Graphical representation of molecules and substructure-search queries in MACCS™

Susan Anderson

Molecular Design Limited, 1122 B Street, Haywood, CA 94541, USA

MACCS™ (the molecular access system) is a graphical interactive database management program for the end-user chemist. This paper describes MACCS's commands and procedures available to describe the atom type, bond type and stereochemistry in a completely defined molecule as well as the variable bond types and atom types available for substructure-search queries. This paper also describes MACCS's new Markush-style options capable of building query structures with variable molecular fragments at variable attachment sites. MACCS was designed for the infrequent as well as the experienced user, allowing both to enter complex molecular information successfully and to conduct query searches.

Keywords: database management, stereochemistry, query structures

received 1 June 1984

MACCS™ (the molecular access system) is one of a series of database programs designed by MDL (Molecular Design Limited) for the end-user chemist. MACCS provides a complete system for the graphical input, storage, retrieval, search and display of molecular information as well as an unlimited number of data fields. REACCS™ (the reaction access system) stores, searches, retrieves and displays both individual molecules and chemical reactions. Both of these systems are designed to be interactive and to allow the chemist to manage molecular information in his own terms, primarily through drawings of molecular structures.

MACCS and REACCS were primarily designed to store information provided by the user, although MDL has developed several commercially available databases (FCD and Aldrich for use with MACCS, ORGSYN and Theilheimer for use with REACCS). A typical user site contains a large corporate database for individual or group use, divisional or project databases and individual databases for particular chemists. Unless the data are already in computer-readable form, the construction of a database requires that the user should enter structural information through interactive graphics and non-structural information through the keyboard. This paper describes the graphical procedures for entering structural data in MACCS. Molecular information is entered in a similar manner in REACCS.

Search queries are also graphically defined with MACCS and can contain sites of free attachment, ambiguous bond or atom types and other unspecified or partially specified features. This paper also describes the use of MACCS's special drawing capabilities for the construction of search queries.

MACCS OVERVIEW

MACCS offers over 120 commands, or options, organized according to function within the framework of six MACCS operating modes. Each operating mode contains one or more graphic menus presenting the options available in that mode. The interrelation of MACCS program modes is described in Figure 1.

The executive mode menu appears at the start of each session and functions as the central passageway through which the user enters and exits other modes and returns control to the operating system. Basic storage and retrieval functions are performed here. Entry into draw mode enables the user to construct a complete molecule graphically. MACCS translates the information drawn on the screen into a connection table ready for an exact-match search or for entry into the database. The user's work in draw mode may also

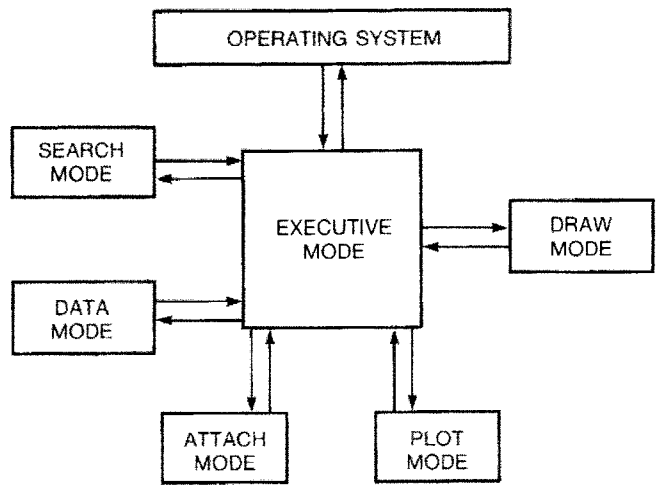

Figure 1. Communication between operating modes in MACCS and between MACCS and the operating system

result in a connection table describing an ambiguous structure (or structures) ready to use in a substructure search.

Attach mode is used to assemble molecules from predrawn fragments. Attach mode provides a rapid way of building structurally related molecules and registering them in the database. Search mode contains the options for initiating both graphic and data searches. In data mode the user may store, retrieve and transfer blocks of data, existing datatypes may be reviewed and new datatypes created. Plot mode is used to create hardcopy pictures of molecular structures.

STORING MOLECULAR INFORMATION IN MACCS

MACCS reads molecular structure information drawn at the terminal and translates that information (as it is drawn) into a connection table. Once the structural diagram (and the connection table) is finished, the user may store the molecule in an external file or register the molecule in a MACCS database. An external file, or molfile, consists of a connection table, *X* and *Y* coordinates, and administrative information stored as text. If the user wishes to store (register) the molecular structure in a MACCS database, MACCS stores the *X* and *Y* coordinates, interprets the connection table and stores the information as a SEMA (stereochemically extended Morgan algorithm) name[1]. The SEMA algorithm generates a unique and non-variant name for each chemical structure. All aspects of a molecule are described by the SEMA name, including atom, bond and stereochemical properties.

Atom and bond properties

MACCS maintains a modified periodic table with the traditional 103 elements as well as the potential for additional 'superatoms' that may be determined at the individual site. MACCS stores an integer representation of an element or superatom according to its position on the table. Isotopic labels are stored as the integral difference from the normal atomic mass, eg carbon-14 would be described as +2. Charged atoms are described by integer code for values from +3 to −3. Free radicals are also identified by code. Bond types are represented by different integers representing single, double, triple or aromatic bonds.

Stereochemistry

MACCS recognizes as a stereo centre any carbon atom where each of four attachments is a chemically different group. The SEMA algorithm describes stereochemistry as an atom property. If a particular stereochemistry is specified at an asymmetric centre, the 'handedness' of the spacial arrangement is assessed and recorded as even or odd. Stereochemistry is also recorded about double bonds between carbon and carbon, carbon and nitrogen, and nitrogen and nitrogen atoms. If the attachments at each carbon or nitrogen atom are different, the SEMA name maintains the *cis* or *trans* geometric isomerism of the diagrammed molecule. Thus, both the *cis* and *trans* forms of a molecule may be entered into a MACCS database.

Using MACCS's graphic options, the user can specify as little or as much stereochemistry as is desired. An asymmetric centre may be left unmarked to indicate either an unknown configuration or a mixture of *R* and *S* enantiomers. A double bond may be marked 'either' again to describe an unknown configuration or a mixture of *cis* and *trans* isomers. A diagram containing at least one specified asymmetric carbon is assumed to represent a racemic mixture unless the compound is specifically labelled 'chiral'. With the chiral label, the diagram represents only the enantiomer on display. Without the chiral label, the diagram represents both the enantiomer and its mirror image.

MACCS can therefore register and retrieve four distinct variations of a compound containing an asymmetric carbon atom: no stereochemistry (none specified), racemic mixture, *R* enantiomer and *S* enantiomer. MACCS can also recognize three variations of a compound containing stereochemistry about a double bond: *cis*, *trans* and 'either', indicating a mixture of *cis* and *trans* forms or an unknown configuration.

Tautomers

New capabilities in MACCS make it possible to identify the tautomeric form of the molecule in current memory.

MACCS can make use of these during registration and retrieval in three ways.

First, in executive mode, MACCS can search for tautomers and the current molecules as part of the registration process (the REGISTER CURRENT option). Second, MACCS will search the database in executive mode for tautomers if an exact-match search (FIND CURRENT) reveals no precise match for the current molecule. Finally a separate option, TAUTOMERIC SEARCH, is available in search mode.

MACCS uses a general tautomer definition to ensure that all possible tautomeric forms are located. MACCS defines as tautomers, or potential tautomers, molecules with the same atom types at corresponding structural positions and parity of changes and isotopes.

Bond types, the exact number and location of charges and isotopes and stereochemical designations may vary.

DRAWING COMPLETE MOLECULES IN MAACS

Interactive communication with MACCS is conducted through menus at a graphics terminal. (This paper describes the graphical use of MACCS at a particular terminal, but MACCS supports various graphics terminals employing a variety of graphics-input devices. See Table 1.) The commands, or 'buttons', needed to draw a complete molecule are presented on the screen. The user draws the desired molecular structure or activates commands directly on the screen with a tablet and stylus.

The user enters MACCS's draw mode and is presented with the draw mode menu. The principal draw mode menu (the draw menu shown in Figure 2) contains the necessary options for drawing a molecule suitable for entry into a database or for an exact-match search of a database. In addition, draw mode accommodates separate query-drawing menus: MACCS's traditional substructure-search drawing menu (Figure 3) and a drawing menu containing Markush-style

Table 1. MACCS will run on these machines and operating systems as of April 1984

Manufacturer	Computer	Operating system
DEC	System-10	TOPS-10
DEC	System-20	TOPS-20
DEC	VAX	VMS
Fujitsu	FACOM	F4, X8, OVIS
Honeywell	68/DPS	MULTICS
IBM	43xx, 30xx	VM/CMS
IBM	43xx, 30xx	MVS/TSO
Prime		PRIMOS

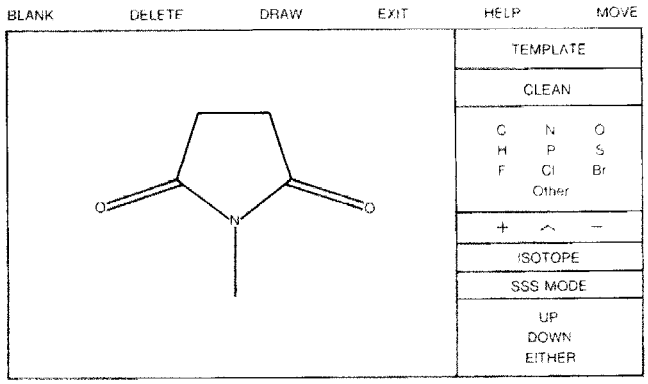

Figure 2. The principal draw menu in MACCS. The basic draw options appear at the top of the menu; other draw options occupy the right side of the screen. The drawing area shows a molecular structure as it would appear when drawn by the user

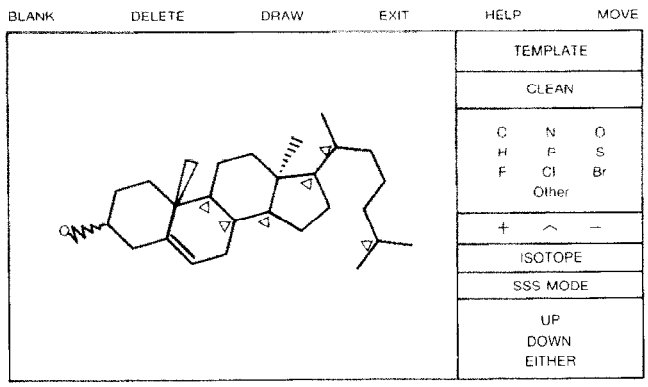

Figure 3. Structural diagram of molecule showing the three types of stereo bonds that may be applied at an asymmetric centre. An open wedge indicates an UP bond, a hashed wedge indicates a DOWN bond and a wedge-shaped wavy line describes an EITHER bond. Potential asymmetric centres with no stereo designations are marked with an arrow

searching capabilities (currently in development). These drawing submenus provide the graphical options used in constructing a query structure with explicitly defined ambiguities.

All draw mode menus maintain the same basic format. Molecular structures are diagrammed on the screen within the drawing area. Graphic options are located at the top of the screen and to the right of the drawing area. The options located at the top of the menu are present on each menu and are used to draw molecular frameworks, to obtain 'help' pages and to correct errors. The options located to the right of the drawing area are different on each menu and perform the functions particular to that menu.

Graphic options are activated by using the stylus and tablet. The user moves the stylus on the tablet in order to position displayed cross hairs over the desired command and then depresses the stylus ('hits' the button). The terminal signals the action by highlighting the command; eg the Envision terminal draws a coloured box around the now-active command.

Drawing the molecular structure

The principal draw menu contains all the options required to describe a complete molecule. The actual drawing of the molecular structure is accomplished with the 'DRAW' option located at the top of all draw mode menus. With the DRAW button activated, the user positions the cross hairs in the drawing area and depresses the stylus (scores a 'hit') at the desired end points of each bond. The line drawn by MACCS between each pair of hits describes a bond with a carbon atom at each end. Double and triple bonds are defined by drawing multiple lines between the two carbon atoms. Aromatic ring systems are represented by alternating single and double bonds. Carbon atoms are implicit at the junctions of bonds and at the ends of bonds.

On the main draw menu a list of heteroatoms appears to the right of the drawing area. Most common heteroatoms are included in the list and may be substituted for carbon atoms in the diagram. The heteroatom buttons are used similarly to the DRAW button. Once a heteroatom button is activated, any hit in the drawing area produces an atom of that type or changes an existing atom to that type. Unlisted heteroatoms are available through the OTHER button; MACCS prompts the user for the atomic symbol and substitutes that element into the diagram on subsequent hits.

Normally, hydrogen atoms are not explicitly added to a MACCS structural diagram. Hydrogen atoms are implied when the valence of an atom is not accounted for by other attachments. Explicit hydrogen atoms are added where it is necessary to describe the stereochemistry and to limit substitution in a substructure-search query.

A charged or radical atom is described with options labelled '+', '−' and '^'. The user applies a charge to an atom in the diagram by hitting the appropriate charge and then hitting the charged atom. Additional hits add to (or subtract from) the charge. The designated atom may carry a charge of from +3 to −3; the applied charge appears on the screen to the right of the atomic symbol. Radicals are defined in a similar manner and are labelled with the symbol '^'. Isotopic information is relayed with the ISOTOPE button; MACCS prompts the user to key in a mass difference, then applies that difference to subsequently hit atoms. The atomic mass of the isotopic atom will be displayed to the upper left of the atomic symbol.

Stereochemistry is added to the molecule through the UP, DOWN and EITHER buttons. MACCS uses traditional graphic symbols to describe stereochemistry. An open wedge is used to indicate a bond coming out of the screen from the pointed end and a hashed

wedge to indicate a bond going into the screen from the pointed end. An asymmetric centre of mixed or undefined configuration may be indicated by a wedge-shaped wavy line. (Figure 3 demonstrates the application of these three bond types at asymmetric centres.) A molecule may be labelled 'chiral' by depressing the keyboard letter 'I'. (MACCS employs several options that are activated through the keyboard. Most keyboard options perform less commonly used functions or present alternative methods for graphic commands.)

Improving picture appearance

The graphic options CLEAN and MOVE, as well as several keyboard options, are used to improve the appearance of the image on the screen, to reorient the structure in the plane of the screen or to enlarge or reduce the image. CLEAN performs a 2D modelling of the molecule, adjusting all bonds to the same length and adjusting bond angles to be tetrahedral, trigonal or linear. MOVE modifies a structure by relocating individual atoms but does not change or delete bonds.

Increasing drawing speed

Several short cuts are available to aid the user in more rapidly drawing a structure. The graphic option TEMPLATE provides access to several predrawn structural fragments, eliminating the need to redraw common groups each time that they appear. Activation of the TEMPLATE option causes a list of the most commonly used groups to appear on the right side of the drawing area. Each template name is an active button which, when hit, causes the appropriate structure to appear on the screen. If the drawing area is blank when the TEMPLATE button is hit, the structure appears in the centre of the screen. However, if a structure is already on display when the TEMPLATE button is activated, then the retrieved structure (the fragment) appears on the screen to the left of the existing structure (the substrate). MACCS will then prompt the user to hit the desired point of attachment in the substrate and subsequently in the fragment. The two fragments are repositioned and a single bond is drawn between the two groups. MACCS then automatically activates the DRAW button to allow additional drawing. (Any stored molfile can be used as a template. The user can store commonly used fragments in a molfile and retrieve the molecule by activating a FILE button located at the top of the template list.)

MACCS provides several keyboard options for the experienced drawer. Pressing the '+' key allows the user to attach templates to the diagrammed structure from the keyboard. Pressing the '=' key allows the user to add heteroatoms to the diagram from the keyboard. The 'C', or continuous-draw, keyboard option aids the user in more rapidly drawing the structural framework of a molecule. Successive pen hits describe a chain of bonds with each pen hit describing the individual carbon atom linking these bonds by scoring a single hit to describe the common endpoint of two adjacent bonds.

The keyboard group-mode option 'G' provides access to a series of suboptions that will manipulate groups of atoms within a structure. Group-mode suboptions will delete atom groups, enlarge or reduce atom groups, rotate groups about a specific bond and move groups within the plane of the screen.

In addition to capabilities in draw mode, a separate mode, attach mode, can be used to build structurally related molecules rapidly. Attach mode, available through executive mode, can also be used to register molecules directly into a database. Attach mode allows the user to attach fragments to a 'parent' molecule, to specify stereochemistry and to register the molecule into a database without changing modes. It then returns the parent molecule so that the user can begin to define another derivative.

Identifying and correcting errors

MACCS provides several options that aid the user in more completely describing the structure and ensuring that the structure is defined as the drawer intended.

Help pages are always available. The HELP option is located at the top of the drawing area on all draw menus. Hitting the HELP option causes a reference page of text to appear in the drawing area and a list of other available help pages to appear on the right. The reference page briefly describes options on the current menu and contains the prompt 'Further help:'. The user may type the page number of a desired help page or type a carriage return to reinstate the current menu and structure.

The BLANK option, also located at the top of the draw menus, clears the entire structure from the screen and erases the molecule from current memory. The DELETE option allows the user to remove an atom and all the bonds connected to it with a single pen hit. Single pen hits with the DELETE button active also remove single bonds, change double bonds to single bonds and change triple bonds to double bonds.

MACCS provides an internal valence check which ensures that normal valence restrictions are applied to each atom. The free-valence keyboard option 'F' disables this checking. With the free-valence option 'on', five bonds of any order may be drawn to any atom. This option might be used to draw more than four bonds to a centre that will later be designated a heteroatom, eg in describing a sulphonate.

MACCS can be asked to verify the diagrammed structure of a molecule before it is registered into a database. The user enters the simple molecular formula, and MACCS will check that formula against the formula calculated from the molecule on the screen. In a similar manner, the keyboard option 'M' causes MACCS to calculate and display the molecular formula and molecular weight of the current structure.

The keyboard option 'U' points out potential asymmetric centres with unspecified stereochemistry. Typing the letter 'U' causes arrows to appear pointing to carbon atoms with at least three non-hydrogen attachments and no stereo markings (Figure 3). Further description of these potential asymmetric centres with UP, DOWN or EITHER arrows eliminates this designation. However, MACCS will register and search using molecules with unspecified centres.

SEARCH QUERIES

Construction of a molecule in draw mode results in a complete fully defined molecule. However, a chemist may need to identify a list of compounds with similar properties and structures. He may wish to search a database for compounds that fulfil certain structural conditions. MACCS can be used to build a query structure that contains explicitly defined ambiguities useful in substructure searching.

Substructure-search queries in MACCS specifically state allowable bond and atom types but are general regarding attachments. Thus, a completely defined molecule drawn without assigning the specific ambiguities available on the substructure-search menu may still usefully function as a substructure-search query. Each atom in a query structure is assumed to have as many free attachment sites as it has remaining valences. A search over a database will locate all molecules containing the query structure, and any attachments are allowed that satisfy the valence requirements.

Substructure-search queries may allow alternative atom or bond types at specific locations and may describe specific stereochemistry. Planned new features in MACCS will allow the construction of queries with alternative molecular fragments at alternative locations. These new MACCS features will provide Markush-style 'generic' searching capabilities and will allow greater flexibility in query design. The formation of a traditional MACCS substructure-search query and of a Markush-style search query is described below. Markush-style searching capabilities are at present under development at MDL.

Substructure-search queries

A separate substructure-search menu in draw mode contains the graphic options that the user needs to construct a query structure. Both the substructure-search menu and the draw menu contain basic draw options. Normally, however, the parent or invariant component of a substructure query is constructed with draw menu options and the substructure-search menu elicited (by activating the 'SSS MODE' button on the draw menu) to add the indeterminate bond and atom types.

The format of the substructure-search menu (Figure 4) is similar to that of the draw menu. The basic draw, help and error options are located above the drawing area. The options used to construct the substructure-search query are displayed to the right of the drawing area.

With the structure diagrammed on the screen, the user is ready to designate permissible bond and atom types. Heteroatoms may be introduced into the structure in several different ways. Permissible atoms may be added to the query by hitting the desired heteroatom button on the screen and then the desired attachment site. Several different atoms may be designated at the same site, creating a list of acceptable elements at that location. Figure 4 shows a structure with an atom list as displayed on the screen. The EXCLUDE option, when applied to an atom or atom list, causes permissible atoms to become non-permissible atoms at that site. The word 'not' appears on the screen to the left of the list of atomic symbols.

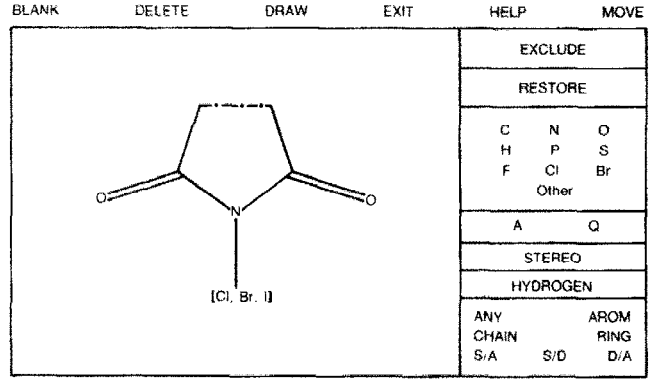

Figure 4. The substructure-search menu in MACCS. The substructure-search query in the drawing area contains the 'parent' structure shown in Figure 2 with 'ANY' bond type and atom list added

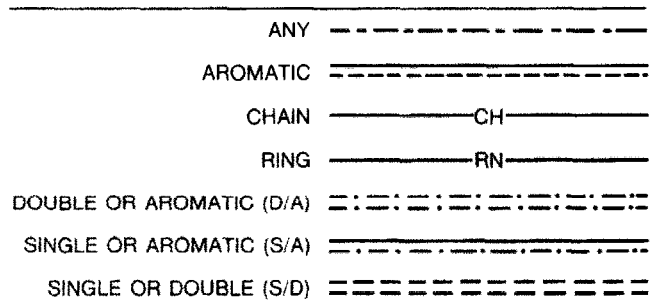

Figure 5. Special bond types available on the substructure-search menu. The chain and ring topology designations may be applied in conjunction with other special bond type designations

The letter 'A' can be introduced into the query structure to mean any atom except hydrogen. Similarly, the letter 'Q' is used to represent any atom except hydrogen or carbon.

The HYDROGEN option allows the user to specify the minimum number of hydrogen atoms permissible at a given site in a substructure-search hit. The user types an integer representing the number of hydrogen atoms required, and that requirement appears as H0, H1, H2, H3 or H4 at the desired site. Use of the HYDROGEN option limits the degree of substitution permitted at that position, effectively blocking out potential attachments.

Specific bond features are added to the query structure by hitting the desired bond option on the screen and then the desired bond. The allowed bond types and the combination of dotted, dashed and solid lines which MACCS uses to represent these bond types are described in Figure 5. Bond topology may be specified as 'chain' (not in a ring) or 'ring'.

Use of the STEREO option ensures that only compounds with the same relative configuration will be matched in a substructure search. The user applies the STEREO option at two or more asymmetric centres or at both ends of a double bond. A small box will then appear on the screen at the designated centres. With the STEREO options appropriately applied, MACCS will distinguish between diastereomers and geometric isomers of the query structure. Valid search hits will

therefore maintain the relative configuration of the query structure. If the STEREO option is not applied, then any isomer of the search query will be considered a valid hit.

The RESTORE option allows the user to remove query features from a query structure, one by one, with successive pen hits. The RESTORE option, when activated, allows the user to change heteroatoms back to carbon atoms and to remove charges, labels and other special atom features introduced into the structure. RESTORE also replaces double and triple bonds, labelled bonds and flexible bond types with a single unlabelled bond.

Non-graphic substructure-search queries

Substructure-search queries may also be formulated using MACCS's search mode options. The non-graphic query structure is described at the keyboard in non-graphic 'query notation' as a series of alternating atom and bond designations, eg 'C—C(=O)—O'. Many of the symbols used in drawing graphic substructure-search queries are used in describing non-graphic queries, eg hyphens indicate single bonds, equals signs indicate double bonds and the symbols 'A' and 'Q' have the same meaning as in graphic queries. In query notation, atom lists are enclosed in square brackets and parentheses enclose atom branches. The notation '@*n*' indicates that atom number *n* (counting from left to right in the query string) is attached to the atom just preceding the '@*n*' symbol. This format is used to describe ring structures. The example below describes the substructure query shown in Figure 4 as it would appear in non-graphic query notation:

Query = O=C—N(—[Cl,Br,I])—C(=O)—CC—@2

Query notation can describe all atom types and most bond types that can be described graphically in draw mode. Query notation cannot, however, describe stereochemistry, and query searches tend to be slower than graphic searches.

Markush-style queries

Markush or generic notation is commonly used to describe a molecule that includes a set of one or more allowed structural fragments at one or more allowed positions on the molecule. Markush notation permits a large number of compounds to be described in a single structural diagram. MDL's new generic searching capabilities will allow the user to construct a complex searching query based on Markush notation. A structure can be defined that specifies a list of allowed molecular fragments and at the same time specifies allowed points of attachment for those fragments.

The Markush-style or R-group query capabilities planned for MACCS are based on standard generic notation conventions. A particular set of allowed molecular fragments is called an R-group. A complete query structure includes a parent or non-variant structure, some number of R-groups attached to the parent molecule and additional conditions limiting the occurrence of the R-groups. An 'assertion' describes one R-group (appearing in generalized notation below):

R*n* = {G1, G2, ..., G*k*}

as well as the occurrence or range of occurrence of that R-group in the parent structure. Multiple assertions may coexist at the same site or sites (as shown in Figure 6, Query 2).

Occurrence counts or ranges of occurrence are specified for each R-group. The user can define the occurrence of an R-group as an integer (which may be zero, specifying 'none'), as a series of integers or as a range of integers. In addition, the user can restrict substitutions in a query through activation of the restH option. This option can ensure that hydrogen atoms will occupy all designated attachment sites not occupied by an R-group fragment.

Markush-style options will be displayed on a separate menu similar to the substructure-search menu. Basic draw menu options will occupy the top of the screen with new graphic commands displayed to the right of the drawing area. The parent or non-variant component of a Markush-style query will normally be construed with the draw menu. Options on the Markush-style or R-group menu will be used to construct and define the assertions.

With the parent structure displayed on the screen, the user will position an 'R*n*' symbol (where *n* is an integer) at the desired site or sites. He will then proceed to draw the molecular fragments of that R-group. MACCS will prompt the user to specify the occurrence or range of occurrence of the R-group and thus to complete the assertion. As an assertion is being defined, the parent structure is displayed in the upper left corner of the drawing area with completed fragments appearing to the right of the parent (Figure 7). If an R-group contains more fragments than can be displayed at once, the user can scroll through the fragment list and view all structures.

A single query structure can combine features of both the substructure-search menu and the Markush-style menu. Careful construction of a Markush-style query structure can result in a complex but highly specific query that will generate a refined hit list with few unwanted compounds, yet an effective query does not require complex logic or complicated construction. A few of the features available on MACCS's new Markush-style menu are described below.

In Query 1 (see Figure 6), a single R-group (R1) contains the fragments Cl, Br and COOH and is affixed at all free attachment sites on the benzene ring. An occurrence of 'exactly 1' has been specified for R1 and, in addition, the graphic option restH has been activated. This assertion ensures that a valid search hit will contain exactly one fragment from R1, but this assertion does not specify the exact location of that fragment (except, of course, that it must be on the ring). The use

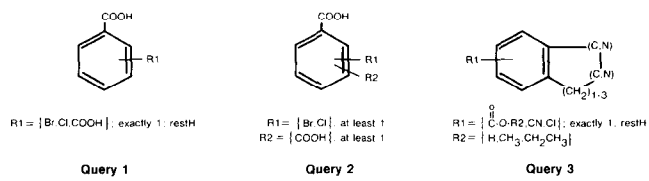

Figure 6. Markush-style queries; (left) Query 1: R1 = {Br, Cl, COOH}; exactly 1; restH; (centre) Query 2: R1 = {Br, Cl}; at least 1, R2 = {COOH}; at least 1 and (right) Query 3: R1 = {C(O)—O—R2, CN, Cl}; exactly 1; restH; R2 = {H, CH_3, CH_2CH_3}

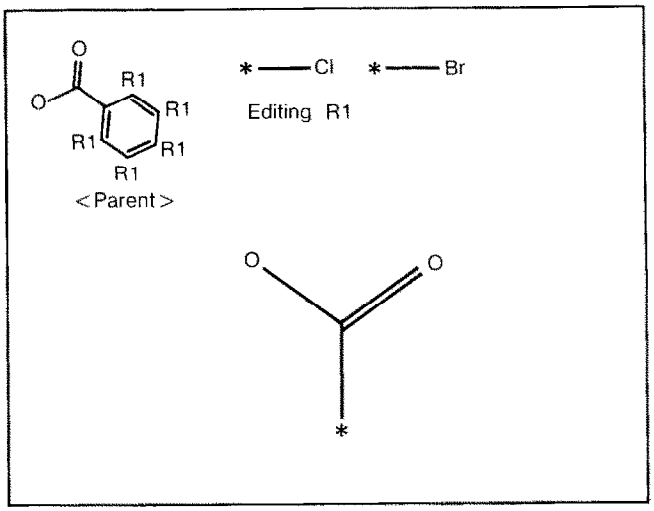

Figure 7. Illustration of the drawing area as an R-group assertion (Figure 6, Query 1) is being defined. The parent structure appears at upper left, and completed fragments appear to its right. An asterisk shows the attachment point on fragments

of the 'restH' option ensures that all other free attachment sites must be occupied by hydrogen atoms.

The second example in Figure 6, Query 2, contains two R-groups, R1 and R2, coexisting on the same ring. An occurrence of 'at least 1' has been specified for each R-group. In this example, at least one R1 fragment and at least one R2 fragment must occur in a valid hit. As in Query 1, the fragments may occur anywhere on the ring. However, in this example no restrictions have been placed on attachments to the ring once the assertion has been satisfied. Any attachments that fulfil the valence requirements (including additional Cl, Br and COOH fragments) are permitted.

Query 3 in Figure 6 provides a more sophisticated use of MACCS's Markush-style query notation and contains elements from both MACCS's substructure-search menu and the R-group menu. The occurrence of R1 is defined as 'exactly 1' with additional substitutions limited to hydrogen atoms. R2 is 'nested' within R1 and functions as a fragment list. The symbols $(CH_2)_{1-3}$ within the five-membered ring describe a variable link node of from one to three carbon atoms. Thus the aliphatic ring may vary in size from five to seven carbon atoms. Either a carbon atom or a nitrogen atom may occur at the aliphatic ring positions having the C, N designation (an 'atom list' from the substructure-search menu). A search using this query will locate compounds with a single fragment from R1 on the aromatic ring (all other attachments will be hydrogen atoms). In addition, each compound must contain an aliphatic ring with 5, 6 or 7 atoms and carbon or nitrogen atoms at the designated positions.

Conducting a search

Once a query structure is complete, the user can initiate the database search. All substructure searches are conducted in search mode. The user exits from draw mode to executive mode, enters search mode and activates the substructure-search option to begin the search. MACCS initially assesses the structure for the presence of certain structural 'keys'. The actual search of the database proceeds in two steps. The first step consists of a rapid key search that locates compounds having the same structural keys as the query. Then MACCS conducts an atom-by-atom search over the compounds identified in the key search. This second search produces an 'active' list of those compounds that satisfy the conditions stipulated by the query structure.

Each time that the atom-by-atom search produces a hit, MACCS increments an active-list counter and displays the molecule name and registry number on the screen. Once the search is complete, the user may view individual compounds on the active list, save the list in a file, send the listed structures to a plotter or use the list as the basis for another search using associated data or a more detailed searching structure. Figure 8 shows some potential hits that might be identified by a substructure search using the query structure shown in Figure 4.

HARDWARE CONSIDERATIONS

MACCS, REACCS and other MDL programs are used by a variety of companies and must be compatible with many different graphics terminals and operating systems. The graphics for MACCS (and all other graphic programs licensed by MDL) are implemented via a high level terminal-independent graphics library written and maintained by MDL. This library, Unigraph, allows graphics terminals with inexpensive raster displays as well as those with vector-refresh displays to function as intelligent display-list devices. The Unigraph library at present supports three raster devices: the VT640 and the DQ650M (Digital Engineering upgrades of the DEC VT100 and VT102 series termi-

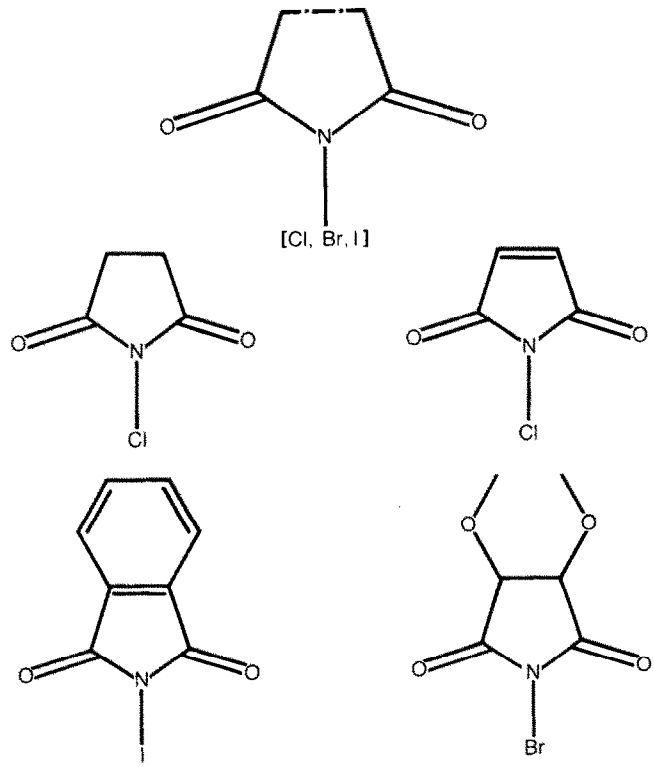

Figure 8. A query structure and hits potentially identified by a substructure search

nals), and the colour (Envision 230). MDL's Unigraph also supports two vector terminals: the Imlac Series II and the Tektronix 4114. The computers and operating systems on which MACCS will presently operate are listed in Table 1.

CONCLUSION

MACCS's drawing capabilities have been used by many companies to construct their own databases. MDL's own database division has built corporate databases of over 100 000 compounds as well as MDL's commercially available FCD and ORGSYN databases. MACCS's drawing options have provided the speed and accuracy necessary to complete such large databases in relatively short periods of time. Using MACCS, MDL's most experienced drawer has constructed over 40 compounds in an hour and 412 compounds in one marathon twelve hour session.

However, MACCS was designed to meet the needs of the naive or infrequent user as well as the experienced user. With graphic prompts, help options and a small number of basic draw options, MACCS allows the inexperienced user to enter complex molecular information successfully and to conduct query searches. MACCS's many graphic and keyboard options allow considerable flexibility, providing the user with a choice of drawing and searching methods and enabling the detailed and sophisticated design of query searches. However, the majority of tasks, the basic drawing and searching functions, can be accomplished with relatively few options and are within the capabilities of most inexperienced users.

REFERENCES

1 **Wipke, W T and Dyott, T M** 'Simulation and evaluation of chemical synthesis. Computer representation and manipulation of stereochemistry' *J. Am. Chem. Soc.* Vol 96 No 15 (1974) pp 4825–4834
Wipke, W T and Dyott, T M 'Stereochemically unique naming algorithm' *J. Am. Chem. Soc.* Vol 96 No 15 (1974) pp 4834–4842