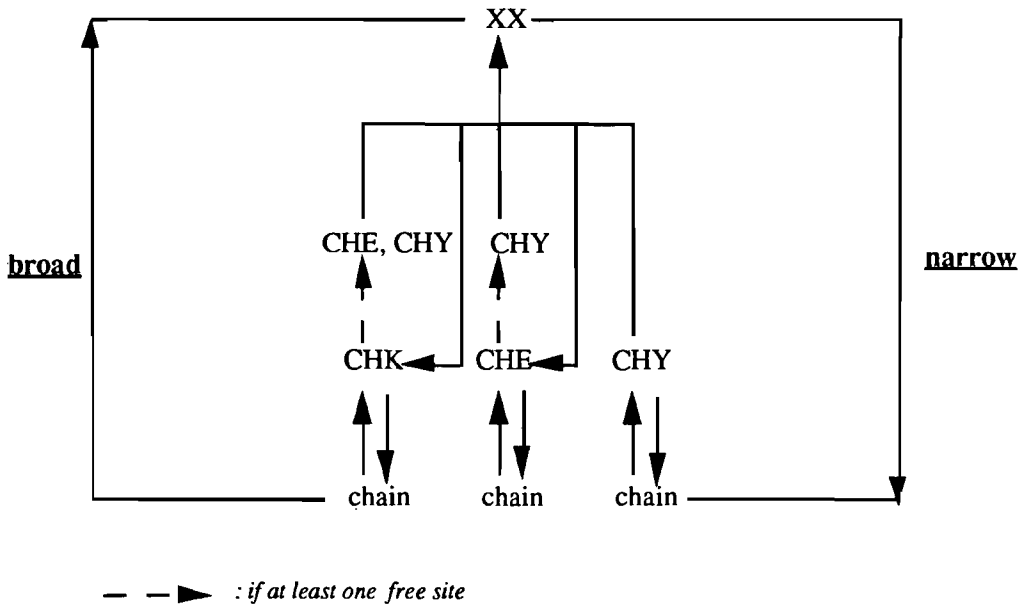
If one free site is applied on two contiguous ring atoms of a query, in addition to Broad Translation on each ring atom, Markush DARC will retrieve the HEF superatom (fused heterocycle) in addition to the monocyclic heterocycle superatoms HEA and HET.

If at least 2 free sites are applied on an ARY (aryl) superatom of a query in addition to Broad Translation, the HEF superatom (fused heterocycle) will be retrieved.

Examples:

Query	Retrieval
ARY$_{BT}$	ARY, XX
ARY$_{BT\ 2*}$	ARY, XX HEF
HEA$_{BT}$	HEA, XX
HEA$_{BT\ 2*}$	HEA, XX HEF
 (Broad Translation on each ring atom)	 HEA XX
	 HEA HEF XX

Carbon chains:

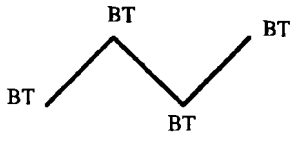
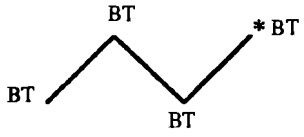
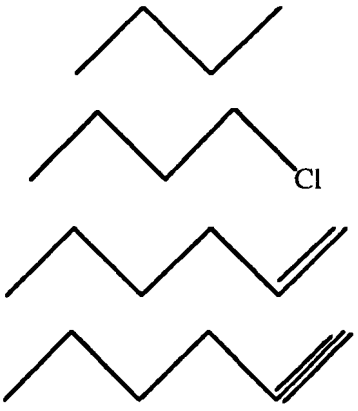


Hierarchy in carbon chains and superatoms within translation

If one free site is applied at the end of a carbon chain of a query in addition to Broad Translation on each carbon atom of the chain, Markush DARC will retrieve the CHK (alkyl) superatom and the CHE (alkenyl) and CHY (alkynyl) superatoms.

If one free site is applied to a CHK superatom in addition to Broad Translation, the CHE and CHY superatoms will be retrieved.

Examples:

Query	Retrieval
CHK_{BT}	CHK XX
CHK_{BT} *	CHK CHE CHY XX
	 CHK XX
	 CHK CHE CHY XX

II.8.6.3. Use of translation attributes

Translation attributes may be used on:

- an atom: example Cl^{BT} (Cl atom with Broad Translation)

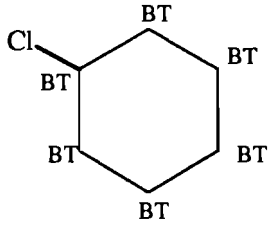
- a superatom: example CHK^{NT} (CHK alkyl superatom with Narrow Translation)

- a shortcut (with BT attribute): example Ph^{BT} (phenyl shortcut with broad translation).

Translation attributes cannot be assigned to a G group. It is necessary in this case to assign the attributes to the various nodes or atoms within the group.

Translation attributes cannot be assigned directly to a ring. It is necessary to assign the attributes to all of the ring atoms.

Example: broad translation required for a cyclohexane ring

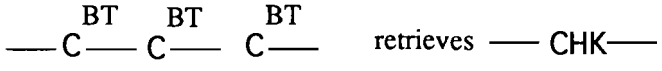


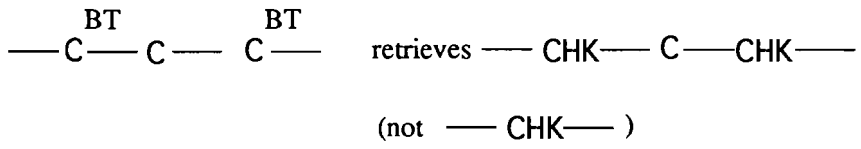
Broad Translation on a cyclohexane ring

If you only put BT on 1 or 2 carbon atoms of the cyclohexane ring, you will not retrieve the CYC (cycloalkyl superatom) linked to a Cl atom but only chlorocyclohexane.

For carbon chains, it is necessary to assign translation attributes on all the atoms of the chain in order to translate the complete chain.

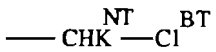
However, if the translation attribute is assigned to only a part of the chain, this chain will be partly translated. The atoms which are not assigned translation attributes will not be translated.





For groups of atoms or superatoms, it is possible to assign a given translation attribute only to a specific atom or superatom of the moiety, or to assign different translation attributes to different atoms of the moiety.

Example:

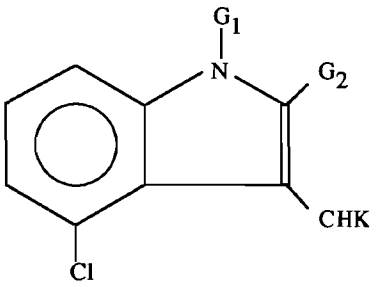
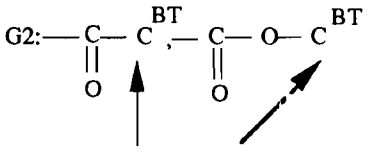
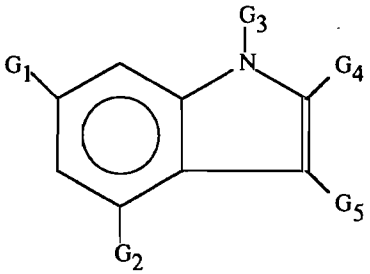


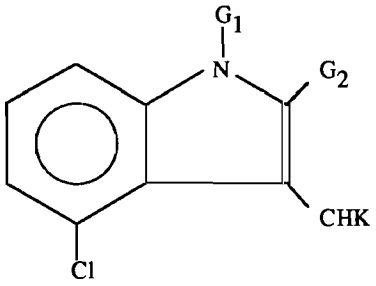
Different translation attributes

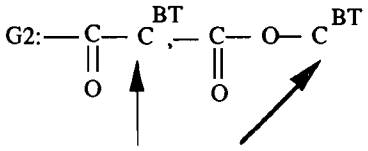
II.8.6.4. Examples of use of translation attributes

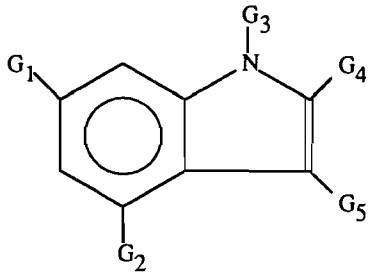
Simple examples are given on the following pages for cases where broad, narrow or any translation attributes are used.

- Broad translation

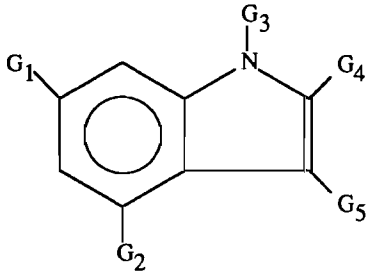
QUERY	MATCH	DATABASE
 G1: H, CHK (alkyl)  BROAD TRANSLATION on the methyl groups	YES	 G1: [H] Cl. N NO2 O G2: H [Cl] N NO2. O G3: [H] C Et G4: H CHK [$-CO_1-CHK$] (alkyl) (alkanoyl) [$-CO_1-O-CHK$] (alkoxycarbonyl) G5: H [CHK] –O–CHK The group values in boxes match the query. In G4, the alkanoyl value matches the acetyl value contained in G2 of the query because the BT attribute translates the methyl group into the CHK superatom. The same mechanism applies for the alkoxycarbonyl value.

<table>
<tr>
<td>


G1: H, CHK (alkyl)


BROAD TRANSLATION
on the methyl groups
</td>
<td>YES</td>
<td>


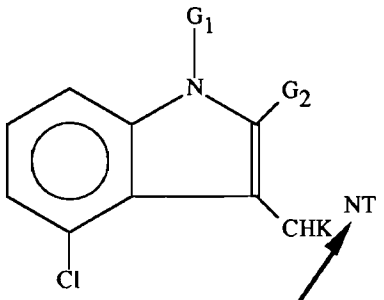
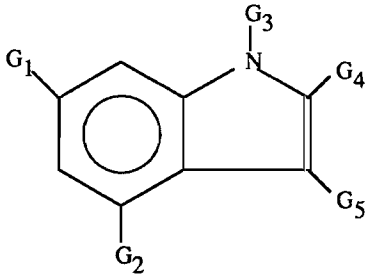
G1: [H] Cl. N NO2 O
G2: H [Cl] N NO2. O
G3: [H] C Et
G4: H [$-CO_1-C$]
G5: H [CHK] $-O-$CHK

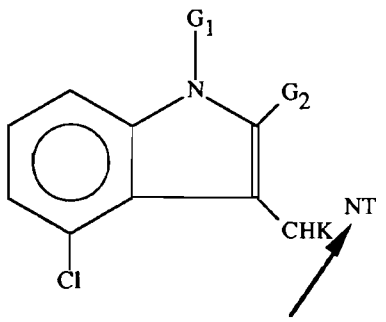
In G4 the acetyl value matches exactly the acetyl value contained in G2 of the query.
</td>
</tr>
<tr>
<td></td>
<td>NO</td>
<td>


G1: H Cl N NO2 O
G2: H Cl N NO2. O
G3: H C Et
G4: H $-CO_1-$Et
G5: H CHK $-O-$CHK

This Markush structure is not an answer to the query because in G4 the ethylcarbonyl value does not match the acetyl value contained in G2 of the query.
</td>
</tr>
</table>

- Narrow translation

QUERY	MATCH	DATABASE
 G1: H, CHK G2: H, C	**YES**	 G1: [H] Cl N NO2 O G2: H [Cl] N NO2 O G3: [H] C Et G4: [H] CHK – CO_r– CHK , (alkyl) (alkanoyl) – CO_r O– CHK (alkoxycarbonyl) G5: H [C] [Et] [iPr] In G5, the methyl, ethyl and isopropyl values match the query because the Narrow Translation attribute applied on the CHK superatom in the main group of the query translates CHK into specific carbon chains.

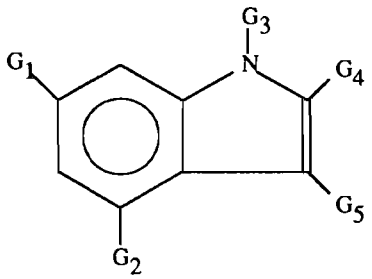


NARROW TRANSLATION
on the CHK superatom

G1: H, CHK

G2: H, C

YES



G1: [H] Cl N NO2 O

G2: H [Cl] N NO2 O

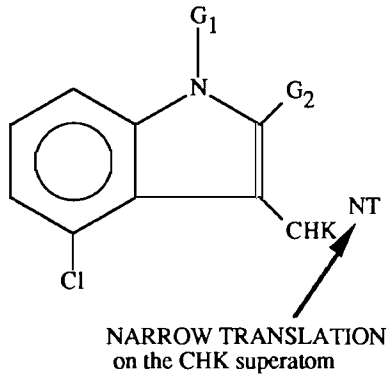
G3: [H] C Et

G4: [H] CHK – CO_{Γ} – CHK ,
(alkyl) (alkanoyl)

– CO_{Γ} O– CHK
(alkoxycarbonyl)

G5: H [CHK]

In G5, the CHK superatom exactly matches the CHK superatom contained in the main group of the query.

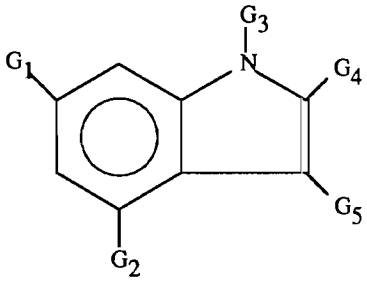


NARROW TRANSLATION
on the CHK superatom

G1: H, CHK

G2: H, C

YES



G1: [H] Cl N NO2 O

G2: H [Cl] N NO2 O

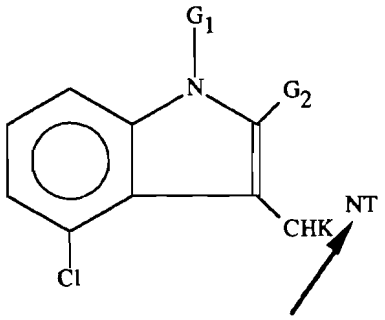
G3: [H] C Et

G4: [H] CHK– CO_{Γ}– CHK,
(alkyl) (alkanoyl)

– CO_{Γ} O– CHK
(alkoxycarbonyl)

G5: H [CHK]

In G5, the CHK superatom exactly matches the CHK superatom contained in the main group of the query.

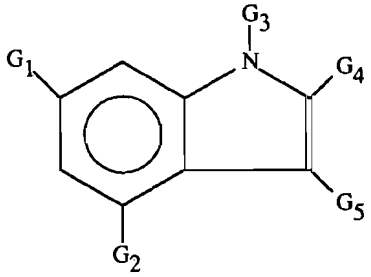


NARROW TRANSLATION
on the CHK superatom

G1: H, CHK

G2: H, C

NO



G1: H Cl N NO2 O

G2: H Cl N NO2 O

G3: H C Et

G4: H CHK — CO_{Γ} — CHK ,
(alkyl) (alkanoyl)

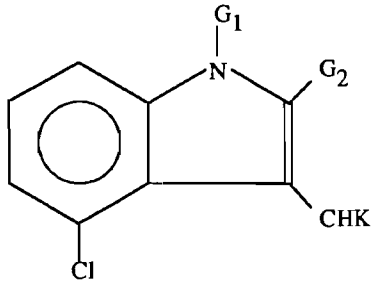
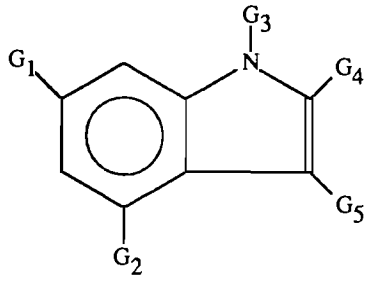
— CO_{Γ} O — CHK

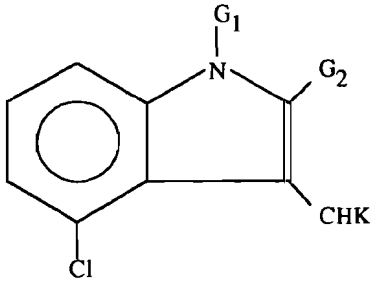
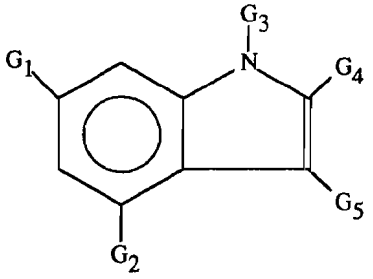
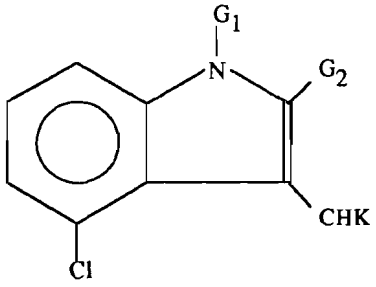
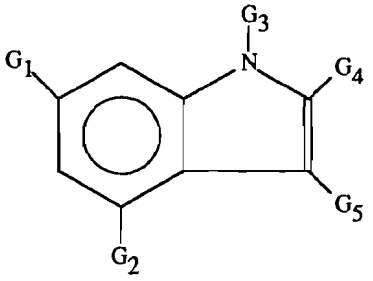
(alkoxycarbonyl)

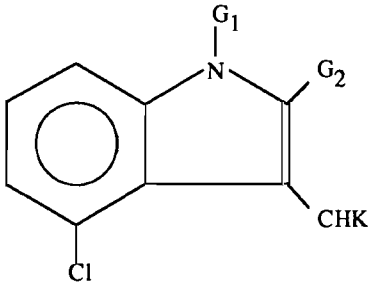
G5: H, XX (any substituent)

This Markush structure is not an answer to the query because G5 does not contain any value which matches the CHK superatom contained in the main group of the query.

- Any translation

QUERY	MATCH	DATABASE
 G1: H, CHK G2: CHK, HAL ↑ ANY TRANSLATION	YES	 G1: [H] Cl N NO2 O G2: H [Cl] N NO2 O G3: [H] C Et G4: H [CHK] — CO_Γ— CHK , (alkyl) (alkanoyl) — CO_Γ O — CHK (alkoxycarbonyl) G5: H [CHK] In G4, the CHK value exactly matches the CHK contained in G2 of the query.

 G1: H, CHK G2: CHK, HAL ↑ ANY TRANSLATION	**YES**	 G1: [H] Cl N NO2 O G2: H [Cl] N NO2 O G3: [H] C Et G4: H [C] G5: H [CHK] In G4, the methyl value matches the CHK contained in G2 of the query because narrow translation is applied.
 G1: H, CHK G2: CHK, HAL ↑ ANY TRANSLATION	**YES**	 G1: [H] Cl N NO2 O G2: H [Cl] N NO2 O G3: [H] C Et G4: H [XX] G5: H [CHK] In G4, the XX superatom (any substituent) matches the CHK contained in G2 of the query because broad translation is applied.

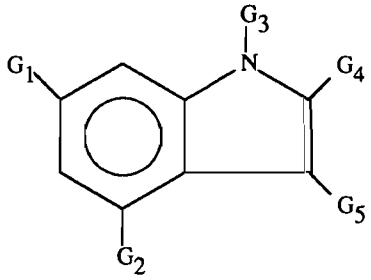


G1: H, CHK

G2: CHK, HAL

ANY TRANSLATION

NO



G1: H Cl N NO2 O

G2: as G1

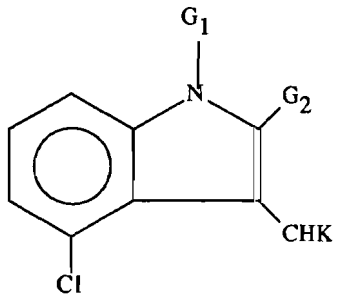
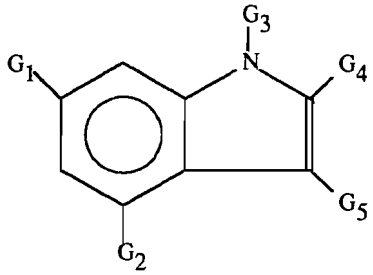
G3: H, C, Et

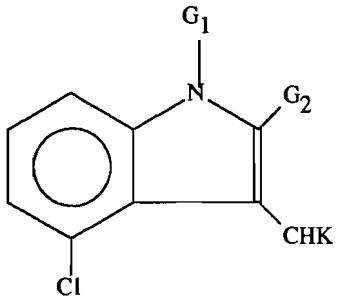
G4: CHE (alkenyl) CHY (alkynyl)

G5: H, CHK

This Markush structure is not an answer to the query because G4 does not contain any value which matches a value contained in G2 of the query.
It would be an answer if a free site was applied on CHK in G2 of the query.

- different translation attributes on a moiety

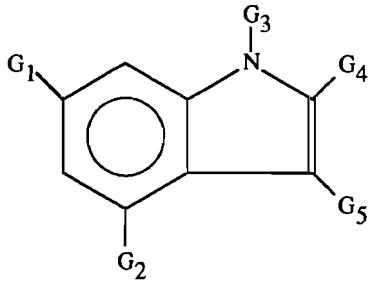
QUERY	MATCH	DATABASE
 G1: H, —CHK(NT)—Cl(BT) G2: H, C	YES	 G1 [H] Cl N NO2 O G2: H [Cl] N NO2 O G3: [—C—Cl] [—C(C)—Cl], [—CHK—HAL] G4: [H] CHK G5: H [CHK] In G3, all values are matching the --CHK--Cl value contained in G1 of the query. For the first value, the methylene group matches the CHK superatom because Narrow Translation is applied. For the second value, the ethylene group matches the CHK superatom because Narrow Translation is applied. For the third value, the CHK superatom exactly matches the CHK superatom of the --CHK--Cl value.



G1: H, —CHK(NT)—Cl(BT)

G2: H, C

NO



G1: H, Cl, N, NO2, O

G2: as G1

G3: -C—I , —C(C)—Br , —CHK—F

G4: H, CHK

G5: H, CHK

This Markush structure is not an answer to the query because G3 does not contain any value matching the --CHK--Cl value contained in G1 of the query.

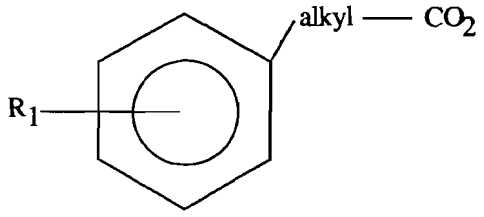
II.8.6.5. Specialised use of translation attributes

II.8.6.5.1. Chain superatom with Narrow Translation connected to a specific carbon atom

Markush DARC offers the possibility of specifying a NT (Narrow Translation) attribute on a chain superatom i.e. CHK (alkyl), CHE (alkenyl) or CHY (alkynyl), which is connected to a specific carbon atom e.g. a carboxyl group (CO2) or a cyano group (CN).

This is illustrated by the following example:

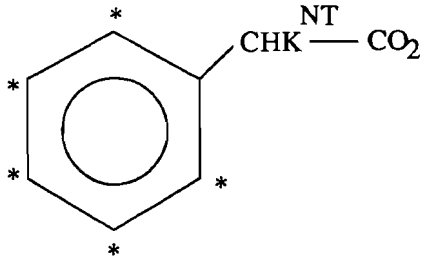
Search for arylalkanoic acids of the following formula:



R1 : any substituent, including hydrogen

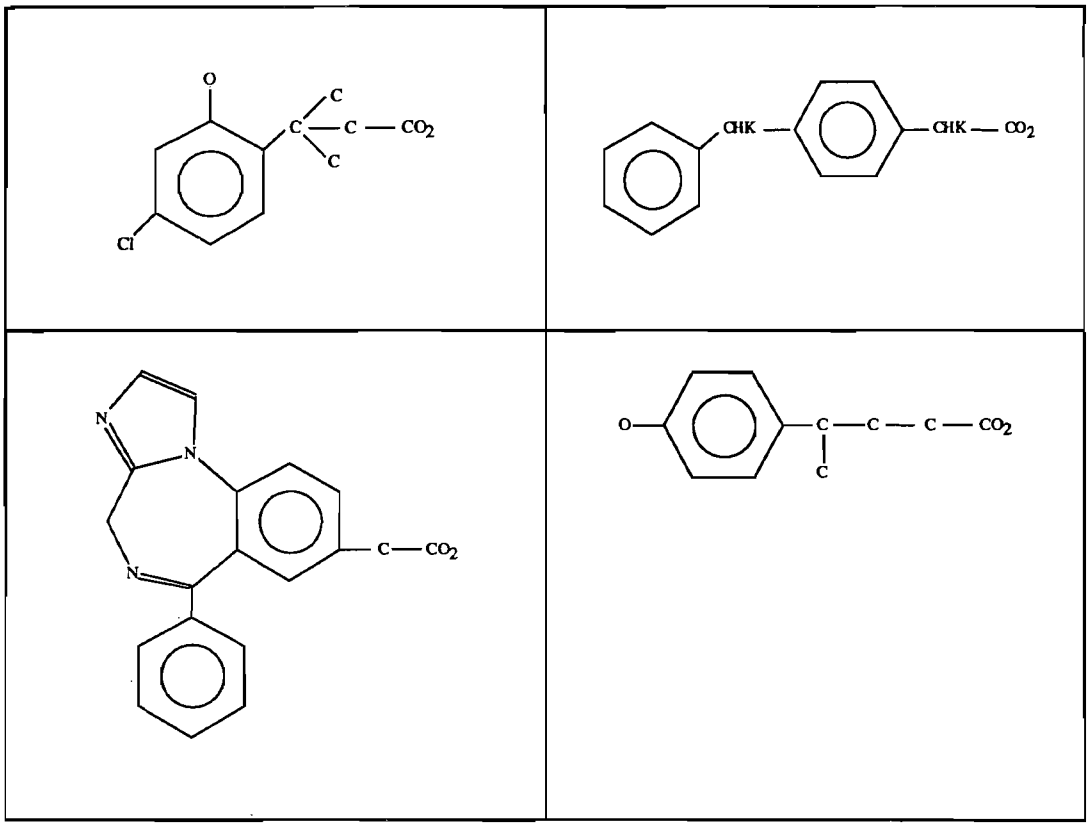
Search problem

This query can be formulated as follows:



Arylalkanoic query

Possible answers are shown below:



II.8.6.5.2. Chain superatom with NT and CR attributes

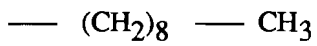
A more precise search can be performed by combining a NT attribute with a Chain and Ring (CR) attribute on a chain superatom.

Let us consider the following query example:

If we consider the chain attribute LO, structures which have this attribute value alone or in combination with other chain attribute values will be retrieved.
However, structures having combinations of CR attributes without LO will not be retrieved.
This is shown in the following table:

retrieved	not retrieved
$CHK^{LO,MID,HI}$	$CHK^{MID,HI}$
$CHK^{LO,MID}$	CHK^{MID}
CHK^{LO}	CHK^{HI}

Specifying a "NT" translation attribute in addition to the "LO" attribute will restrict the answers to those structures which have a specific carbon chain containing from 1 to 6 carbon atoms. The following structure would not be retrieved:

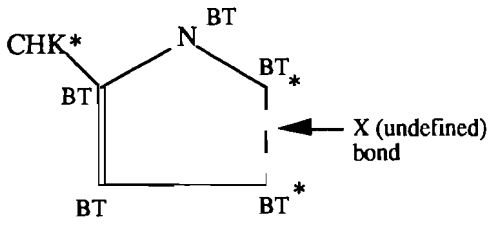
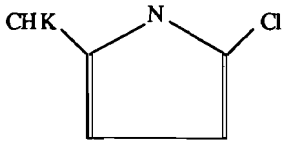
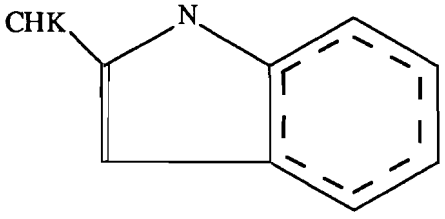


II.8.6.5.3. Free sites and Broad Translation on a specific ring

If free sites are specified on two contiguous ring atoms of a ring carrying BT (Broad Translation) on all of the ring atoms, the system will retrieve both monocyclic and fused corresponding superatoms.

If only one free site on only one ring superatom is specified, the system will retrieve only the corresponding monocyclic superatom.

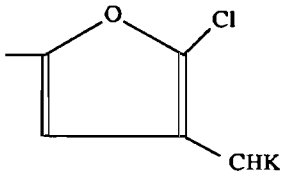
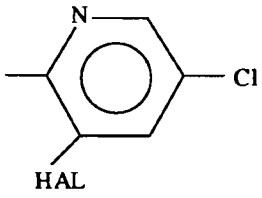
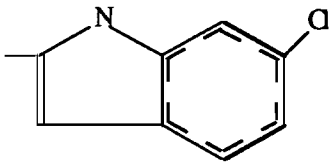
This is illustrated by the following example:

Query	Possible answers
	
	
	O — CHK — HEA — Cl
	O — CHK — HEF — Cl

II.8.6.5.4. Free sites and Narrow Translation on a ring superatom

If free sites are specified on a monocyclic ring superatom in addition to NT (Narrow translation), the system will retrieve corresponding monocyclic rings but will not retrieve fused rings.

This is illustrated by the following example:

Query	Match	Database
2* NT —— HEA —— Cl	YES	
	YES	
	NO	

II.8.6.5.5. Free sites and ANY translation on a specific ring

If free sites are specified on two contiguous ring atoms of a monocyclic ring carrying ANY translation on all the ring atoms, the system will retrieve the corresponding monocyclic and fused superatoms and also the specific monocyclic and fused rings corresponding to those superatoms.

II.8.6.5.6. Free sites and ANY translation on a ring superatom

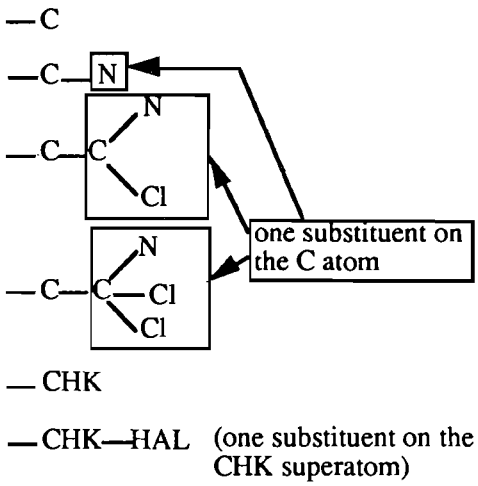
If two or more free sites are specified on a monocyclic ring superatom in addition to ANY translation, the system will retrieve the corresponding monocyclic and fused superatoms and also the specific monocyclic and fused rings corresponding to those superatoms.

II.8.6.5.7. Free sites and Narrow Translation on carbon chain or ring superatoms

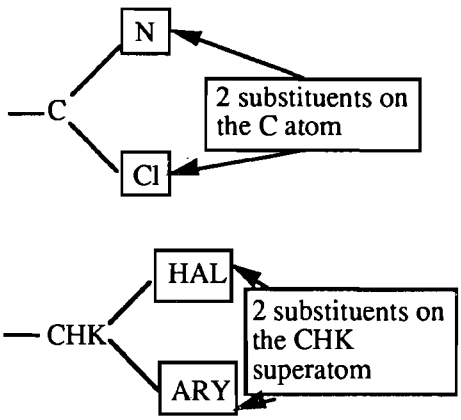
If one free site is applied on a carbon chain superatom (CHK, CHE or CHY) in addition to the NT attribute, specific carbon chains with more than one substituent may be retrieved. This is illustrated by the following example:

The following query:

could retrieve:



but would not retrieve:

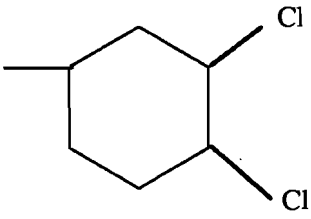


If two free sites are applied to a carbon ring superatom (CYC or ARY) in addition to the NT attribute, specific fused carbon rings with more than two substituents may be retrieved. This is illustrated by the following example:

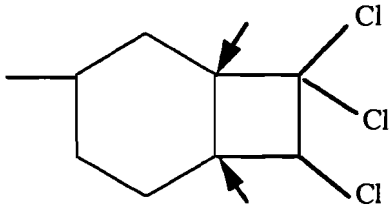
The following query:

—CYC $^{2*\ NT}$

could retrieve:

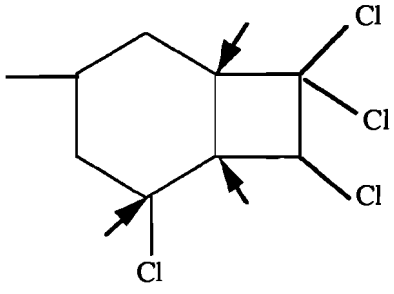


(2 substituents on the cyclohexane ring)

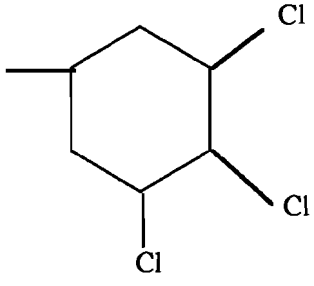


: 2 substituents on the cyclohexane ring

but would not retrieve:



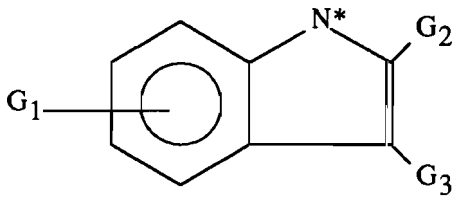
: 3 substituents on the cyclohexane ring



(3 substituents on the cyclohexane ring)

II.9. Attachment points

Attachment points (AP) are the atoms of any generic group attached to the nodes in the parent group.
In the following example, attachment points for each group value are the atoms of the group value carrying the attachment bond.



Generic group G1 G1 : —H , —N , — NO_2

double or normalized bond

Generic group G2 G2 : —H , —CHK* , — C(—G_4)(—G_5)

single or normalized bond

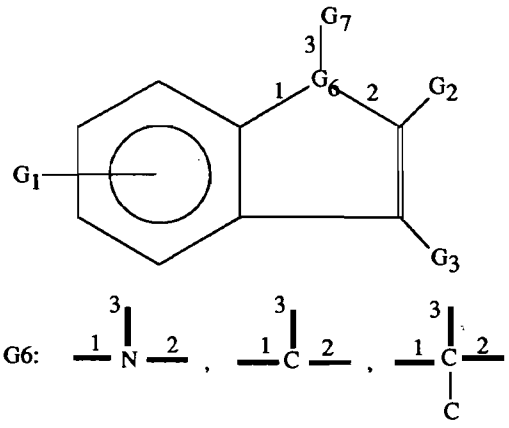
Generic group G3 G3 : —H , — HET 2* , — CHK*

Generic group G4 G4 : — O , — N 2*

Generic group G5 G5 : — H , — N* , — CHK*

Markush DARC query

A G group may have up to eight attachment points. In the above example, if N* is replaced by a group G6 linked to a group G7, the values of G6 have three points of attachment:



G6:

3 points of attachments

G7: H, C, Et

Example of a query with several attachments in a G group

II.10. Variable attachments

II.10.1. Definitions and limits

The variable point of attachment (VP) capability may be used in Markush DARC queries where a substituent may be attached to one or several sites.

This capability allows faster input of the query and more efficient searching, because it alleviates the need to create multiple structure queries.

With this capability, it is possible to select multiple sites of attachments for:

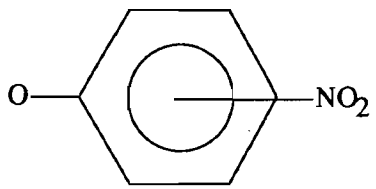
- an atom (e.g. Cl)
- a group of atoms (e.g. NO2)
- a superatom (e.g. CHK)
- a G group

on:

- a single ring (e.g. a benzene ring)
- a ring system (e.g. naphthalene)
- a chain (e.g. butadiene)

Variable points of attachment can be defined in the main group G0 or in a G group.
Up to ten atoms, superatoms, shortcuts or G groups may be involved in variable attachments in a Markush DARC query, with a total of fifty sites of attachment in the whole query structure.

Below is an example of a Markush DARC query with a variable attachment:



Markush DARC query with variable attachment

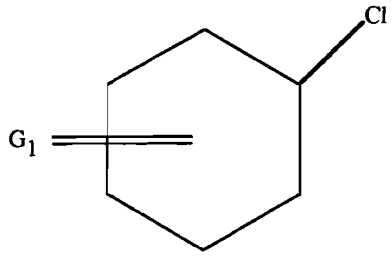
II.10.2. Types of variable attachments

II.10.2.1. One ring, single bond

The classic type concerns variable attachments of a single group on a single ring through a single bond: this is the case in the above example.

II.10.2.2. One ring, bond other than single

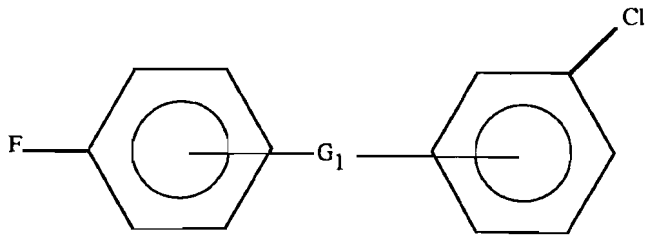
The variable attachment may be through a bond other than single e.g. a double bond:



G1: $=O$, $=S$

Variable attachment through a double bond

II.10.2.3. Two rings



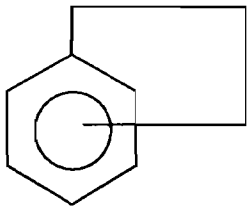
G1: O, S, $-N-CO_1-N-$

Variable attachments on two rings

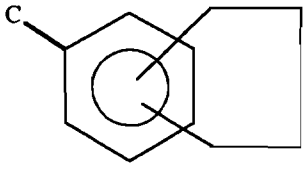
II.10.2.4. Variable attachment of a fused ring

It is possible to specify a variable attachment of one or several fusion points which are part of a fused ring system. One or two sites of variable attachment are possible.

Examples of both possibilities are shown on the following pages. In such a case the bonds of the fused ring must be specified in the query as being cyclic to ensure that only ring systems are retrieved.



Variable attachment of a fused ring (one site of variable attachment)



Variable attachment of a fused ring (two sites of variable attachment)

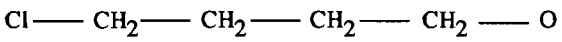
II.10.2.5. Variable attachment on a chain

It is possible to specify a variable attachment on a chain.

The following example illustrates how the variable attachment capability can be used in a search strategy to avoid multiple queries.

Let us consider the following search question:

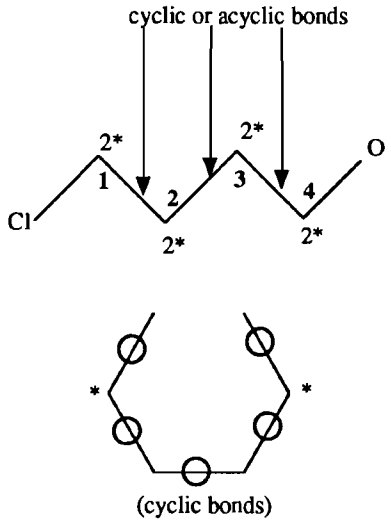
Search for the following structure:



where one or more carbon atoms is part of a saturated carbocyclic ring having from 5 to 8 carbon atoms.

Search problem with variable attachment on a chain

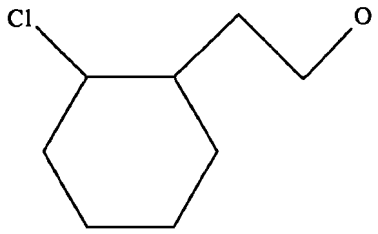
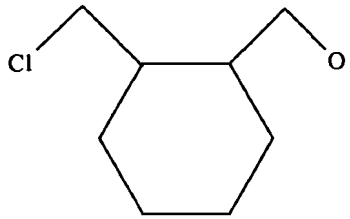
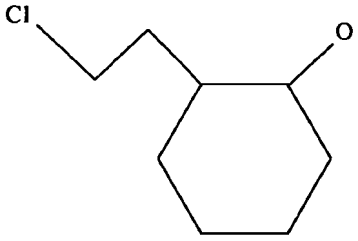
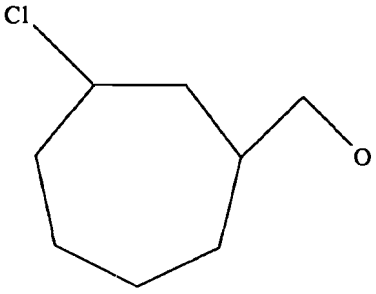
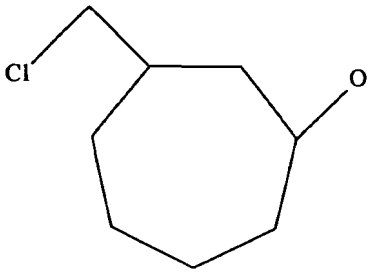
This question can be searched with a single query, as follows:



Query with variable attachment on a chain

In this query, each carbon atom of the chain carries two free sites (2*) to permit the possibility of being included in the carbocyclic ring. Two free sites are useful to allow a spiro ring.
The bonds linking the carbon atoms of the chain are specified as cyclic or acyclic.
For the ring portion which is drawn under the carbon chain, the bonds are specified as cyclic and the terminal carbon atoms carry one free site.
The possible sites of attachment of the ring portion on the carbon chain are shown by numbers (1, 2, 3 and 4).

Possible answers to this query are listed in the following table:

6-membered ring	
	
	
7-membered ring	
	

III. Query text input procedure

III.1. ST and QT command levels

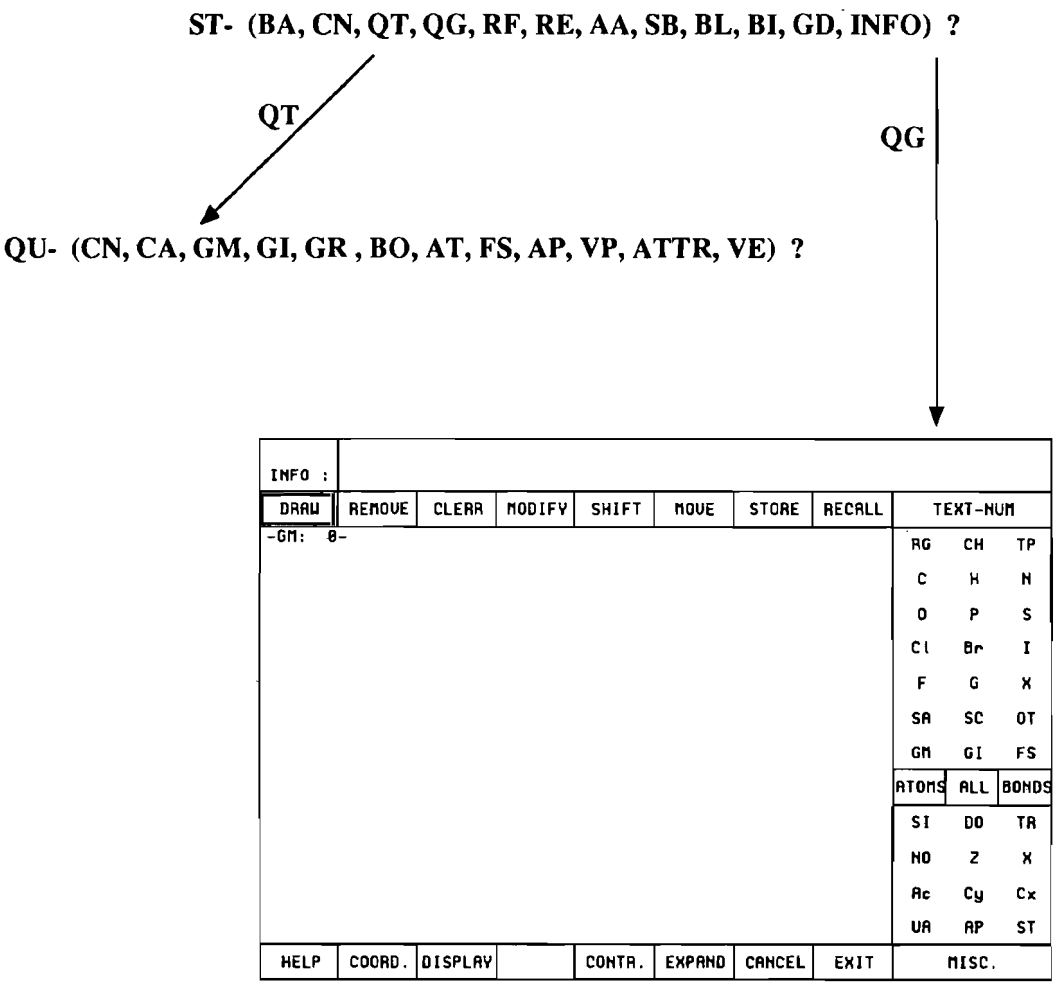
On connecting to Markush DARC, the user receives the initial Markush DARC prompt which starts with "ST". This prompt is referred to as the ST Command Level. The various commands which appear in this prompt are explained on the next pages.

For input of the structure query, Markush DARC offers both Graphic and Alphanumeric (Text) input methods. The method of query input is selected at the ST Command Level.

For Text input, select QT (Query Text). The system then provides a second prompt level for query definition (the QU Command Level). The use of commands at the QU Command Level is explained in the remainder of this Chapter.

For Graphic input, select QG (Query Graphic). The system then provides a graphic screen onto the structure query can be drawn directly.

This is illustrated by the following diagram:



To go from the QT level to the ST level, type FI when query input is finished.

The ST command level comprises the following commands:

COMMANDS	DEFINITION
BA	BAse (Specify or change structure database)
CN	Compound Number Search
QT	Query Text (Alphanumeric input of query)
QG	Query Graphic (Graphic input of query)
RE	REtrieve Candidates (Perform fragment or screen search)
AA	Atom-by-Atom Search (Perform atom-by-atom search on candidates)
SB	Search Bit screens
BL	Boolean Logic (Perform Boolean logic operation on answer sets)
BI	BIbliographic system (Switch to bibliographic system)
GD	Switch to Generic DARC
INFO	INFOrmation (Display news or information)
FI	FInish (Log off the system)
OP	OPtions (change terminal options, language etc)
OS	Other Specifications (add other specifications to query)

DISPLAY COMMANDS	
LI	LIst Compounds numbers
VI	VIew structure answers (main group) one at a time
VI FO	View FOcus the parts of the answers which match the query structure
VI MAX	VIew MAXimum the whole structure answer (all the groups in succession)

QUERY HANDLING COMMANDS	
SV	SaVe query or answer set
RF	Recall file (recall saved query or answer set)
ER	ERase (erase previously saved query or answer set)
HI	HIstory (list history of saved queries or answer sets)
QUERY HANDLING COMMAND PARAMETERS	
QU	QUery
CN	Compound Numbers
DN	Darc Numbers
QP	Questel-Plus query ("join" set)
BT	BaTch search query or answer
PC	PC query (query input on a PC using Darc-Chemlink)

The QT command level comprises the following commands:

COMMANDS	DEFINITION
CN	Recall Compound Number of a specific compound in the database to use as a base for query input
CA	CAncel entire query, a G group or attributes
GM	Group Markush (Specify Markush group being defined)
GI	Group Identical (Create a new group identical to an existing group)
GR	GRaph (Specify the skeleton of the query)
BO	BOnds (Specify the bonds in the query)
AT	AToms (Specify the atoms in the query)
FS	Free Sites (Specify desired substitution in query)
AP	Attachment Points (Define attachments of G groups to parent group)
VP	Variable Points of attachment (Define points of variable attachment)
ATTR	ATTRibutes (Define attributes on atoms in the query)
VE	VErify (Graphic verification of the query)
FI	FInish query input (returns to ST prompt level)

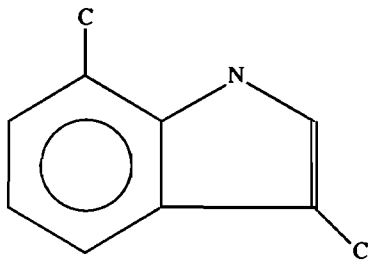
III.2. Types of queries

There are three basic types of queries:

- fully specified structure query

This is the case for fully defined queries where no substitutions are allowed.

Example:

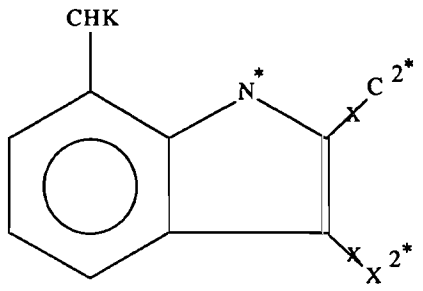


Example of a full structure query

- substructure query

This is the case for queries where possible substitutions are allowed using free sites and bonds and/or atoms may be undefined.

Example:

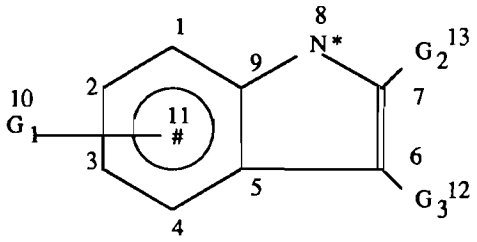


Example of a substructure query

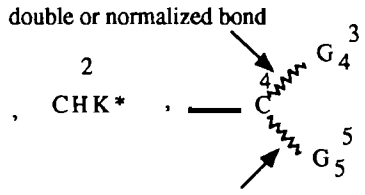
- Markush query

This is the case for queries with variable groups, free sites, bonds etc...

Example:



G1: 1 H , 2 N , 3 NO_2



G2: 1 H , 2 CHK* , C

G3: 1 H , 2 HET 2* , 3 CHK*

G4: 1 O , 2 N 2*

G5: 1 H , 2 N* , 3 CHK*

Markush DARC query with nodes numbered

III.3. Query formulation

For fully specified structure queries or substructure queries, proceed as follows:

1- Define query
2- Define (sub)structure of interest
3- Input query

For Markush queries:

1- Define query
2- Define main (father) group
3- Define the generic groups

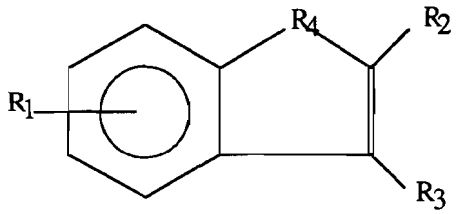
For each group:

- define graph (nodes of structure)
- specify atoms
- specify bonds
- specify free sites (substitutions)
- specify attributes (if desired)
- specify attachment points

4- Input the query

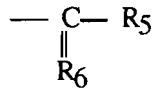
Example:

The following query problem:



R1: hydrogen, amino or nitro

R2: hydrogen, optionally monosubstituted alkyl or:



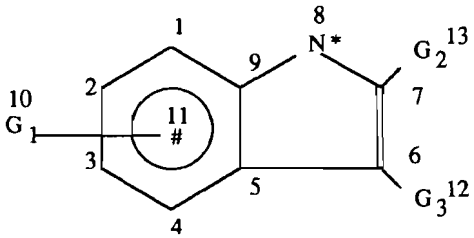
where R5 is hydrogen, optionally monosubstituted amino group or optionally substituted alkyl
and R6 is oxygen or nitrogen

R3: hydrogen, optionally monosubstituted alkyl or monocyclic non-aromatic heterocycle optionally substituted by 1 or 2 substituents.

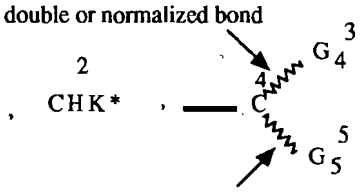
R4: optionally monosubstituted nitrogen.

Query problem

could be expressed as the following query:



G1: 1 H , 2 N , 3 NO_2



G2: 1 H , 2 CHK* ,

G3: 1 H , 2 HET 2* , 3 CHK*

G4: 1 O , 2 N 2*

G5: 1 H , 2 N* , 3 CHK*

Markush DARC query with nodes numbered

III.4. Query input procedure

III.4.1. Define graph (GR command)

It is necessary to enter the graph of the query before any other command.
The GR command allows the user to enter the graph of the query.

The system assumes by default that all atoms are carbon and all bonds are single.
All unoccupied valencies are assumed to have hydrogen atoms attached (no free sites are assumed).

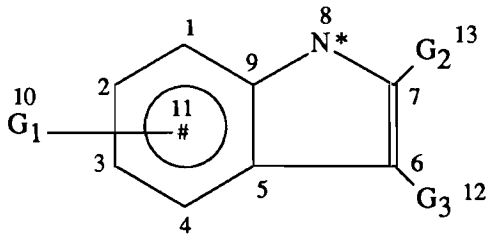
To define the graph, proceed as follows:

- Draw the structure on paper and number all non-hydrogen atoms.
The numerical values are arbitrary and do not have to be consecutive or to correspond to any nomenclature rules.

- Input the graph using:

- (hyphen)	to show a link between two atoms
, (comma)	to separate atoms in the input which are not directly linked
: (colon)	to define links between a series of consecutive atoms
/ (slash)	to remove a portion of the graph that has already been defined

In our Markush query example, the numbering of the nodes could be as follows:



Node numbering for the main group of a Markush DARC query

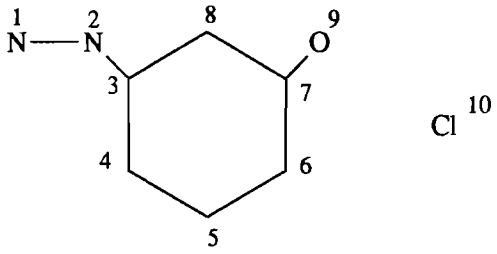
which would be input as follows:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? QT

***  QT     ***

-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  GR
*GRAPH
?  1:9-1,5-9,10-11,6-12,7-13
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  _
```

Input of the main graph using the GR command

The / (slash) may be used to remove a portion of the graph that has already been defined, using the GR command.

In the following example:

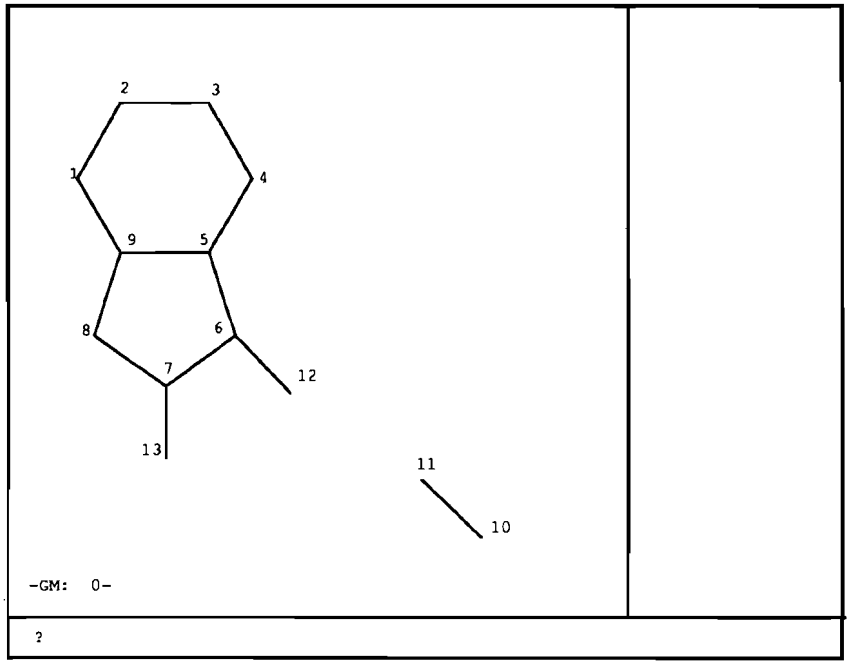


we could remove portions of the graph, as follows:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,VE)  ?  GR
*GRAPH
?  /3-2  removes bond 3-2 only (-N-N- remains as an isolated 2-atom fragment)
?  /7-9  removes bond 7-9 and removes the oxygen atom
?  /10   removes the Cl atom
```

III.4.2. Graphic verification (VE command)

The system takes the input it has received and draws it as a chemical structure to simplify checking of the input.
The VE command, which can be used at any stage of the query definition, allows the graphic verification of the query which is being input.



Verification of the graph of the main group

You can also use the VE TX command (text mode verification) if:

- you do not have a graphics emulator

- the query contains more than a certain number of atoms and therefore cannot be displayed by the Markush DARC software (e.g. teicoplanin)

- the display is not clear (e.g. lots of overlapping labels on atoms).

Please refer to the *Alphanumeric verification* chapter further on this section for more details on this command.

III.4.3. Atom specification (AT command)

This command is used to define:

- all non-carbon atoms
- superatoms
- generic groups ("G" groups)

The syntax of the AT command is as follows:

- specify the kind of node using:

- the element symbol, e.g. N
- a "G" group symbol, e.g. G1
- a shortcut, e.g. NO2
- a superatom, e.g. CHK
- the X symbol to represent any atom except hydrogen
- the # symbol to specify a variable attachment: more details on the use of the # symbol are given further on this section in the *Variable attachment specification* paragraph

- type a space

- specify the position(s) of the atom, using the same numbering as for the graph definition.

When the atom occurs in several positions use a comma (,) between the different positions.

It is not possible to specify several positions for a G group symbol. The system sends the following message:

```
SYNTAX ERROR - COMMAND IGNORED
```

A G group symbol may be used only once in the query.

Examples:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AT
*ATOMS
?  N 1,3
?  G1 4
?  G2 6
?  NO2 5
?  CHK 2,7
?  X 8
?  # 11
```

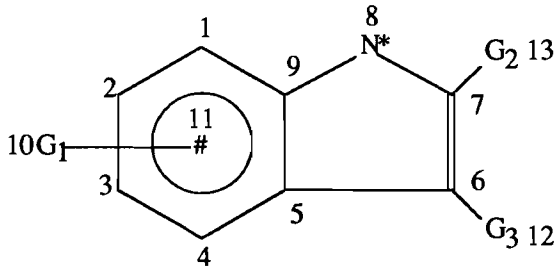
Examples of node specification using the AT command

If you want to change all the nodes to a particular value, just type the value, e.g.:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AT
*ATOMS
?  N
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?
```

AT command: typing a particular value changes all the atoms of the query to this value

If we consider our query example, we would specify the non-carbon atoms of the main group as follows:



Markush DARC query main group with nodes numbered

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AT
*ATOMS
?  N 8
?  G1 10
?  G2 13
?  G3 12
?  * 11
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

Atom specification for the main group using the AT command

The graphic verification can be used at this stage to verify the atoms.
If an error has been made for one atom, use the AT command again and specify the atom and the correct position. It is not necessary to re specify the other atoms.
The C element symbol can be used to replace an atom which has been wrongly entered.

In our example, if we want a N atom in the 4 position instead of the 8 position, we could type the following modifications, using the AT command:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AT
*ATOMS
?  N 4
?  C 8
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

Atom modification for the main group using the AT command

We do not need to re specify the other atoms.

III.4.4. Bond specification (BO command)

The syntax of the BO command is as follows:

- specify the nature of the bond using the following symbols:

SI	:	single
DO	:	double
TR	:	triple
NO	:	normalized (aromatic or tautomeric)
PEP	:	peptide bond
X :		undefined value
Z :		list of alternative bond values, in the format:
		Z alternative 1 alternative 2 Example: Z SI DO

- type a space

- optionally, specify the type of bond

AC	:	acyclic
CY	:	cyclic
CX	:	undefined type (cyclic or acyclic)

- type a space

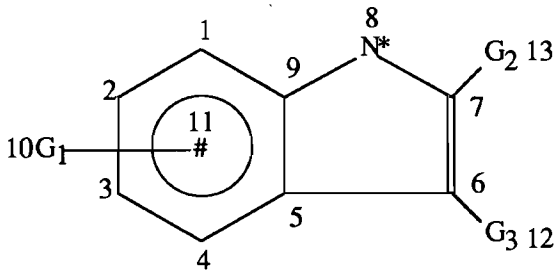
- specify the node numbers which are the bond termini, separated by a hyphen.

The colon and the hyphen can be used in the same way as for the input of the graph. A colon can be used to specify a range of consecutively numbered bonds. Hyphens can be used to specify consecutive bonds.

Example: NO 1:5-9-1

If the bond type is not specified, the system automatically generates the bond types in accordance with the graph input. The cyclic type is assigned to all ring bonds, and the acyclic type to all other bonds.
Regarding the value of the bond, all bonds are considered as single, unless specified otherwise.

In our query example, we could specify the bonds of the main group as follows:



Markush DARC query main group with nodes numbered

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  BO
*BONDS
?  NO 1:5-9-1
?  DO 6-7
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?
```

Bond specification for the main group using the BO command

The system assumes that bonds 1-2-3-4-5-6-7-8-9, 1-9 and 9-5 are cyclic and 7-13, 6-12 and 10-11 are acyclic unless specified otherwise.

III.4.5. Free site specification (FS command)

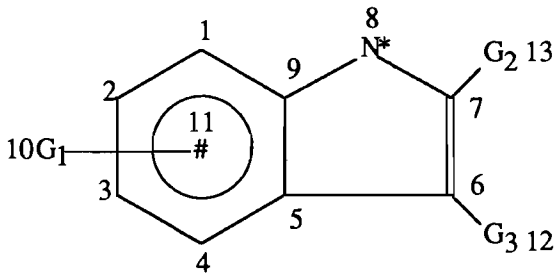
The syntax of the FS command is as follows:

- specify the maximum number of optional attachments allowed for any position

- type a space

- specify the node numbers.

In our query example, we could specify free sites in the main group as follows:



Markush DARC query main group with nodes numbered

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  FS
*FREE SITES
?  1 8
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

Free site specification for the main group using the FS command

III.4.6. Variable attachment specification (# symbol and VP command)

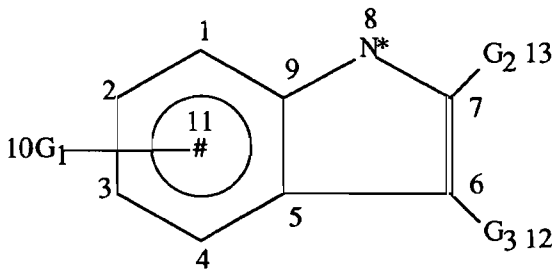
If the main group of the query contains variable attachments, it is necessary to specify these variable points of attachment before defining the values of the groups included in the main group.

To specify a variable attachment of a group, atom, superatom or shortcut, proceed as follows:

- specify the group, atom, superatom or shortcut which is variably attached by linking it to the # symbol: the node of the # symbol must be defined with the GR command and linked to the atom which is variably attached. If the definition of this node has been omitted at the stage of the graph input, this can be done even after the atom, bond and free site specification

- specify the positions of variable attachment, using the VP command: positions must be separated by commas.

In the query, we could specify the variable attachment of group G1 on the benzene ring as follows:



Markush DARC query main group with nodes numbered

We first specify that we want G1 to be variably attached by linking G1 to the # symbol. We have defined in the graph nodes 10 and 11, linked with a single bond and we have specified G1 in position 10 and # in position 11, using the AT command:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AT
*ATOMS
?  N 8
?  G1 10
?  G2 13
?  G3 12
?  # 11
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

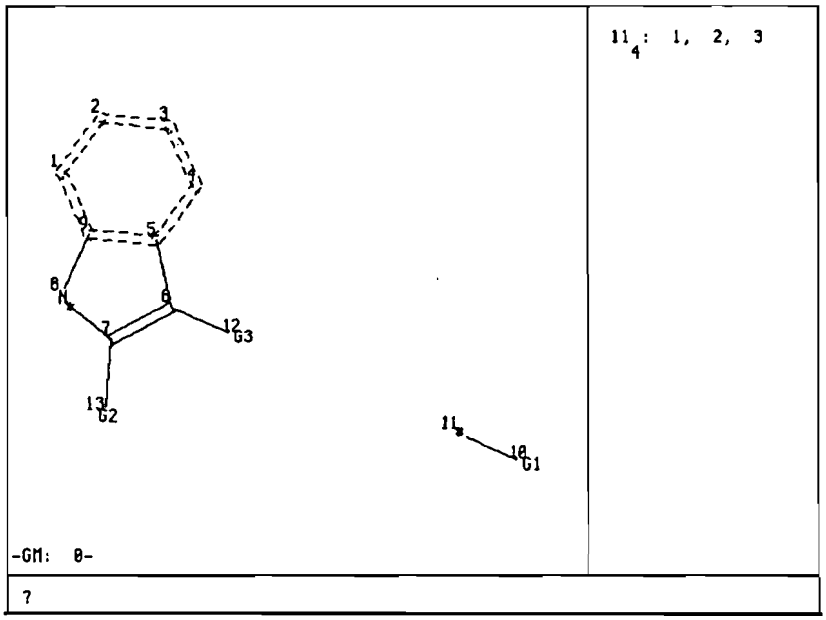
Atom specification for the main group using the AT command

We then specify the positions of variable attachment, using the VP command:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  UP
* VARIABLE POSITIONS *
NODE   11 :
?  1,2,3,4
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

Variable attachment specification, using the VP command

We have now finished inputting the main group of the query, and we can ask for a graphic verification, using the VE command.



Graphic verification of the main group, including variable attachments

When the group comprises a variable attachment, it is indicated in the upper right hand corner of the display.

Typing VP displays the positions implied in the variable attachment.

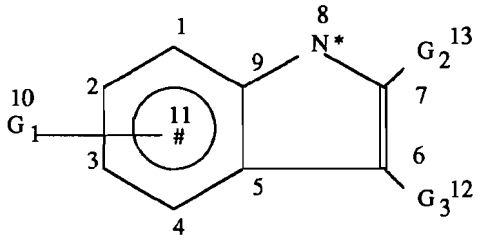
Here "`11 : 1, 2, 3, 4`" indicates that atom 11 represents a variable attachment at positions 1, 2, 3 and 4.

III.4.7. Group definition (GM command)

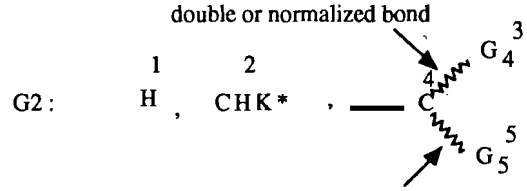
To define the values of a group:

- specify the GM command at the QT level
- enter the number of the group
- define the individual graphs of the values, using the GR command, in the same way as for the main group
- define the atoms (AT command), bonds (BO command), free sites (FS command)
- define the attachment points (only for multi-node fragments), using the AP command

In our query example, we could specify the values of group G1 as follows:



G1: 1 H , 2 N , 3 NO_2



G2: 1 H , 2 CHK* ,

G3: 1 H , 2 HET 2* , 3 CHK*

G4: 1 O , 2 N 2*

G5: 1 H , 2 N* , 3 CHK*

Markush DARC query with nodes numbered

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  GM
GROUP TO DESCRIBE ?  1
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  GR
*GRAPH
?  1,2,3
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AT
*ATOMS
?  H 1
?  N 2
?  NO2 3
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

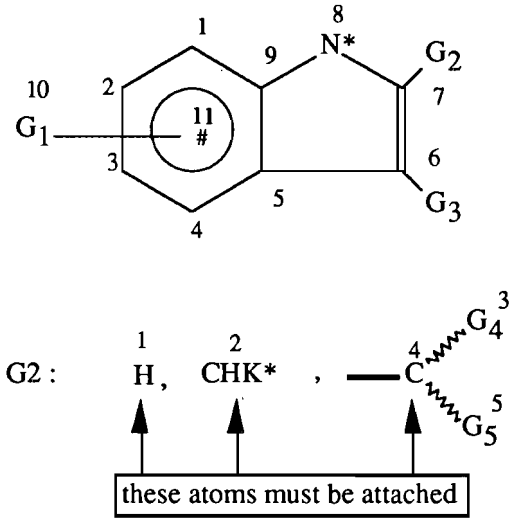
Input of group G1

III.4.8. Attachment points specification (AP command)

If a group definition includes one or more multi-node value, it is necessary to specify attachment points for all values of that group, using the AP command.

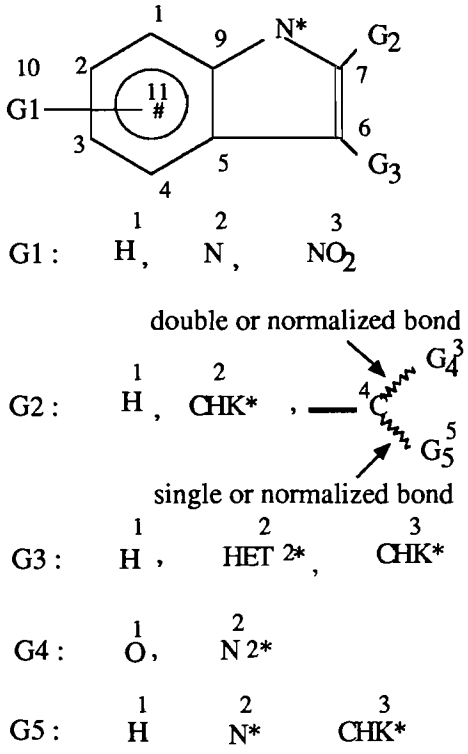
The positions of attachments are separated by commas.

In our query example, group G2 contains one multi-node value (the value including G4 and G5): it is necessary to specify the attachment points of all the values of group G2 as follows:



Main group and group G2

The input of group G2, including attachment points, is carried out as follows:




```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  GM
GROUP TO DESCRIBE ?  2
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  GR
*GRAPH
?  1,2,3:5
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AT
*ATOMS
?  H 1
?  CHK 2
?  G4 3
?  C 4
?  G5 5
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  FS
*FREE SITES
?  1 2
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  BO
*BONDS
?  Z DO NO 3-4
?  Z SI NO 4-5
?
-QU- (CN,CA,GM,GI,GR,BD,AT,FS,AP,UP,ATTR,UE)  ?  AP
ATTACHMENT POINTS
ATT. : ATOM   7 OF GROUP NUMBER  0
NB ATOMS ? 1,2,4
-QU- (CN,CA,GM,GI,GR,BD,AT,FS,AP,UP,ATTR,UE)  ?  _
```

Input of group G2

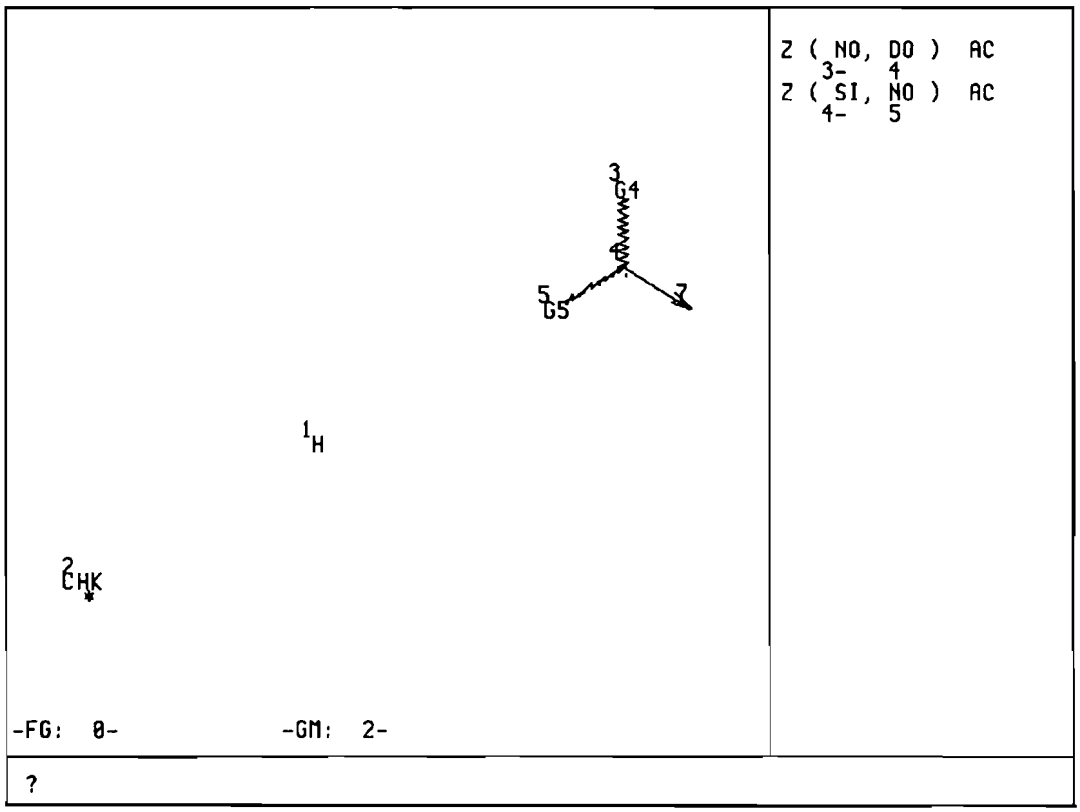
In group G2, it is necessary to specify Z bonds for bonds 3-4 and 4-5, in order to take into account the possible normalization of these bonds if G4 and G5 are both N.

Bond 3-4 should be specified as normalized in addition to double: a Z bond should be used, with the DO (double) NO (normalized) list of alternative values.

Bond 4-5 should be specified as normalized in addition to single: a Z bond should be used, with the SI (single) NO (normalized) list of alternative values.

For more details on normalization of bonds, please refer to the *Chapter in Section 4 - Structure representation conventions - .*

The graphic verification (VE) of group G2 gives the following display:



Graphic verification of group G2 (VE command), including Z bonds

The lists of possible alternatives for the two Z bonds are provided automatically in the upper right hand corner of the display.

III.4.9. Identical group definition (GI command)

When two groups comprise identical or slightly different values, it is possible to define the second group by copying the first group (origin group), using the GI command.

This command works as follows:

- type GI at the QT level

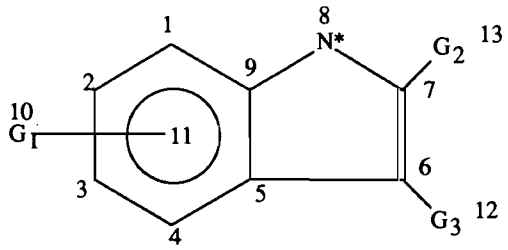
- specify the group to be described

- specify the group to be copied (origin group)

- modify the values if desired

- if there are multi-node values, specify the attachment points using the AP command.

If we consider the following query:



1 2 3 4 5
G1 : H, N, Cl, CHK*, NO2
1 2 3 4
G2 : H, CHK*, — O — CHK

G3 : H, CHK*, — O — CHK

Query with two identical groups (G2 and G3)

We could input group G3, which is identical to group G2, as follows:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UA,ATTR,VE)  ?  GI
GROUP TO DESCRIBE ?  3
GROUP TO COPY     ?  2
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UA,ATTR,VE)  ?  AP
ATTACHMENT POINTS
ATT. : ATOM   6 OF GROUP NUMBER  0
N0 ATOMS ? 1,2,3
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UA,ATTR,VE)  ?  _
```

Use of the GI command to input an identical group

III.4.10. Attributes specification (ATTR command)

Attributes on atoms or superatoms are specified using the ATTR command at the QT level.

The following table summarises all possible attributes, and the corresponding input syntax.

attribute	symbol in the ATTR menu	input syntax	possible values
charges	CH	sign/magnitude/space/atom number(s) separated by commas or x (any charge)/space/atom number(s) +1 4,5 x 4	-9 to +9 or x
delocalized charges	CH	D/sign/magnitude/space/atom number(s) D+1 4,5	-9 to +9
abnormal mass	AM	mass value/space/atom number(s) or x (any mass)/space/atom number(s) 99 5,6	2 to 1023 or x
abnormal valency	AV	valency/space/atom number(s) or x (any valency)/space/atom number(s) 5 2,3	0 to 255 or x
deuterium and tritium	DT	D or T/value/space/atom number(s) D2 5	0 to 15
free sites	FS	number of substitutions/space/atom number(s)	0 to 15

chain/ring attributes (on chain and ring superatoms)	CR	attribute alternatives separated by commas/space/atom number(s) LO,MID,BRA 1,2 FU,UNS 4	chains: LO, MID, HI, STR, BRA rings: MON, FU, SAT, UNS
multiplier attribute (on atom or superatom)	MU	specific number/space/atom number(s) or min. value/colon/max. value/space/atom number(s) 2 5,6 1:5 4	1 to 16
D, L or DL amino acid configuration (on peptide superatom)	DL	list of alternatives/space/superatom number(s) D,L 2,3	D, L or DL
position of substitution on the amino acid	SP	position number/space/atom number(s) 3 5	see amino acid numbering in *Appendix 1*
roles of atoms or superatoms in a polymer (on atoms, shortcuts or superatoms)	PA	attribute value/space/atom number(s) EG 2	MC, XL. EG, DE or GM
translation attributes (on atoms, superatoms or shortcuts)	TRA	attribute value/space/atom number(s) BT 1,2,3,6	EQ, NT, BT or ANY

Example:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,VE)  ?  ATTR
-ATTR- (FS,CH,AV,AM,CR,MU,PA,DT,SP,TRA)  ? CH
*CHARGES
?  +1 1,3
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,VE)  ?  ATTR
-ATTR- (FS,CH,AV,AM,CR,MU,PA,DT,SP,TRA)  ? CR
*CCHAIN/RING
?  LO,MID 1,2
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,VE)  ?  _
```

Attribute specification, using the ATTR command

When a group comprises several atoms carrying the same type of attribute, you may want to cancel this attribute on one of the atoms. This can be done by specifying the value zero for the FS, CH, DT and AV numerical attributes.

Example 1:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,VE)  ?  ATTR
-ATTR- (FS,CH,AV,AM,CR,MU,PA,DT,SP,TRA)  ? CH
*CHARGES
?  0 3
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,VE)  ?  _
```

Charge cancellation

to cancel the charge on atom 3.

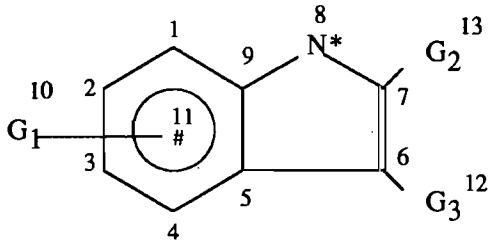
Example 2:

Let us suppose that we have a query with two G groups G1 and G2 having the same definitions, but with less charges in group G2. We could first input group G1, then use the GI command to copy the definition of group G1 into group G2 and finally cancel the charges that we do not wish in G2 on e.g. nodes 3, 5 and 7.

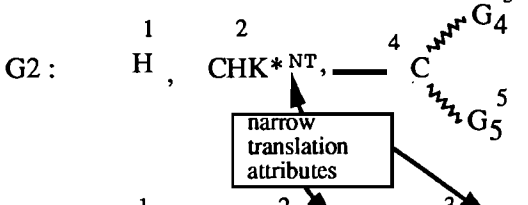
```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,VE)  ?  ATTR
-ATTR- (FS,CH,AV,AM,CR,MU,PA,DT,SP,TRA)  ? CH
*CHARGES
?  0 3,5,7
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,VE)  ?  _
```

Free site cancellation

The input of translation attributes for our query example could be performed as follows:



G1 : 1 H , 2 N , 3 NO_2



G3 : 1 H , 2 HET 2* NT , 3 CHK* NT

G4 : 1 O , 2 N 2*

G5 : 1 H 2 N* 3 CHK*

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  GM
GROUP TO DESCRIBE ?  2
PREVIOUS INPUT ? (Y/N) Y
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  ATTR
-ATTR- (FS,CH,AV,AM,CR,MU,PA,DT,SP,TRA)  ? TRA
*TRANSLATION ATTRIBUTES
?  NT 2
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  GM
GROUP TO DESCRIBE ?  3
PREVIOUS INPUT ? (Y/N) Y
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  ATTR
-ATTR- (FS,CH,AV,AM,CR,MU,PA,DT,SP,TRA)  ? TRA
*TRANSLATION ATTRIBUTES
?  NT 2,3
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?
```

Translation attributes specification

Note:

Inputting a translation attribute on an atom, superatom or shortcut already carrying a translation attribute will add this new attribute to the already existing attribute.
It will not replace the existing attribute.
If a replacement is required rather than an addition, it is necessary to first modify the existing attribute, to EQ and then apply the desired translation attribute.

More details on attributes searching can be found below in the *Specialised search capabilities* chapter.

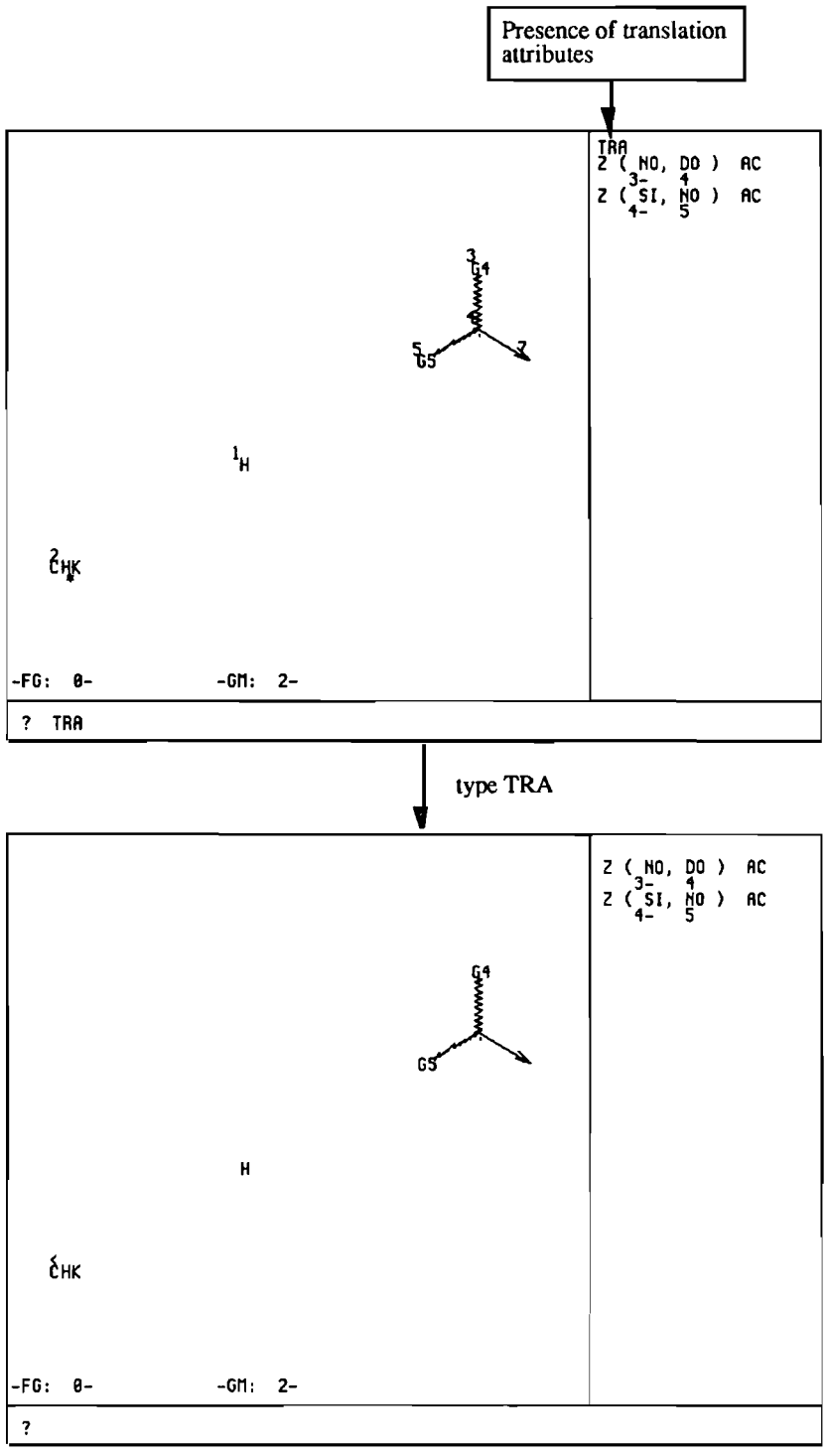
III.4.11. Graphic verification of attributes (VE command)

Attributes input can be checked using the VE command.
The presence of an attribute in a group is indicated by the code for this attribute in the upper right hand corner of the display.
Typing the code of the attribute will give the same display, with the addition of the attribute(s) on the atoms carrying this attribute.

Translation attributes (type TRA at the question mark prompt if TRA is indicated in the upper right hand corner of the screen) are displayed as follows:

type of translation attribute	attribute code	display
equal attribute	EQ	no display
narrow attribute	NT	<
broad attribute	BT	>
any attribute	ANY	x

Example: the display of translation attributes in group G2 of our query example would be as follows:



Example of translation attribute (TRA) graphic verification

III.4.12. Alphanumeric verification (VE TX command)

An alphanumeric verification of the input is possible in addition to the graphic verification, at any stage of the input, using the VE TX command.

If we consider our example, the alphanumeric verification of group G2 would be as follows:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,VE)  ?  VE TX
** GROUP N0 :  2 **
GR
3- 4- 5, 1, 2

AT
C
H    1
CHK  2
G4   3
G5   5

BO
Z ( NO, DO )  AC  3- 4
Z ( SI, NO )  AC  4- 5

AP
ATT. : NODE   7 OF GROUP  0 :   1, 2, 4

FS
0
1    2

TRA
EQ
NT   2

-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,VE)  ?
```

Alphanumeric verification (VE TX) of group G2

III.4.13. Cancel (CA command)

The CA command may be used at any stage of query input to cancel:

- the whole query
- the current group
- attributes

The CA command works as follows:

To cancel the whole query:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  CA
ALL,GM,(ATTR) ?  ALL
STRUCTURE CANCELLED
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

Cancellation of the whole query structure using the CA command

To cancel the current group:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTA,UE)  ?  CA
ALL,GM,(ATTR) ?  GM
CURRENT GROUP CANCELLED
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

Cancellation of the current group using the CA command

Caution: if the current group is the main group G0, the whole structure will be cancelled.

To cancel an attribute type in a group:

- specify the group using the GM command

- type CA

- enter the type of attribute to cancel, e.g. FS (free sites):

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  GM
GROUP TO DESCRIBE ?  3
PREVIOUS INPUT ? (Y/N) Y
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  CA
ALL,GM,(ATTR) ?  FS
"FS  " ATTRIBUTES CANCELLED
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

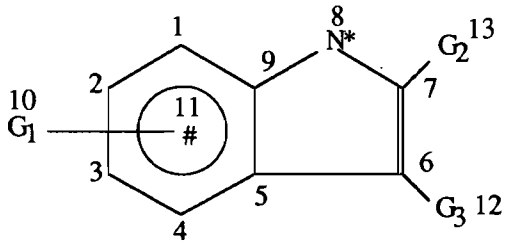
Cancellation of the free sites attributes in group G3 using the CA command

This cancels all free sites of group G3.

It is not possible to cancel translation attributes using the CA command because the presence of this type of attribute is mandatory on all the nodes of the query.
The default attribute i.e. the attribute which is applied to all the nodes of the query unless specified otherwise, is "equal" (EQ).

Thus, cancelling a given translation attribute other than EQ, is equivalent to modifying it with the value EQ.

Let us consider our query: if we want to cancel the NT attribute on the CHK superatom in group G3, we will specify the EQ value for the position of CHK.



G3 : H, CHK 1 2 * NT , HET, 3 2* NT

to cancel the NT attribute, modify it to EQ

The modification is carried out as follows:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  GM
GROUP TO DESCRIBE ?  3
PREVIOUS INPUT ? (Y/N) Y
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  ATTR
-ATTR- (FS,CH,AV,AM,CR,MU,PA,DT,SP,TRA)  ? TRA
*TRANSLATION ATTRIBUTES
?  EQ 2
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

Modifying translation attributes to EQ

If you want to change all the translation attributes to EQ at once, just type EQ.

III.4.14. Principal error messages during query input

III.4.14.1. Attachment points not defined

If you want to input the next group using the GM command and you have not defined the attachment points of the current group (which contains multi-atom fragments), you will get the following message, inviting you to define the attachment points before going to the next group:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  GM
ATTENTION - "AP" NOT SPECIFIED
DESCRIPTION ? (Y/N)  Y
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  AP
ATTACHMENT POINTS
ATT. : ATOM   6 OF GROUP NUMBER  0
N0 ATOMS ? 1,2,3,4
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  _
```

Error message if attachment points have not been defined

III.4.14.2. Wrong number of attachment points

If, while specifying the attachment points using the AP command, you specify too few or too many atom numbers, you will get the following message:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  AP
ATTACHMENT POINTS
ATT. : ATOM   6 OF GROUP NUMBER  0
N0 ATOMS ? 1,2,4,5
WRONG NUMBER OF ATTACHMENT POINTS
N0 ATOMS ? 1,2,3,4
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  _
```

Error message if wrong number of attachment points

If you specify an atom number that does not exist, you will get the following message:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  AP
ATTACHMENT POINTS
ATT. : ATOM   2 OF GROUP NUMBER  0
N0 ATOMS ? 1,2,5
WRONG NUMBER - COMMAND IGNORED
N0 ATOMS ? 1,2,3,4
WRONG NUMBER OF ATTACHMENT POINTS
N0 ATOMS ? 1,2,3
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  _
```

III.4.14.3. Groups not all defined

If you want to exit from the query input level (QT level), you must type FI (FInish).
But if some groups have not been defined, you will get the following message

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  FI
NON-DEFINED GROUP(S) - NO VALIDATION POSSIBLE
GM :    5
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

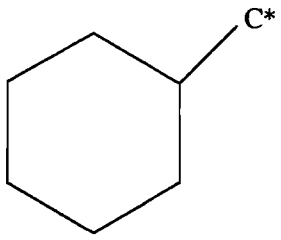
Error message if all groups have not been defined

III.4.15. Other specifications after query input

When the query has been defined, "FI" is entered at the "QU" prompt level.
The system then prompts for

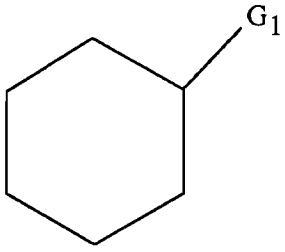
```
OTHER SPECIFICATIONS (Y/N) ?
```

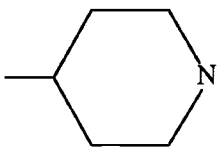
The "other specifications" capability is a mechanism for restricting (or limiting) the answers which are accepted as the results of a search.
The "other specifications" are used to forbid the presence of carbon atoms, heteroatoms or of certain types of rings in the answers.
They are also used to specify limits.
If you use "other specifications", records where permutations contain that particular feature will be excluded. You may retrieve answers which contain the restriction.
This is illustrated by the following example:



No heterocycles

will retrieve



G1 =  , Cl , Br

Other specifications example

In this example, some values of G1 satisfy the restriction "no heterocycle", even though the value which matches the query does contain a heterocycle.

The two possibilities for "other specifications" are summarised in the following tables:

forbid the presence of:
atoms: carbon, nitrogen, oxygen, phosphorus, sulphur or halogens
heteroatoms
heteroatoms other than N, S, O, P and halogens
rings of any kind
heterocycles
carbocycles
unsaturated rings
saturated rings

limits
simple molecule restriction
atom number limitation (minimum and/or maximum)
component number limitation (minimum and/or maximum)
file segments

When the other specifications capability is used, the system analyses the query and prompts for only those other specifications which would be applicable to the query. For example, if the query contains a heterocycle, the system does not provide the prompt to forbid heterocycles in the answer set.

If the response to the other specifications prompt is N(o), no restrictions are placed upon the acceptable answers.
The system will then provide the following prompt

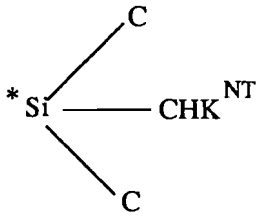
```
FILE SEGMENTS(Y/N) ?
```

to allow the application of this restriction if desired (see next paragraph, for details on file segments).

If the response to the other specifications prompt is Y(es), the system analyses the query and provides prompts for those possible restrictions which are or could be relevant to the query.
If any of the relevant restrictions is desired, enter Y(es) to the prompt.
If the restriction needs further definition, such as component number limitation, the system will prompt for the minimum and maximum number which is acceptable.

Example:

With the following query:



the following "other specifications" are possible:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  fi
OTHER SPECIFICATIONS (Y/N) ? y
ON COMPONENT NO.  1 DO YOU FORBID THE PRESENCE OF :
   N  (Y/N) ?  N
   O  (Y/N) ?  N
   P  (Y/N) ?  N
   S  (Y/N) ?  N
   HALOGEN(S)  (Y/N) ? Y
   RING(S)(Y/N) ?  N
   HETEROCYCLE(S)(Y/N) ? N
   CARBOCYCLE(S)(Y/N) ?  N
   UNSATURATED RING(S)(Y/N) ?  N
   SATURATED RING(S)(Y/N) ?  N
COMPONENT NUMBER LIMITATION (Y/N) ?  N
SIMPLE MOLECULES LIMITATION (Y/N) ?  N
FILE SEGMENTS(Y/N) ? N

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

III.4.16. File segments

There are thirty file segments, falling logically into a number of classes, which represent a class of features which a record, or the structure contained in the record may possess. These are denoted by single letters A - Z and 1 to 8
A record may have more than one file segment.

The reasons for having file segments are:

- to distinguish between representations which are similar graphically although quite different chemically.

- to allow the user to pre select a part of the database and expedite the search.

Not all file segments are used in some of the databases. Please check the *Database Appendices* of this manual to determine which file segments are used by the database producers.

The available file segments are the followings:

CPI SECTIONS		
A	Section A	
B	Sections B or C	
E	Section E	
GENERAL		
Y	Mixture	
Z	Salt	
1	Registry	
POLYPEPTIDES		
P	Polypeptide	
NON-POLYMER COMPOUNDS		
C	Coordination Compound / Complex	
L	Oligomer	
W	Extended structures	
M	Metals and Alloys	
V	Ordinary organic chemicals	
7	Inorganic	
POLYMERS		
F	Any polymer	
	Backbone type	
	H	Homopolymer / homocondensate
	S	Simple binary condensate
	J	Alternating copolymer / condensate
	K	Block copolymer / condensate
	R	Random copolymer / condensate
	N	Natural polymer
	Q	No backbone
	Number of components in backbone	
	2	2 monomers / condensates
	3	3 monomers / condensates
	4	4 monomers / condensates
	Modification	
	X	Crosslinked polymer
	D	Derivatized polymer
	G	Grafted polymer
	T	End modified polymer
	5	Surface modified polymer
	U	Unmodified polymer

File segments may be specified after query input has been completed, after the "other specifications" prompt, or as part of the "other specifications".

When query input has been completed, the system automatically prompts for "Other Specifications."
After a response to this prompt has been entered, the system then prompts for "File Segments." If you wish to specify file segments, enter a "Y" (yes) response to this prompt.
The system then prompts for the segments which are to be combined with the various Boolean operators "AND," "OR," or "NOT."
The file segments to be combined with each type of logic are input on a single line with no separations.

Example:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UA,ATTR,UE)  ?  FI
OTHER SPECIFICATIONS (Y/N) ? N
FILE SEGMENTS(Y/N) ? Y
  SEGMENTS LIST :
   AND ? EC
   OR  ?
   NOT ?
   AND : CE
   OR  :
   NOT :
CONFIRM(Y/N) ? Y
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,INFO) ?
```

File segment specification

III.5. Specialised search capabilities

III.5.1. Attribute searching

If no standard attributes are specified in the query, Markush DARC searches for all forms - with or without an attribute.

If an attribute is specified, Markush DARC searches for only those structure records with the designated attribute.

For Charge, Abnormal Mass and Abnormal valency attributes, Markush DARC can also search for these attributes without specifying the actual sign or magnitude of the attribute in question.

Attributes may be specified directly at the QT level or at the ATTR level, as shown in the following example:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AM
*ISOTOPS
?  99 1
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  ATTR
-ATTR- (FS,CH,AV,AM,CA,MU,PA,DT,SP,TRA)  ? AM
*ISOTOPS
?  99 1
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

Attribute searching at the QT or ATTR level

III.5.1.1. Searching for abnormal masses

If no mass is specified in the query, Markush DARC will search for the elements with any mass.
If you wish to specify an abnormal mass for an atom, this can be done by using the ATTRibute Command during query input.
You may also ask to restrict the retrieval to only those structures which have an abnormal mass of any kind on atoms of interest. To do this, you use "x" to describe the mass.
Note that this will not retrieve records where the atom has the normal (default) mass.

To specify an Abnormal Mass, enter "ATTR" at the "QU" prompt. The system then prompts for the type of attribute which you wish to specify. For Abnormal Mass, we input "AM". The Abnormal Mass desired is then specified at the "?" prompt using the following format:

- mass value desired
- a space
- the atom number(s) that you wish to have this value, separated by commas.

ABNORMAL MASS	
C	will retrieve C or any isotope
^{13}C	will retrieve only ^{13}C
^{x}C	will retrieve only labelled isotopes of C

Example:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  ATTR
-ATTR- (FS,CH,AU,AM,CR,MU,PA,DT,SP,TRA)  ? AM
*ISOTOPS
?  13 2
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

Example of AM attribute searching

III.5.1.2. Searching for abnormal valencies

If no specification of valency is used in the query, Markush DARC will search for the elements with any of their possible valencies.

If you wish to specify an abnormal valency for an atom, this can be done using the ATTRibute Command during query input. You may also ask to restrict the retrieval to only those structures which have an abnormal valency of any kind on atoms of interest. To do this, you use "x" to describe the valency.

A table listing the Normal Valencies of elements is given in this Section of the manual - *Query input basics Chapter - element symbol Paragraph - .*

To specify an Abnormal Valency, enter "ATTR" at the "QU" prompt. The system then prompts for the type of attribute which you wish to specify. For Abnormal Valency, we input "AV".

The Abnormal Valency which is desired is then specified at the "?" prompt using the following format:

- valency value
- a space
- the atom number(s) that you wish to take this value, separated by commas.

ABNORMAL VALENCY	
Tc^{6*}	will retrieve any Technetium atom with any valency (normal Tc valency is 7)
Tc^{6*} AV=6	will retrieve only Technetium atoms with an abnormal valency of 6
Tc^{6*} AV=x	will retrieve only Technetium atoms with an abnormal valency (it will not retrieve Tc with a valency of 7, the normal valency)

Example:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  ATTR
-ATTR- (FS,CH,AV,AM,CR,MU,PA,DT,SP,TRA)  ? AV
*VALENCIES
?  6 2
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

Example of AV attribute searching

III.5.1.3. Searching for charges

If no specification of charge is used in the query, Markush DARC will search for the elements with any charge.

If you wish to specify a charge for an atom, this can be done using the ATTRibute Command during query input. You may also ask to restrict the retrieval to only those structures which have a charge of any kind on atom(s) of interest. To do this, you use "x" to describe the charge.

To specify a Charge, enter "ATTR" at the "QU" prompt. The system then prompts for the type of attribute which you wish to specify. For Charge, we input "CH". The Charge desired is then specified at the "?" prompt using the following format:

- sign of the desired charge (+ or -)
- magnitude of the charge
- a space

- the atom number(s) that you wish to take this value, separated by commas.

When searching for charge of any kind (x value), the sign of the charge is not input:

- letter "x"
- a space
- the atom number(s) that you wish to have this value, separated by commas.

CHARGES	
N	will retrieve any N, N^+ or any charged nitrogen
N^+	will retrieve only N^+
N^x	will retrieve any charged nitrogen (+ or -)

Example:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  ATTR
-ATTR- (FS,CH,AU,AM,CR,MU,PA,DT,SP,TRA)  ? CH
*CHARGES
?  +1 2
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?
```

Example of charge searching

III.5.1.4. Searching for chain and ring (CR) superatom attributes

A search for a superatom without specification of an attribute in the query will retrieve all occurrences of the superatom - with any attribute.

If a superatom attribute is specified in the query, the search will retrieve only the specified superatom with the designated attribute.

Attributes for the chain superatoms CHK, CHE, and CHY fall into two classes: LO MID HI and STR BRA. Specifying an attribute in one class does not restrict the attributes in the other.

Likewise, the ring superatoms have two classes of attributes.

Specifying an attribute in one class does not restrict the attributes in the other class.
For example, attributes for the CYC superatom fall into two classes: MON, FU and SAT, UNS. Specifying an attribute in one class, MON, for example, does not restrict the retrieval on the basis of saturation or unsaturation. This is illustrated by the following table:

CHAIN/RING (CR) SUPERATOM ATTRIBUTE SEARCHING	
Specified attribute	**Retrieves**
None	LO, MID, HI, STR, BRA
LO	STR, BRA
MID	STR, BRA
HI	STR, BRA
BRA	LO, MID, HI
STR	LO, MID, HI

Caution: the use of combinations of attributes on the same superatom may result in a more restricted retrieval than was intended.
For example, a search for CHK $^{LO\ MID}$ will retrieve only those CHK superatoms which have both LO and MID specified as attributes.

To search for those structures which have either CHK ^{LO} or CHK ^{MID} at the specified position, formulate the query to include both CHK ^{LO} and CHK ^{MID} as alternatives in a G group at that position.

Search	Retrieval
CHK	all CHK
CHK^{LO}	CHK^{LO}, $CHK^{LO\ MID}$ etc...
$CHK^{LO\ MID}$	$CHK^{LO\ MID}$ not CHK^{LO} or CHK^{MID}

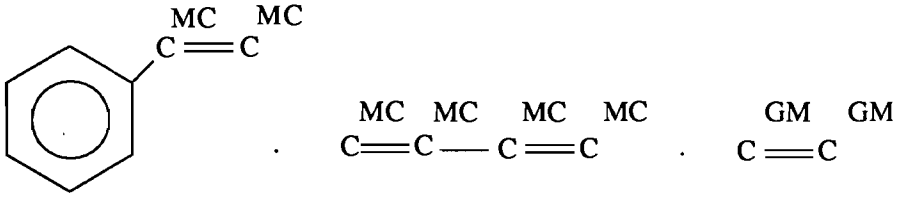
III.5.1.5. Searching for polymer attributes (PA)

Polymer attributes are used to distinguish the role that a particular atom or group of atoms has in the polymer. Polymer attributes can be applied to atoms or to superatoms.

When a polymer attribute is assigned to an atom of the component of the polymer, all atoms and superatoms of that component also are assigned that polymer attribute.

Example of use of polymer attributes:

Consider a styrene-butadiene copolymer which has ethylene grafted onto it. This polymer would be represented as three disconnected structures: styrene and butadiene, both with the MC attribute on all atoms, and ethylene with the GM attribute on all its atoms.



Styrene-butadiene copolymer, grafted with ethylene

Searching without specification of polymer attributes will retrieve all forms of the substance specified. To use a polymer attribute as a search term, specify that attribute on at least one atom of the component of interest.

Example:

specified attribute	retrieval
C=C	ethylene with any attribute
MC MC C=C	only ethylene as monomer
MC C=C	only ethylene as monomer

III.5.2. Simple molecule limitation

Simple Molecules in this connotation denote small molecules with fewer than 13 atoms. This includes Markush formulae containing permutations which have fewer than 13 atoms.

If your query and all the answer structures you want contain only a few atoms, it is necessary to restrict the search to just those records in the database which contain simple molecules. You can specify that the answers must fit within one or more of the categories listed below:

1 atom
2 atoms
3 atoms
4 atoms
5 atoms
6 atoms
7-12 atoms

Some records retrieved may contain more than 13 atoms in the matching structure; one or more permutation of each of these records will contain less than 13 atoms.
It is essential to specify one of these screens if you have a very small query (less than 3 atoms and/or less than 2 bonds) as it is necessary to have at least two or three connected atoms in the structure for the RE screens to be used. The system will be able to handle these small queries much better if the Simple Molecule Limitation is used.

III.5.3. Peptide searching

A search for a peptide structure or substructure with Markush DARC can be carried out with:

- peptide superatoms
- peptide bonds
- peptide attributes: configuration attributes (D,L or DL) and position attributes (SP).

Please refer to the *Databases Appendices (Appendices 3 and 4)* of this manual for details on the use of these features in the Markush DARC databases, particularly for the position attributes feature.

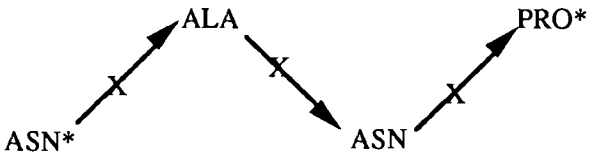
The following examples illustrate the peptide searching characteristics.

Example 1:

Search problem:

Search for polypeptides containing the NANP tetrapeptide fragment typical of a protein of Plasmodium Falciparum.

Query:



Peptide query 1

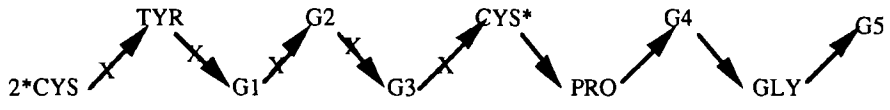
It is necessary to specify a CX (cyclic or acyclic) type of bond in order to retrieve both linear and cyclic polypeptides containing this fragment.

Example 2:

Search problem:

Search for vasopressin analogues

Query:



Peptide query 2

In this query:

- specifying a CX type of bond for the 5 bonds of the fragment containing the two Cys amino acids and a free site on each CYS superatom allows to retrieve both reduced forms and forms with a disulphide bridge between the two Cys amino acids.

- an additional free site on the terminal CYS superatom allows the retrieval of vasopressin analogues having possible substitutions on the N-terminal position.

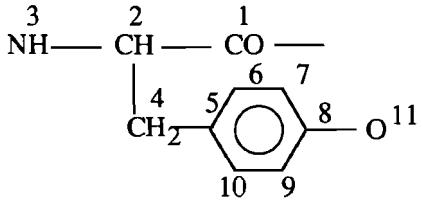
Example 3:

Search problem:

Search for peptides containing a O-Me tyrosine amino acid i.e. tyrosine substituted by a methyl group on the oxygen atom.

1) In MPHARM

According to the amino acid numbering in Markush DARC, the oxygen atom is at position 11 (see *Appendix 1: Markush DARC shortcuts, superatoms and attributes*).



Tyrosine with atom numbering

If the oxygen atom is substituted by a methyl group, this group is a substituent of position 11.
The query should be formulated as follows:
Query:

```
      2*      SP 11
TYR —— C
```

2) In WPIM

The O-Me tyrosine should be drawn in full.

Section 4
Structure representation conventions

I. Basic principles

Conventions in structural representation are created to impose some form of standardisation on the form of the structures.

Many compounds can be drawn in more than one way:

- chemists have several ways of expressing chemical structures, each tailored to a certain class of compounds. For some cases, more than one approach may seem appropriate.

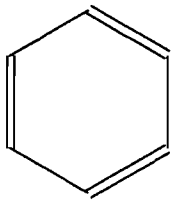
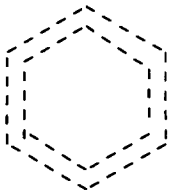
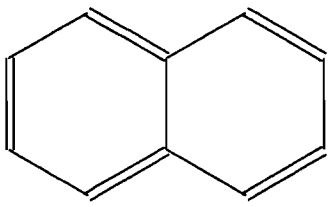
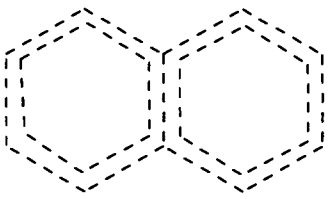
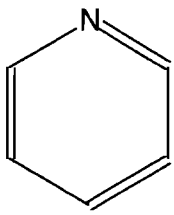
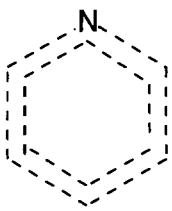
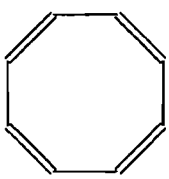
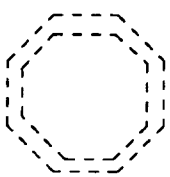
- even using the most appropriate way of depicting a structure, there may be more than one equally valid representation. The most important example of this is tautomerism.

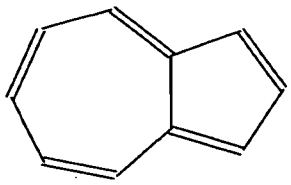
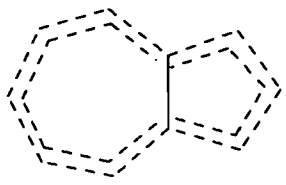
The following chapters explain the main rules used to represent structures in Markush DARC. More details are given in the *database appendices*.

II. Normalisation of bonds in rings

In all cyclic structures where there are closed cyclic paths comprising 2n atoms with n double bonds alternating with single bonds, the single and double bonds are replaced by *normalised bonds*.

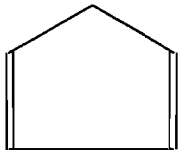
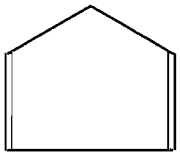
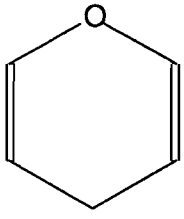
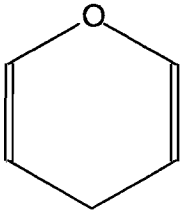
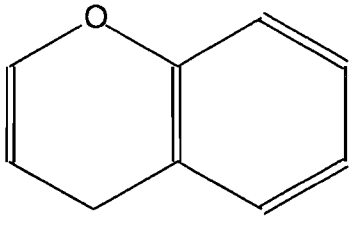
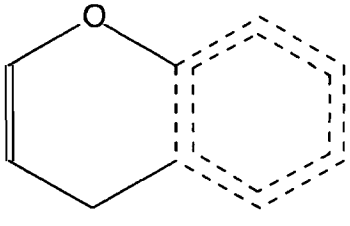
The following table lists examples of rings or ring systems where alternating single and double bonds are replaced by normalised bonds.

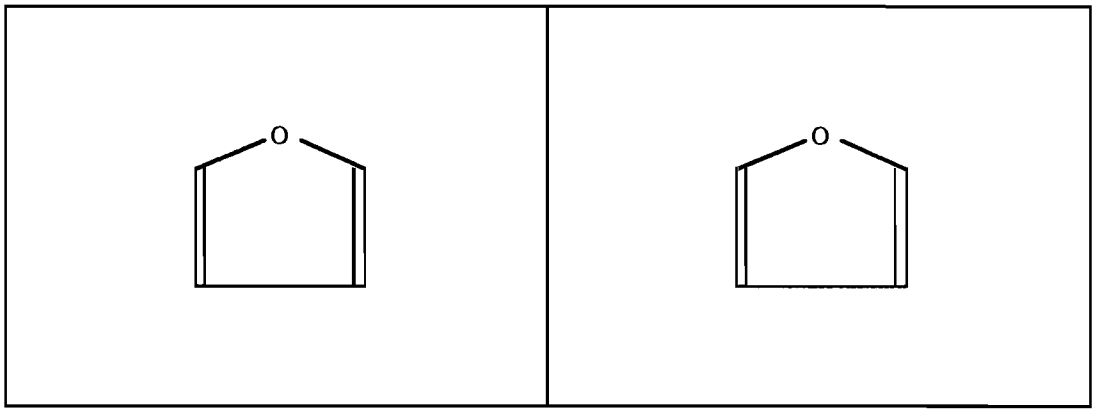
Ring or ring system	Markush DARC representation
	
	
	
	

Note that for the last ring of the above table, the bond common to the two rings of the ring system is not normalised.

The next table gives examples where single and double bonds are not normalised because they are not alternating.

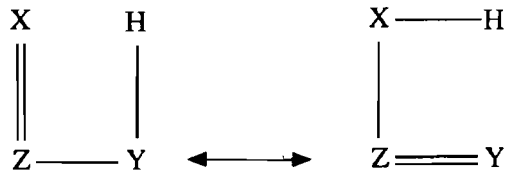
Ring or ring system	Markush DARC representation
	
	
	


O
O

III. Tautomerism

III.1. General case

The general case of tautomerism covers compounds of the type:

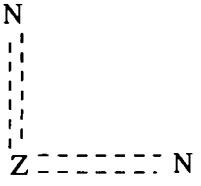


where Z is B, C, Si, N, P, As, S, Se, Te, F, Cl, Br, I

X and Y are O, S, Se, Te, N

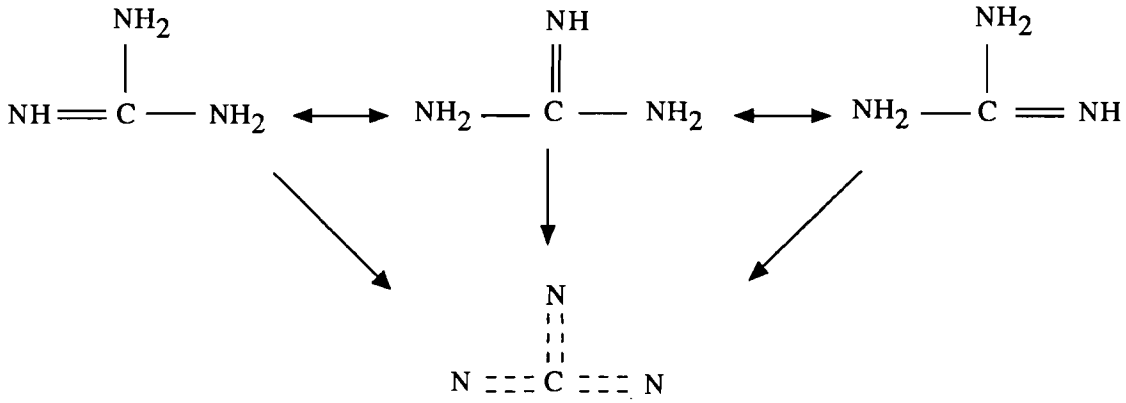
III.1.1. X and Y are both N

If X and Y are both N, normalised bonds are used in place of the single and double:



The following examples illustrate this rule:

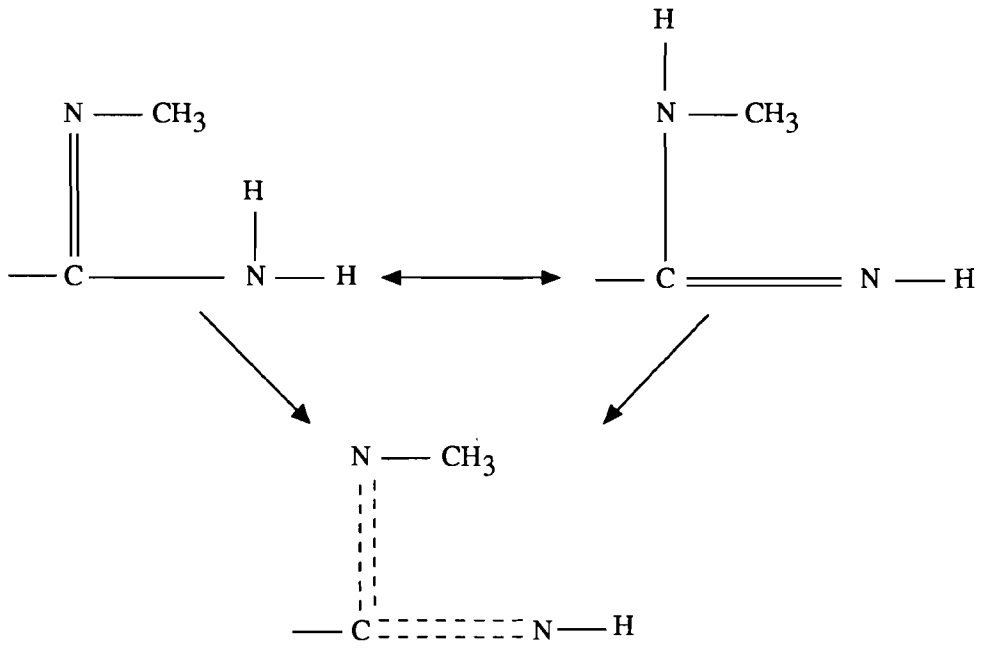
- normalisation of bonds in guanidine:



(normalisation)

Normalisation of bonds in guanidine

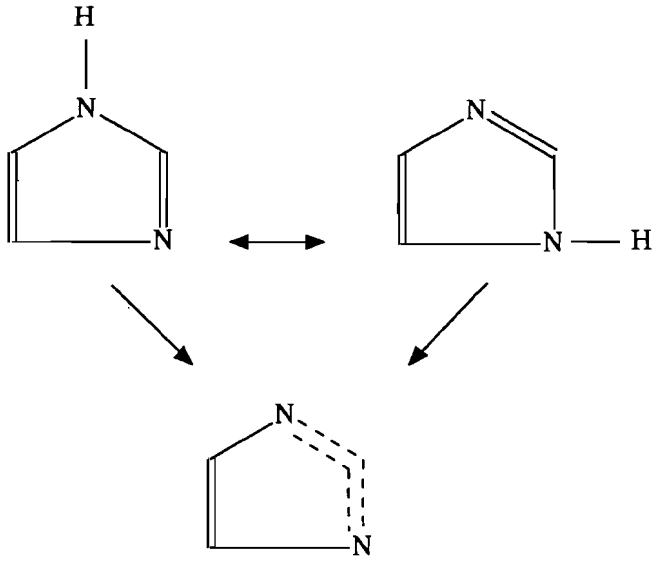
- normalisation of bonds in the N-methyl amidino group:



(normalisation of bonds)

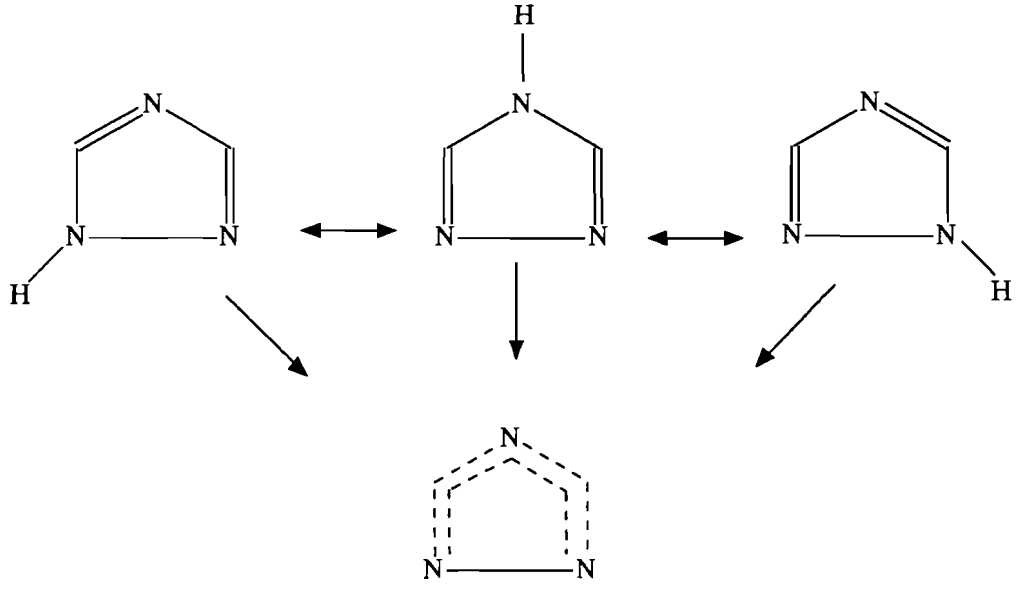
Normalisation of bonds in N-methylamidino

- normalisation of bonds in cyclic structures:



(normalisation)

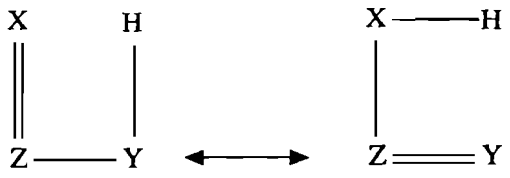
Normalisation of bonds in imidazole



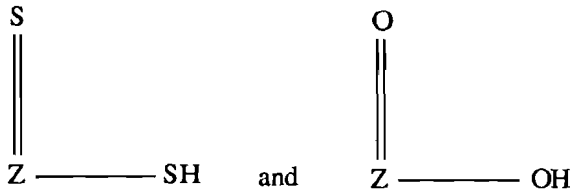
(normalisation)

Normalisation of bonds in 1,3,4-triazole

III.1.2. X and Y are the same but not N



If X and Y are the same but not N, the representation uses one single and one double bond:

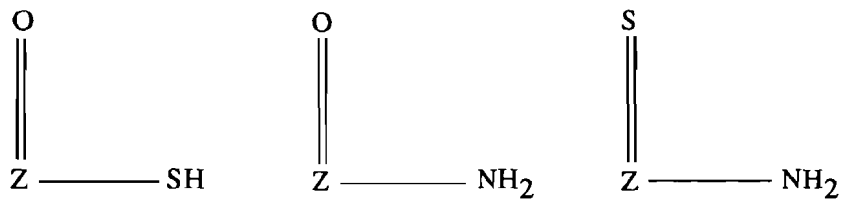


III.1.3. X and Y are different

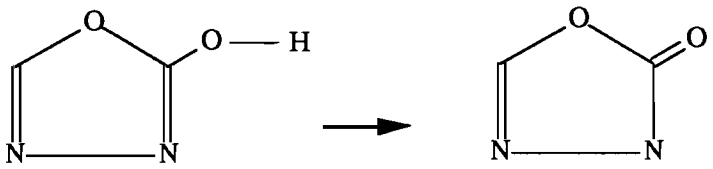
If X and Y are different, the double bond is placed preferentially on the first atom in the sequence:

O>S>Se>Te>N

as shown below:



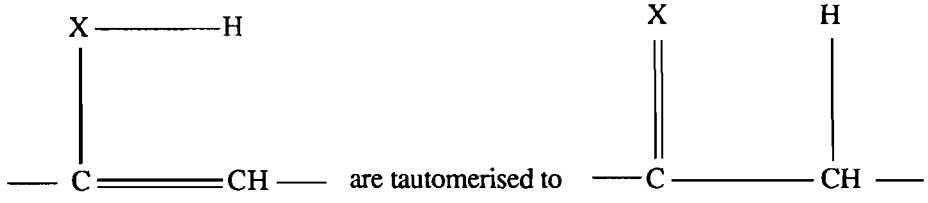
Example:



(the double bond is placed on the oxygen atom which is first in the sequence O>S>Se>Te>N)

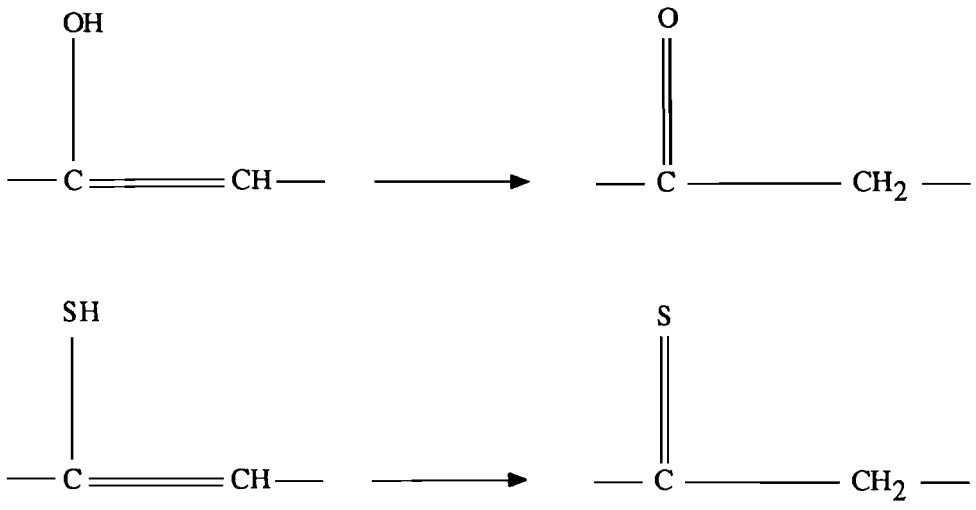
III.2. Keto-enol tautomerism

For unsaturated organic compounds containing a group capable of keto-enol tautomerism, the rule is to give preference to the keto form.

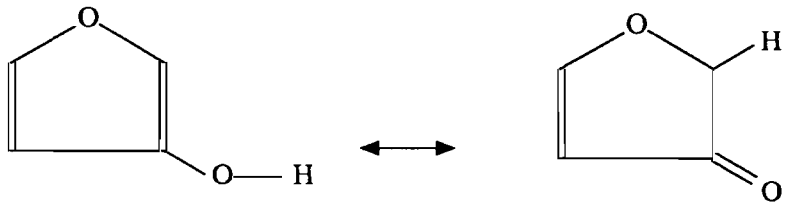


(where X is O, S, Se, Te) (keto form)

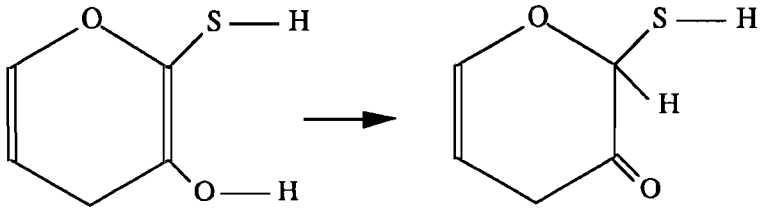
Examples:



Examples of keto-enol tautomerism



Example of keto-enol tautomerism in a cyclic structure

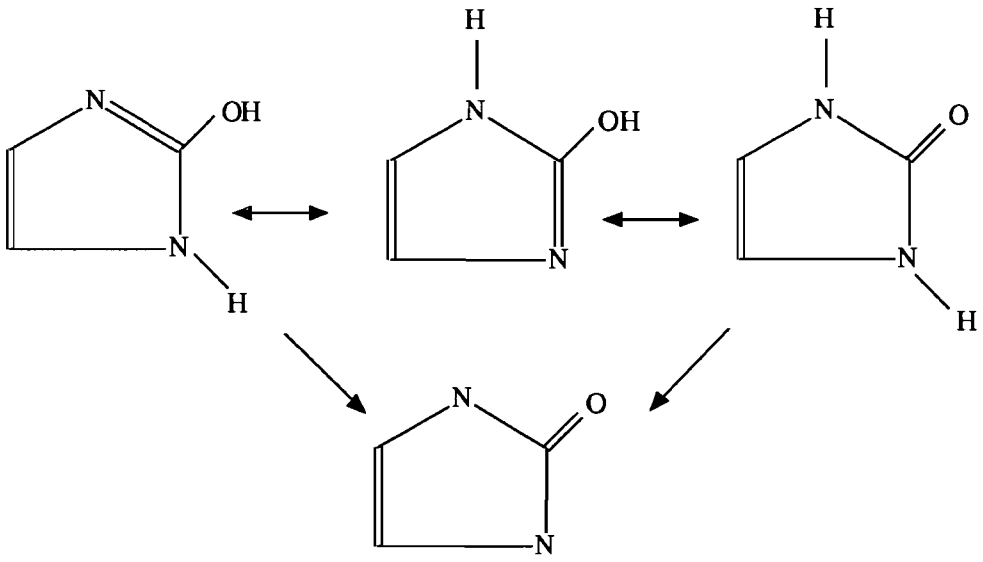


(tautomerisation to the keto form)

III.3. Conflict between tautomeric forms

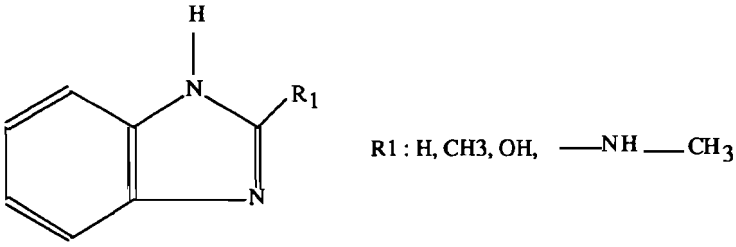
Sometimes, particularly in cyclic structures containing substituents, there may be a conflict between contiguous tautomeric forms or variable substituents which may cause different representations.

In the following example (2-hydroxy-imidazole), two sets of three atoms are involved in tautomeric representations: -N=C-O and -N=C-N-.

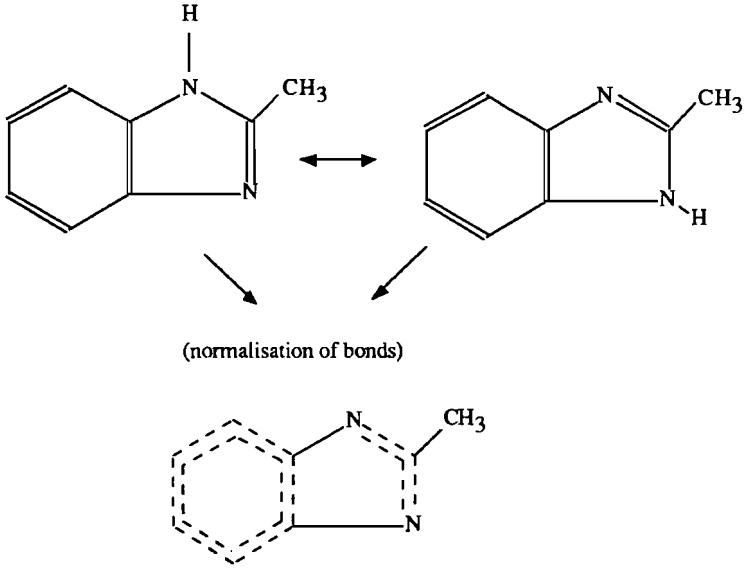


(Markush DARC representation)

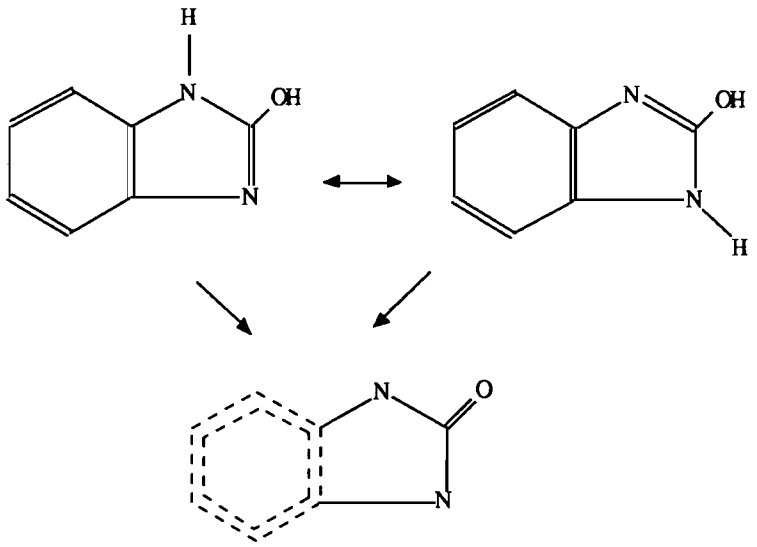
The next example shows how variable substituents may change the value of bonds (single, double or normalised):



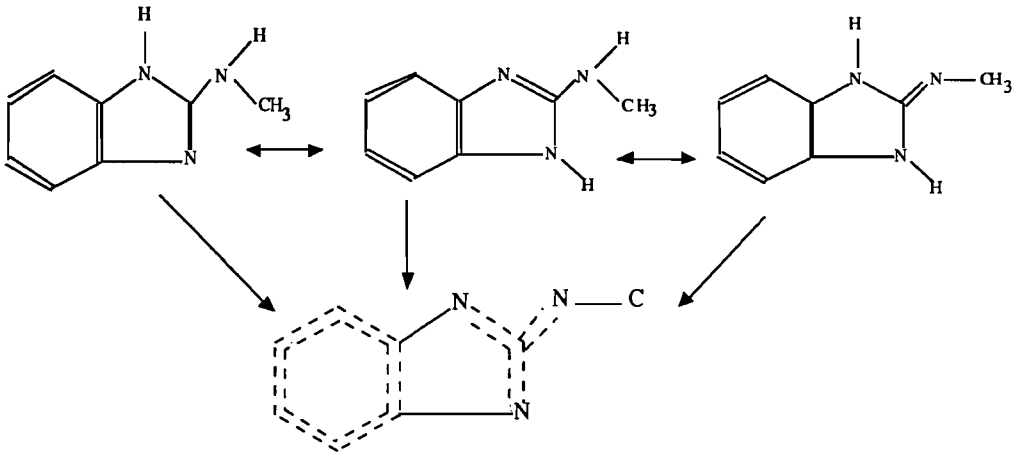
If R1 is CH3, the representation will be as follows:



If R1 is OH, the representation will be as follows:

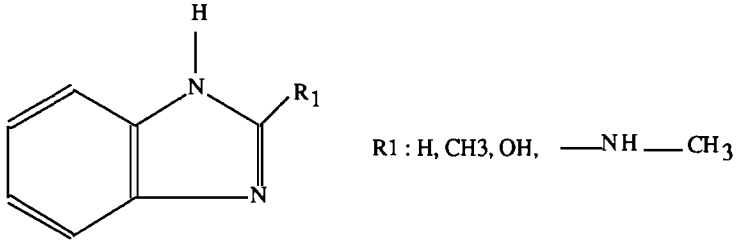


If R1 is -N-CH3, the representation will be as follows:

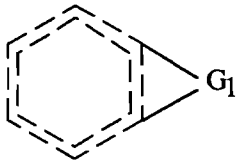


A query structure of this type would be represented as follows:

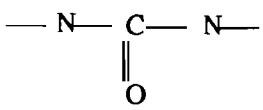
Query:



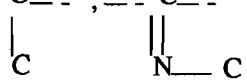
Query:



G1: — N = : G2 = : N — ,



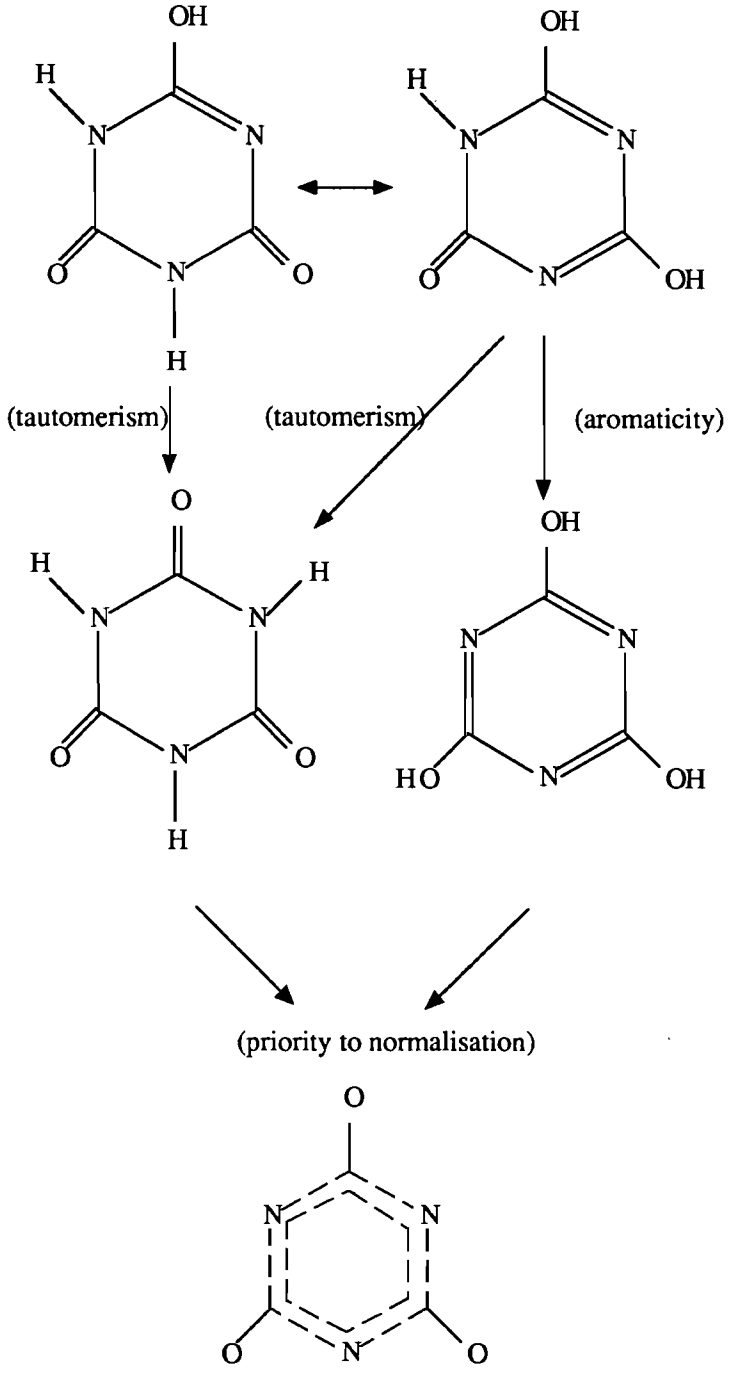
G2: = C = , = : C = : , = : C = :



IV. Priority of convention rules: preference for normalised representations

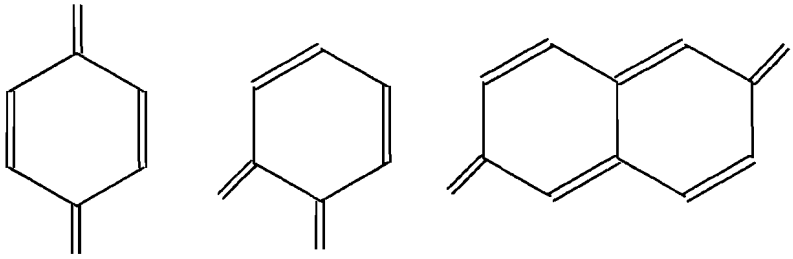
Normalisation of bonds in cyclic paths of 2n atoms containing n alternating multiple bonds has priority over the tautomerism conventions.

This is illustrated by the following example:

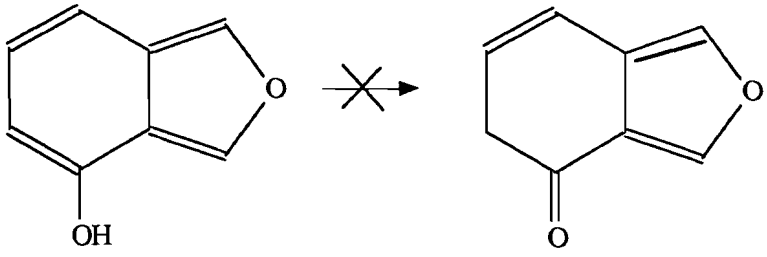


V. Quinonoid systems

Quinonoid systems are not normalised, unless there are further fused rings which make it possible.

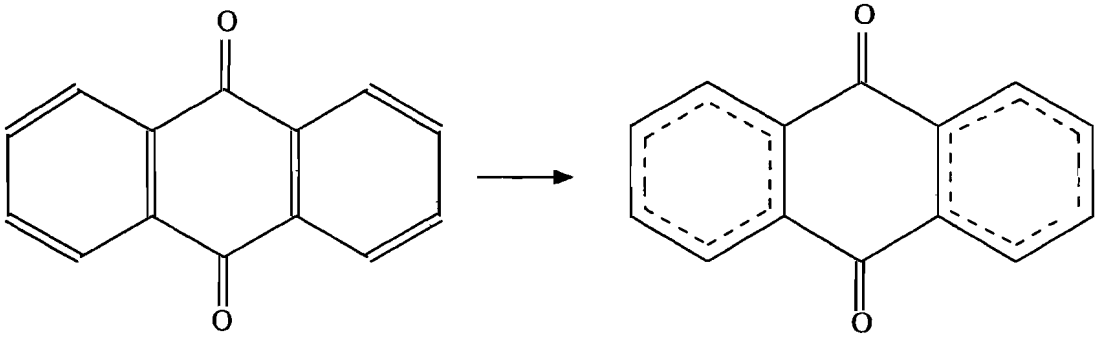


Example:



Representation of 4-hydroxy benzofuran (left)

Other example:



Representation of anthraquinone

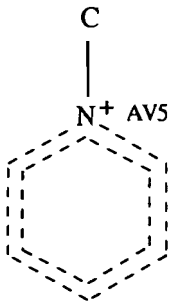
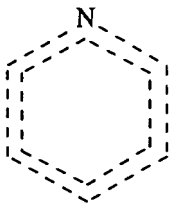
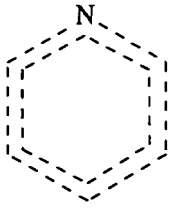
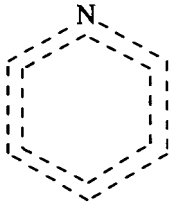
VI. Representation of salts

VI.1. General case

A multifragment representation is used for salts. The anionic part is disconnected from the cationic part: the two parts are represented as separate fragments.

According to the nature (organic or inorganic) of the anionic and cationic parts, and depending on the presence or absence of hydrogen, the anionic and cationic parts are represented in the charged or uncharged forms.

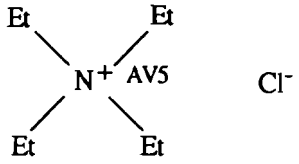
This is illustrated by the following examples, where the abnormal valency (AV) is shown if applied:

Cation/ anion	Inorganic acid (anion)	Organic acid (anion)
Inorganic base without hydrogen (cation)	Na^+ Cl^-	Na^+ CH_3CO_2
Inorganic base containing hydrogen (cation)	N Cl	N CH_3CO_2
Organic base without hydrogen (cation)	 Cl^- (1:1 , 1:1)	 CH_3CO_2 (1:1 , 1:1)
Organic base containing hydrogen (cation)	 Cl (1:1 , 1:1)	 CH_3CO_2 (1:1 , 1:1)

VI.2. Onium salts

Onium salts are represented with the charge located on the atom where it is normally drawn. Both ions are shown with charges unless the anion is derived from an organic acid.

Example:



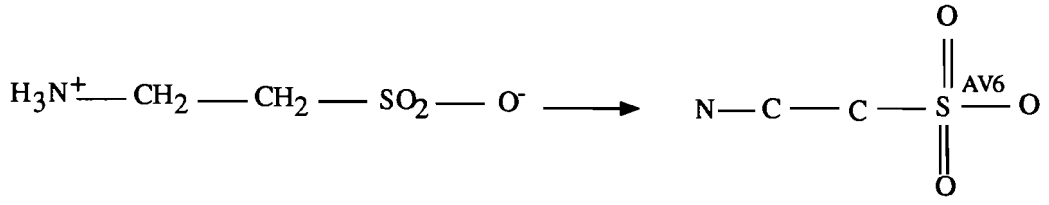
(1:1 , 1:1)

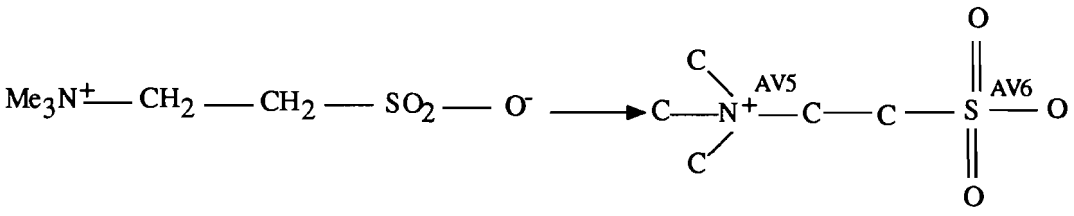
Representation of tetraethylammonium chloride

VI.3. Zwitterionic or ampholytic compounds

In zwitterionic compounds, if a H can be shifted from the positive part to the negative part of the structure to yield a neutral molecule, this is done.

Examples:



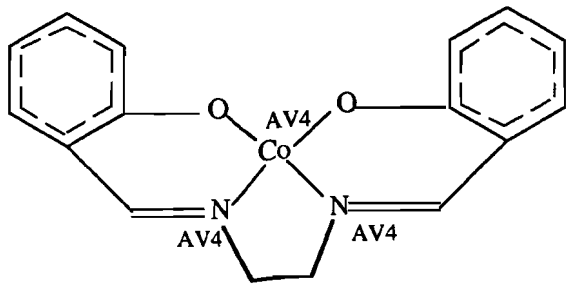


VII. Metal complexes and coordination compounds in Markush DARC databases

VII.1. Metal complexes

When the structure of the complex is known, it is represented with the metal bonded to the ligand.

Example:



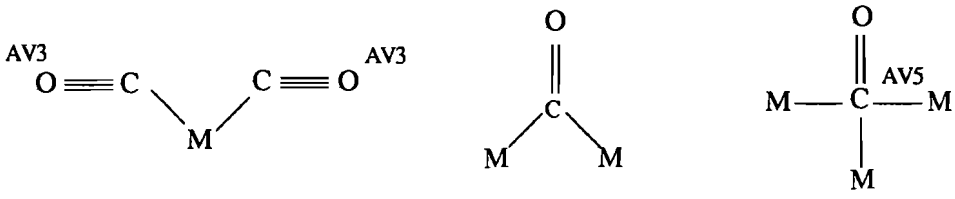
Cobalt complex of bis(salicylidene)ethylene diamine

If this is not the case, multifragment representation is used.
Thus, when the ligands are organic and the structure is not known, the representation is essentially the same as for simple organic salts.
The metal and the organic acid are represented as disconnected fragments, with the appropriate charge on the metal and no charges on the organic part.

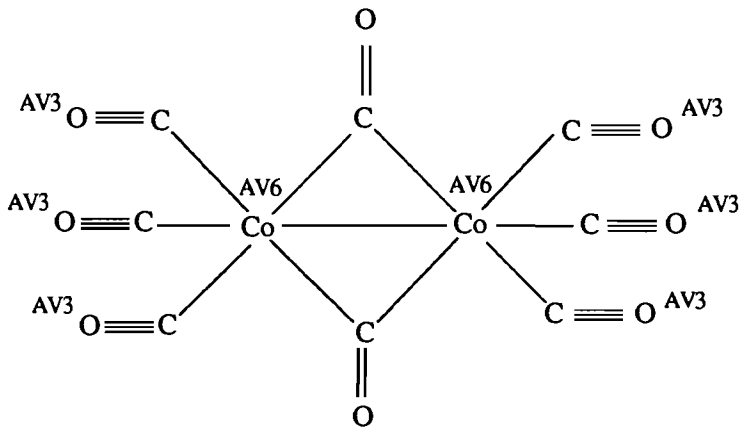
VII.2. Metal carbonyls

In metal carbonyl complexes where carbonyl groups are terminal, a single bond is shown from the metal to the carbonyl carbon.
The carbonyl oxygen is triply bonded to the carbon and has an abnormal valency of 3.
Bridging carbonyls are shown with single bonds between the carbonyl carbon and the two metal atoms, and the oxygen is doubly bonded to the carbon which has the normal valency.
Apical carbonyls will have three metal-carbon bonds, carbon with an abnormal valency of 5, and oxygen doubly bonded to the carbon (normal valency).

This is summarised below:



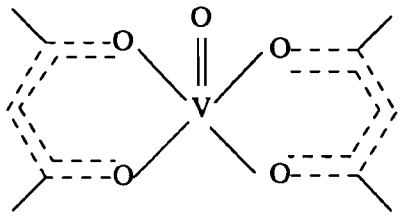
Example:



Representation of dicobalt octacarbonyl

VII.3. Acetylacetone and related complexes

In acetylacetone and related complexes, bonds are normalised between the carbonyl carbons and the carbon between them.
This is illustrated by the following structure:

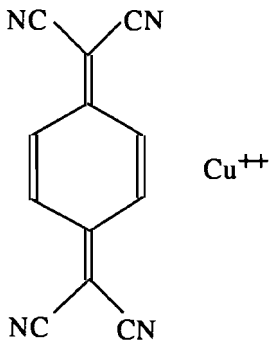


Representation of VO(acac)2

VII.4. TCNQ complexes or salts

Tetracyanoquinodimethane complexes (TCNQ complexes) are represented in salt-like form, with an appropriate charge on the metal, and no charge on the organic acid.

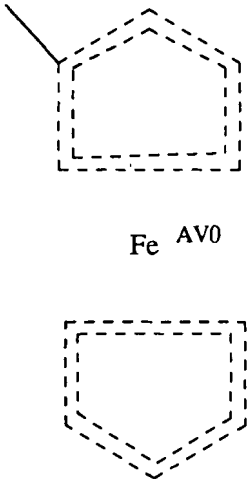
Example:



Representation of copper II TCNQ complex

VIII. Representation of multicentre Pi bonding

Dewar type Pi bonding found in metallocenes such as ferrocenes is handled by multifragment representation, as shown in the following example:



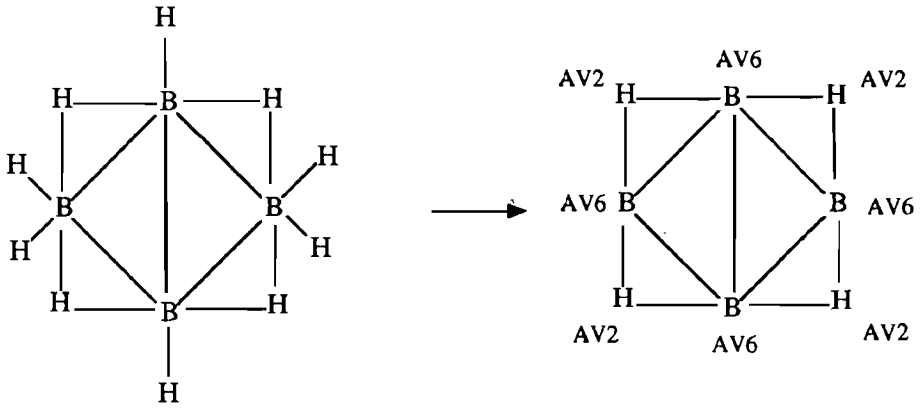
Representation of methyl ferrocene

Note that the Fe atom has an abnormal valency of zero, as it is disconnected from the cyclopentadiene rings.

IX. Boranes and carboranes

In boranes and carboranes, bonds between boron atoms and between boron and carbon are single.
Bridging hydrogen atoms are shown explicitly, with an abnormal valency of 2 and single bonds to each of the atoms joined to them.

Example:



B_4H_{10}

Section 5
Searching

I. Markush DARC searching capabilities

I.1. Compound Numbers

Markush structures and specific compounds are recorded in the Markush DARC structure databases with a numeric identifier, the Compound Number (abbreviated to CN).
Each Compound Number is unique and corresponds to only one Markush structure or specific compound.
Compound numbers make the link between the structure databases and the corresponding bibliographic files which contain the bibliographic patent information.

I.2. Searching capabilities

Markush DARC provides both Compound Number and structure search capability.

Compound number searching is useful to call up and display the structure record of a Compound Number which has been identified in another source, such as in the bibliographic files.

Compound number searching can also be used to identify the structure records associated with particular publication periods.

Structure searching is useful to identify species which contain either the full structure or substructure which is of interest.

The following structure searching options are provided with Markush DARC:

- RE search: fragment search (also called screen search) to retrieve candidate answers

- AA search: atom-by-atom search on the results from the RE search

    - on-line atom-by-atom search
    - batch atom-by-atom search

- SB search: bit-screen searching, used to perform searches for queries with few defined atoms and/or bonds

- BL search: boolean logic search to combine answer sets with boolean operators (and, or, not).

II. Compound Number searching

II.1. Format of Compound Numbers

Each database producer specifies the format used for the Compound Numbers in their database. Both WPIM and MPHARM use Markush structure record Compound Numbers and Specific Compound Numbers.

The formats of the Compound Numbers in the two databases are as follows:

	Markush Records	Specific Compounds
WPIM format	**YYWW-CCCSS** **YY**= last two digits of year **WW** = Derwent week number **CCC** = document identifier **SS** = integer in the range 01 to 99	**Rnnnnn** **nnnnn** = integer starting at 00001
MPHARM format	**YYMMXXXX-NN** **YY** = last two digits of year **MM** = month of patent publication **XXXX** = sequential number **NN** = integer in range 01-99	**RYYMMXXXX-NN** definitions same as for Markush Compound Numbers

II.2. Compound Number searching procedure

Compound Number searching can be done for either a single compound record or for a range of compound records. Compound number searching is performed by specifying the CN command at the ST prompt level.

The procedure to search for a single Compound Number or for a series of single Compound Numbers is:

- specify the CN command at the ST prompt level
- specify the Compound Number of interest at the system "CN ?" prompt
- specify a second Compound Number at the next "CN ?" prompt
- enter a carriage return at the "CN ?" to terminate Compound Number input.

The system then automatically performs the search and places the structures of the desired Compound Number(s) in the answer file.

These results can then be displayed using the answers display commands described in the next Section (*Section 6: Answer display and interpretation of the results*).

The Compound Number must be in the format used by the producer of the database being searched.

- WPAT example:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? CN

***  CN    ***

CN ? 8701-C6402
CN ? 8705-02803
CN ?


- CN / WPAT     -   R1 :      2 answer(s)


-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Example of a Compound Number search in WPAT

- IPAT example:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? CN

***  CN    ***

CN ? 87020024-01
CN ? 87080214-01
CN ?


- CN / IPAT     -   R1 :      2 answer(s)


-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Example of a Compound Number search in IPAT

II.3. Compound Number range searching

The searching procedure for a range of Compound Numbers is:

- Specify the CN command at the ST prompt level
- Specify the first (lowest) Compound Number in the range of interest at the system supplied "CN ?" prompt
- Enter a colon (:)
- Specify the second (highest) Compound Number in the range followed by a carriage return
- Specify the next single Compound Number or range of interest at the next "CN ?" prompt
- Enter a carriage return at the "CN ?" to terminate Compound Number input.

The system then automatically performs the search and places the structures of the desired Compound Number(s) in the answer file.

It is not necessary that the lowest and the highest Compound Numbers in a range actually exist in the database.

- WPIM example:

Find all structure records in Derwent week 8701 (first CN is 8701-A1501):

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? CN

***  CN    ***

CN ? 8701-A1501:8701-99999
CN ?


- CN / WPIM     -   R1 :    508 answer(s)


-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Example of a Compound Number range search in WPIM

- MPHARM example:

Find all structure records in MPHARM which are associated with patents published in January of 1993:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? CN

***  CN    ***

CN ? 93010001-01:93010999-99
CN ?


- CN / MPHARM   -   R1 :    283 answer(s)


-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ?
```

Example of a Compound Number range search in MPHARM

Searching with truncated CN's is also available in Markush DARC, as shown in the following example:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? CN

***  CN    ***

CN ?..8701+
CN ?


- CN / WPIM     -   R1 :    508 answer(s)
```

Truncated CN search

III. Structure searching

III.1. RE and AA search

III.1.1. Search mechanism

The Markush DARC structure search software employs a two step search process.

- RE (REtrieve) search (first step):
    - automatic generation and ranking of fragments
    - retrieval of candidate answers by fragment search
- AA search (second step): Atom-by-Atom search on the results from the RE search
    - on-line atom-by-atom search
    - batch atom-by-atom search

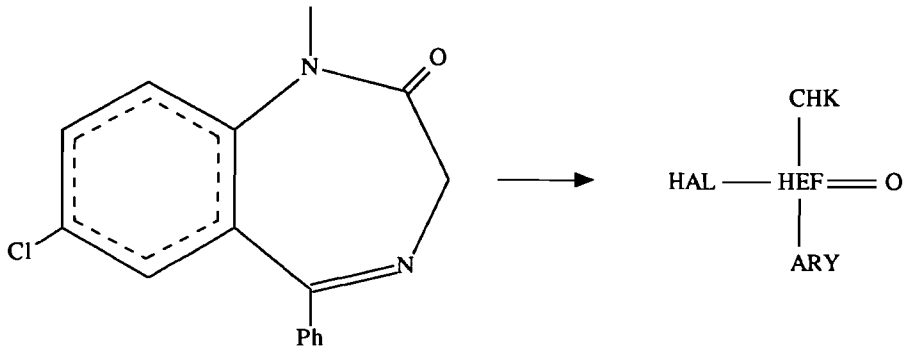
```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? re

***  RE    ***

RESULT    :        2548
NEXT LIST :        2633
CONTINUE ? (Y/N) y
RESULT    :        2515
NEXT LIST :       20559
CONTINUE ? (Y/N) y
RESULT    :        2462
NEXT LIST :       24264
CONTINUE ? (Y/N) y


- RE / WPIM     -   R0 :    123 answer(s)


-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? aa

***  AA    ***


-   AA    - NUMBER OF ANSWERS          :     9
- FILE RX - NUMBER OF CANDIDATES       :     0
- CANDIDATES REMAINING TO BE PROCESSED :     0

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ?
```

Search example

The RE search is fragment (or FREL) based screening step to eliminate those candidates that are not at least similar to the target compound.
Fragments are generated by the system "looking" at an atom and its' neighbours and constructing fragments in this way.
So long as an atom is connected by a defined bond to at least one neighbour, fragments can be generated and compared with those stored in the database.

The rarest fragments (i.e. those with the least number of "postings") are compared first, to eliminate as many candidates as possible in the first step.
After the RE search is completed, an automatic Bit Screening search is carried out, where bits are set according to type and number of atoms and bonds present.

In order with the FRELs and therefore the RE to work with the translation capability and the variable attachments feature, each structure is collapsed into its corresponding superatom(s) before the FRELs are worked out.
So a benzene ring would collapse to an ARY superatom, a Cl atom would become a HAL superatom and so on as in the following example.



In practice, in order to make the RE as selective as possible, the more functional groups that can be included in the structure, the better.

The answers are automatically stored in the R0 set and remain in this set until the next RE search.
It is possible to display the results of an RE search (see *Section 6: Answer display and interpretation of the results*).
The results of a RE search can be saved for further processing using the boolean logic search capabilities (see the *Boolean logic chapter* in this section).

III.1.2. RE limitations

III.1.2.1. Specificity of the query

If the query does not contain at least one fragment, i.e. a group of three defined atoms linked by defined bonds, the RE search is not possible. In this case, a SB search is automatically performed by the system when the RE command is entered.
If your query has at least three atoms, but these atoms are not defined (e.g. X atoms) and/or are linked with undefined bonds (X bonds) or Z lists of bonds, you should either modify your query and make it more specific or carry out an SB search.

III.1.2.2. Limitation on the number of candidates

There is a limit on the number of candidates which can be kept as a result of an RE search in the R0 file for further AA processing.
This limit has been set up at 30000 candidates, for the present.

If the number of candidates is over this limit, the following message will be sent by the system:

```
***  RE    ***

SEARCH INTERRUPTED - MORE THAN 30000 ANSWERS
DO YOU WANT TO KEEP YOUR ANSWERS (Y/N) ?
```

Limitation on the number of candidates

III.1.3. AA search

The Atom-by-Atom search (AA) search is performed on the candidate answers resulting from the RE/SB search.

The AA search operates progressively and automatically on sets of candidates.

Example:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? AA

***  AA    ***

     28 ANSWERS FOR    575 CANDIDATES
     67 ANSWERS FOR   1362 CANDIDATES
```

Note:

- The sets of candidates which are processed are related to Darc Numbers (DN's, i.e. DARC internal numbers) and not to Compound Numbers (CN's).

It may happen that the total number of candidates which are processed through the AA search is higher than the actual result of the RE search, which is related to Compound Numbers.

This occurs since Markush DARC databases may contain structures, particularly those having variable points of attachment, which are identified with one "displayable CN" (i.e. the one you actually see when you display the answers) and several searchable CN's (i.e. the ones which are actually compared with the query during the search process, and which are identified by corresponding DN's).

- The actual number of results from the AA search is normally less than the number of results from the RE search.
This occurs since the RE search identifies candidates which contain various fragments which are in the query, but the RE search has no mechanism to check the relationship between the fragments.
The AA search requires that all elements in the query be in the answer and that the elements have the same relationship as specified in the query.

III.1.4. AA limitations

III.1.4.1. CPU time limits

CPU time limits have been set up for the AA search process.

A first limit concerns the CPU time needed for processing one compound.
If the CPU time exceeds a certain limit, the compound is placed in a response file called "File RX".
It may or may not be an answer to the query.
The contents of the RX file can be viewed using the VI RX command, saved using the SV RX command or processed in batch mode.

A second limit concerns the overall CPU time for the AA search process.
If this overall CPU time exceeds a certain limit, the AA process is interrupted.
The remaining candidates from the RE search can be processed in a batch search if there are less than 5000 remaining candidates, or a further AA can be carried out.
If there are more than 5000 remaining candidates to be processed, the following message is sent by the system:

```
-   AA    - NUMBER OF ANSWERS           :   873
- FILE RX - NUMBER OF CANDIDATES        :     1
- CANDIDATES REMAINING TO BE PROCESSED : 15017

BATCH IMPOSSIBLE - MORE THAN   5000 CANDIDATES

CONTINUE AA (A) OR CANCEL (C) ?
```

The following table lists the possible messages which are given by the system according to the different situations.

CPU limit for ONE comp-ound	CPU limit for overall AA	Message
not reached	not reached	- AA - Number of answers :X - File RX - Number of candidates :0 - Candidates remaining to be processed :0
reached	not reached	- AA - Number of answers :X - File RX - Number of candidates :Y - Candidates remaining to be processed :0 Perform batch search on file RX (Y/N) ?
not reached	reached	- AA - Number of answers :X - File RX - Number of candidates :0 - Candidates remaining to be processed :Z ≤5000 - Continue AA(A),Perform batch search(B),Cancel(C): ?
not reached	reached	- AA - Number of answers :X - File RX - Number of candidates :0 - Candidates remaining to be processed :Z >5000 Batch impossible - more than 5000 candidates - Continue AA(A) or Cancel(C): ?

After the message:

```
- Continue AA(A),Perform batch search(B),Cancel(C): ?
```

- typing A will restart the AA search with the same CPU limits; the following message is displayed by the system:

```
*** AA *** continued
```

- typing B will request a batch search for all candidates from RE (a maximum of three batch searches per day is allowed); a message such as the one shown below is given:

```
BATCH SEARCH REQUESTED, PLEASE ENTER A NAME ( 8 CHAR.) ? CEPHALO

03/05/93 19*16*59 BATCH SEARCH * :    1, NAME : CEPHALO
```

Batch search request

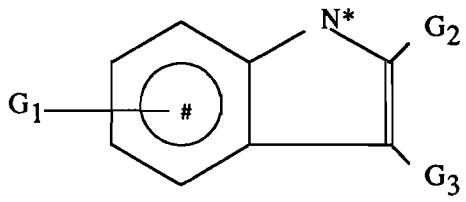
- typing C will cancel the search.

III.1.4.2. Limitation on the number of answers

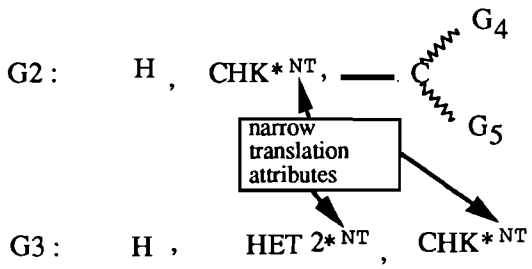
The AA search is interrupted if there are more than 5000 answers.

III.1.5. Search example

Let us consider the query that we have been using from the beginning of this manual.



G1 : H , N , NO_2



G4 : O , N 2*

G5 : H , N* , CHK*

After the query input step, we have the possibility of imposing some restrictions on the answers, using the *Other Specifications* capability. We also have the option of specifying *File Segments*. We can then run the RE search.

With our query example, we do not impose any restriction.

Let us first select the WPAT database. The search is performed as follows:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? re

***  RE    ***


- RE / WPAT     -   R0 :    273 answer(s)


-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? aa

***  AA    ***

-   AA    - NUMBER OF ANSWERS          :    12
- FILE RX - NUMBER OF CANDIDATES       :     0

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

RE/AA search in the WPAT database

There are 12 answers, and all the candidates have been processed.

Let us switch now to the IPAT database, using the BA command:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? BA

****    LAST SELECTED DATA BASE :    WPAT      ****
        ELAPSED TIME       :      3.03


***   AVAILABLE DATA BASES   ***

           DATA BASE  :   WPAT      (  1 )
           DATA BASE  :   IPAT      (  2 )
           DATA BASE  :   MPHARM    (  3 )
           DATA BASE  :   WPIM      (  4 )

***     4 DATA BASE(S) AVAILABLE   ***

DATA BASE # ?  2
****    BASE IPAT              -   31/03/93    ****
      1520 COMPOUNDS - LAST CN : 87110081-01
```

Switching from the WPAT to the IPAT database using the BA command

and perform the search in this file:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? RE

***  RE    ***


- RE / IPAT     -   R0 :    442 answer(s)


-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? AA

***  AA    ***


-   AA    - NUMBER OF ANSWERS          :    17
- FILE RX - NUMBER OF CANDIDATES       :     0

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

RE/AA search in the IPAT database

There are 17 answers, and all the candidates have been processed.

III.2. Batch search

III.2.1. Batch search process

The batch search itself works as summarised in the following table:

	Batch process	CPU limits for the batch search	Number of batch searches per day
Candidates in file RX	on file RX candidates only	upgraded CPU time limit for the processing of one Markush	maximum of 3 batch searches per day above this limit the batch process on file RX is not possible `You cannot process your file RX in batch. You have reached your maximum number of batch requests`
Candidates remaining to be processed	on all candidates of the RE search again*	- for the processing of one Markush (higher than on-line) - for the overall batch AA process (higher than on-line)	maximum of 3 batch searches per day above this limit the batch search is not proposed any longer `- Continue AA(A) or Cancel(C): ?`

*: if a batch search is required when there are remaining candidates to be processed, the AA search is carried out on all the candidates resulting from the RE search, and not only on the candidates remaining to be processed.

III.2.2. Batch search results

At the first connection to Markush DARC, the status of the batch search requests is given by the system.

The following table shows all the possible parameters which can be listed:

NAME	STATUS	RE/RX	DATABASE	CAND	ANSWERS	RX	DATE	TIME
XXXX	COMPLETED	RE	WPIM	nnnn	nnnn	nn	JJ/MM/AA	HH*MM*SS
"	WAITING	RX	"	"	0	0	"	"
"	INTERRUPT	SB	"	"	0	0	"	"
"	MAX.RX		"	"	"		"	"
"	MAX.CPU		"	"	"		"	"

The status of the batch search can be as follows:

COMPLETED:

The search has been completed, results are available.

WAITING:

The batch search request has not been processed yet (or is in the stage of being processed).

INTERRUPT:

Processing has been interrupted for operating reasons.
The search will be processed later.

MAX.RX:

Too many compounds have exceeded the CPU time limit set up for the batch processing of one Markush (RX file).
The maximum number of Markush structures allowed in the RX file is 400.

MAX.CPU:

The CPU time limit set up for the overall batch AA process has been exceeded.

The following operations are possible with batch search results:

- you can recall the results of a batch search at the ST level, using the RF BT command.

  Answers of a recalled batch search are automatically stored in the current answer set R1 (see the following chapter - *Boolean logic* - for details on available answer sets).

- you can recall the query corresponding to a batch search at the ST or QU levels, using the RF QB command.

- a history of batch search requests can be displayed at ST or QT levels.
The list displayed is the one that appears on connection to Markush DARC.

- you can erase a batch search at the ST or QU levels.

The system asks the user to confirm the deletion of the batch request.
The confirmation message is preceded by one of the following messages indicating the status:

```
- BATCH RESULT:XXXX NOT USED
- BATCH SEARCH:XXXX NOT COMPLETED
- BATCH SEARCH:XXXX INTERRUPT, TOO MANY RX REJECTS
- BATCH SEARCH:XXXX INTERRUPT, TIME LIMIT EXCEEDED
- BATCH SEARCH:XXXX INTERRUPT
```

There are two cases where erasing is not possible:

```
- NO BATCH SEARCH NAMED:XXXX
- BATCH SEARCH:XXXX ON HAND
```

Candidates in the RX file are stored in an answer set labelled RX that can be used for display and saving purposes.

III.3. SB search

The RE search requires that you have at least two defined bonds and three defined atoms in your query.
In some specific cases, you may want to use a query which has fewer atoms and bonds than this minimum for RE search.
In Markush DARC it is possible to perform such searches using the Bit Screen search capability (SB command).

The SB command will perform a Bit Screen search to retrieve candidates.
This SB search is used in place of the RE search.
The SB search must normally be followed by an atom-by-atom (AA) search.

Caution:

- the query must be selective enough to produce a reasonable number of candidates.
- the use of "Other Specifications" (see the *other specifications* paragraph in the *Query input procedure* chapter) is highly recommended to reduce the number of candidates before using the SB command.

Example:

Search for platinum compounds in IPAT using the following query:

Pt6*

The search using the SB command is performed as follows:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  GR
*GRAPH
?  1
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AT
*ATOMS
?  PT
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  FS
*FREE SITES
?  6
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  FI
OTHER SPECIFICATIONS (Y/N) ? N
FILE SEGMENTS(Y/N) ? N
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? SB


***  SB    ***



- SB / IPAT      -   R0 :    307 answer(s)


-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? AA

***  AA    ***


-   AA    - NUMBER OF ANSWERS          :    19
- FILE RX - NUMBER OF CANDIDATES       :     0
- CANDIDATES REMAINING TO BE PROCESSED :     0

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Example of SB searching

III.4. Summary of RE, SB, AA and batch search limitations

The limits which have been set up for the RE, SB, AA and batch search processes are summarised as follows:

- maximum number of candidates kept as a result of a RE or SB search: 30000

- maximum number of candidates processed in batch: 5000

- maximum number of candidates in file RX (AA or batch search): 400

III.5. Boolean logic search

III.5.1. Markush DARC answer sets

The Markush DARC boolean logic search capability has three answer sets, named R1, R2 and R3, for storing the answers of boolean operations.

The results of a current search (RE + AA) are automatically stored in answer set R1 unless specified otherwise.

The results of a search can be saved for further recalling and processing using the SV CN command.

Example:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? SU CN
NAME ( 8CHAR.) ? INDOL
      NAME   TYPE NUMBER    ORIGIN        DATE   SPACE  DATABASE
 1  INDOL    CN       19   MARKUSH DARC  09/09/93   1    IPAT

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Saving Compound Numbers (CN) using the SV CN command

Note:

The same operations can be performed using the DARC Numbers (DN), which are internal numbers assigned by Markush DARC to every structure stored in the database in addition to the Compound Numbers (CN).

The boolean logic search capabilities are the following:

- answer sets can be combined with the boolean logic operators AND, OR or NOT
- stored answer files can be recalled and combined with boolean operators
- the current answer set can be combined with any recalled answer file previously stored
- the results of the current search (which are automatically stored in the R1 file unless specified otherwise) can be stored in R2 or R3
- the candidate answers contained in the current R0 or RX files can be stored in one of the R1, R2 or R3 answer sets.

Examples:

R3 = R1 OR R2	combines the set of answers stored in R1 with the set of answers stored in R2 using OR logic and stores the results of the boolean operation in set R3
R3 = R2 NOT R1	removes the answers stored in set R1 from the set of answers stored in R2 and stores the results in set R3
R2 = * AND R1	combines the answers from the current search (*) with the answers stored in set R1 using AND logic, and stores the results in set R2
R3 = *	stores the results of the current search (*) in set R3
R2 = R0	stores the results of the last RE search (automatically stored in set R0) in set R2

III.5.2. BL search procedure

The boolean logic search can be performed using the BL command at ST level.

Each time you select the BL command, the system will display a list of the answer sets R1, R2, and R3 in the order in which the searches were performed. This display is provided as a quick review of the contents of each answer set.

Example:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? BL

***  BL    ***


** LIST OF VALID ANSWER FILES **
   R1  : AA                                                  15 ANSWER(S)
   R3  : BL      R3  =   *                                    2 ANSWER(S)
   R2  : BL      R2  =   R1  NOT *                            9 ANSWER(S)

EQUATION ?
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ?
```

BL search: list of valid answer files

In this example:

- the first search was an AA search, with a result of 15 answers, automatically stored by default in answer set R1

- the next search was a BL search: a BL search was carried out in order to store the results of the search with the current query (*) in the R3 answer set

- the last search was a BL search: a BL search was carried out in order to combine the results of the current search with the results stored in R1, using the NOT boolean operator.

The operation to be performed is specified after the prompt " EQUATION".

Examples:

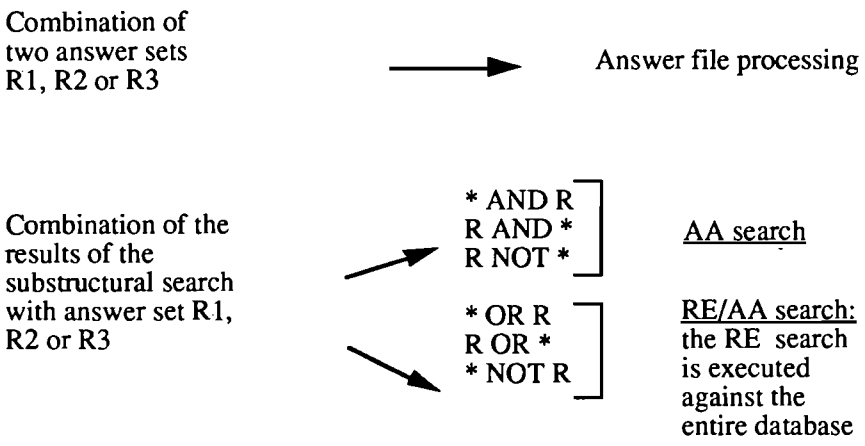
```
EQUATION ? R2 = R1 NOT *

EQUATION ? R3 = R1 OR R2
```

BL search: examples of possible equations

The first equation eliminates the results of the current search (*) from the R1 answer set and stores the results in the R2 answer set.

The second equation combines the R1 and R2 answer sets with the OR boolean operator and stores the result in the R3 answer set.

After the equation has been entered, the system automatically executes the search as follows :



BL search mechanism

Note:

- For the OR logic, the system gives the choice for either a RE/AA search or a SB/AA search.

```
EQUATION ? R2 = * OR R1

POSSIBLE SOLUTIONS  :
   RE   AA   OR   (1)
   BS   AA   OR   (2)
?  1
```

Possible solutions for a BL search with OR logic

- When phrasing an equation such as R1 = * NOT R3, the search is performed against the full database first and then the system eliminates the compounds of the R3 answer set.

- When phrasing an equation such as R1 = R3 NOT *, the system eliminates from the R3 answer set the answers of the sub-structural search.

III.5.3. Storing the current answer file

When performing the usual RE + AA search, the results are automatically stored in answer set R1.

RE + AA ----> R1

If you want the results of the current search to be stored in a different answer set (R2 or R3), use the following procedure:

- select the BL command
- specify the equation:

R2 = *

The result of the current search will then be stored in the R2 answer set.

III.5.4. Recalling a previously stored answer file for use

When recalling a saved answer file, the RF CN command automatically stores the answers in answer set R1.

RF CN -----> R1 (by default)

If you want to store the recalled answers in the R2 or R3 answer set, you can specify the answer set desired with the RF CN command.

Example:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? RF CN R2
NAME ( 8CHAR.) ? INDOL
      NAME   TYPE NUMBER    ORIGIN        DATE   SPACE  DATABASE
 1  INDOL    CN      19   MARKUSH DARC  09/09/93   1    IPAT


- RF / IPAT     -   R2 :     19 answer(s)


-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Recalling Compound Numbers (CN) in the R2 answer set using the RF CN command

III.5.5. Using answer sets R1, R2 and R3 for saving and displaying answer files

The system automatically performs a temporary save (which is kept until the following Saturday) of the R1, R2 and R3 answer sets upon logoff.

The R1, R2 or R3 answer sets can be displayed or saved at any time.

To determine the content of the R1, R2 and R3 answer sets, use the BL command.

The last listed answer set, R2 in the example, is the last set of answers retrieved.

To return to the ST prompt, press Return after the "EQUATION ?" prompt.

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? BL

***  BL    ***


** LIST OF VALID ANSWER FILES **
   R1  : AA                                          15 ANSWER(S)
   R3  : BL      R3  =   *                            2 ANSWER(S)
   R2  : BL      R2  =   R1  NOT *                    9 ANSWER(S)

EQUATION ?
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ?
```

BL search: list of valid answer files

The LI, VI, VI MAX, VI FO, and SV commands apply by default to the last answer set which was retrieved (R2 answer set in the above example).

The R1, R2 and R3 parameters can be specified directly with the LI, VI or SV commands (see *Section 6: Answer display and interpretation of the results* for more details on the LI and VI commands and *Section 7: Cross-file searching* for more details on the SV command).

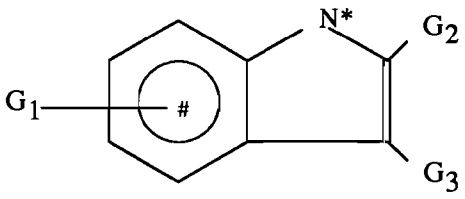
The R0 and RX parameters can also be specified.

Examples:

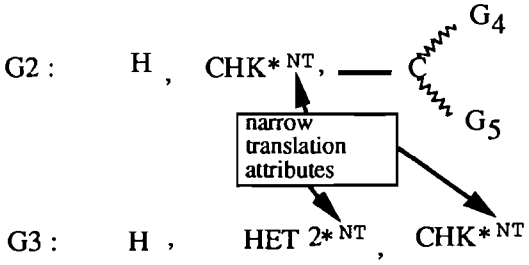
LI R1	lists the CNs of answer set R1
VI	displays the last retrieved answer set
VI R3	displays the R3 answer set
VI RX	displays the RX answer set
SV DN	saves the DNs of the last retrieved answer set
SV CN R1	saves the CNs of answer set R1
SV DN R0	saves the DNs of the R0 answer set

III.5.6. BL search examples

Let us consider the example that we have been using from the beginning of this manual:



G1 : H , N , NO_2



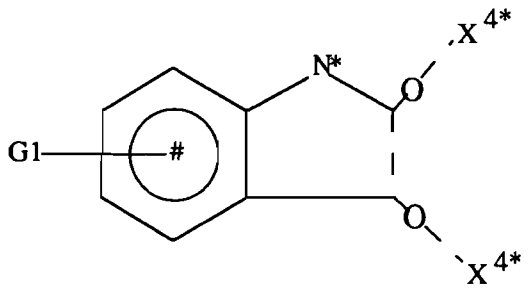
G4 : O , N 2*

G5 : H , N* , CHK*

We will call this query "main query" hereafter in this chapter.

III.5.6.1. OR logic

Let us now consider the following query:



G1 : H, N, NO2

We want to combine this query with the main query where we did not allow the possibility of a fused ring between positions 2 and 3 of the indole ring.

In the above query:

- the use of atom X (any atom excluding hydrogen) in positions 2 and 3 of the indole ring, linked with cyclic X bonds to the ring, and

- the use of an X bond (undefined nature) between the positions 2 and 3 of the indole ring

will allow the retrieval of fused indole rings.

We will use the BL command and combine the results of the previous query, which are stored in R1, with the above query (current query, labelled * in the BL search equation), using the OR boolean operator.
We will thus add the results of the current query to the results of the previous query.

We will ask the system to store the results in the R2 answer set.

This BL search in the IPAT database is performed as follows:

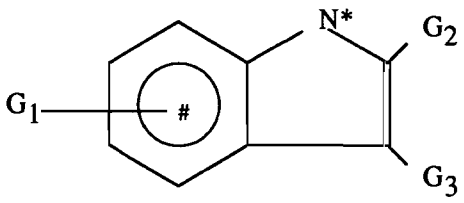
```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? BL

***  BL    ***


** LIST OF VALID ANSWER FILES **
   R0  : RE                                              442 ANSWER(S)
   R1  : AA                                               17 ANSWER(S)

EQUATION ? R2 = * OR R1

POSSIBLE SOLUTIONS  :
   RE   AA   OR   (1)
   BS   AA   OR   (2)
?  1


- BL / IPAT     -   R2 :     24 answer(s)


-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Example of a BL search with OR logic

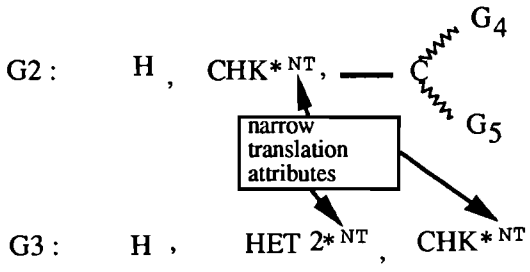
Note that in this case the system offers the choice for either a RE/AA or a SB/AA search. As our query contains more than three atoms, we choose RE/AA.

III.5.6.2. AND logic

Let us still consider the main query:



G1 : H , N , NO_2

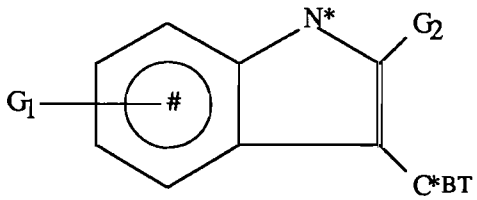


G4 : O , N 2*

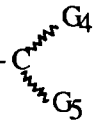
G5 : H , N* , CHK*

Suppose that, among the 17 answers obtained with this query, we are interested only in compounds having an optionally monosubstituted carbon substituent in position 3 of the indole ring.
We will use the following query, where we have replaced G3 of the main query with a C atom carrying one free site and a BT (Broad Translation) attribute.

We will combine this query with the results of the main query, using the AND boolean operator.



G1 : H, N, NO2
G2 : H, CHK*, NT
G4 : O, N 2*
G5 : H, N*, CHK*



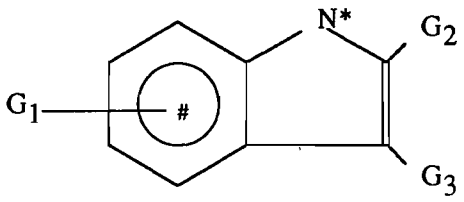
(In this query, translation attributes in group G2 are the same as in the main query).
The results of the search with the main query (17 answers) are still in the R1 answer set.
The results of the BL search with the OR logic are in the R2 answer set.
We will place the results of the new BL search ("BL AND" search) in the R3 answer set, as follows :

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? BL

***  BL    ***


** LIST OF VALID ANSWER FILES **
   R1  : AA                                          17 ANSWER(S)
   R2  : BL      R2  =   *   OR  R1                  24 ANSWER(S)

EQUATION ? R3 = * AND R1


- BL / IPAT     -   R3 :     10 answer(s)


-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

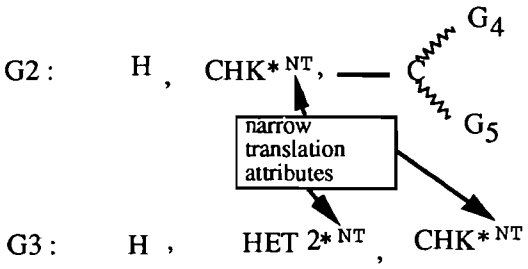
Example of a BL search with AND logic

III.5.6.3. NOT logic

Suppose now, still considering the main query, that we are interested only in tryptophane derivatives.



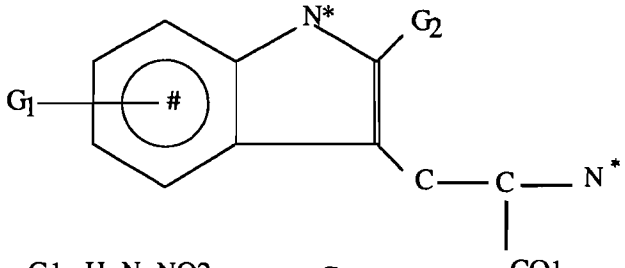
G1 : H , N , NO_2



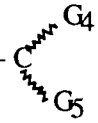
G4 : O , N 2*

G5 : H , N* , CHK*

We will modify the main query, by replacing G3 as follows:



G1 : H, N, NO2



G2 : H, CHK* NT,

G4 : O, N^{2*}

G5 : H, N*, CHK*

(In this query, translation attributes in group G2 are the same as in the main query).

The results of the search with the main query (17 answers) are still in the R1 answer set.
The results of the BL search with the OR logic (24 answers) are in the R2 answer set and those with the AND logic (10 answers) are in the R3 answer set.
We will place the results of the new BL search ("BL NOT" search) in the R2 answer set, as follows:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? BL

***  BL    ***


** LIST OF VALID ANSWER FILES **
   R1  : AA                                              17 ANSWER(S)
   R2  : BL      R2  =   *   OR  R1                     24 ANSWER(S)
   R3  : BL      R3  =   *   AND R1                     10 ANSWER(S)

EQUATION ? R2 = R1 NOT *


- BL / IPAT     -   R2 :     14 answer(s)


-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Example of a BL search with NOT logic

Section 6
Answer display and interpretation of the results

I. Answer display commands

Markush DARC provides two types of answer display:

- the List Command (LI command) which lists the Compound Numbers of the answers
- the View Command (VI command) which displays the structural diagrams of the answers.

The View Command has three variations:

VI FO	View Focus which highlights the portion of the structure which resulted in it being retrieved
VI	View which displays the main group (group 0) of each answer
VI MAX	View Maximum which displays group 0, followed by display of each of the G groups before moving to the next answer

The List and View Commands are entered at the ST prompt level:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ?
```

The List and View Commands assume that you want to see the results of the last search, which can be in the following answer sets:

- R1 if your search was a AA search
- R1, R2 or R3, if your search was a BL search, depending on the equation you specified.

Thus, unless otherwise specified, the LI and VI commands will list or view the results contained in one of these files.
It is however possible to add to LI and VI commands one of the R1, R2, R3, R0 or RX parameters, in order to specifically list or view the answers contained in these files.

Parameter	Signification	Example
R1, R2 or R3	to list or view answer set R1, R2 or R3	LI R1 VI R3
R0	to list or view the results of the last RE or SB search	LI R0 VI R0
RX	to list or view the answers in the RX file	LI RX VI RX

I.1. LI command

The LI command lists the Compound Numbers of the selected set, or of the answers of the last search if no parameter is specified.

The LI command is used at the ST prompt level. Upon entry of the LI command, the system displays a message indicating that you are looking at the results of the last search and indicates the number of answers in that set followed by a question mark (?) prompt .

The answers which are to be listed should be specified at this question mark prompt. You can list a single answer, a group of individual answers or a range of answers. The formats are shown with the examples listed below:

Format	Signification
2	lists the Compound Numbers for answer number 2
2,5,8	lists the Compound Numbers for answers 2, 5 and 8
1-11	lists the Compound Numbers for answers 1 through 11
2,5-9,11	lists the Compound Numbers for answers 2, 5, 6, 7, 8, 9 and 11

After displaying the Compound Numbers for the indicated answers, the system provides another question mark prompt. At this point, the searcher can enter additional answer numbers which are to be listed or a carriage return to exit the LI Command.

Examples:

1- Listing all the answers:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? LI

**   LI   **
** LAST SEARCH **    15  ANSWER(S)**
?  1-15
     1    CN = 8705-67201
     2    CN = 8705-66701
     3    CN = 8705-29401
     4    CN = 8705-09601
     5    CN = 8705-08002
     6    CN = 8705-08001
     7    CN = 8705-05401
     8    CN = 8705-02201
     9    CN = 8705-B1503
    10    CN = 8704-65201
    11    CN = 8704-17701
    12    CN = 8704-08901
    13    CN = 8701-24401
    14    CN = 8701-15701
    15    CN = 8701-10101
?
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Listing answers of the last search with the LI command

2- Listing answers number 1, 5 and 10:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? LI

**   LI   **
** LAST SEARCH **    15  ANSWER(S)**
?  1,5,10
     1    CN = 8705-67201
     5    CN = 8705-08002
    10    CN = 8704-65201
?
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Listing answers 1, 5 AND 10 of the last search with the LI command

3- Shortcut for listing answers

It is possible to input all the commands on a single line, as shown in the following example:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? LI 1,5,10-20

**   LI   **
** LAST SEARCH **   508  ANSWER(S)**
     1    CN = 8701-71301
     5    CN = 8701-65201
    10    CN = 8701-04802
    11    CN = 8701-02702
    12    CN = 8701-E1905
    13    CN = 8701-C6501
    14    CN = 8701-C2201
    15    CN = 8701-B8901
    16    CN = 8701-B1002
    17    CN = 8701-A4501
    18    CN = 8701-A4401
    19    CN = 8701-A4202
    20    CN = 8701-A1602
?

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

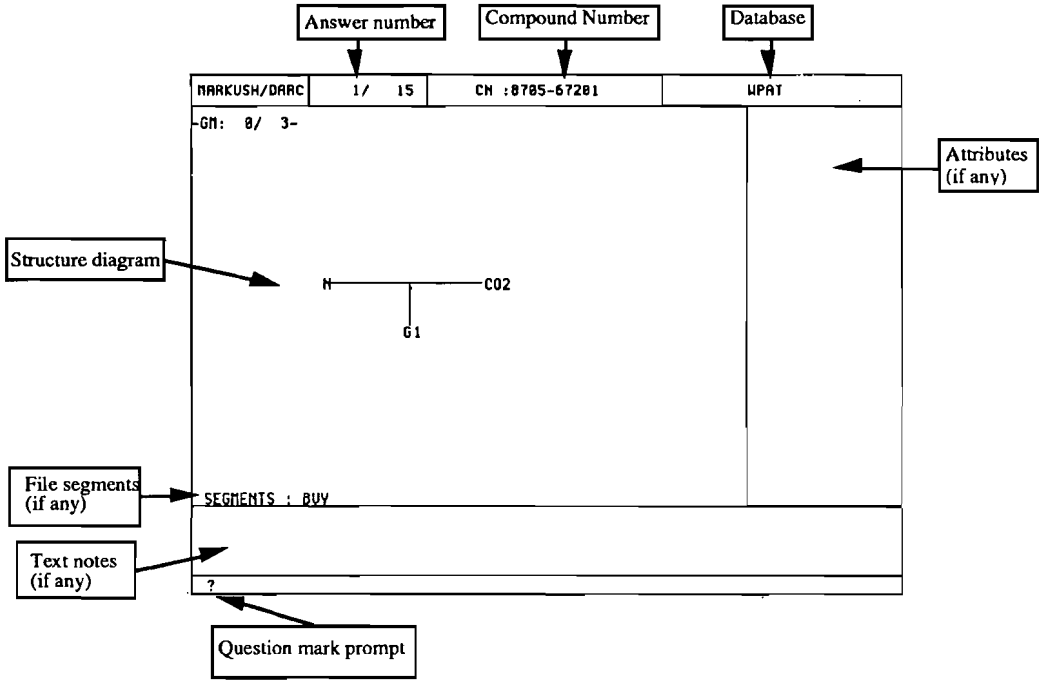
I.2. View commands

The View Commands are entered at the ST prompt level.

For all view commands, the system first displays Group Zero (G0) of the first answer in the current answer set.

The following information is displayed for group G0:

- the answer number followed by the total number of answers in the answer set.
- the Compound Number of this answer
- the database in which the search was done
- the group number being displayed followed by the total number of groups in this answer.
- the structural diagram of group G0
- the file segments (if any) associated with this answer.
- the question mark prompt (?) where the user can indicate what is to be displayed next.



Display of the main group G0 with View commands

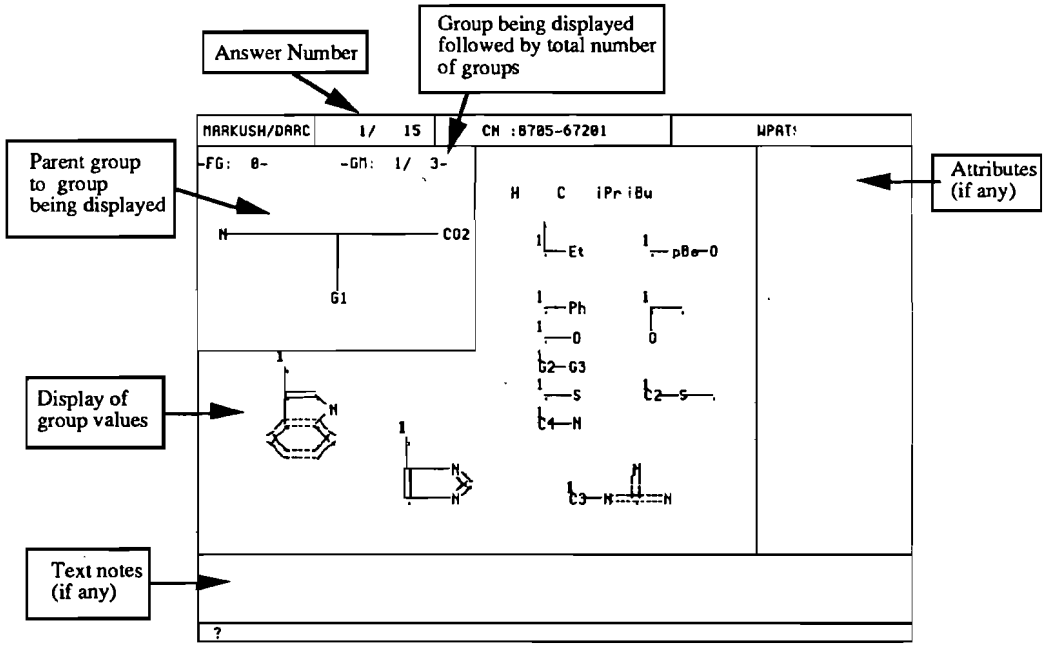
The next display which occurs automatically when you enter a carriage return will vary according to which View option was selected (VI, VI FO or VI MAX) at the ST prompt level: you will see either the main group of the next answer or the first subgroup of the current answer.
You can see groups by entering the group number, or other answers by typing the answer number.

For all view commands the display of subgroups provides the same type of information as for the main group G0 and the following additional information:

- the structural diagram of the father group of the subgroup being displayed, in a window located in the upper left hand corner of the screen, with the number of this father group

- the subgroup number being displayed followed by the total number of G groups in this answer, in the same window

- the structural diagrams of the subgroup values



Display of a subgroup with View commands

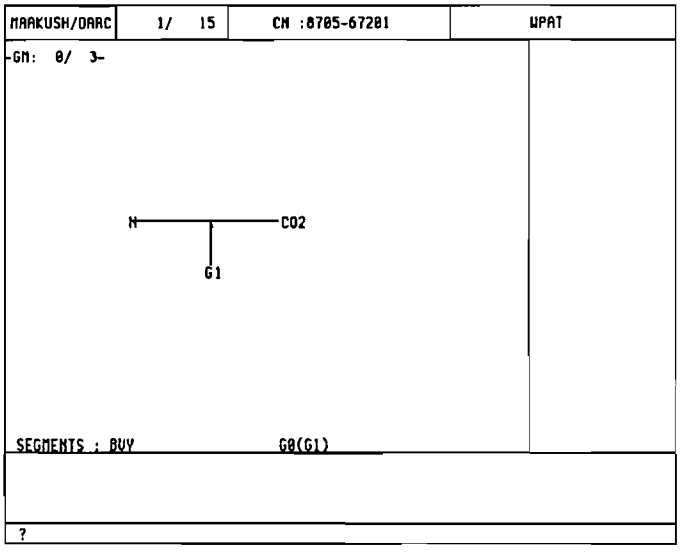
I.2.1. VI FO command

The VI FO command allows you to automatically display the portions of the structure which match the structure query.
This is particularly useful when the structure contains a large number of groups.

If you want to use VI FO on an answer set other than the current one, the query must be recalled as well as the answer set (see Section 7: - *Answer Handling and Cross-file searching* - for details on the recalling commands).

With this command, the system will display one after the other the groups containing the portions of the structure which caused the retrieval, with, for each group displayed, the highlighting of the first value of the group which caused the retrieval.

As for all other view commands, the system automatically displays first the main group G0 of the first answer, as shown in the following example:



Display of the main group G0 with the VI FO command

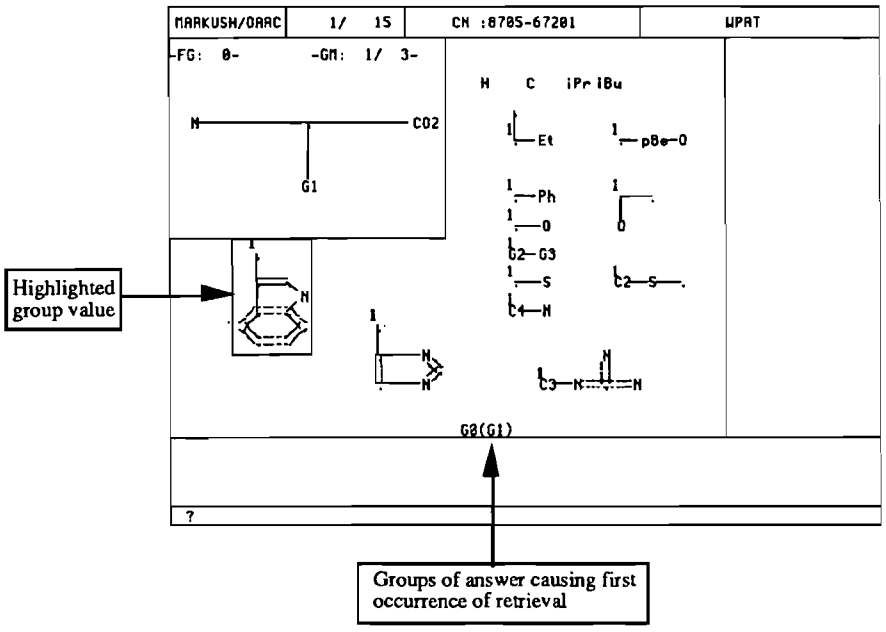
In addition to the information provided for all the view commands, the system indicates the group numbers which are involved in the retrieval.
This combination of groups corresponds to the first combination of group values of the answer which matches the query.
G0(G1) means G1 is within G0.
G0(G1(G8)) means G8 is within G1 which is within G0.
The groups are displayed in numerical order: G0(G1(G8)G2(G6)) will display G0, G1, G2, G6 then G8.

The display of subgroups involved in the retrieval using VI FO command comprises the same information as the VI command, with the following additional information:

- group numbers which are involved in the retrieval (as for G0) are shown

- the first group value involved in the retrieval is highlighted with a box

This is shown in the following example:



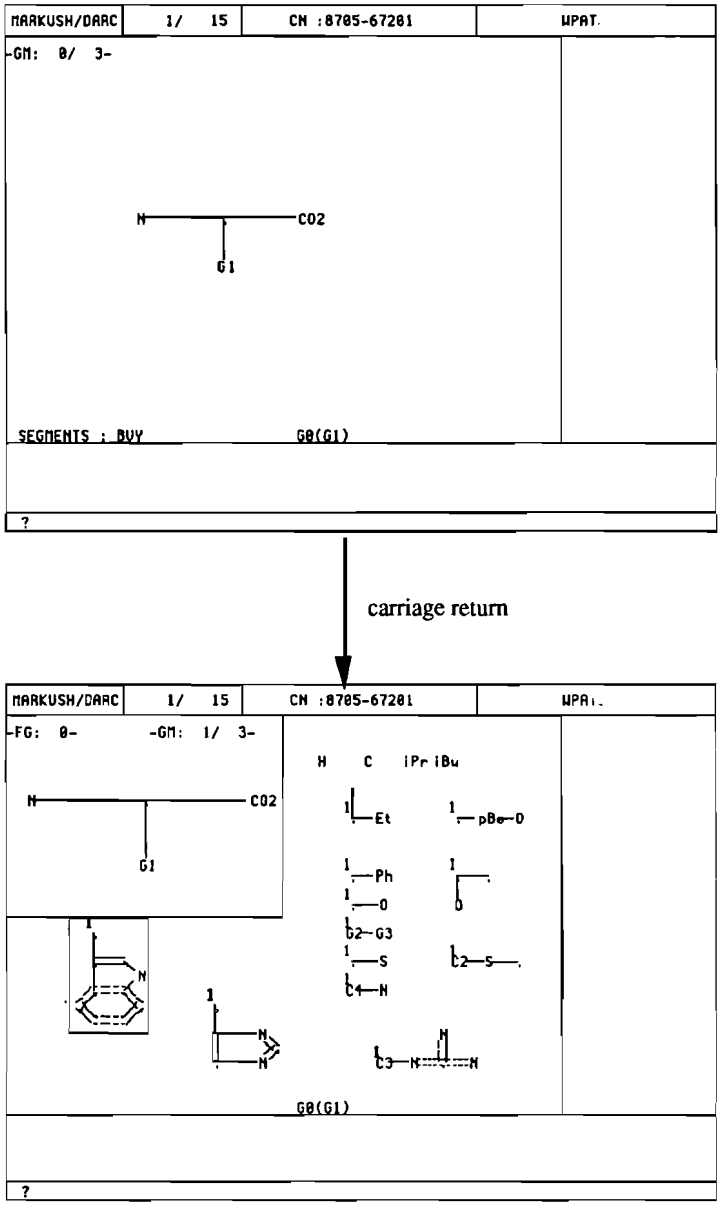
VI FO command: display of the next group involved in the retrieval

The View Focus (VI FO) command options for display which can be entered at the question mark prompt are:

Option	Displays
carriage return	next group of current answer involved in the retrieval of the answer if the current group is not the last group of the relevant combination of groups or group G0 of the next answer if the current group is the last group of the relevant combination of groups
a G number	displays the specified G group of the current answer
an answer number	displays group G0 of the specified answer
an attribute type	values for the specified attribute

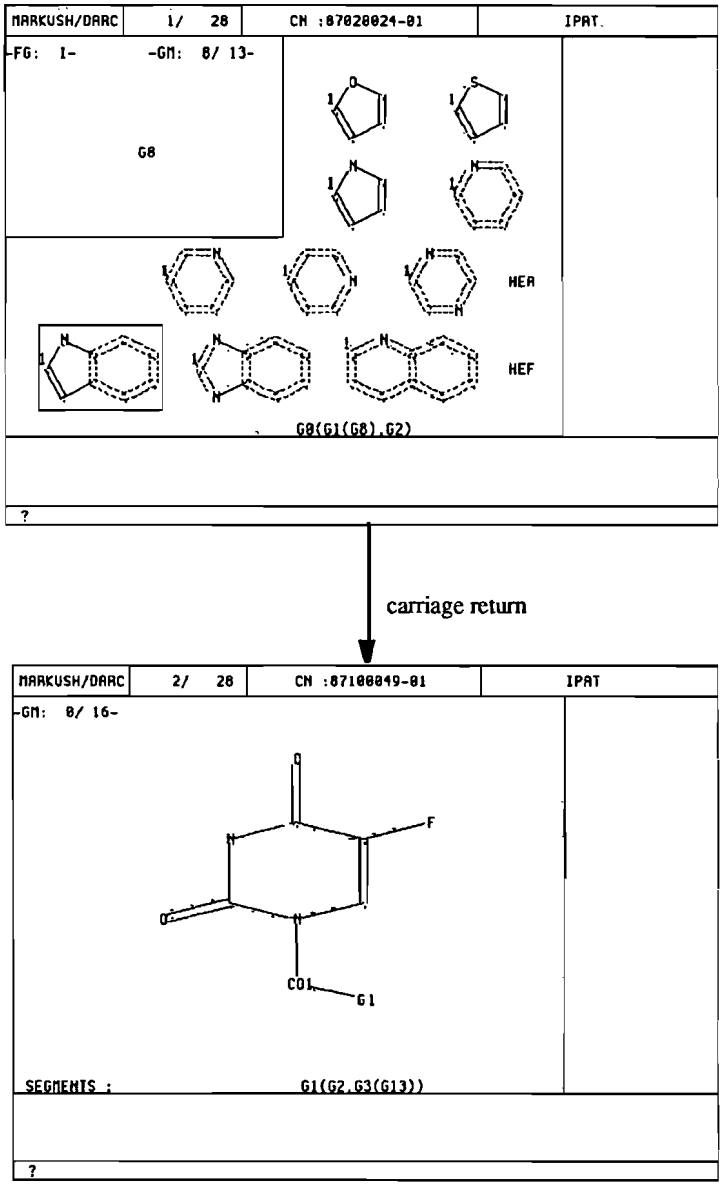
The use of these options is illustrated by the following examples:

1- Carriage return if the current group is not the last group of the relevant combination of groups:



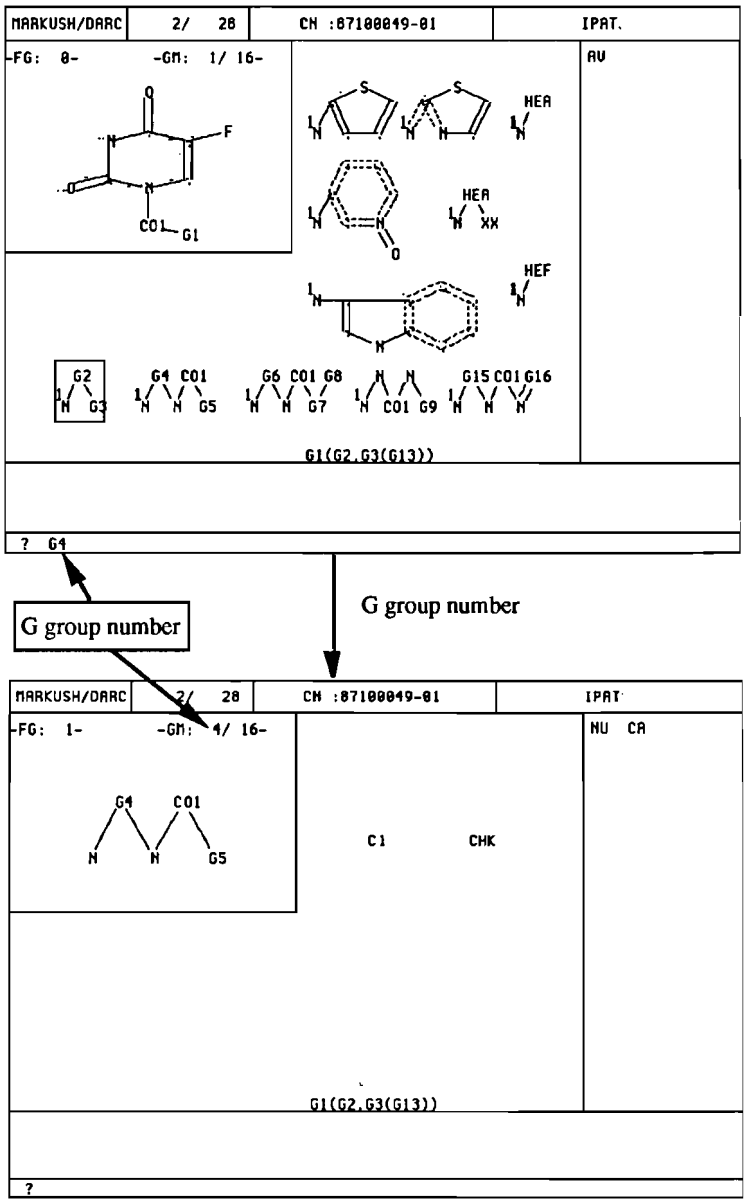
VI FO command: a carriage return in G0 displays the next group involved in the retrieval

2- Carriage return if the current group is the last group of the relevant combination of groups:



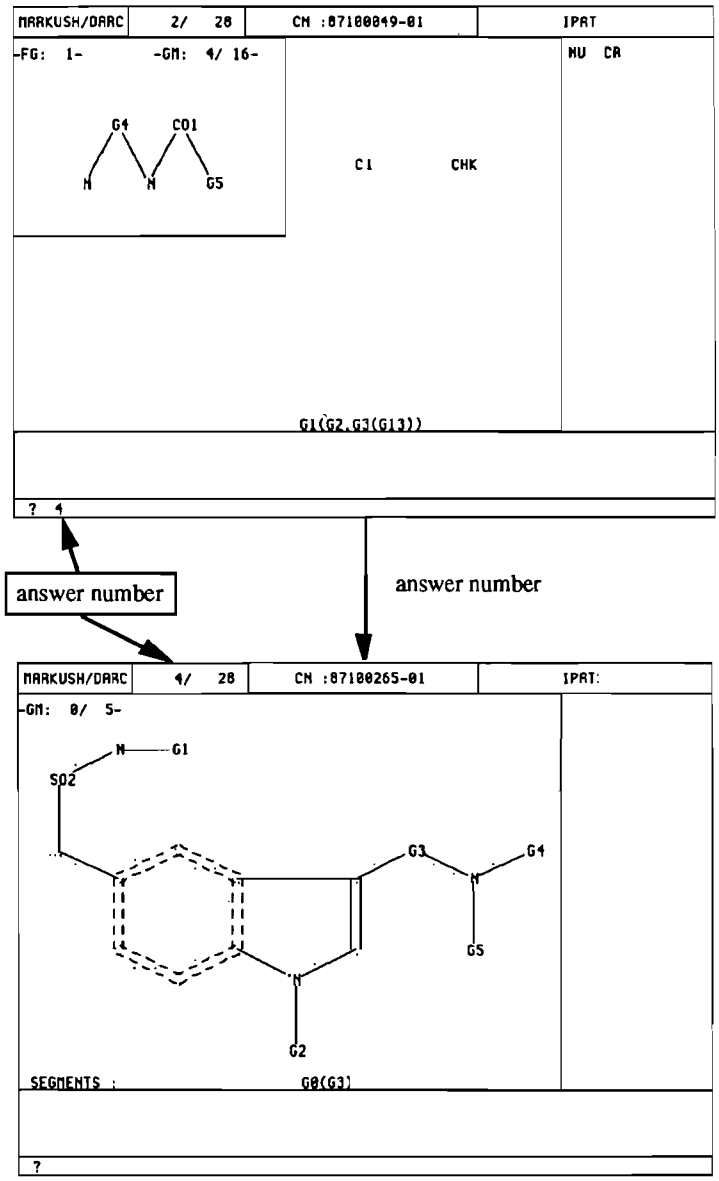
VI FO command: a carriage return in the last group involved in the retrieval displays G0 of the next answer

3- Specification of a G group number:



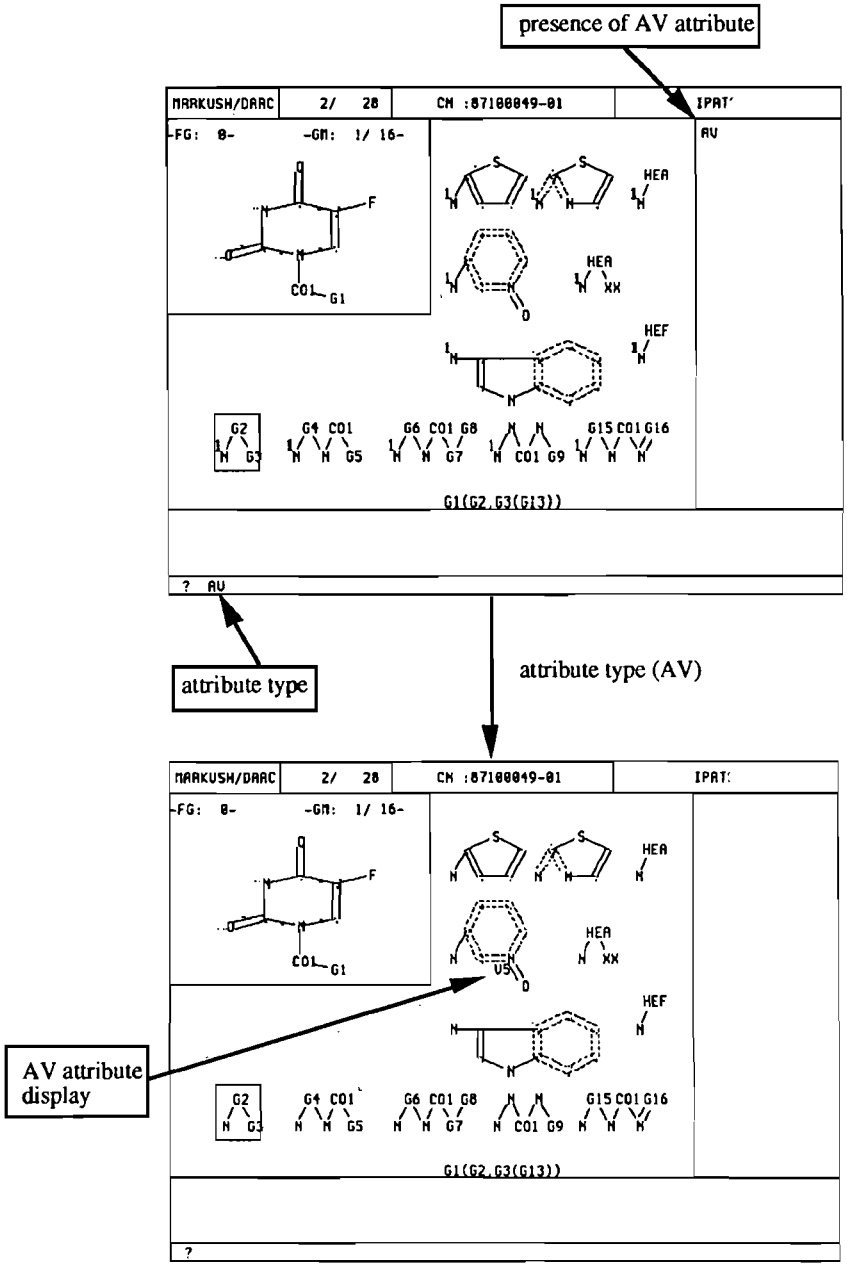
VI FO command: enter a G group number at any stage to display the specified group

4- Specification of an answer number:



VI FO command: typing an answer number at any stage displays group G0 of the specified answer

5- Specification of an attribute type:



VI FO command: typing an attribute type displays the specified attribute

I.2.2. VI command

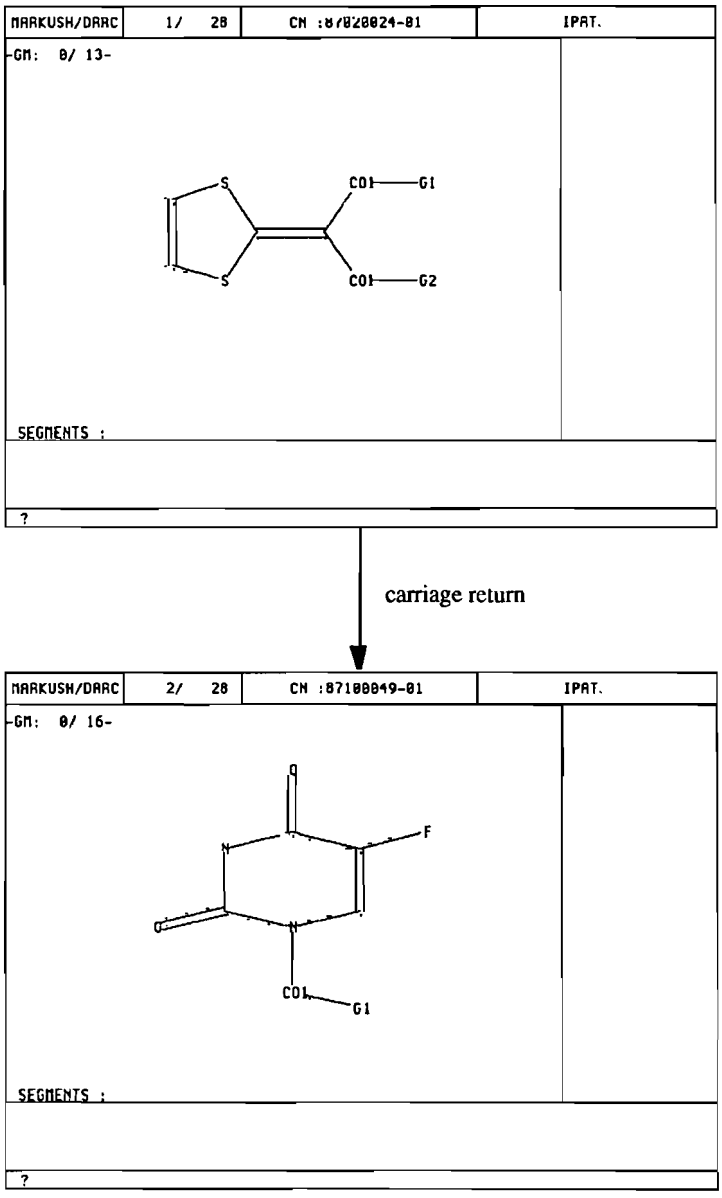
The VI command displays the main group G0 of each answer in succession unless specified otherwise.

The VI command options for display which can be entered at the question mark prompt are:

Option	Displays
carriage return	group G0 of the next answer
a G number	displays the specified G group of the current answer
an answer number	displays group G0 of the specified answer
an attribute type	values for the specified attribute

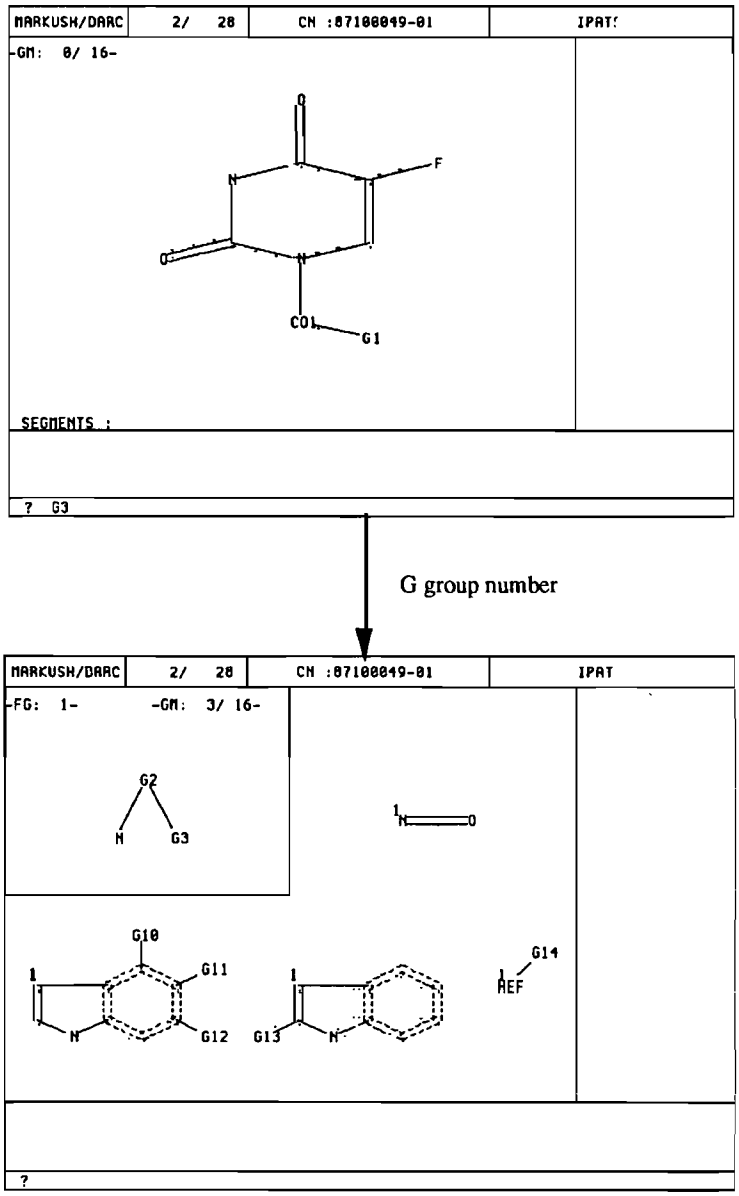
The use of these options is illustrated by the following examples:

1- Carriage return:



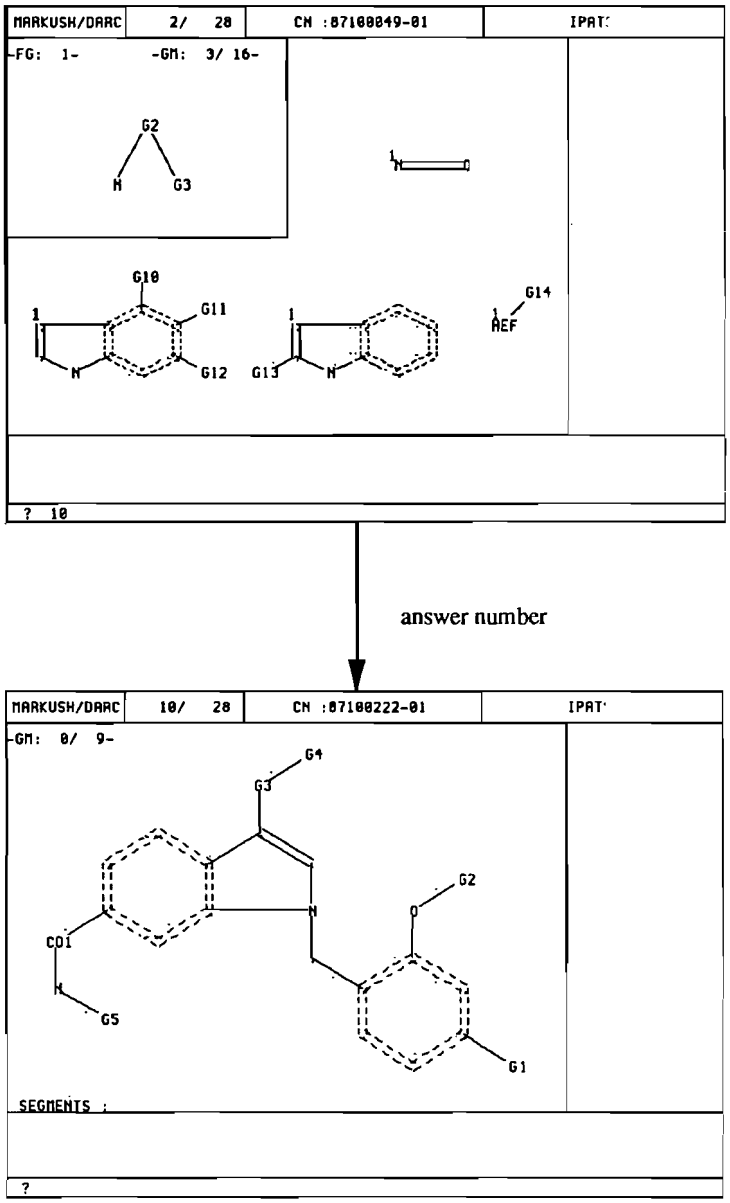
VI command: a carriage return displays the main group of the next answer

2- Specification of a G group number:



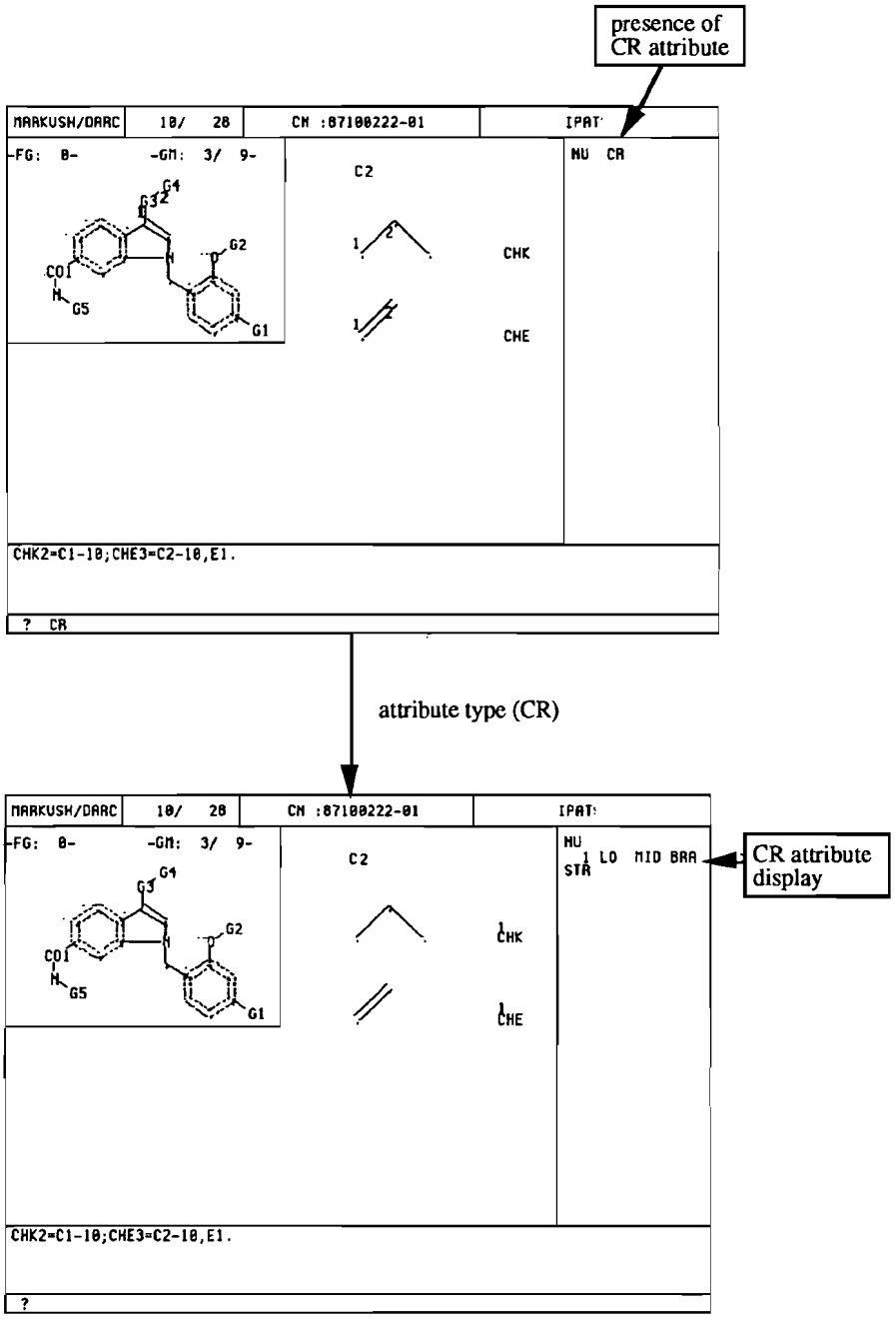
VI command: enter a G group number at any stage to display the specified group

3- Specification of an answer number:



VI command: typing an answer number at any stage displays group G0 of the specified answer

4- Specification of an attribute type:



VI command: typing an attribute type displays the specified attribute

I.2.3. VI MAX command

The VI MAX command displays each group of an answer in succession unless specified otherwise.

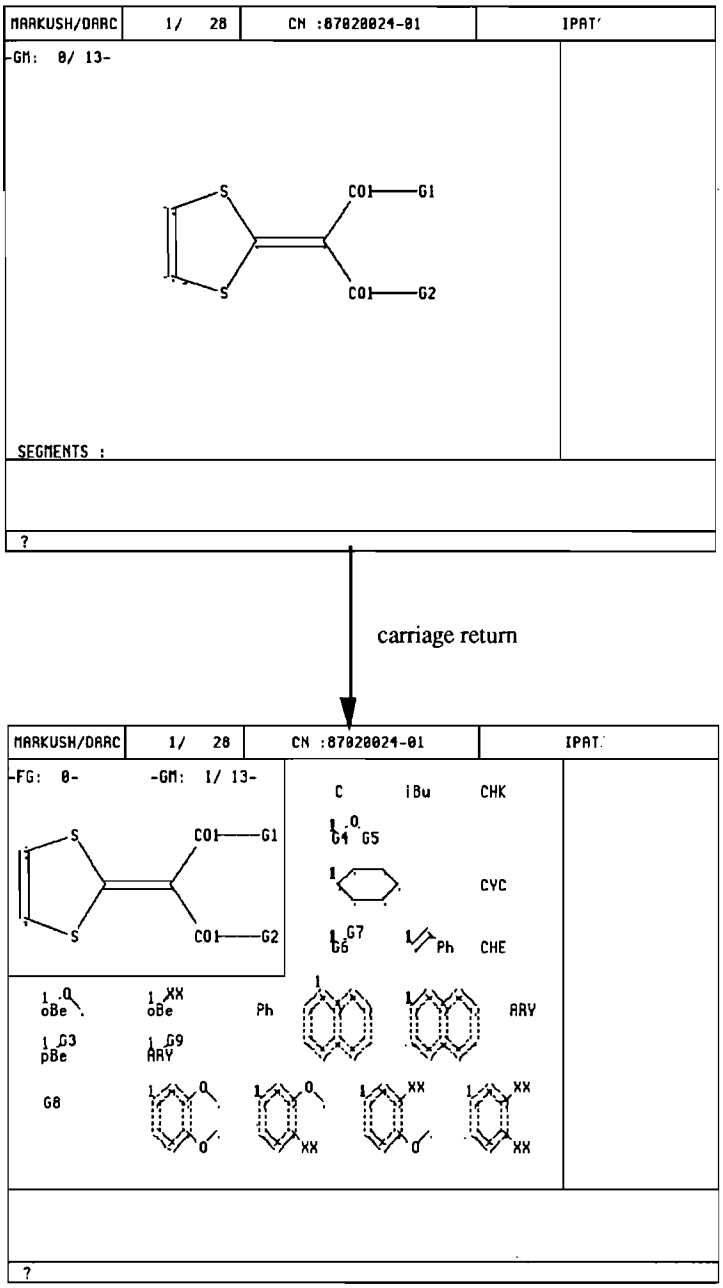
When "VI MAX' is entered at the ST prompt level, the system displays the main group G0 of the first answer in the current answer set.

The VI command options for display which can be entered at the question mark prompt are:

Option	Displays
carriage return	next group of the current answer
a G number	specified G group of the current answer
an answer number	displays group G0 of the specified answer
an attribute type	values for the specified attribute

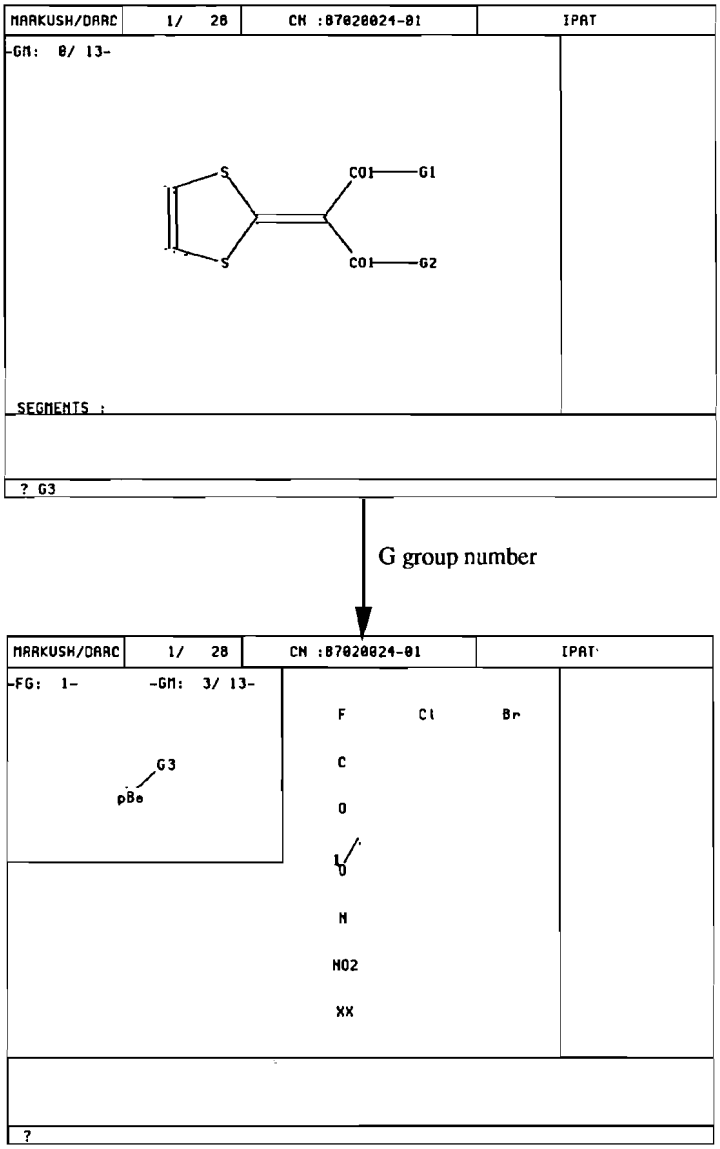
The use of these options is illustrated by the following examples:

1- Carriage return:



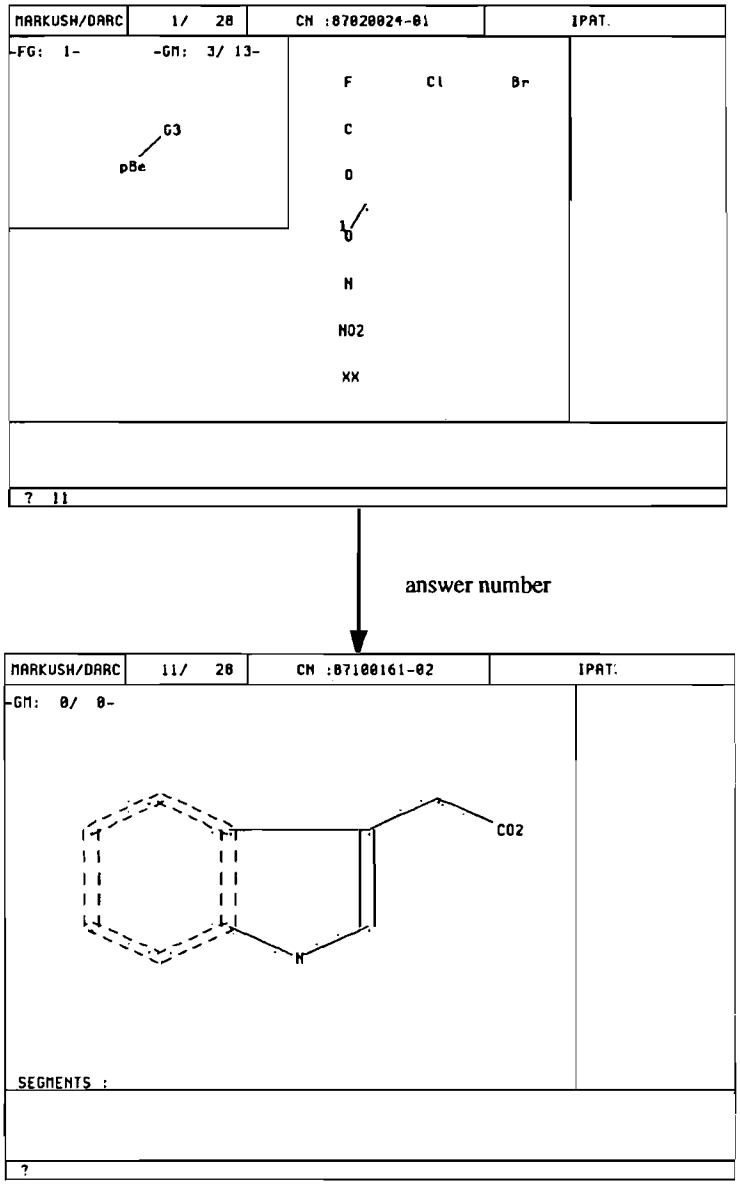
VI MAX command: a carriage return displays the next group of the current answer

2- Specification of a G group number:



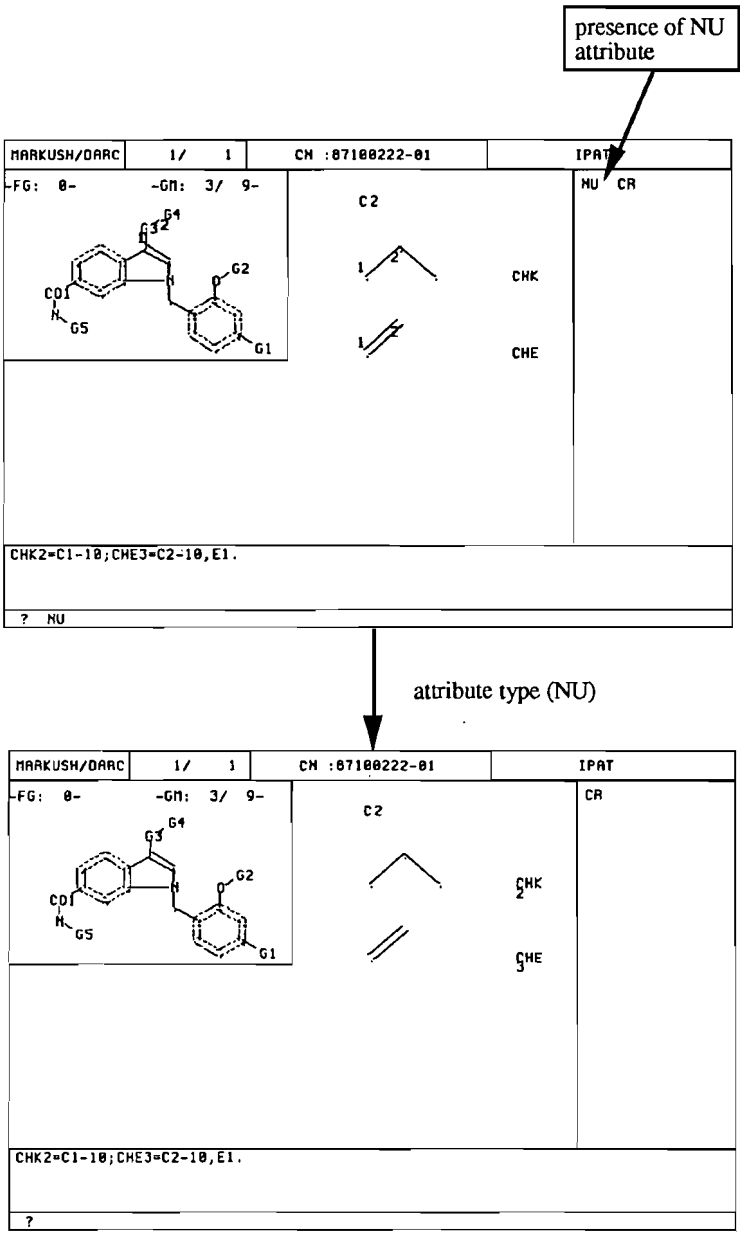
VI MAX command: enter a G group number at any stage to display the specified group

3- Specification of an answer number:



VI MAX command: typing an answer number at any stage displays group G0 of the specified answer

4- Specification of an attribute type:



VI MAX command: typing an attribute type displays the specified attribute

I.2.4. Attribute display

When an attribute is present in a group, it is indicated in the upper right hand corner of the display of the group being displayed, by an abbreviation representing the type of attribute. Several types of attributes may coexist in the same group. In this case, all the type codes are listed in the upper right hand corner.

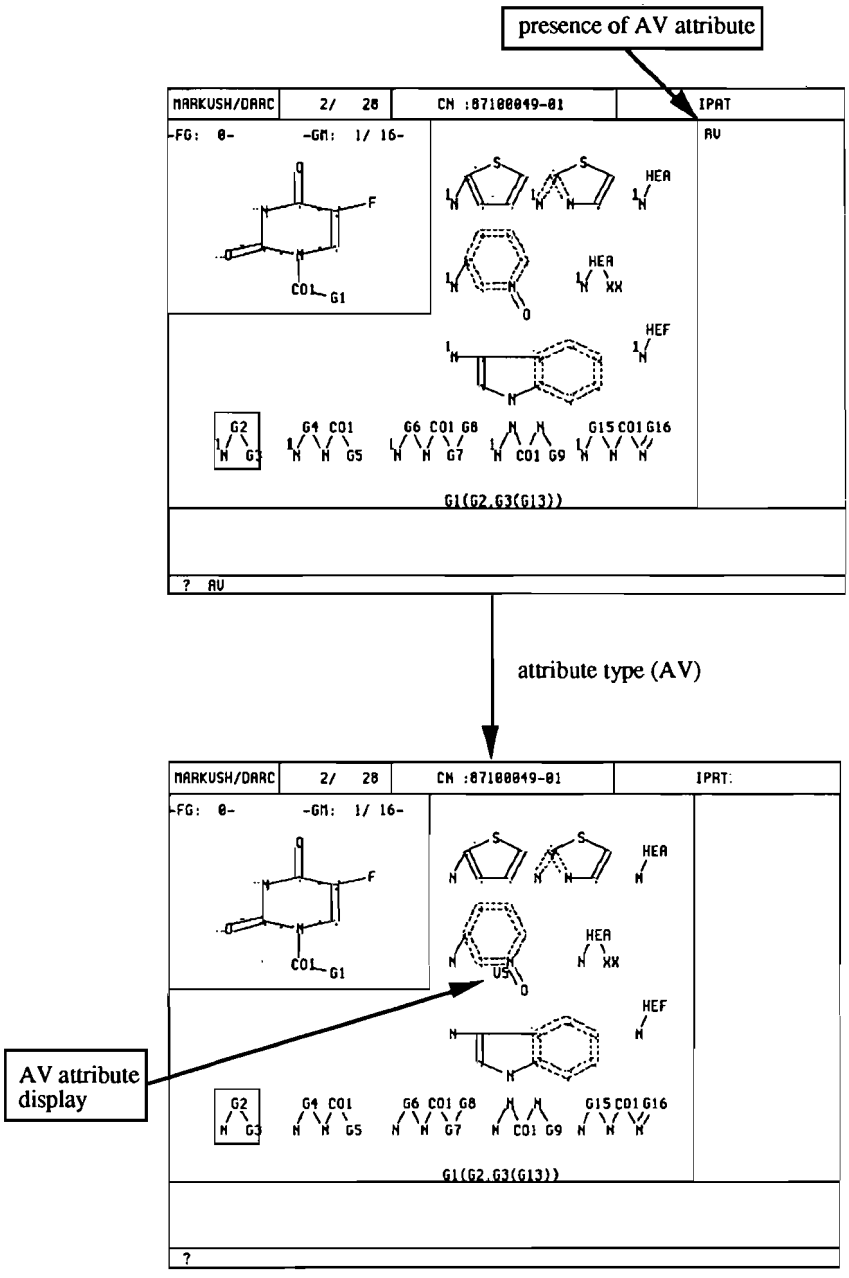
Attributes can be displayed by typing the attribute type at the question mark prompt, within any of the view commands (VI FO, VI or VI MAX).

The system re displays the current group with the attribute included in the display.

After viewing the attribute display, enter the appropriate response for what you want require displayed next.
This response depends upon which view command your are using at the time (please refer to the above paragraphs for the display options).

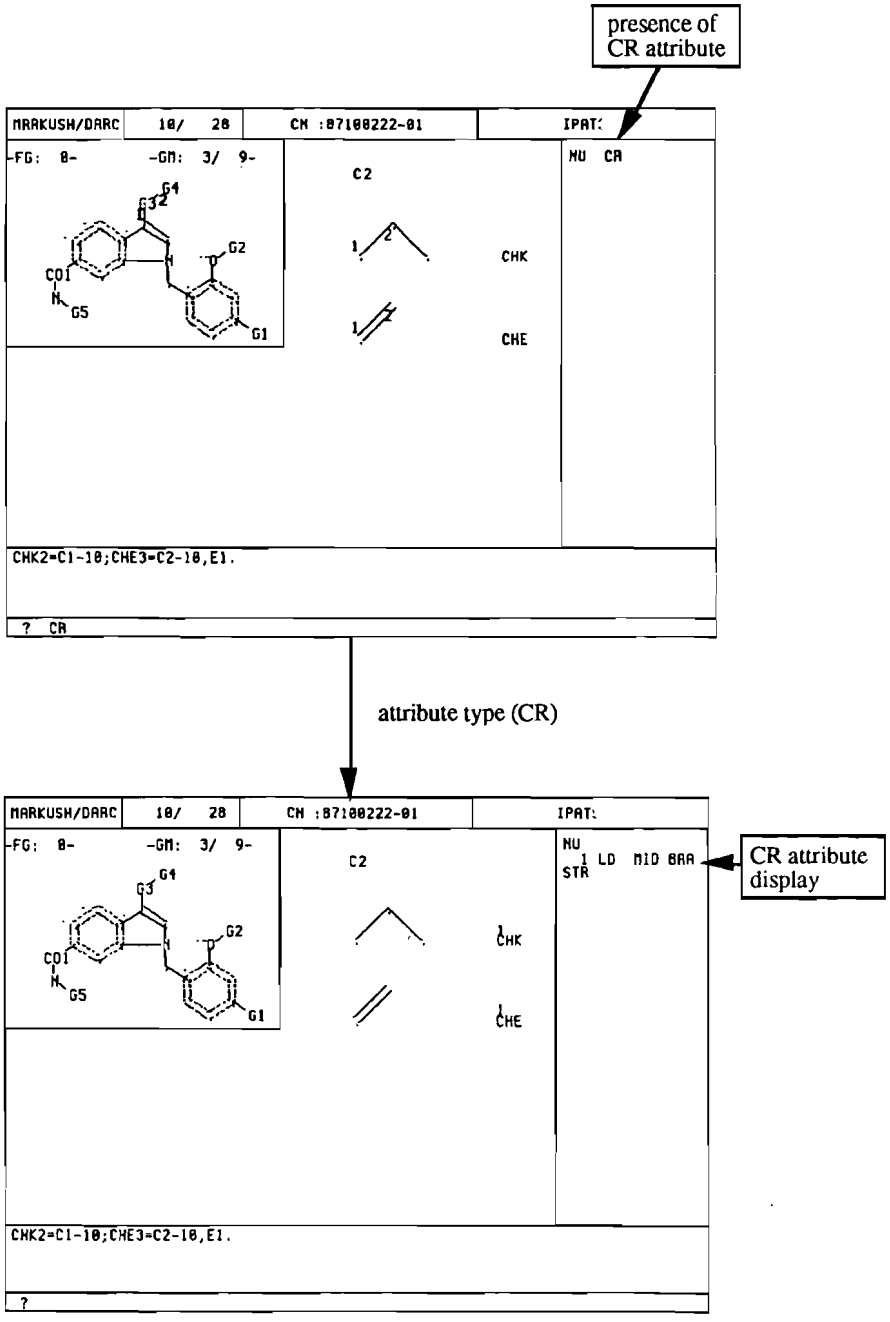
Below are some examples of attribute displays:

1- AV (abnormal valency) attribute:



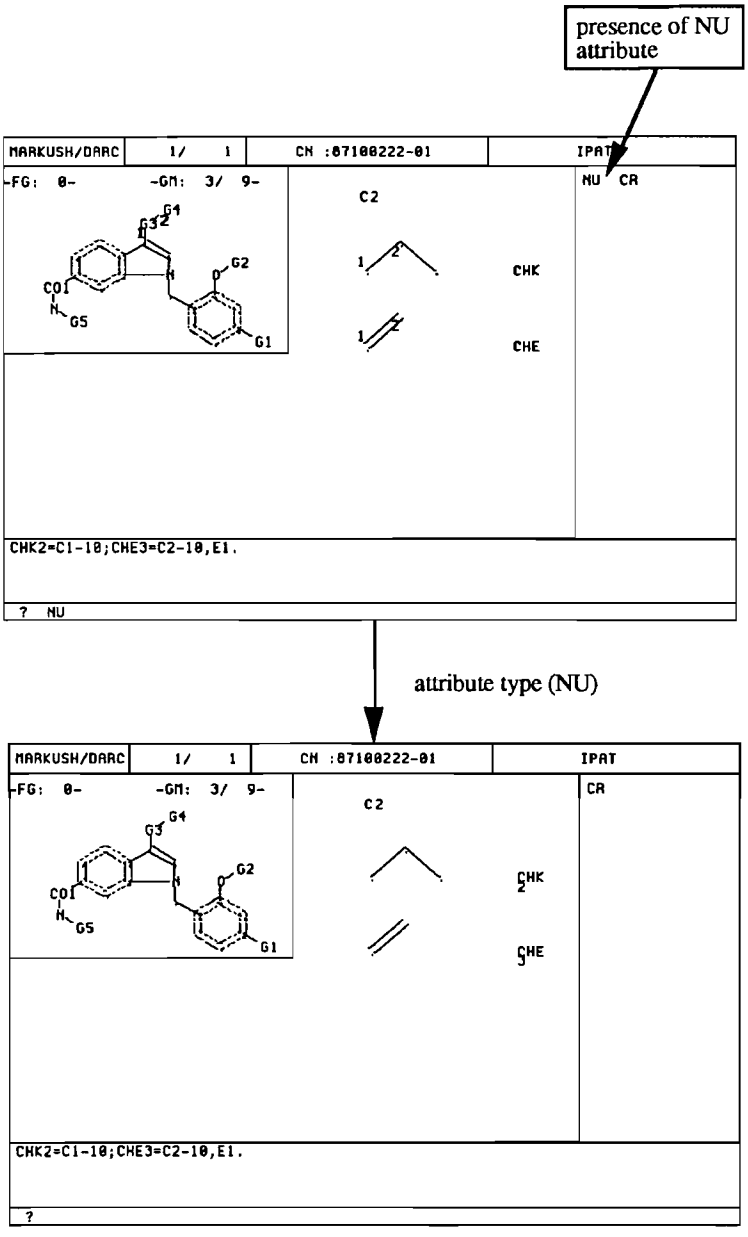
VI FO command: typing an attribute type displays the specified attribute

2- CR (chain/ring superatom) attribute:



VI command: typing an attribute type displays the specified attribute

3- NU (numbering attribute):



VI MAX command: typing an attribute type displays the specified attribute

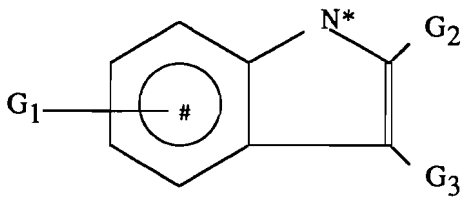
II. Interpretation of the results

Interpreting the results of a search, i.e. examining the answers to see why they are relevant, is a very important step in all types of chemical structure searching.

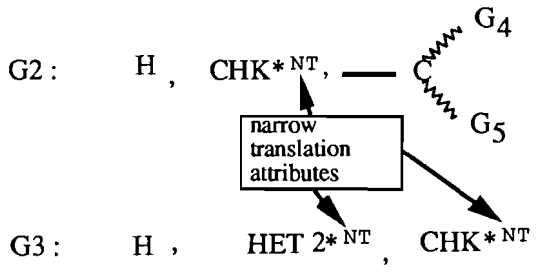
For the most complex type of searching, i.e. searching Markush structures using Markush queries, which is possible with the Markush DARC system, there is a need for efficient display capabilities. This is because it is very tedious and time-consuming to inspect all the groups of a Markush answer (there may be up to 100 groups!) to see which parts of the answer match the query.

The VI FO display command in the Markush-DARC system is a very convenient aid to the interpretation of the results, because it indicates the groups of the Markush answer which have permitted the match with the query and, within these groups, it highlights the matching values.

Let us consider the query example that we have used throughout this manual.



G1 : H , N , NO_2

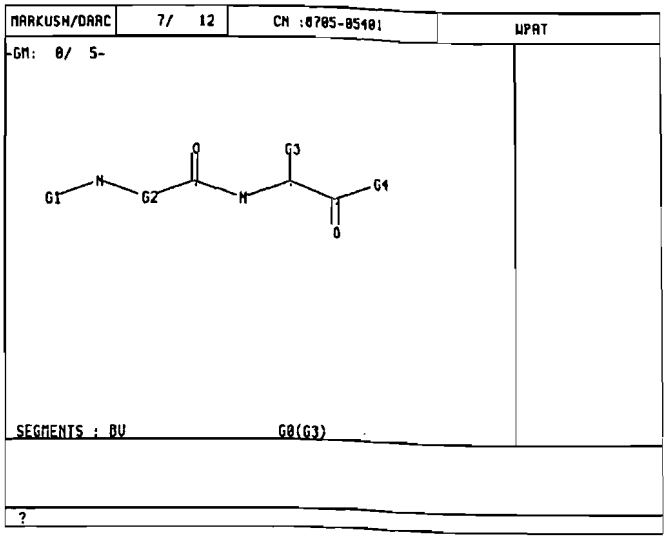


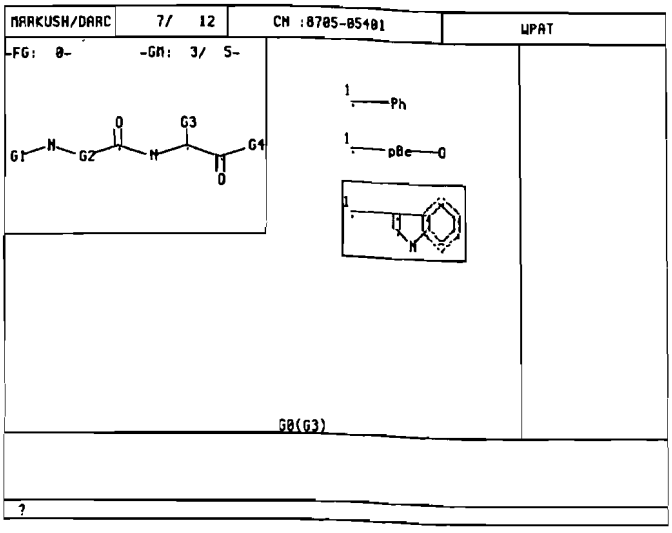
G4 : O , N 2*

G5 : H , N* , CHK*

There are 12 answers in the WPAT database.
Let us see, for some of these answers, why they have matched the query.

Let us use the VI FO command to see how the answers have matched the query, with answer number 7 first:

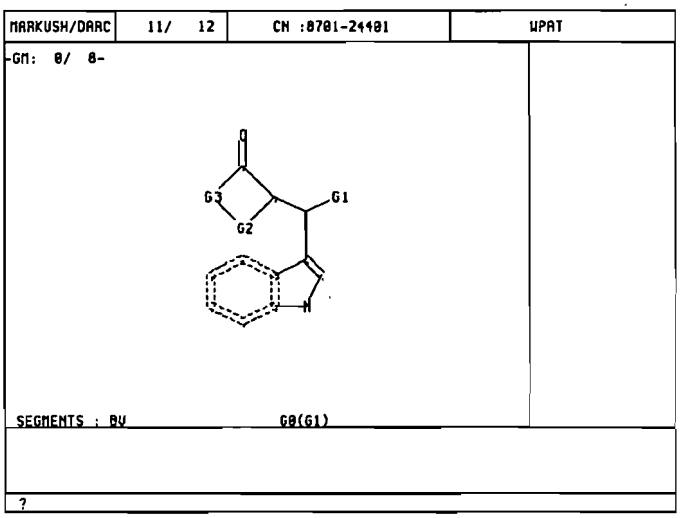


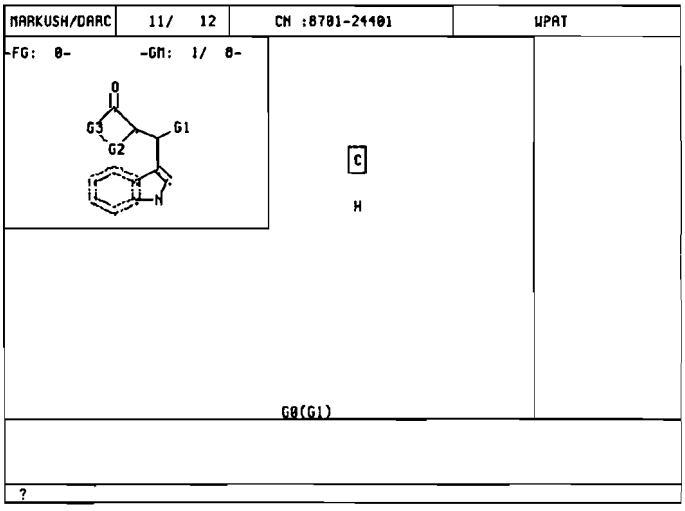


The "View Focus" command indicates that the query structure is contained in G0 and G3.
The above screens (main group G0, followed by group G3) show that position 3 of the indole ring is substituted by a carbon atom, itself substituted by another carbon atom..

This matches the query, where in G3 we have specified an alkyl superatom with one free site and Narrow Translation (CHK*^{NT}). The CHK superatom of the query matches the carbon atom contained in the indolylmethyl value in G3 of the answer.
The free site on the CHK allows one substitution. The carbon atom of the indolylmethyl value corresponds to the "translated" CHK.
It is substituted by a carbon atom (contained in G0), which corresponds to the possible substitution allowed by the free site on the CHK.

Let us now consider answer number 11, still using the VI FO command:





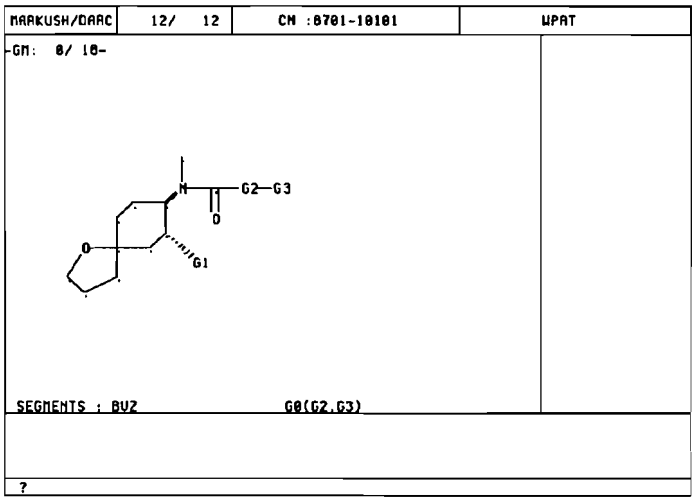
The "View Focus" command indicates that the query is contained in G0 and G1.

Position 3 of the indole ring is substituted by a substituted carbon chain.

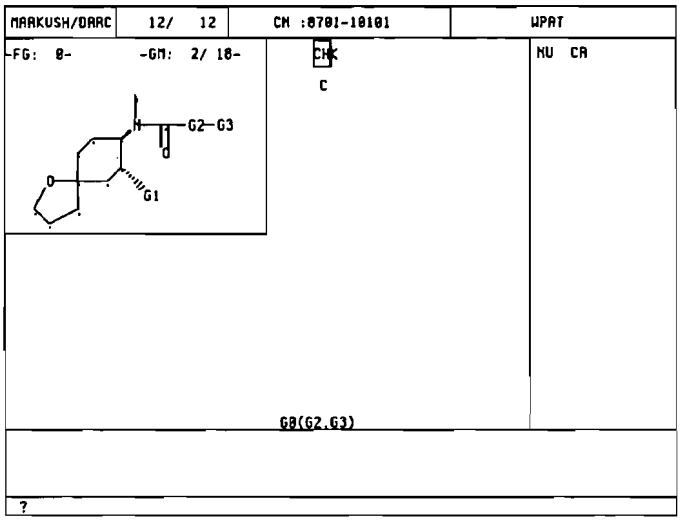
This matches the query, where in G3 we have specified an alkyl superatom with one free site and Narrow Translation (CHK*^{NT})

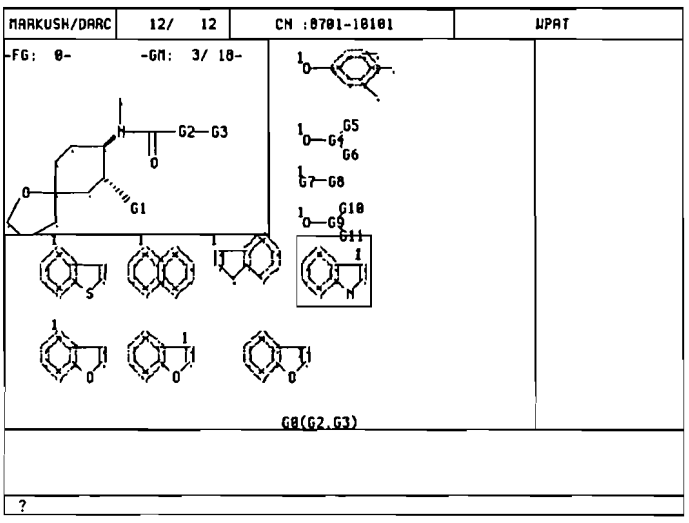
Group G1 of this answer contains two values: C and H. The C value is highlighted, because, with the VI FO command, the system highlights the first value that is found that matches the query. The H value also matches the query.

Let us consider the next answer (answer number 12) :



The VI FO command indicates that the answer is contained in groups G0, G2 and G3. Groups G2 and G3 are displayed successively, as shown on the following page:





The above screen shows that the indole ring is contained in group G3.
It is attached to group G2 through its 3 position.
Group G2 (previous screen) contains two values: CHK, which is highlighted, and C.
The CHK value is matching the CHK contained in G3 of the query.
The C value is also matching the query, because a narrow translation was applied on the CHK contained in G3 of the query.

Section 7
Answer Handling
and
Cross-file searching

I. Query Saving

Having completed a search in a Markush DARC structure file you may want to save the strategy you have created to re-use at a later date. Similarly, you may wish to save a list of Compound or DARC Numbers corresponding to structures you have retrieved.in a search session. The save command (SV) is used to save a current query or answer files at the ST command level The SV command can also be used at the QU command level exclusively to save search queries and not Compound or DARC number lists.

Queries or answer files saved using the SV command will be saved indefinitely until they are deleted using the erase (ER) command.

Each time the SV command is used the type of save operation - a search query or number list , must be indicated. This is done using parameters which are entered with the SV command. The following parameters are used:

QU permanent saving of a query

CN permanent saving of a list of compound numbers for further processing in the corresponding bibliographic files. on Questel Plus

DN permanent saving of a list of DARC numbers for further processing in Markush DARC.

Please note that a list of DARC Numbers (DNs) can be generated from a saved list of Compound Numbers (CNs). In addition, a list of Compound Numbers (CNs) can be generated from a list of DARC Numbers (DNs).

I.1. Saving Limits

A maximum of 20 queries can be saved per user logon. For answer files containing either CNs or DNs, a maximum of 20 answer files can be saved per user logon.

If the list being saved requires more space than is available the system will issue the following warning message:

```
NO MORE SPACE AVAILABLE, ADDITIONAL SPACE NEEDED; "X" TO FREE SOME SPACE,
USE HISTORY (HI) AND DETERMINE WHAT TO ERASE (ER)
```

I.2. Save Command

The example below illustrates the SAVE command used alone or with a parameter indicating the type of save being executed.

When the SV command has been entered alone with no parameters the system prompts the user to specify the type of save operation required - a query or answer list.

Each saved search must be assigned a name consisting of a maximum of eight alphanumeric characters.

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? SV
QU / DN,CN / ? CN
NAME ( 8CHAR.) ? INDOLE
      NAME    TYPE NUMBER    ORIGIN        DATE    SPACE  DATABASE
 1  INDOLE    CN       15   MARKUSH DARC  05/05/93    1    WPATSA

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Save command without parameters

Other examples illustrate the use of the SV command with the parameters QU, CN or DN. Since the parameter indicating the type of save search is entered directly with the SV command the system immediately prompts the user to enter the name of the save search.

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? SV DN
NAME ( 8CHAR.) ? CEPH
      NAME    TYPE NUMBER    ORIGIN        DATE    SPACE  DATABASE
 1  CEPH      DN       15   MARKUSH DARC  05/05/93    1    WPATSA

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ?
```

Save command with DN parameter

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? SV CN
NAME ( 8CHAR.) ? PENICILL
      NAME    TYPE NUMBER    ORIGIN        DATE    SPACE  DATABASE
 1  PENICILL CN       15   MARKUSH DARC  05/05/93    1    WPATSA

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Save command with CN parameter

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? SV QU
NAME ( 8CHAR.) ? CEPHALO
      NAME    TYPE NUMBER    ORIGIN        DATE    SPACE  DATABASE
 1  CEPHALO   QU        1   MARKUSH DARC  05/05/93    1

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Save command with QU parameter

Access to three current answer sets, R1, R2 and R3 is permitted in a single session. The three answer sets can be saved by entering the parameters R1, R2 or R3 with the SV command. Where an answer set number is not specified the system automatically saves the last answer set generated in the current search.

The results of a RE or SB search can be saved using the SV R0 command. Similarly to save the RX file use the SV RX command.

When lists of CNs or DNs are being saved the system saves the numbers in groups of 170 numbers. After 170 numbers have been saved the system displays a message confirming that a group of 170 numbers has been saved. This continues until all the numbers in the answer set have been saved.

When lists of CNs or DNs are being saved the system saves the numbers in groups of 170. containing more than 170 numbers

II. Recalling Queries

The recall file command (RF) is used to recall saved queries or saved DARC Number (DN) lists. When the RF command is used to recall a saved DN or CN list, the list of DNs is transfered into the current answer file. The file can then be used in any Markush DARC operation including Boolean Logic.

The RF command can be used to recall a query at the ST or QU command levels. At the QU command level, only saved queries (QU) or queries corresponding to batch requests (QB) can be recalled.

At the ST command level, saved queries, saved DN lists or batch search request lists (BT) can be recalled.

The example below illustrates the use of the RF QU command used at the ST command level. On entering the command the system prompts the user for the name of the query to be recalled. In this example the recalled query is not modified and the user enters 'No' to the system prompt 'Other Specifications.'

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? RF QU
NAME ( 8CHAR.) ? INDOLE2G
      NAME    TYPE NUMBER    ORIGIN        DATE    SPACE  DATABASE
 1  INDOLE2G QU       1    MARKUSH DARC  01/03/93    2

OTHER SPECIFICATIONS (Y/N) ? n
FILE SEGMENTS(Y/N) ? n
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Recalling a query

The use of the RF command to recall a saved query places the recalled query in the active query file. This query can then be searched as is or modified before running a search.

A query phrased with graphic input (QG) can be recalled using both text input (QT) and graphic input (QG).

If you should happen to recall in graphic mode a query which was saved in text mode the system will issue the message:

```
USE TEXT INPUT
```

A Markush DARC query cannot be recalled in Generic DARC, nor can a Generic DARC query be recalled in Markush DARC.

III. Erasing Queries

The Erase (ER) command is used to delete saved queries, or saved CN or DN lists. The ER command can be used at either the ST or QU command levels.

The ER command can be entered alone or with a parameter specifying the type of saved item to be deleted.

The example below illustrates using the ER command alone or with the parameter QU. If you enter the ER command alone the system prompts the user to enter the required parameter specifying the type of item to be erased. It should be indicated whether the saved item is a saved query or list of CNs or DNs. The name of the saved query or answer list must then be entered. Before the query or answer list is finally erased the system prompts the user to confirm the execution of the ER command.

Where the ER command is simultaneously issued with the parameter specifying the type of item to be erased the system will proceed directly to prompting the user for the name of the saved search or list.

It is always advisable to check the contents and names of previously saved searches before proceeding to erase them. This is done using the History (HI) command.

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? HI QU
      NAME    TYPE NUMBER    ORIGIN        DATE    SPACE  DATABASE
 1   INDOLE2G QU       1    MARKUSH DARC  01/03/93    2
 2   INDOLE2T QU       1    MARKUSH DARC  01/03/93    2
 3   CHKCO2   QU       1    MARKUSH DARC  26/03/93    1
 4   UACHAIN  QU       1    MARKUSH DARC  09/04/93    1
 5   INDOLEQT QU       1    MARKUSH DARC  21/04/93    2
 6   BLOR     QU       1    MARKUSH DARC  21/04/93    1
 7   BLAND    QU       1    MARKUSH DARC  21/04/93    2
 8   BLNOT    QU       1    MARKUSH DARC  21/04/93    2
 9   CEPHALO  QU       1    MARKUSH DARC  05/05/93    1

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? ER QU
NAME ( 8CHAR.) ? CEPHALO
      NAME    TYPE NUMBER    ORIGIN        DATE    SPACE  DATABASE
 1   CEPHALO  QU       1    MARKUSH DARC  05/05/93    1

CONFIRM ER   (Y/N) ? Y
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Erasing a query

IV. History Command

The History (HI) command is used to list the names of items which have been saved. The command can be used at either the ST or QU command levels.

The command can be entered with or without parameters. When the HI command is entered alone with no parameter the system automatically prompts the user to identify the type of saved item list to be displayed - saved search (QU), batch search (BT) or saved CN or DN list. When the type of list has been specified the system proceeds to display the list of required saved items.

Entering the HI command together with a specific parameter results in the required list being displayed immediately.

The example below illustrates the use of the HI command entered with and without a specific parameter.

```
-ST- (BA,CN,QT,QG,AF,RE,AA,SB,BL,BI,GD,INFO) ? HI
QU / DN,CN / QP / BT / ? CN
      NAME    TYPE NUMBER    ORIGIN        DATE    SPACE  DATABASE
 1  INDOLEQT CN       15   MARKUSH DARC  31/03/93    1     WPATSA
 2  INDOLE   CN       15   MARKUSH DARC  05/05/93    1     WPATSA
 3  PENICILL CN       15   MARKUSH DARC  05/05/93    1     WPATSA
 4  INDO     CN      227   MARKUSH DARC  05/05/93    2     MPHARMSA

-ST- (BA,CN,QT,QG,AF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

History command without parameters

V. Cross - File searching

A Markush DARC search provides access to structure records contained in Markush structure databases. During the indexing of a patent document Markush structures are entered into the Markush structure database. At the same time bibliographic information pertaining to the patent document is entered into a companion bibliographic database. The link between a Markush structure in a Markush structure database and its companion bibliographic database is its unique Compound Number. During indexing the Compound Number (CN) of each structure associated with a patent document is entered into the bibliographic record for that patent document. These bibliographic patent databases are searchable using the Questel Plus software.

Cross-file searching permits the running of a search in more than one file without having to re-enter information each time.

It is possible to cross-file search between structure databases on Markush DARC and companion bibliographic databases on Questel Plus. For example, a cross-file search between WPIM and WPIL or MPHARM and PHARM.

Cross-file searching between structure files on Markush DARC is also possible. For example, between MPHARM and WPIM.

Cross-file searching between bibliographic files on Questel Plus is also possible using the ..MEMORY and ..MEMSORT commands. Details of how this is done are given in the Questel Plus User Manual.

A structure file to structure file capability allows the phrasing of a single structural query on Markush DARC and the subsequent execution of the query in both the WPIM and MPHARM. A structure to structure cross-file search is described in this section

Processing the results of a structure search in a corresponding bibliographic database is also described in this section.

VI. Cross-file searching between WPIM and MPHARM structure files

The example below shows how a query created on WPIM can later be run on the MPHARM database.

The search is first of all run in the database in which the query was input. The results of the search are saved as a list of Compound Numbers (CNs). The saved list of CNs is assigned a unique name. In the example given below the saved list is given the name WPIM1.

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? BA

****    LAST SELECTED DATA BASE :    WPIM      ****
        ELAPSED TIME       :      3.58


***   AVAILABLE DATA BASES   ***

           DATA BASE  :   WPAT      (  1 )
           DATA BASE  :   IPAT      (  2 )
           DATA BASE  :   MPHARM    (  3 )
           DATA BASE  :   WPIM      (  4 )

***     4 DATA BASE(S) AVAILABLE   ***

DATA BASE # ?  3
****    BASE MPHARM            -   28/04/93    ****
     73943 COMPOUNDS - LAST CN : 93030281-01

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? RE_


***  RE    ***

RESULT    :         550
NEXT LIST :        2628
CONTINUE ? (Y/N) Y


- RE / MPHARM   -   R0 :    406 answer(s)


-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? AA

***  AA    ***


-   AA    - NUMBER OF ANSWERS          :    69
- FILE RX - NUMBER OF CANDIDATES       :     0
- CANDIDATES REMAINING TO BE PROCESSED :     0

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? _
```

Executing in MPHARM a search created in WPIM

VII. Cross-file searching between structure and bibliographic files

During the indexing of a patent document both structure and bibliographic data are recorded in structural and bibliographic databases respectively. Each Markush structure entered into a Markush DARC database is assigned a unique Compound Number (CN). This number is recorded in both structure and bibliographic databases. It is the Compound Number (CN) of a structure that links the structural and bibliographic records and allows cross-file searching between Markush DARC structural and companion bilbliographic databases.

There are two ways of cross-file searching between Markush structural and bibliographic databases. Namely, starting with a structure search in Markush DARC and transfering the results of the search to Questel Plus. Or, starting with a bibliographic search in Questel Plus and transfering the results of the search to Markush DARC. Both types of cross-file searching are described in this section.

VII.1. Cross-file searching from a structural to a bibliographic database

The procedure followed to obtain the bibliographic records for patent documents which in turn contain Markush structures recorded in Markush DARC databases is given below:

- Save the results of a structure search as a named list of Compound Numbers (CNs)
- Connect to Questel Plus
- Select the appropriate bibliographic database
- Process the list of Markush DARC Compound Numbers
- Refine the search with other elements if required
- Display the bibliographic records in the required display format

Example:

```
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? SU CN
NAME ( 8CHAR.) ? MPHARM1
      NAME   TYPE NUMBER    ORIGIN        DATE   SPACE  DATABASE
 1  MPHARM1  CN      69   MARKUSH DARC  05/05/93   1    MPHARM

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? BI
```

On leaving Markush DARC a summary of the cost estimation is given before connection is established with Questel Plus.

```
 * SEE YOU LATER *
(C) QUESTEL-1992
  QUESTEL Plus 9201                              05/05/93 17*19*13
  Last connection: 04/05/93 15*10*58

   A SINGLE FILE FOR DERWENT PATENTS NOW AVAILABLE: WPIL.
   NEW : US Patents with claims now available online !
   - Enter ..FI IFIPAT or ..FI IFIUDB
  ..FILE / ..INFO / ..GUIDE

?..FI PHARM


  QUESTEL      - Time (in hours) :   0,008
  Selected file: PHARM

  The PHARM file and its structural associated file MPHARM cover :
  US patents : 1985 week 28 to 1993 week 03
  EP and FR patents : 1985 week 36 to 1993 week 11
  DE and GB patents : 1993 week 02 to 1993 week 03
  PCT patents       : 1993 week 03

  Search statement  1

?*MDARC MPHARM1/ALL

  ** SS 1: Results 49

  Search statement  2

?..FO

  1/49 - (C) INPI
  AN  - 93020019
  CN  - 93020019-01; 93020019-02; 93020019-03; 93020019-04; 93020019-05;
        93020019-06; 93020019-07; 93020019-08; 93020019-09; 93020019-10;
        ***93020019-11***; 93020019-12; 93020019-13; 93020019-14; 93020019-15

  More: M / Repeat max: R / Keep: K / None: N

?M

  2/49 - (C) INPI
  AN  - 72068478
  CN  - ***72068478-01***; ***72068478-02***

  More: M / Repeat max: R / Keep: K / None: N

?M
```

VII.2. Cross-file searching from a bibliographic to a structural database

Having carried out a bibliographic search in the PHARM or WPIL databases on Questel Plus it is possible to transfer the results of the search to the Markush DARC system and to search for corresponding structural records in the MPHARM of WPIM databases.

This is done using the ..JOIN command which is a Questel Plus command used to transfer Compound Numbers retrieved from a search on Questel Plus to the Markush DARC system.

The ..JOIN command can be used to save up to 5000 Compound Numbers (CNs). The syntax of the command is as follows:

..JOIN {SS n} {i-j} TO {Markush filename} VIA *name*

{SS n} - search statement number. Default: last search statement

{i-j} - record numbers from which Compound Numbers should be extracted e.g; 1-20 records 1 to 20. Default: all records

{Markush DARC filename} - WPIM; MPHARM, IPAT, WPAT

name - Compound List name (maximum 8 characters)

A search can be refined using other parameters with the ..JOIN command. These parameters include the Derwent role qualifiers (with the Derwent databases). For example, the role qualifier, N which indicates new compound , S which indicates starting material etc... Several role qualifiers may be entered at the same time. To limit the extraction to Markush Compound Numbers only, the parameters the GEN parameter is used. To retrieve specific compound the SPE parameter is used. Role identifiers and parameters are entered as follows:

..JOIN {SS n} {i-j} TO {Markush filename} VIA *name* {ROLE x, y} {GEN or SPE}

Once a name list has been saved connection is made to the Markush DARC system. The WPIM or MPHARM database is selected. The list of saved Compound Numbers is recalled using the RF and QP commands at the ST Command level. The structure search can be refined with other criteria if required. The search is carried out. The final results can be displayed.

Example:

```
?/TI QUATERNARY AND AMMONIUM

  ** SS 1: Results 1.251

  Search statement  2

?..JOIN TO WPIM VIA AMMON GEN

  Total number of terms extracted:  137

  Search statement  2

?..ST MDARC

  Temporary save search: Y / N

?n
```

A search is carried out in the WPIL database. The Compound Numbers (CNs) contained in the retrieved records are extracted and saved using the ..JOIN command. Only those CNs pertaining to Markush structures are extracted - using the GEN parameter.
To connect to Markush DARC from Questel Plus, the command ..ST MDARC is entered. On leaving Questel Plus a summary of the costs incurred when searching in Questel Plus is given. The CNs saved in the list named AMMON are recalled using the RF QP command.

```
****    LAST SELECTED DATA BASE :    WPIM      ****
****    BASE WPIM              -   30/04/93    ****
    187615 COMPOUNDS - LAST CN : 9226-74105

-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? li
NO VALID ANSWER FILE
-ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? RF QP
NAME ( 8CHAR.) ? AMMON
      NAME   TYPE NUMBER    ORIGIN        DATE   SPACE  DATABASE
 1  AMMON    CN     137   QUESTEL-PLUS  05/05/93   1    WPIM


- RF / WPIM     -   R1 :    121 answer(s)

    -ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? LI

    **   LI   **
    ** LAST SEARCH **   121  ANSWER(S)**
    ?  1-8
         1    CN = 9222-F1902
         2    CN = 9222-F1901
         3    CN = 9218-52501
         4    CN = 9218-21801
         5    CN = 9208-18001
         6    CN = 9203-E6301
         7    CN = 9151-D9902
         8    CN = 9151-D9901
    ?
    -ST- (BA,CN,QT,QG,RF,RE,AA,SB,BL,BI,GD,INFO) ? VI
```

To erase a previously created list of CNs created using the ..JOIN command in Questel Plus use the ER QP command in Markush DARC together with the name of the list.

Appendix 1
Markush DARC superatoms, shortcuts and attributes

I. Markush DARC superatoms

I.1. Standard superatoms

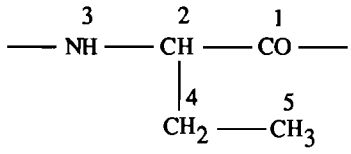
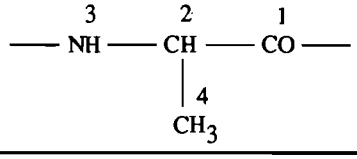
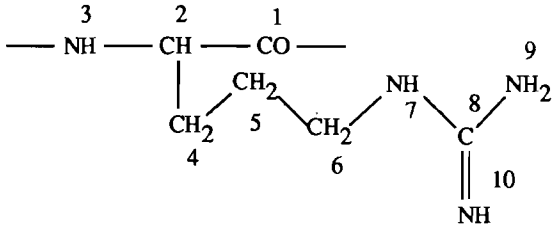
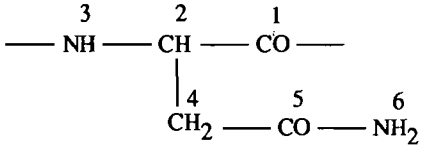
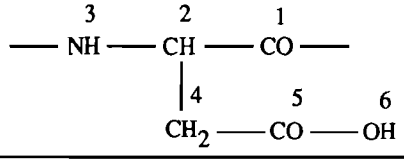
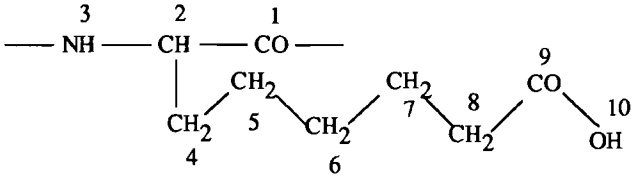
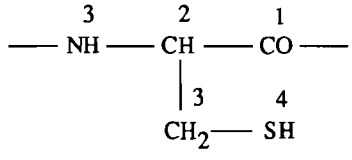
ACYCLIC SUPERATOMS		
CHK	**CHE**	**CHY**
alkyl, alkylene	alkenyl, alkenylene	alkynyl, alkynylene
CYCLIC SYSTEMS		
ARY	**CYC**	**HEA**
- in WPIM: monocyclic or fused carbocyclic system containing at least one benzene ring - in MPHARM: monocyclic or fused carbocyclic system containing an even number of C atoms and the same number of normalized bonds	cycloaliphatic monocyclic or fused, non-aromatic carbocycle optionally substituted by acyclic hydrocarbons	monocyclic heteroaryl, i.e. 5-membered heterocycle with 2 double bonds or 6-membered heterocycle with 3 double bonds alternating with 3 single bonds
HET	**HEF**	
non-aromatic monocyclic heterocycle	fused heterocycle	
ELEMENTS		
MX	**AMX**	**A35**
any metal	Alkali and Alkaline earth metals	Group III A-VA metals

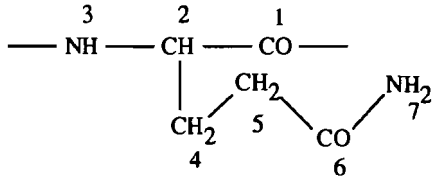
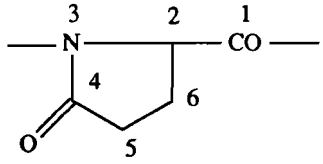
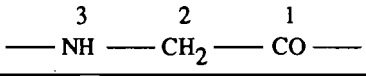
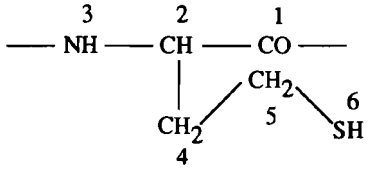
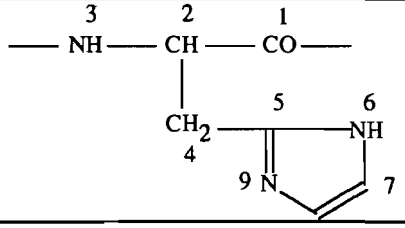
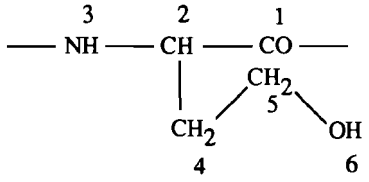
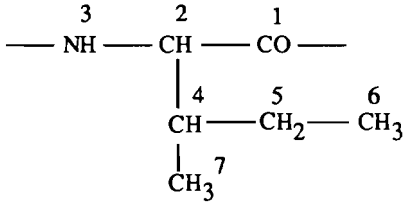
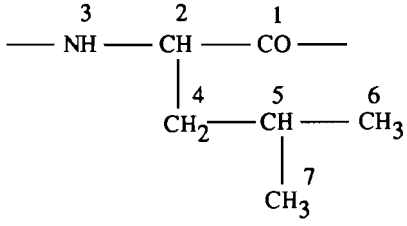
TRM	LAN	ACT
Transition metals excluding Lanthanum	Lanthanides (including Lanthanum)	Actinides (including Actinium)
HAL		
halogen		
OTHER SUPERATOMS		
ACY	**DYE**	**POL**
acyl (i.e. residue left after removal of one or more OH groups from an acid)	chromophore or fluorescent group (including dye residue)	polymer, polypeptide residue
PEG	**PRT**	**XX**
polymer end group	protecting group	any atom or group excluding hydrogen
UNK		
undefined group		

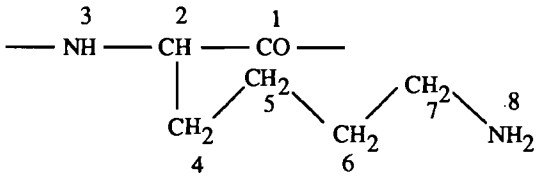
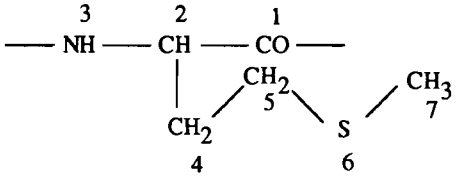
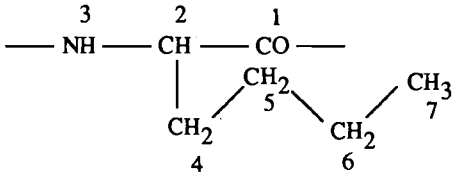
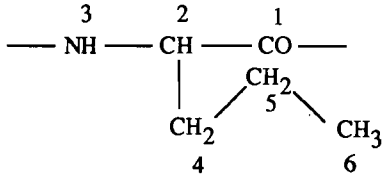
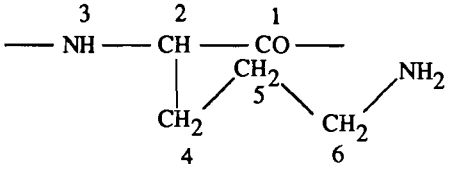
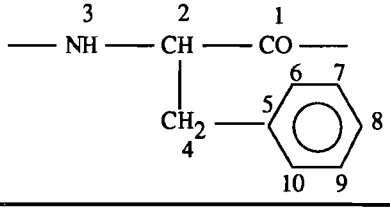
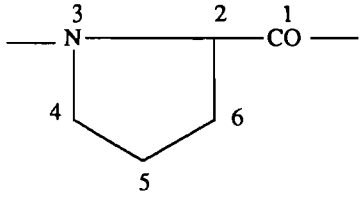
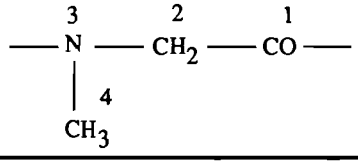
I.2. Peptide superatoms

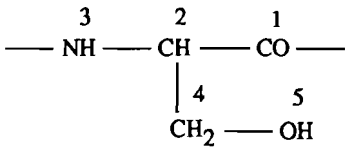
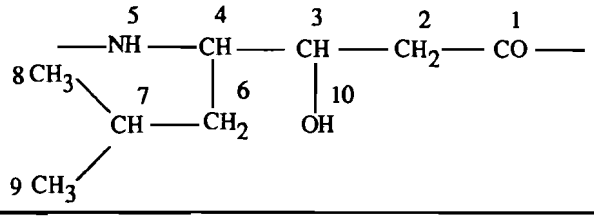
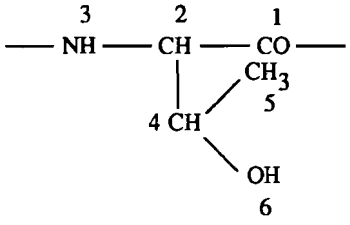
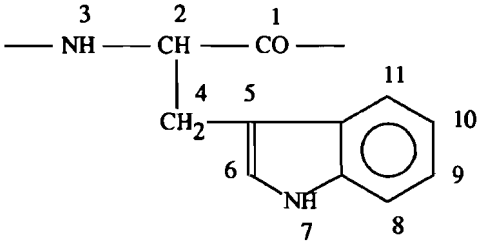
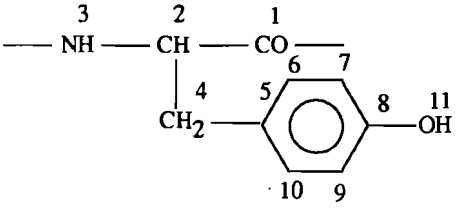
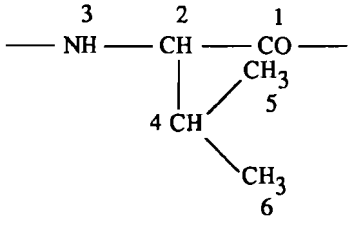
ABU	aminobutyric acid	**LEU**	leucine
ALA	alanine	**LYS**	lysine
ARG	arginine	**MET**	methionine
ASN	asparagine	**NLE**	norleucine
ASP	aspartic acid	**NVA**	norvaline
ASU	aminosuberic acid	**ORN**	ornithine
CYS	cysteine	**PHE**	phenylalanine
GLN	glutamine	**PRO**	proline
GLP	pyroglutamic acid	**SAR**	sarcosine
GLU	glutamic acid	**SER**	serine
GLY	glycine	**STA**	statine
HCY	homocysteine	**THR**	threonine
HIS	histidine	**TRP**	tryptophane
HSE	homoserine	**TYR**	tyrosine
ILE	isoleucine	**VAL**	valine

I.3. Peptide superatoms with position attributes

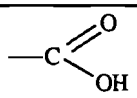
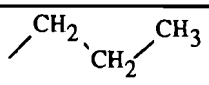
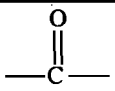
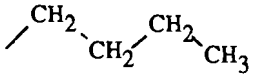
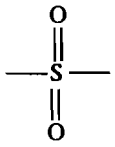
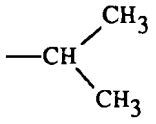
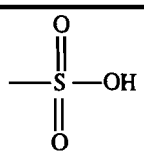
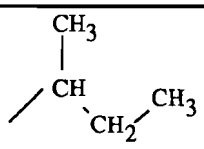
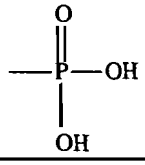
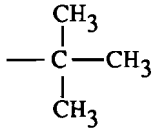
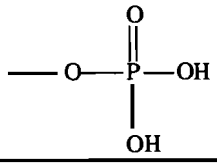
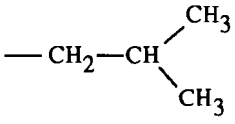
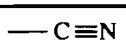
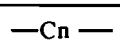
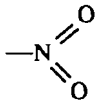
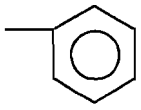
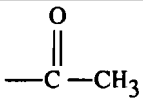
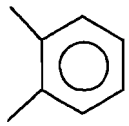
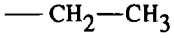
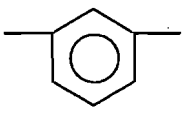
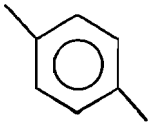
Superatom	Amino acid name	Amino acid structure with position attributes
ABU	amino-butyric acid	
ALA	alanine	
ARG	arginine	
ASN	asparagine	
ASP	aspartic acid	
ASU	amino-suberic acid	
CYS	cysteine	

GLN	glutamine	
GLP	pyrogluta-mic acid	
GLY	glycine	
HCY	homo-cysteine	
HIS	histidine	
HSE	homoserine	
ILE	isoleucine	
LEU	leucine	

LYS	lysine	
MET	methionine	
NLE	norleucine	
NVA	norvaline	
ORN	ornithine	
PHE	phenyl-alanine	
PRO	proline	
SAR	sarcosine	

SER	serine	
STA	statine	
THR	threonine	
TRP	tryptophane	
TYR	tyrosine	
VAL	valine	

II. Markush DARC shortcuts

--CO2		--nPr	
--CO1		--nBu	
--SO2--		--iPr	
--SO3		--sBu	
--PO3		--tBu	
--PO4		--iBu	
--CN		--Cn (CHA)	
--NO2		--Ph	
--Ace		--oBe--	
--Et		--mBe--	
		--pBe--	

III. Markush DARC searching attributes

III.1. General table

MARKUSH DARC ATTRIBUTES		
ATOM ATTRIBUTES	charges	CH
	delocalized charges	CH
	abnormal mass	AM
	abnormal valency	AV
	deuterium	D
	tritium	T
	free sites	FS
SUPERATOM ATTRIBUTES	chain/ring attributes	CR
	free sites	FS
	multiplier attribute (non-searchable, only displayable)	MU
PEPTIDE ATTRIBUTES	D, L or DL amino acid configuration (on peptide superatom)	D, L, DL
	position of substitution on the amino acid	SP
POLYMER ATTRIBUTES (on atoms or superatoms)	roles of atoms or superatoms in a polymer (on atoms, shortcuts or superatoms)	PA

III.2. Markush DARC chain/ring superatom attributes

MARKUSH DARC CHAIN/RING (CR) SUPERATOM ATTRIBUTES			
SUPERATOMS	**CHAIN 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)
SUPERATOMS	**RING ATTRIBUTES**		
ARY (aryl)	MON (monocyclic)	FU (fused)	
CYC (cycloalkyl)	MON (monocyclic)	FU (fused)	
	SAT (saturated)	UNS (unsaturated)	
HEF (fused heterocycle)	SAT (saturated)	UNS (unsaturated)	
HET (non-aromatic mono-cyclic heterocycle)	SAT (saturated)	UNS (unsaturated)	
HEA (heteroaryl)	no attributes on HEA		

Appendix 2:

Markush DARC and Questel Plus commands summary

I. Markush DARC commands

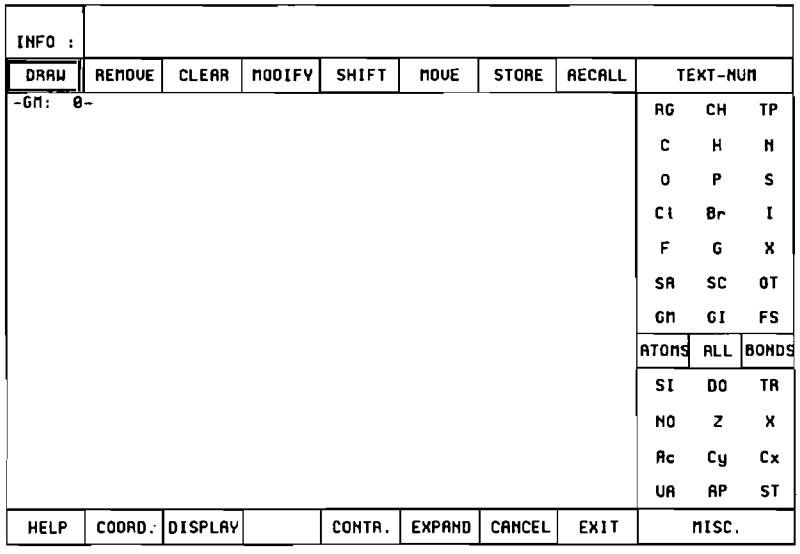
I.1. Basic command levels

ST- (BA, CN, QT, QG, RF, RE, AA, SB, BL, BI, GD, INFO) ?

QT

QG

QU- (CN, CA, GM, GI, GR , BO, AT, FS, AP, VP, ATTR, VE) ?



I.2. ST command level

COMMANDS	DEFINITION
BA	BAse (Specify or change structure database)
CN	Compound Number Search
QT	Query Text (Alphanumeric input of query)
QG	Query Graphic (Graphic input of query)
RE	REtrieve Candidates (Perform fragment or screen search)
AA	Atom-by-Atom Search (Perform atom-by-atom search on candidates)
SB	Search Bit screens
BL	Boolean Logic (Perform Boolean logic operation on answer sets)
BI	BIbliographic system (Switch to bibliographic system)
GD	Switch to Generic DARC
INFO	INFOrmation (Display news or information)
FI	FInish (Log off the system)
OP	OPtions (change terminal options, language etc)
OS	Other Specifications (add other specifications to query)

DISPLAY COMMANDS	
LI	LIst Compounds numbers
VI	VIew structure answers (main group) one at a time
VI FO	View FOcus the parts of the answers which match the query structure
VI MAX	VIew MAXimum the whole structure answer (all the groups in succession)

QUERY HANDLING COMMANDS	
SV	SaVe query or answer set
RF	Recall file (recall saved query or answer set)
ER	ERase (erase previously saved query or answer set)
HI	HIstory (list history of saved queries or answer sets)
QUERY HANDLING COMMAND PARAMETERS	
QU	QUery
CN	Compound Numbers
DN	Darc Numbers
QP	Questel-Plus query ("join" set)
BT	BaTch search query or answer
PC	PC query (query input on a PC using Darc-Chemlink)

I.3. QT command level

COMMANDS	DEFINITION
CN	Recall Compound Number of a specific compound in the database to use as a base for query input
CA	CAncel entire query, a G group or attributes
GM	Group Markush (Specify Markush group being defined)
GI	Group Identical (Create a new group identical to an existing group)
GR	GRaph (Specify the skeleton of the query)
BO	BOnds (Specify the bonds in the query)
AT	AToms (Specify the atoms in the query)
FS	Free Sites (Specify desired substitution in query)
AP	Attachment Points (Define attachments of G groups to parent group)
VP	Variable Points of attachment (Define points of variable attachment)
ATTR	ATTRibutes (Define attributes on atoms in the query)
VE	VErify (Graphic verification of the query)
FI	FInish query input (returns to ST prompt level)

II. Questel Plus commands

II.1. Questel Plus basic commands

..FI	Database (**FI**le) selection
..FI NAME	Specific database (**FI**le) selection
..LI	**LI**st answers online
..PR	**PR**int answers online
.. MEM	**MEM**ory to extract terms from an answer set
..MEMS	**MEM**ory Sort
..JOIN	To transfer compound numbers from Questel Plus to Markush DARC
..ST	**ST**op to logoff
..POP	Permanent **OP**tions

II.2. Truncation symbols

+	Open (unlimited) truncation
?	Limited (zero or one character per ?) truncation
#	Mask (exactly one character per #)

II.3. Questel Plus operators

AND	Intersection of sets
OR	Union of sets
NOT	Exclusion of sets
W	Adjacent terms in order specified
xW	Proximity in order specified
D	Adjacent terms in any order
xD	Proximity in any order
S	In the same Sentence
P	In the same Paragraph
L	Link
F	In the same Field
NOTS	**NOT** in the same Sentence
NOTP	**NOT** in the same Paragraph
NOTF	**NOT** in the same Field

Appendix 3:

WPIM database

1 Chemical Indexing in Derwent's Patent Files

1.1 Chemical Indexing in WPIL

The World Patents Index file is Derwent's file of patent information. It contains bibliographic data, titles and abstracts, and indexing generated by Derwent. This file is available on Dialog as files 350 and 351, on Orbit as files WPI and WPIL (containing pre-1981 data and data from 1981 onwards respectively) and also as the combined file WPAT, and on Questel as WPIL (previously it was split into files WPI and WPIL). Beginning with the start of the Farmdoc service, covering pharmaceutical and veterinary patents, in 1963, chemical structure information has been indexed in WPIL. Coverage of chemical structures in WPIL has increased over time, with the service being extended to cover agricultural patents through the Agdoc service in 1965, and general chemistry patents (Chemdoc) in 1970.

Structures have been indexed in WPIL using a Fragmentation coding system, which has undergone several modifications over time. These codes are used to index structural information, and there are also non-structural codes used to index information such as biological activity, chemical properties etc. Searching for structures using the Fragmentation code results in high recall, but relevance is often low, resulting in a large number of false drops.

In addition, ring systems have been indexed since 1972 using a system based on the Patterson Ring Index to assign Ring Index Numbers (RINs) to the 'less common' systems that do not have a unique Fragmentation code.

Since 1981, Derwent Registry Numbers have also been applied to specific compounds of importance to the chemical and pharmaceutical industries. There are about 2000 of these Registry Numbers, which are assigned whenever one of these compounds appears in the patent, subject to the coverage rules in Section 1.2.3

1.2 Chemical Indexing in WPIM

Beginning in Derwent Week 8701, Markush compounds and specific structures from patents have also been graphically indexed in the World Patents Index Markush file (WPIM).

Each structure record in WPIM is identified with a Compound Number, which also appears in the bibliographic file WPIL.

1.2.1 Coverage in WPIM - Derwent selection

Structures from patents in Derwent Sections B, C and E are indexed in WPIM.

Only structures from patents published in those countries considered by Derwent to be 'Major countries' are indexed. Structures from 'Equivalent' patents are not indexed by Derwent if the parent 'Basic' patent has already been indexed in WPIM or using the Fragmentation code. For more information on Derwent's selection policy, see the Derwent Instruction Manuals.

1.2.2 Coverage in WPIM over time

Indexing in WPIM began in Derwent Week 8701. At first, coverage was not complete: in particular, some of the more complex structures (known as 'nasties') were not indexed graphically. Structures which have been graphically indexed in WPIM are identified by a special code M904 in the corresponding WPIL record. This allows the searcher to use Markush DARC to search for those records with graphics indexing, and then to search using the less specific Fragmentation code for those records with no graphics indexing. More details on how to do this are given later in this section.

All chemical structures from Major country Basic patents in Derwent sections B, C and E are now graphically indexed in WPIM (subject to the coverage rules - see 1.2.3).

Coverage in WPIM over time is shown in the following table:

	B	C	E
8701 - 8920	structures except oligosaccharides, polypeptides & polymers, and 'nasties'	as for B	structures except 'nasties'
8920 - 9101	all structures except oligosaccharides, polypeptides & polymers	as for B	"
9101 - 9116	all structures except polymers & oligosaccharides	as for B	"
9116 - 9125	"	as for B	all structures
9125 - present	all structures	all structures	all structures

The database is currently updated weekly, level or slightly ahead of the Fragmentation code.

1.2.3 Depth of coverage in WPIM

In WPIM, all of the specific examples and all of the information contained in the patent claims are included as a minimum. Generic terms (such as 'alkyl') in the claims are represented using superatoms (such as CHK). Specific examples of these generic terms are also indexed, so a group G1 described as alkyl with an example ethyl would be indexed as G1 = CHK, Et. Information in the Wider Disclosure is also indexed, at least generically.

All structures in the claims, examples and disclosure covered in the categories below are indexed in WPIM:

1. All compounds described as new
2. Products of new processes, including materials purified in new ways
3. Compounds which are removed, when this is important to the novelty of the invention, e.g. removal of pollutants from waste streams
4. Compounds used to effect removal when this is important
5. Compounds that are analysed or detected, when this is the novelty of the invention
6. Compounds used in analysis or detection, which are important to the invention
7. Catalysts that are new
8. Important ingredients of compositions

Registry Numbers are also assigned to starting materials (including catalysts) which are not new, but these are not indexed in WPIM

2. Searching WPIM and WPIL

2.1 Compound Numbers

As mentioned earlier, the link between WPIM and the WPIL bibliographic file is the Compound Number (CN). Each structure record in WPIM has a unique Compound Number. This Compound Number also appears in the WPIL record for that patent. Compound Numbers appear in WPIL in their own field, (the CN field), and also in the relevant M1 to M6 Fragmentation coding fields.

There are two types of Compound Number: Markush Compound Numbers and Specific Compound Numbers.

2.1.1. Markush Compound Numbers

These are of the general format:

YYWW-CCCSS

YY is the year and WW is the Derwent week number for the document. CCC is a three character identifier unique to a document in any given week, and SS is an integer from 01 to 99. One or more of these Markush Compound Numbers may be assigned to any document containing compounds represented as Markush (generic) structures. Compound Numbers in this format may also be assigned to a single compound if this compound is claimed as new.

2.1.2. Specific Compound Numbers

These are of the general format:

RNNNNN

NNNNN is a sequential number from 00001 (the Specific Compound Number for formaldehyde).

A number of patents concern individual compounds and may not contain a Markush formula. In these cases, if the specific compounds are not described as new, a Specific Compound Number for each compound is applied to that document. If a Specific Compound Number for a compound does not already exist, a new one is created and added to Derwent's internal database.

For example, if a patent claims a new process for the production of para-xylene, a Specific Compound Number for p-xylene would be used to index the compound. The number is re-used whenever p-xylene is again disclosed specifically (i.e. not as a possibility of a Markush).

Some patents may disclose a Markush formula, and also other unrelated single compounds that fall within Derwent's coverage rules. Thus a given patent may be indexed with a mixture of Markush and Specific Compound Numbers.

Important Note:

Specific Compound Numbers are **not** applied when the compound is defined within a Markush structure.

2.2 Searching CNs in WPIL

As mentioned earlier, Compound Numbers (CNs and SCNs) provide the link between the structures indexed in WPIM and the bibliographic records in WPIL. Normally, the result of a Markush DARC search in WPIM will be a list of Compound Numbers. If you are searching on Questel, these are saved using the SV CN command, and then the list is searched in WPIL using the *MDARC command, to retrieve the bibliographic records corresponding to these

Compound Numbers. When accessing through Orbit, the PRT SEL and SEL commands are used instead of the SV CN and *MDARC commands, while on Dialog, the MSELECT or MSAVE commands are used.

As we have seen, each bibliographic record may have more than one Compound Number associated with it, and, conversely, a Specific Compound Number may be applied to more than one patent. Therefore, the number of bibliographic records when the CNs are transferred to the bibliographic file is unlikely to be the same as the number of Compound Numbers you started with.

It is also possible to search for Compound Numbers in WPIL by entering them directly. They can be searched in the /CN field or the appropriate M1-M6 field.

2.2.2 Roles

Compound Numbers in WPIL have Role Qualifiers associated with them, indicating the role of that particular compound in the patent. Each CN may have one or more roles. The roles are as follows:

A Substance analysed/detected
C Catalyst
D Detecting agent
M Component of a mixture
N New compound
P Known compound produced
Q Product defined by starting materials
R Removing/purifying agent
S Starting material
U Use of a single compound
X Substance removed
Z Miscellaneous

Role qualifiers can be applied to a CN list when you cross-file into WPIL, to restrict your search to one or more of these categories. The format is:

***MDARC SEARCH1.J.K.L** on Questel

MSEL *(followed by the name of the answer set when prompted)* on Dialog
S S1 (S) RL=(J OR K OR L)

where SEARCH1 is the name of the saved CN set, and J, K and L are the appropriate Roles, which are ORed together.

For example, to restrict the search to structures which are used as **either** a detecting agent (D) **or** a removing/purifying agent (R), you could enter:

***MDARC SEARCH1.D.R** on Questel

MSEL on Dialog
S S1 (S) RL=(D OR R)

2.2.3 LINKing and ANDing Compound Numbers with other terms

Compound Numbers entered directly or searched as *MDARC lists may be ANDed with other search terms, such as Patentee names or date ranges. For example, if you had carried out a search in WPIM and saved the Compound Numbers as SEARCH1, and wanted to retrieve only those patents with a publication date in 1991, you could enter:

***MDARC SEARCH1 AND 91/PD**

Compound Numbers may also be searched in the appropriate M1-M6 field, where they can be LINKed to Fragmentation codes or Ring Index Numbers. For example, to retrieve those patents where the structure of interest was described as an antifungal in SUBS M2, you could enter:

/M2 *MDARC SEARCH1 L P241

On Dialog and Orbit, the terms can be LINKed or ANDed to the results of the MSEL or SELECT operations in a similar way.

It is also possible to carry out a search in WPIL, and transfer the results to WPIM using the Join command (described in Section VII.1.1) to carry out a structure search on this set, rather than searching the whole WPIM database.

2.2.4 Searching WPIM and the Fragmentation code in WPIL

As mentioned in an earlier section, a special code M904 is applied to those structures in WPIL which have graphics indexing in WPIM. This means that it is possible to carry out a Markush DARC search on those structures which have graphics indexing, and then carry out a Fragmentation code search on only those structures which do not have graphics indexing.

a. On Questel:

Carry out the search on WPIM first, and save the Compound Numbers using the SV CN command. Cross-file to WPIL, and carry out the Fragmentation code search. Finish the code search with a line to exclude those structures which have an M904 code. Finally, find the patent records corresponding to the results of your Markush DARC search. These results should then be ORed with the results of the Fragmentation code search. For example, after a 6 line strategy, lines 7-9 should read:

7 **/M2/M3 (6 NOTL M904)**
8 ***MDARC SEARCH1**
9 **7 OR 8**

b. On Dialog:

Either: carry out the WPIM search, and then carry out the Fragmentation code search, finishing with the lines:

7	**S (S6 (NOTS) M2,M3=M904)**	
8	**MSEL**	*(followed by the answer set, usually R1)*
n	**S S7 OR Sn-1**	*(where n is the number of the search statement)*

or: carry out the WPIM search and use the MSAVE command to save the CNs. Then carry out the Fragmentation code search, finishing with the lines:

7 **S (S6 (NOTS) M2,M3=M904)**
8 **EX SEARCH1**
9 **S7 OR S8**

Remember to erase the MSAVE set when you have finished.

c. On Orbit:

Carry out the WPIM search, and use the PRINTSEL command to save the CNs. Then carry out the Fragmentation code search, finishing with the lines:

```
7    6+<M904
8    SEL SEARCH1 1-n          where n is the number of CNs
9    7 . 8
```

Note that M904 is linked to the codes which correspond to the (Markush) structure, rather than applied to the whole patent record. This means that if one compound in a patent has been graphically indexed in WPIM, and another has not, only the structure without graphics indexing will be included in the Fragmentation code search.

3. Conventions & Indexing policy in WPIM

There are a few minor differences in some of the indexing conventions between the Markush DARC databases. Those conventions which are common to both databases are documented in Section 4 of this manual. The following is a summary of those conventions which apply only to WPIM.

3.1 Definition of superatoms

A complete list of the superatoms used in WPIM is given below. The only difference between WPIM and MPHARM here is the definition of the ARY superatom.

ACT Actinide (including actinium)
ACY Acyl (i.e. residue after removing 1 or more OH groups from an acid)
AMX Alkali(ne Earth) Metal
ARY Aryl - Carbocyclic, optionally fused, containing at least 1 benzene ring
A35 Group IIIa-Va Metal - Al, Ga, In, Tl, Ge, Sn, Pb, Sb, Bi
CHE Alkenyl, Alkenylene
CHK Alkyl, Alkylene
CHY Alkynyl, Alkynylene
CYC Cycloaliphatic - Carbocyclic, optionally fused
DYE Dye group or residue (chromophore, fluorophore)
HAL Halogen
HEA Heterocyclic - Aromatic*, Monocyclic (5 or 6 ring atoms)
HEF Heterocyclic - Fused, optionally aromatic
HET Heterocyclic - Non-aromatic, Monocyclic
LAN Lanthanide (including Lanthanum)
MX Metal, any
PEG Polymer End Group
POL Polymer, Polypeptide residue
PRT Protecting group
TRM Transition Metals excluding Lanthanum
UNK Undefined group
XX Any atom or group excluding hydrogen

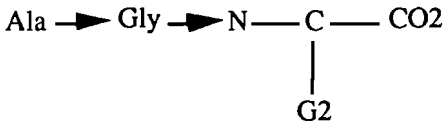
* *A monoheterocycle is considered to be 'aromatic' if:*
a. it contains 6 atoms and is fully normalised, e.g. pyridine
or b. it contains 5 atoms and maximum unsaturation, e.g. furan

3.2 Peptides

In WPIM, peptide chains of 4 or more and up to 31 amino acids are indexed using peptide superatoms to represent the amino acid units. These superatoms are used only for unmodified amino acid units. Modified amino acid units are indexed as for other chemical structures.

In WPIM, the SP attribute, which specifies the attachment position of a substituent on a peptide superatom, can only have a value of 3, to indicate alpha-amino substitution; it is therefore usually not applied to records in the database, and should not be used in searching. Amino acid units modified by substitution on other positions should be searched using the explicitly drawn structure.

If an explicitly drawn nitrogen atom is attached to a chain of peptide superatoms, a peptide bond is used (rather than a single bond), for example, in the following structure:

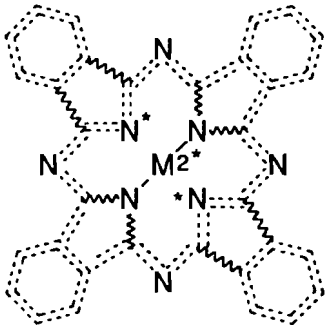


The peptide configuration attributes D, L and DL are applied to peptide superatoms where this information is given in the patent, and can be searched. Searching without these attributes will retrieve superatoms with or without attributes.

3.3 Special cases

3.3.1 Phthalocyanines

In WPIM, most metal phthalocyanines have been indexed with all the bonds around the ring normalised; however, some have been indexed with some single bonds. To search for all metal phthalocyanines, use a structure like that shown below, with free sites on each of the two N atoms not bonded to the metal, and z lists for the bonds (shown ~~~~) which could be single or normalised:



Non-metallised compounds have been indexed fully normalised. To search for metallised and non-metallised structures, use a query structure with variable bonds like that shown above, but without the metal and with free sites on the other two nitrogen atoms.

3.4 Polymers in WPIM

Polymers which are classified under Derwent Section B or C are graphically indexed in WPIM as the monomers. If the monomers are not given and are not deducible from the patent, the structural repeating unit may be indexed, qualified by a numbering attribute and associated text (which is not currently searchable). Polymers in Sections B and C are only indexed if they appear in examples and are fully defined, unless a polymer is the essential feature of an invention.

There are two ways to search for a polymer in Markush DARC: by monomer, or by the structural repeating unit. For full retrieval, you may need to search in both ways.

3.4.1 Searching by monomer

Draw the monomer as a structure query in Markush DARC. If you want to search for a copolymer of two or more different monomers, search for each monomer as a separate structure query. Save the CN lists, and cross-file into WPIL. Then search for the CNs in WPIL with the role Q (see section 2.2.2)., LINKing them to the code for a polymer, V742 or V743. If you are searching for a copolymer and want to specify more than one monomer, the CN lists obtained as a result of each Markush DARC search should be LINKed to one another as well as to the V7 code.

3.4.2 Searching by Structural Repeating Unit

Ordered polymers (such as nylon 6,6), of the general formula:

E1—$[A]_m$—$[B]_n$—E2 *where A and B are the repeating units and E1, E2 are end groups*

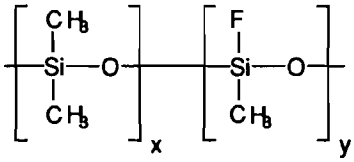
are indexed in the form:

E1—A—B—A—B—E2 (with appropriate text notes)

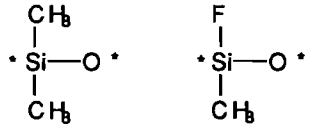
and can therefore be searched in any of the following forms:

A—B *B—A* E1—A—B* *A—B—E2 * A—B—A* etc.

More complex polymers which contain more than two different repeating units, in unknown order, should be searched by drawing the separate units in a single structure query, with free sites. For example, to search for a polymer of the type:



construct a single search query containing two fragments on the same screen:



3.5 File segments

The following File Segments are applied in WPIM. See Section III.4.16 for details on how to search using File Segments.

CPI Sections

A	Section A
B	Sections B or C
E	Section E

General

Y	Mixture
Z	Salt
1	Specific Compound Number

Polypeptides

P	Polypeptide

Non-polymer compounds

C	Coordination Compound/Complex
L	Oligomer (3-10 units or 'oligomer' in patent)
W	Extended Structure (e.g. zeolites)
M	Metals and Alloys
V	Ordinary organic chemicals
7	Inorganic

Polymers

F	Any polymers
N	Natural polymer

4. Support

Derwent holds training classes which cover all aspects of searching Derwent's databases, including WPIM. For details, please contact Technical and Online Services at your local Derwent office.

The Derwent Help Desk will provide advice and assistance with any aspect of searching in complete confidence, and can be contacted by 'phone on:

+44 71 344 2999 in the UK
800 451 3551 (toll free) at Derwent Inc. in the US
1 03 3581 7711 via NGB in Japan

Derwent also provides a Bureau Search service, which can carry out searches of any of the Derwent databases, including WPIM.

5. Markush TOPFRAG

5.1 Overview

Construct queries offline

Markush TOPFRAG is a microcomputer program which allows you to draw chemical structures on your PC offline, and converts them to strategies for searching the WPIM database, and also, if you wish, the Fragmentation Code on WPIL on Questel, Orbit and Dialog.

The program is based on the Windows PSIDOM chemical structure drawing interface, which enables you to draw structures quickly and easily, using a mouse and/or the keyboard. Drawing tools and atoms and bonds can be selected from a palette on the screen, and menu commands allow you to define variable G groups and place free sites to define generic queries. Templates are provided for many common structural groups, and can be customised and added to.

Structure checks

Once you have drawn your structure query, strategies for searching WPIM and/or WPIL are generated by selecting a menu command. Markush TOPFRAG will automatically detect tautomeric structures, and apply the correct structural conventions where possible, or warn you if your query needs modification.

Strategies

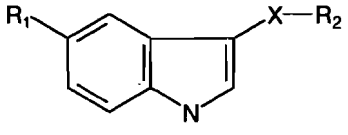
A strategy for searching WPIM on Markush DARC is generated in seconds. The strategy consists of a simple text file, which you can upload to Markush DARC using a suitable communications package. If you have a graphics terminal emulator, you can view the structures on-screen; if not, you can save the Compound Numbers and retrieve the corresponding bibliographic records in WPIL.

Searching both WPIM and the Fragmentation code

Markush TOPFRAG will generate strategies for both Markush DARC and the Fragmentation code from a single structure query.

5.2 Example

5.2.1 Structure query:



The ring system is drawn using templates (there are templates for most common ring systems, and you can add to the collection simply by drawing a structure and saving it in the template directory). The CHK group is selected from a menu of chain superatoms, and attached to the ring using the pencil tool controlled by the mouse. The chain length is limited to 1-4C by selecting from the attributes menu. Since answers containing specific lower alkyl groups (such as ethyl, n-butylene) are required as well as those containing 'alkyl', the NT translation attribute is applied to the CHK. The nitrogen atom is selected

from the Common Atoms palette and attached to the ring using the pencil tool, and G1 is attached in a similar way.

G1 is defined as O, S or N, with free sites on each. Each atom is selected from the Common Atoms palette and placed on the screen. An attachment point is placed on each by clicking. Free sites are applied to these three atoms and to the nitrogen attached to the CHK by choosing a command from a menu.

5.2.2 Generating the strategy

The query structure is now complete, so another menu command is used to generate the strategy or strategies. If just the Markush DARC strategy is required, this only takes a second or two. The program automatically checks for tautomeric structures, and in this case, normalises the benzene ring. The strategy is displayed in the MTFEDIT window; from here, it can be edited using the standard Windows commands such as Copy and Paste, printed with or without the structure, and saved as a text file.

If the Fragmentation code strategy as well as the Markush DARC strategy is required, the strategy generation procedure takes a few minutes rather than seconds. Since there are free sites in the structure, the program asks for further information about the type of substitution permitted at these sites, so that the code strategy can be made as specific as possible. Extra commands are available in the MTFEDIT window when the code strategy is displayed, allowing you to see definitions of the codes, and providing help with editing and checking the strategy.

5.2.3 Uploading the strategy

The strategies can now be uploaded, using your normal communications software to send the text files. To send the Markush DARC strategy, log on to WPIM in the usual way, and at the –ST– prompt, type QT. Send the Markush DARC strategy at the –QT– prompt. Once the strategy has finished uploading, you can verify the query if you are using a suitable graphics terminal emulator by typing VE at the –QT– prompt. If you do not have a graphics emulator, it is possible to do a textual verification by typing VE TX.

Then type FI to leave –QT– level, and select any Other Specifications and File Segments. Then type RE (or SB) at the –ST– prompt, followed by AA in the usual way.

When the search has finished, save the answers using the SV CN command, and enter a name for the saved answer set (e.g. SEARCH1). If you are using a graphics emulator, you can view the structures retrieved, by typing VI FO (or VI or VI MAX) at the –ST– prompt. Leave Markush DARC and enter WPIL by typing BI at the –ST– prompt, followed by ..FI WPIL (for alternative procedures through the Dialog and Orbit Gateways, see Section 2.2.4 or Dialog/Orbit literature). If you also want to search the Fragmentation code on WPIL, upload the Code strategy before using the *MD command. When the Fragmentation code search has finished, remove any records from the answer set which have Markush DARC indexing. Then find the records corresponding to the results of your Markush DARC search, and OR the result with the result of the code strategy, as described in Section 2.2.4.

Appendix 4:

MPHARM database

I. Pharmsearch database overview

PHARMSEARCH is a patent database produced by I.N.P.I., Institut National de la Propriété Industrielle, which is the French Patent and Trademark Office.

PHARMSEARCH covers all patent publications of European (including Euro-PCT) and International applications (PCT), American publications (US) , French (FR), United Kingdom (GB) and German (DE) applications in the field of pharmaceutical chemistry and biology.

When complete, PHARMSEARCH will include

European, PCT and US patents from 1978,
French patents from 1961,
British and German patents from 1992.

The information about the exact period coverage is displayed online at the connection to the database. Most of these patents are available a mere six weeks after the day of their publication. The database is updated twice a month.

Example of the online message at the connection the 11th of november 1993:
The PHARM file and its structural associated file MPHARM cover:

US patents	: from 1985 week 27 to 1993 week 31
FR and EP patents	: from 1985 week 35 to 1993 week 39
DE and GB patents	: from 1992 week 36 to 1993 week 31
PCT patents	: from 1993 week 01 to 1993 week 31

BSM (Specific French Drug Patents): complete collection

PHARMSEARCH is composed of two files.

MPHARM, the structure file, contains Markush structures (80%) and single compounds (20%).

PHARM, the bibliographic file, contains bibliographic patent information, indexing information, specially written english abstracts and images (only displayable) from the patent.

The "link" between MPHARM and PHARM is the Compound Number (CN). Compound Numbers identify structure records in MPHARM and are searchable in their own field (CN field) of the corresponding bibliographic records of the PHARM file.

TRAINING FILES for how to use PHARMSEARCH.

ZPHARM file, the companion training file of PHARM includes updates from January 1987 to October 1987 (3595 references) **IPAT file**, the companion training file of MPHARM contains 1520 structures, corresponding to the structural information of a part of the ZPHARM file.

II. Patent coverage

PHARMSEARCH covers patents which are classified in at least one of the following classes of the International Patent Classification: **A01N, A61B-005/14, A61B-010/00, A61K, A23L-001/236, A23L-001/237, C07** and **C12 "classes".**

Since only pharmaceutical patents are covered in the PHARMSEARCH database, patents comprised in these classes are not all included in the database; patents are selected according the following criteria:

For the **A01N** subclass (biocides), only patents related to pharmaceutical activity are included.

For the **A61K-005/14** subclass (devices for taking samples of blood) and **A61K-010/00** (methods or instruments for diagnosis) subdivisions , only documents which cover in vitro diagnostics are covered since 1993.

For the **A61K** subclassm (Preparations for medical, dental and toilet purposes), only patents claiming preparations for medical purposes are included in the database; these patents are classified in the A61K-009/00, A61K-031/00, A61K-033/00, A61K-035/00, A61K-037/00, A61K-039/00, A61K-041/00, A61K-043/00, A61K-045/00, A61K-047/00 and A61K-049/00 subclasses.

Patents comprised in the **A23L-001/236** (artificial sweetening agents) and **A23L-001/237** (diuretic salt substitutes) subdivisions are all included in the database.

Patents contained in the **C07** (organic chemistry) and **C12N, C12P** and **C12Q** (microorganisms and enzymes) subclasses are included in the database when they cover chemical or biological substances possessing at least one of the following properties: therapeutically active agents (human and veterinary medicine), sweetening agents, dietetic salt substitutes, sunscreens and in vivo testing agents.

For the **G01N-033** (investigating or analysing materials by specific methods not covered by preceding groups) subdivision only documents which cover in vitro diagnostics are covered since 1993.

III. MPHARM: Pharmsearch structure file

III.1. Structural coverage

Chemical compounds are recorded in the MPHARM structure file in the following conditions:

1) New compounds claimed as active compounds of pharmaceutical compositions;

2) New compounds claimed as intermediates in preparation processes (synthesis and purification processes) of pharmaceutically active compounds;

3) Known compounds which are active compounds of new pharmaceutical compositions;

4) Known pharmaceutically active compounds with new preparation or purification processes or with new pharmaceutical activities.

III.2. Markush structures and single compounds in MPHARM

Two types of records (two types of Compound Number) can be found in the MPHARM file: Markush structures and single compounds.

III.2.1. Markush structures

Compounds which are described in a patent document as a possibility of a Markush structure are recorded in the MPHARM file as one or more Markush structures covering all the specific examples and corresponding generic information extracted from the claims.

Markush structures are recorded in the MPHARM file with a compound number having the following format:

YYMMXXXX-NN

in which

YY = last two digits of year of publication of the patent
MM = month of publication
XXXX = sequential number identifying the patent in a given month
NN = integer in the range 01 to 99

If several Markush structures are derived from a patent document, the first Markush is assigned the YYMMXXXX-01 Compound Number, the second the YYMMXXXX-02 Compound Number, etc...

III.2.2. Single compounds

Chemical compounds are recorded in the MPHARM file as single compounds when they are described specifically in the patent document (i.e. when they are not described as a possibility of a Markush structure).

This may occur in the following cases:
- single compounds claimed as new,
- single compounds with new processes of preparation or new pharmaceutical activities,
- single compounds as active ingredients of new pharmaceutical compositions.

When a single compound is recorded for the first time in the MPHARM file, it is assigned a compound number of the following format:
RYYMMXXXX-NN *(Compound Number format preceded by R)*

The same compound number is reused when this single compound is again described specifically in another patent document.

Thus, the YY, MM and XXXX identifiers relate respectively to the year, the month of publication and the sequential number within the month of the first patent document included in PHARMSEARCH that disclosed the single compound.

It may occur that both Markush structures and single compounds are derived from the same patent document and that the single compounds are recorded for the first time in MPHARM. In this case the n Markush structures will be assigned the YYMMXXXX-01, YYMMXXXX-02,... YYMMXXXX-n Compound Numbers and the m single compounds the RYYMMXXXX-n+1,... RYYMMXXXX-n+m Compound Numbers.

III.3. File segments used in MPHARM

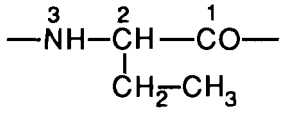
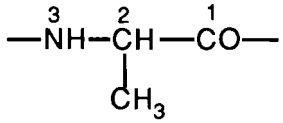
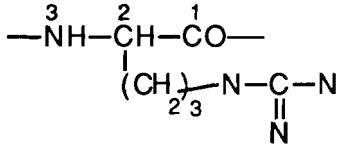
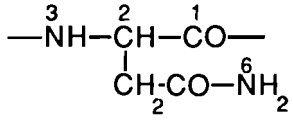
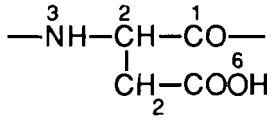
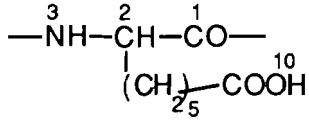
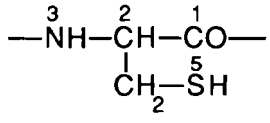
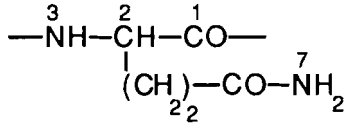
File segment 1 is assigned to all specific compounds records (with a CN in the form RYYMMXXXX-NN).

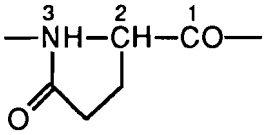
File segment P is assigned to polypeptides (at least two amino acids superatoms coupled together in a molecule).

III.4. Peptide searching in MPHARM

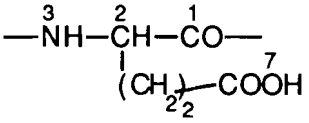
In MPHARM the amino acid superatoms are used as soon as there are two amino acids coupled together in a molecule through a peptide bond provided that these amino acids are either unmodified or substituted on a function COOH, SH, NH2, CONH2, OH or on their N- or CO1- terminal fonction.

III.4.1. Amino acids superatoms and their substituable positions

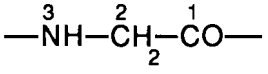
SUPERATOM	STRUCTURE	SUBSTITUABLE POSITION
ABU		1/3
ALA		1/3
ARG		1/3
ASN		1/3/6
ASP		1/3/6
ASU		1/3/10
CYS		1/3/6
GLN		1/3/7



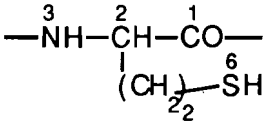
GLP 1/3



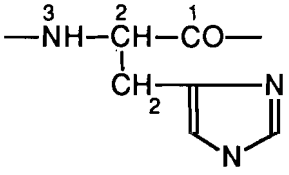
GLU 1/3/7



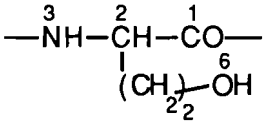
GLY 1/3



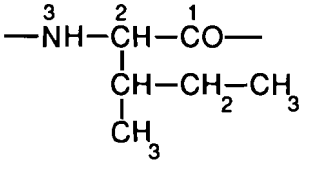
HCY 1/3/6



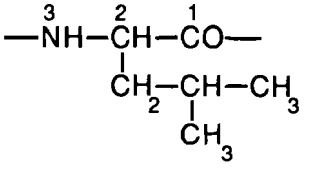
HIS 1/3



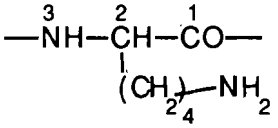
HSE 1/3/6



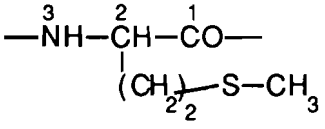
ILE 1/3



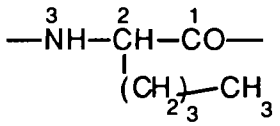
LEU 1/3



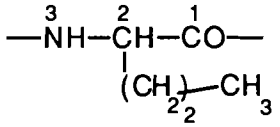
LYS 1/3/8



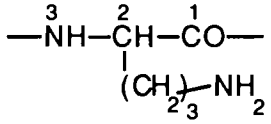
MET 1/3



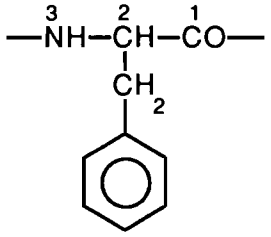
NLE 1/3



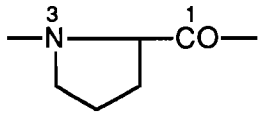
NVA 1/3



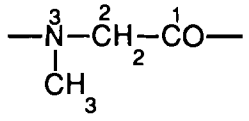
ORN 1/3/7



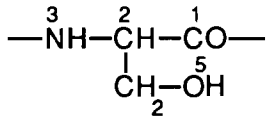
PHE 1/3



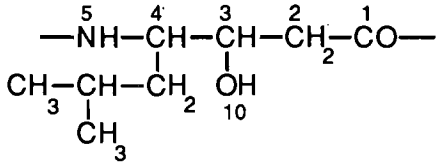
PRO 1/3



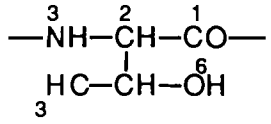
SAR 1/3



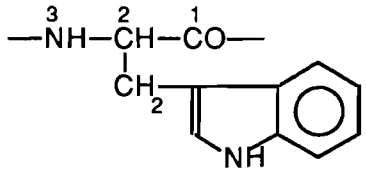
SER 1/3/5



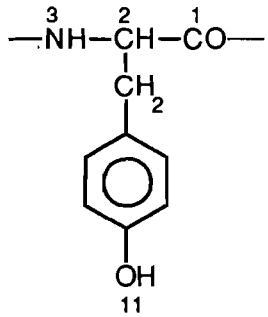
STA 1/5/10



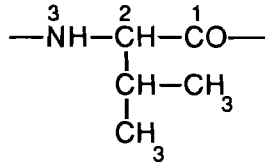
THR 1/3/6



TRP **1/3**



TYR **1/3/11**

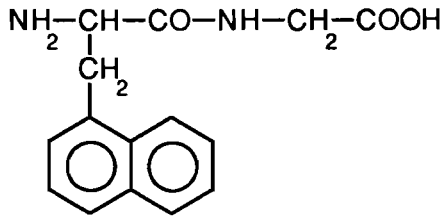


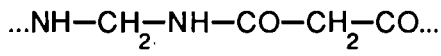
VAL **1/3**

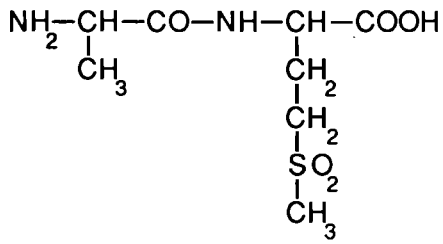
III.4.2. Modified Peptides

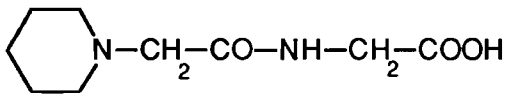
When substituted on their side chain (except on the above-mentioned functions), deeply modified, reduced, oxided, retroinverted or with their N-terminal involved in a tautomeric bond, the amino acids are indexed as ordinary chemicals. When a peptide structure does not contain two or more adjacent amino acids which are unmodified or substituted on a functional group, they are indexed as ordinary chemicals.

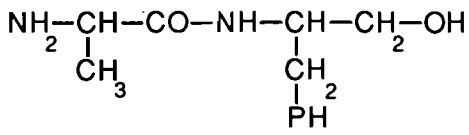
Examples:

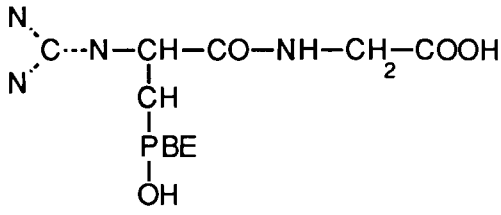










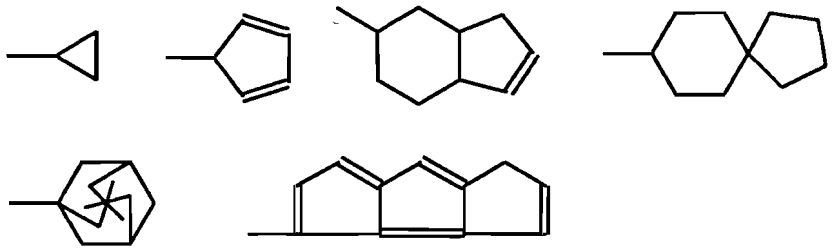


III.5. Ring Systems and Superatoms

III.5.1. Cyclic hydrocarbons

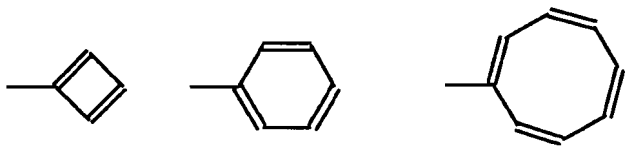
III.5.1.1. CYC

saturated or unsaturated, mono- or polycyclic system containing no aromatic rings, including bridged and spiro hydrocarbons.
ex:



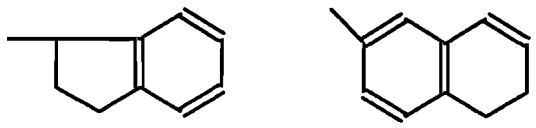
III.5.1.2. ARY

1. **ARY(MON):** monocyclic hydrocarbon system comprising an even number of atoms and alternating single and double bonds.
ex:

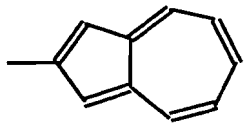


2. **ARY (FU):**
2.1. fused hydrocarbon system comprising at least one aromatic monocyclic `ring`.

ex:



2.2. fused hydrocarbon system comprising an even number of atoms and alternating single and double bonds on the cycle periphery.
ex:

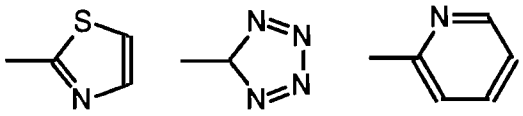


III.5.2. Heterocycles

III.5.2.1. HEA

monocyclic aromatic heterocycle comprising 5 ring members and 2 double bonds or 6 ring members and 3 double bonds.

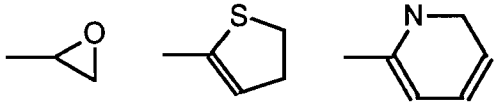
ex:



III.5.2.2. HET

saturated or unsaturated, monocyclic non-aromatic hetrocycle.

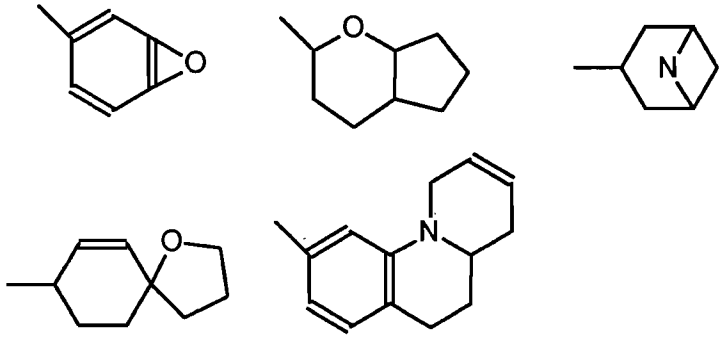
ex:



III.5.2.3. HEF

saturated or unsaturated, fused aromatic or non-aromatic heterocycle, including bridged and spiro heterocycles.

ex:



Appendix 5:

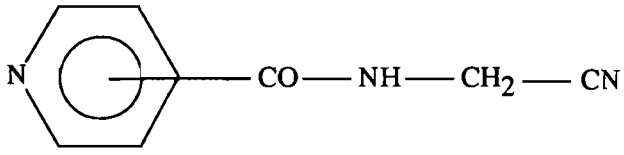
Search examples

I. WPIM examples

I.1. Example with variable points of attachment

I.1.1. Question

Find in WPIM all pyridine derivatives with the following structure:



The pyridine may optionally be substituted by another group.
The methylene group may optionally carry a substituent.

I.1.2. Query formulation

The text input can be carried out as follows:

- Graph of the query:

```
 -QU-  (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)   ?   GR
* GRAPH
?    1:5
?    6:11-6
?
 -QU-  (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)   ?   -
```

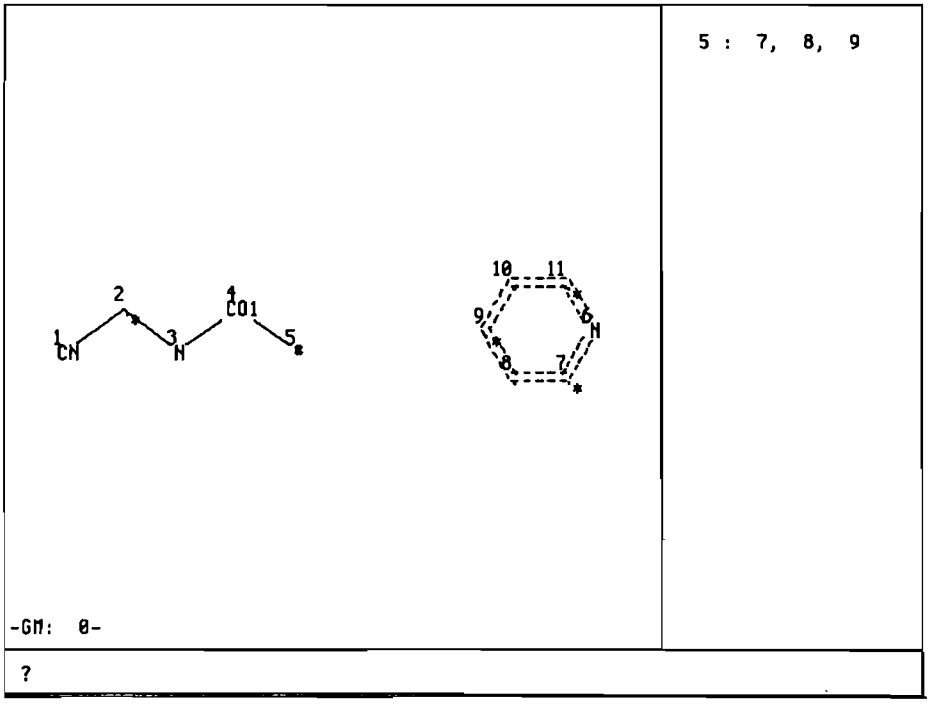
- Atoms, bonds and free sites:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)   ?   AT
*ATOMS
?   N 3,6
?   CN 1
?   CO1 4
?   # 5
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)   ?   BO
*BONDS
?   NO 6:11-6
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)   ?   FS
*FREE SITES
?   1 2,7,9,11
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)   ?   _
```

- Variable points of attachment:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)   ?   VP
* VARIABLE POSITIONS *
NODE     5 :
?   7,8,9
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)   ?   _
```

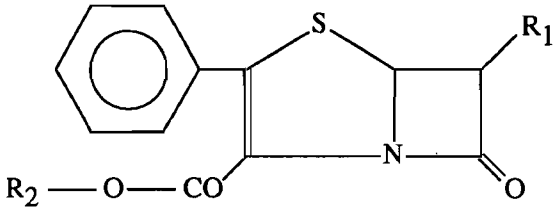
- Graphic verification:



I.2. Examples with generic variables

I.2.1. Question 1

Find in WPIM all penem derivatives having the following generic structure:



R1: optionally substituted alkyl

R2: hydrogen or ester group

I.2.2. Query 1 formulation

The text input can be carried out as follows:
- Graph of the query:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  GR
*GRAPH
?  1:8-2
?  4-8
?  3-9
?  5-10-11
?  12:17-12
?  6-12
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

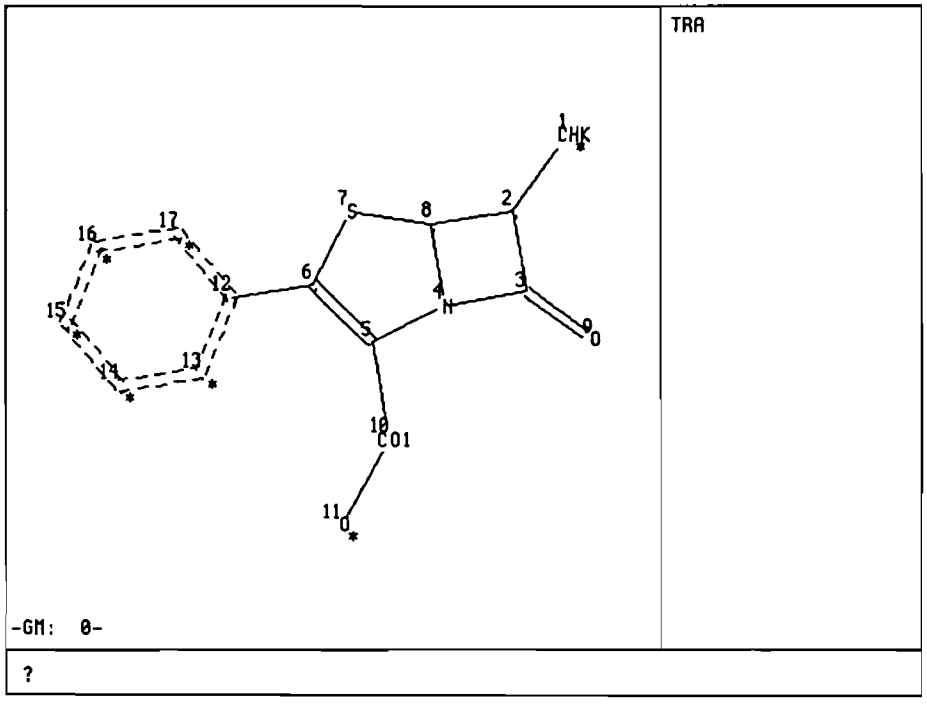
- Atoms and bonds:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AT
*ATOMS
?  CHK 1
?  O 9,11
?  S 7
?  N 4
?  CO1 10
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  BO
*BONDS
?  NO 12:17-12
?  DO 5-6,3-9
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

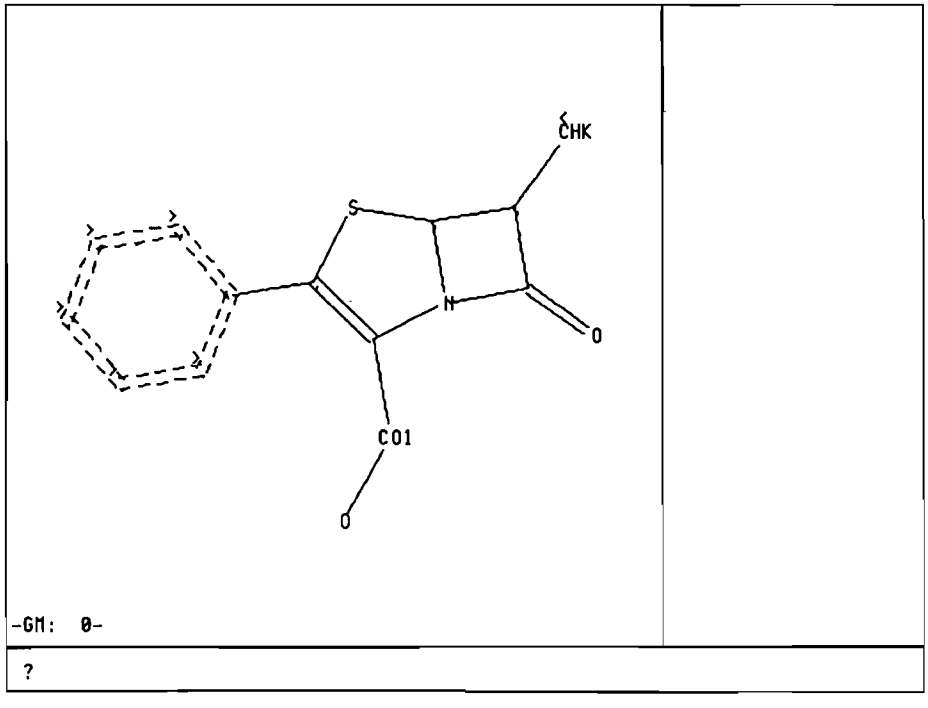
- Free sites and translation attributes:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  FS
*FREE SITES
?  1 1,11,13,14,15,16,17
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  TRA
*TRANSLATION ATTRIBUTES
?  NT 1
?  BT 13,14,15,16,17
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?
```

- Graphic verification:

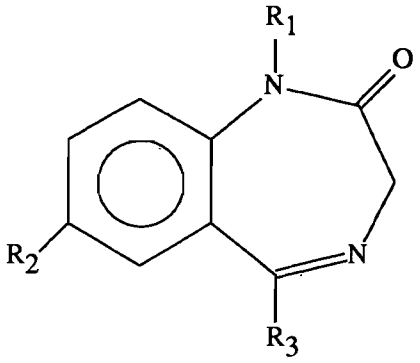


- Graphic verification including translation:



I.2.3. Question 2

Find in WPIM all benzodiazepine derivatives having the following generic structure:



R1: alkyl group

R2: halogen

R3: optionally substituted phenyl ring

I.2.4. Query 2 formulation

The text input can be carried out as follows:
The graph is input as follows:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  GR
*GRAPH
?  1:11-1
?  3-18
?  5-11
?  6-17
?  10-19
?  12:17-12
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AT
```

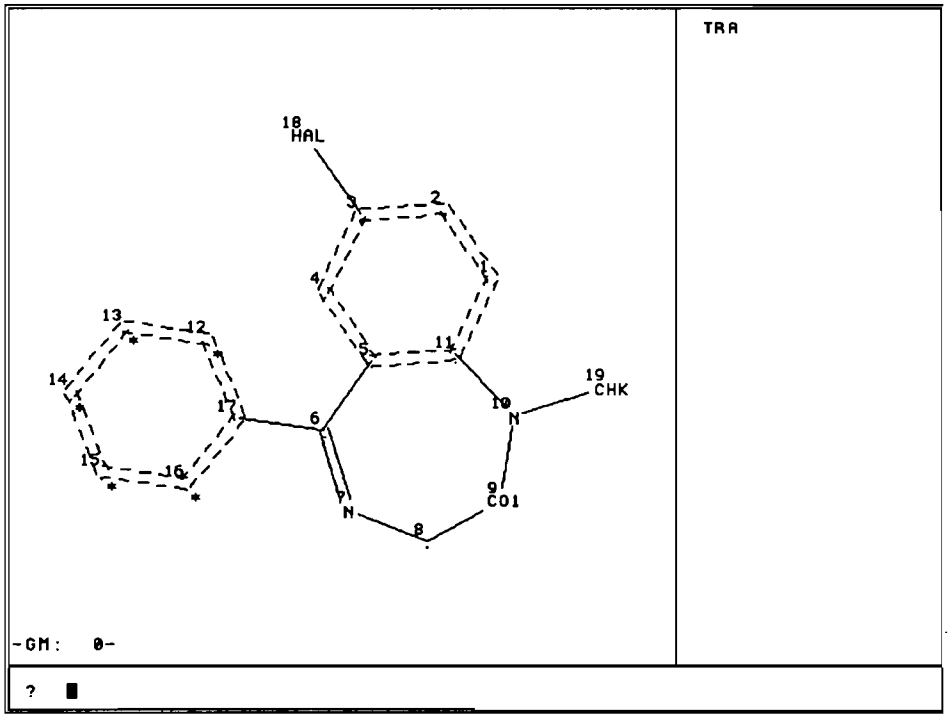
Then the atoms and bonds are specified:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AT
*ATOMS
?  N 7,10
?  HAL 18
?  CO1 9
?  CHK 19
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  BO
*BONDS
?  NO 1:5-11,1-11,12:17-12
?  DO 6-7
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

Free sites and translation attributes are input as follows:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  FS
*FREE SITES
?  1 12,13,14,15,16
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  ATTR
-ATTR- (FS,CH,AV,AM,CR,MU,PA,DT,SP,TRA)  ? TRA
*TRANSLATION ATTRIBUTES
?  NT 18,19
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

Graphic verification:

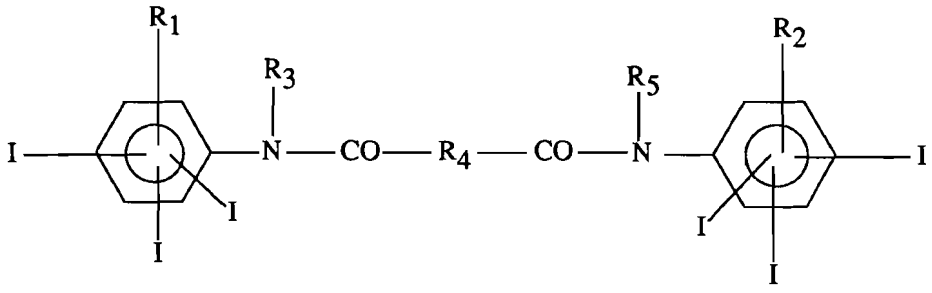


II. MPHARM examples

II.1. Example 1: variable points of attachment and generic variables

II.1.1. Question

Search for the following X-ray contrast agents:



R1, R2: hydroxy or ether group, or optionally mono- or disubstituted amino.

II.1.2. Query formulation

The text input can be carried out as follows:

- Graph of the query:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  GR
*GRAPH
?  1:6-1
?  7-8
?  9-10
?  11-12
?  13:15
?  16:26
?  1-16
?  21-26
?  27-28
?  29-30
?  31-32
?  33:35
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  _
```

- Atoms and bonds:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AT
*ATOMS
?  N 16,20
?  CO1 17,19,14,34
?  CHK 18
?  I 7,9,11,27,29,31
?  * 8,10,12,15,28,30,32,35
?  G1 33
?  G2 13
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  BO
*BONDS
?  NO 1:6-1, 21:26-21
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

- Free sites and translation:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  FS
*FREE SITES
?  1 16,20
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  TRA
*TRANSLATION ATTRIBUTES
?  NT 18
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

- Chain attribute:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  ATTR
-ATTR- (FS,CH,AU,AM,CR,MU,PA,DT,SP,TRA)  ? CR
*CCHAIN/RING
?  HI 1
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?
```

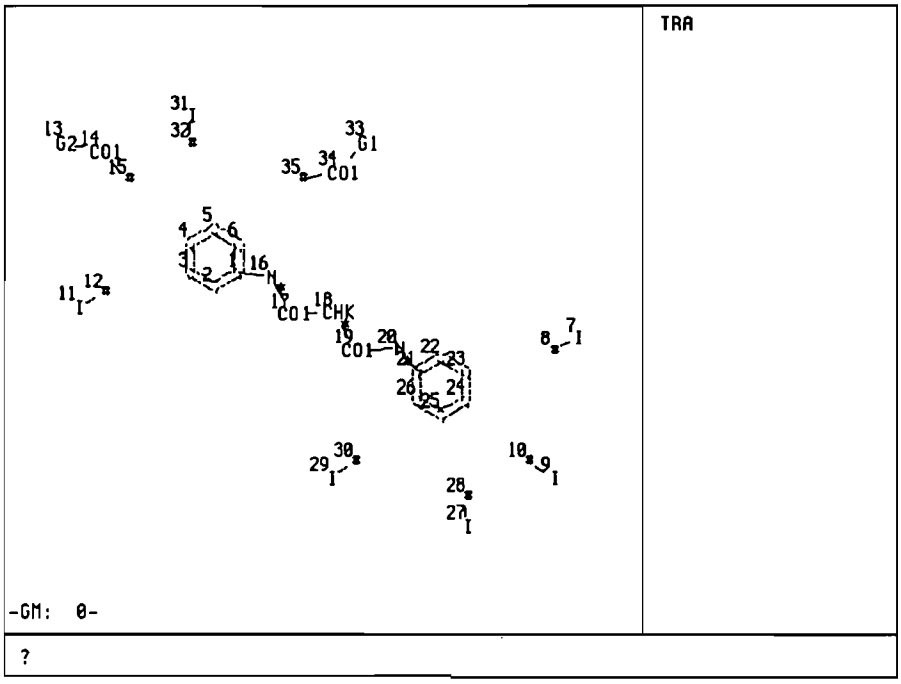
- Groups:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  GM
GROUP TO DESCRIBE ?  1
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  GR
*GRAPH
?  1,2
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  AT
*ATOMS
?  O 1
?  N 2
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  FS
*FREE SITES
?  1 1
?  2 2
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  GI
GROUP TO DESCRIBE ?  2
GROUP TO COPY     ?  1
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

- Variable points of attachment

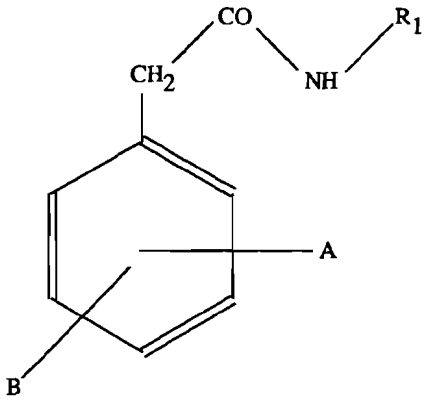
```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  UP
* VARIABLE POSITIONS *
NODE    8 :
?  2,3,4,5,6
NODE   10 :
?  2,3,4,5,6
NODE   12 :
?  2,3,4,5,6
NODE   15 :
?  2,3,4,5,6
NODE   28 :
?  22,23,24,25,26
NODE   30 :
?  22,23,24,25,26
NODE   32 :
?  22,23,24,25,26
NODE   35 :
?  22,23,24,25,26
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?
```

- Graphic verification:



II.2. Example 2: variable points of attachment and generic variables

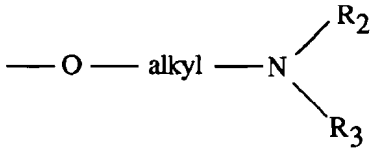
II.2.1. Question



R1: alkenyl with more than 10C

A: alkoxy on the ortho or the meta position of the benzene ring

B: aminoalkoxy of the following structure on the meta or para position:



R2, R3: hydrogen, alkyl or alkenyl

II.2.2. Query formulation

The text input can be carried out as follows:

-Graph of the query:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  GR
*GRAPH
?  1:10-5
?  11:13
?  14:18
?  17-19
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,UP,ATTR,UE)  ?  _
```

- Atoms:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  AT
*ATOMS
?  CHE 1
?  N 2,17
?  O 12,15
?  CHK 13,16
?  CO1 3
?  * 11,14
?  G1 18
?  G2 19
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  _
```

- Bonds and translation:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  BO
*BONDS
?  NO 5:10-5
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  TRA
*TRANSLATION ATTRIBUTES
?  NT 1,13,16
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?
```

- Chain attribute:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  ATTR
-ATTR- (FS,CH,AV,AM,CR,MV,PA,DT,SP,TRA)  ? CR
*CCHAIN/RING
?  HI 1
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?
```

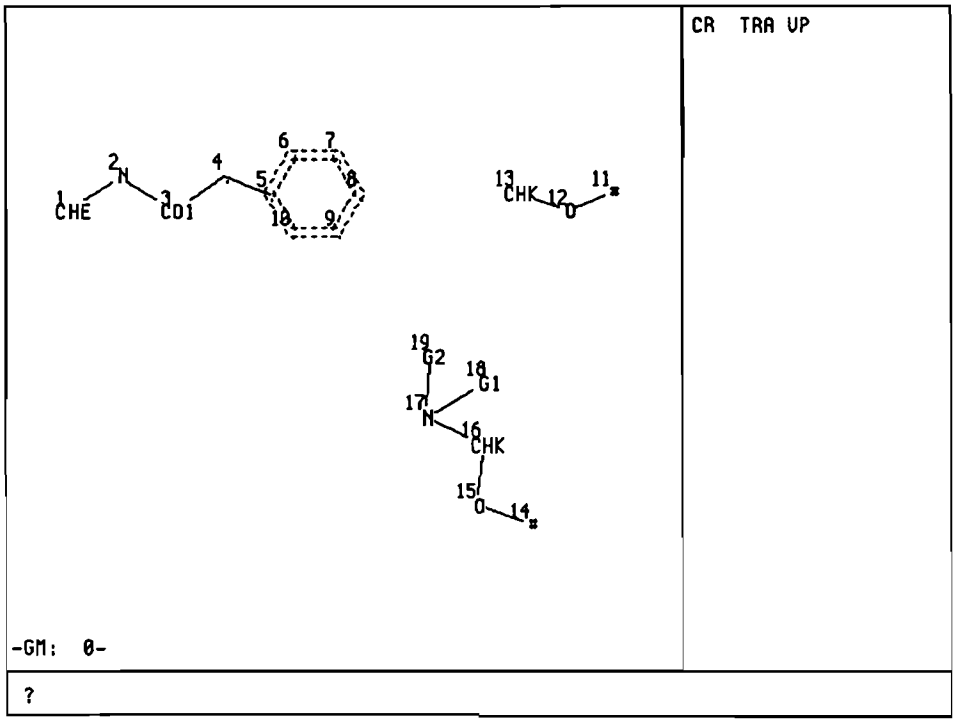
- Groups:

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  GM
GROUP TO DESCRIBE ?  1
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  GR
*GRAPH
?  1,2,3
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  AT
*ATOMS
?  H 1
?  CHK 2
?  CHE 3
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  TRA
*TRANSLATION ATTRIBUTES
?  NT 2,3
?
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  GI
GROUP TO DESCRIBE ?  2
GROUP TO COPY     ?  1
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  _
```

- Variable points of attachment

```
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  VP
* VARIABLE POSITIONS *
NODE   11 :
?  9,10
NODE   14 :
?  7,8
-QU- (CN,CA,GM,GI,GR,BO,AT,FS,AP,VP,ATTR,VE)  ?  _
```

- Graphic verification:



Index