one will be able to determine relatively quickly whether CA covers the literature needed by these researchers.

ACKNOWLEDGMENT

I would like to thank Lilian Peng and Kathy Brown who looked up a number of references in Chemical Abstracts to determine coverage.

REFERENCES AND NOTES

(1) Baldinger, E. L.; Nakeff-Plaat, J. P. S.; Cummings, M. S. "An Experimental Study of the Feasibility of Substituting Chemical Abstracts On-Line for the Printed Copy in a Medium-Sized Library". *Bull. Med. Libr. Assoc.* **1981**, *69*, 247-251.

(2) Fox, D. J. "The Reseach Process in Education"; Holt, Rinehart and Winston: New York, 1969.

(3) Garfield, E. "Significant Journals of Science" *Nature (London)* **1976**, *264*, 609-615.

(4) Lowry, O. H.; Rosebrough, N. J.; Farr, A. L.; Randall, R. J. "Protein Measurement with the Folin Phenol Reagent". *J. Biol. Chem.* **1951**, *193*, 265.

(5) "Subject Coverage and Arrangement of Abstracts by Sections in Chemical Abstracts", 1975 ed.; Chemical Abstracts Service: Columbus, OH, 1974.

(6) Sengupta, I. N. "The Literature of Microbiology". *Int. Libr. Rev.* **1974**, *6*, 353-369.

Molecular Substructure Searching: Computer Graphics and Query Entry Methodology

W. J. HOWE* and T. R. HAGADONE*

The Upjohn Company, Kalamazoo, Michigan 49001

Received July 27, 1981

The increased availability of interactive graphics hardware has enabled the construction of true end-user oriented substructure search (SS) systems. However, because of the complexities inherent in the structural definition of substructure search queries, the design of a graphical query module should be accompanied by a careful consideration of SS query methodology. This paper discusses some of the important design considerations and then describes features and capabilities of the SS query definition module of Upjohn's COUSIN system. Techniques are presented for graphical entry of the core substructure of interest, for indicating variable atom, group, or fragment attachments, for specifying indefinite positioning of groups, and for tightening or relaxing search constraints on individual bonds. The R-group notational method for variable attachments and indefinite positioning also appears to have utility outside the context of a computerized retrieval system.

INTRODUCTION

Interactive graphics is a tool which has been used quite effectively as a full structure input medium in a variety of chemistry-based computer systems.[1] Even more effectively, graphics can be used to permit direct end-user specification of substructure search queries; yet relatively little work has been done in this area.[2]

The nature of substructure searching is such that the design of a graphical front-end should recognize and be guided by two important factors: (1) the data to be input is usually very complex and (2) the end users of an SS facility may run searches as often as daily or as infrequently as once a year. The graphical SS query module must be designed, therefore, with the relatively computer-naive user in mind. This is not easy to do since, on the one hand, the graphical controls must be sophisticated enough to permit reasonably complex queries to be entered and executed with a single search, eliminating (as much as possible) the need for post-search manipulations of result files. Among other things, the graphical controls should therefore provide a capability for indicating variable structural units that are allowed at a particular location in the substructure, indefinite positioning of groups, and variable constraints on bond-type specificity. Yet on the other hand, characteristics of the user population require that the system be as simple to use and error tolerant as possible. This can be accomplished through a variety of means, the most effective of which is continuous system monitoring of the "chemical reasonableness" of the query as it is being constructed; the degree of handholding and the level of detail in messages suggesting ways to fix a problem should increase automatically as the user gets into more complex regions of the query.

With the preceding goals in mind we began several years ago to develop a graphical substructure search module for Upjohn's COUSIN system.[4]

OVERVIEW OF COUSIN'S GRAPHICAL SS QUERY MODULE

COUSIN is an interactive graphics-based chemical and biological information system which currently operates on a data base of 65 000 compounds. It is end-user oriented and provides capabilities for not only substructure searching but also structure retrieval and display, compound registry, full structure searching, retrieval and display of biological screening data, file manipulation, report generation, and searches keyed on a variety of textual and numeric data types associated with each compound. The following discussion will deal only with the portion of COUSIN that is used to define a substructure search query. Other parts of the system, including the part that actually carries out a substructure search once the query has been entered, will be described elsewhere.[5]

Method of Interaction. To interact with the COUSIN system, the user may either type commands on a keyboard (the typed characters appear on the graphics screen) or draw structure diagrams and associated symbols using a graphics tablet and stylus (the drawing appears on the screen). Examples of the former are commands to display particular compounds, combine files, activate the drawing controls, and so on. On the other hand, the tablet is used when the data to be entered is pictorial, such as structures to be registered or SS queries to be defined. Either the tablet or the keyboard is active at a given time.

When the user indicates via a typed command that he wishes to enter a substructure search query, the system responds by displaying the drawing controls pictured in Figure 1 (details of the drawing controls are shown in Figures 2-4). The stylus and tablet are also activated. The stylus is held like a pen, with the tip touching the horizontal surface of the tablet which sits in front of the display. As the user moves the stylus across the tablet, a cursor or "tracking cross" follows the motion on

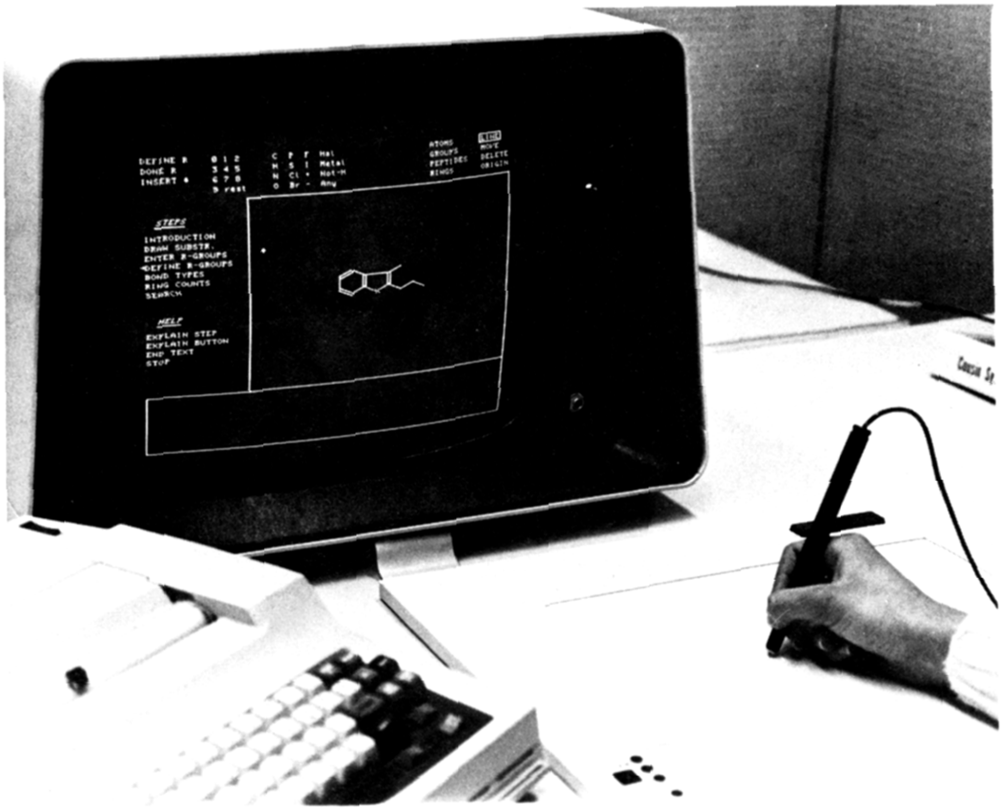


Figure 1. COUSIN system user interface showing SS query entry display. Hardware consists of graphics screen, keyboard, graphics tablet, and stylus.

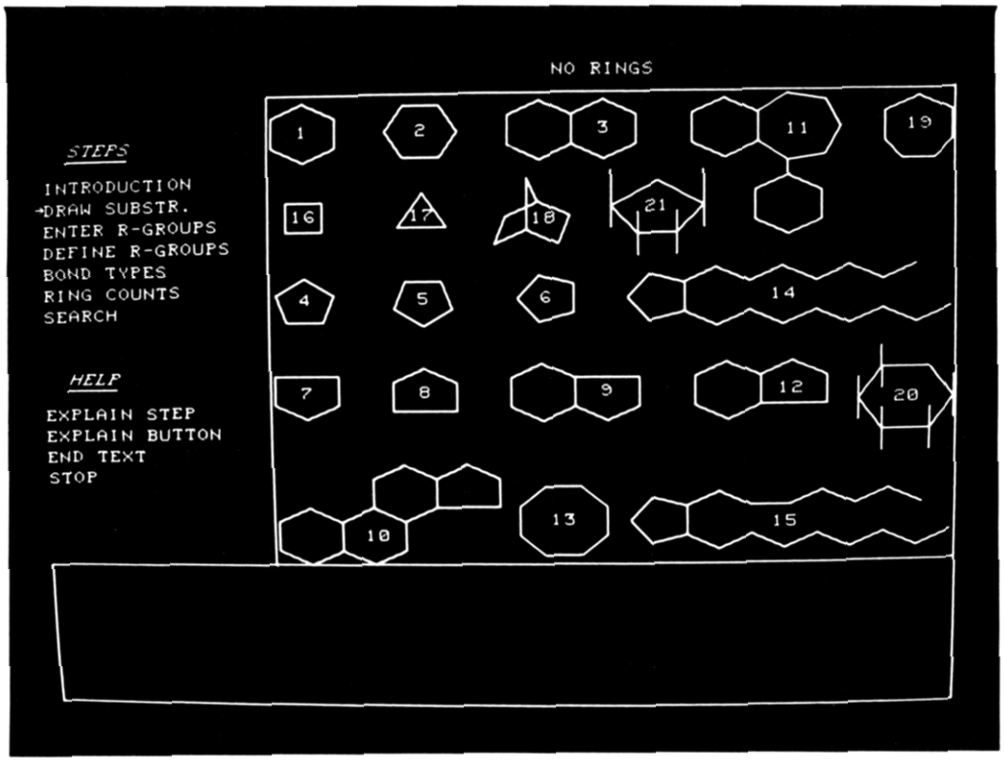


Figure 2. Details of the STEP and HELP buttons. The arrow beside the Draw Substructure step button indicates that it is the currently active step. When the RINGS button is hit (see Figure 3A), the user's query and all the drawing control buttons are temporarily replaced by a set of predrawn rings that can be used to speed up the drawing operation. The ring templates are also available in the "Define R Groups" step.

the screen. Many of the words and symbols which appear on the screen are graphical controls or "buttons"; to select or "hit" a button, the user moves the stylus to superimpose the tracking cross on the desired symbol and then depresses the stylus.

The SS query entry display is divided into five regions. All drawing operations occur inside the large rectangle in the center of the screen. This is the "drawing box". Below that is an elongated rectangle (the "message box") inside which appear messages in response to the user's request for help or if COUSIN detects a problem with the query. Along the left side of the display are seven STEP buttons (see detail in Figure 2) that can be selected by the user and which correspond to

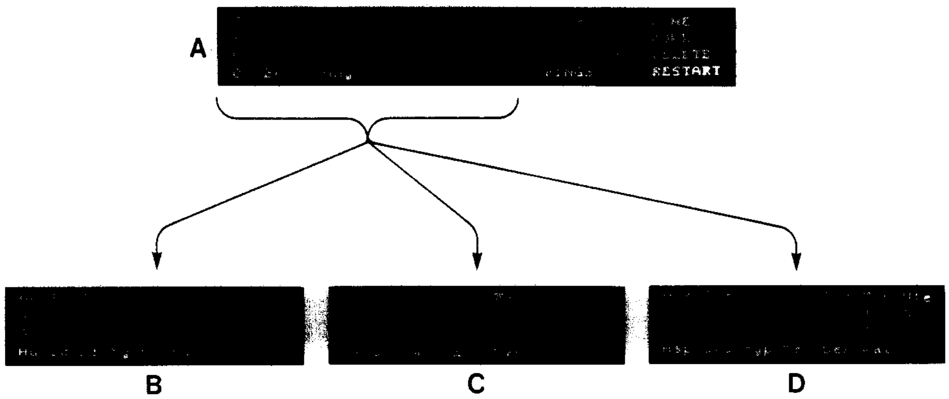


Figure 3. Drawing controls available in "Draw Substructure" step. (A) Normal set of controls, containing cluster of common atom symbols, some generalized pseudoatoms, and positive and negative charge buttons, (B) cluster of less common atoms appears in that region when ATOMS button is hit, (C) common functional groups appear when GROUPS button is hit, and (D) collection of amino acid symbols appears in that region when PEPTIDES button is hit.

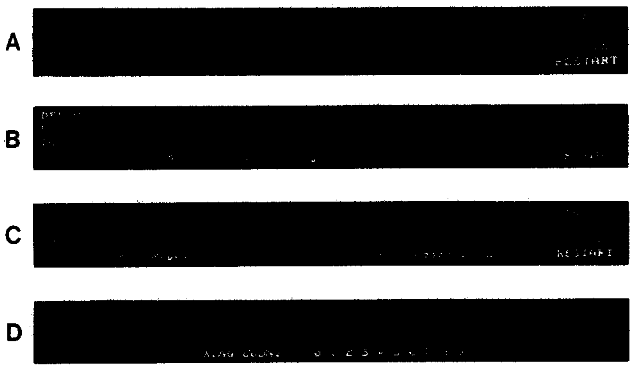


Figure 4. Drawing controls available in remaining steps: (a) "Enter R Groups" step, (b) "Define R Groups" step, (c) "Bond Types" step, and (d) "Ring Counts" step.

various stages of query entry. Below the STEP buttons are four HELP buttons, and along the top of the display are the controls that enable the user to draw the query. A new set of drawing controls (Figures 2–4) appears with each STEP button that is selected by the user. The drawing controls will be discussed shortly.

Built-In User Assistance. The HELP buttons on the display (bottom left, Figure 2) allow the user to request detailed information on all aspects of the SS query entry process. Whenever the user changes steps by hitting a STEP button on the display, a descriptive paragraph automatically appears in the message box which gives a brief outline of what is expected in the current step. However, the user can also request additional information on the currently active step by hitting the EXPLAIN STEP button. A complete description of the function of that particular query entry step is then displayed in the message box. It also provides a general discussion of the drawing control buttons (located above the drawing box) which are associated with that step and structural assumptions or defaults used by the system. Since the complete message is normally larger than the message box will hold, it is shown a page at a time. When the user depresses the stylus anywhere on the tablet the next page is displayed. If the user has read enough and does not wish to page through to the end of the message, the END TEXT button may be used to clear the message box and terminate the message paging. If the user is simply interested in how a particular button works, he can hit the EXPLAIN BUTTON button. A message will then ask him to "point to" (depress the pen on) the button he wishes explained. A detailed description of the selected button will then appear in the message box.

The EXPLAIN BUTTON and EXPLAIN STEP buttons provide assistance only when the user requests it. There are, however, three sets of circumstances which will result in automatic message feedback: (a) an improper action (such as an attempted valence violation) will cause a message to appear which explains why the requested drawing operation was not accepted; (b) query inconsistency detected later in the query input will generate a detailed message pointing out the location of the problem and suggesting ways to fix it; (c) guidance messages automatically appear when the system detects that the user is moving into the more complicated aspects of query building, such as defining R groups (discussed below).

Stepwise Approach to Query Entry. The query entry operation has been divided into steps for two reasons. First, it makes it easier for the user to concentrate on one aspect of the entry operation at a time. Second, there are many graphical control buttons; the display would be cluttered if all the buttons were visible at the same time. As it is, only those buttons which are used in a particular step are visible.

Although seven step control buttons appear below the STEPS heading on the display (figure 2), the first and last do not correspond to true entry operations. The "Introduction" step is for beginning users. If the user hits the EXPLAIN STEP button when the Introduction step is active, an overview of the entire SS query entry process is provided in the message box. No drawing controls appear at the top of the screen in this step. The last step is the "Search" button. This begins the search.

It is in steps 2–6 that the query is contructed. The user has complete control of the order in which he uses the steps to define his query, and not all steps are even needed for some queries. Many queries can be completely defined with the controls in just the "Draw Substructure" step, for example. While the user may define his query in whatever order he wants, beginners are encouraged to select the steps in the order that they are listed on the screen.

To illustrate the stepwise construction of a query, the following discussion will refer to the sample sequence in Figure 5.

(*a*) *Draw Substructure Step.* When the user selects the "Draw Substructure" step button, the graphical controls that appear at the top of the screen are those pictured in Figure 3A. It is in this step that the user draws the major portion of the substructure, which will be referred as the "fixed fragment". Multiple fixed fragments are also allowed; a compound must contain all of them to be retrieved in a search.

The most commonly used button is the LINE button which is used to draw bonds. After hitting LINE the user depresses the pen inside the drawing box and moves the pen, causing a bond to be drawn on the screen following the motion of the pen. The interaction is immediate. New bonds can be created, old bonds can be made multiple, and rings can be drawn with

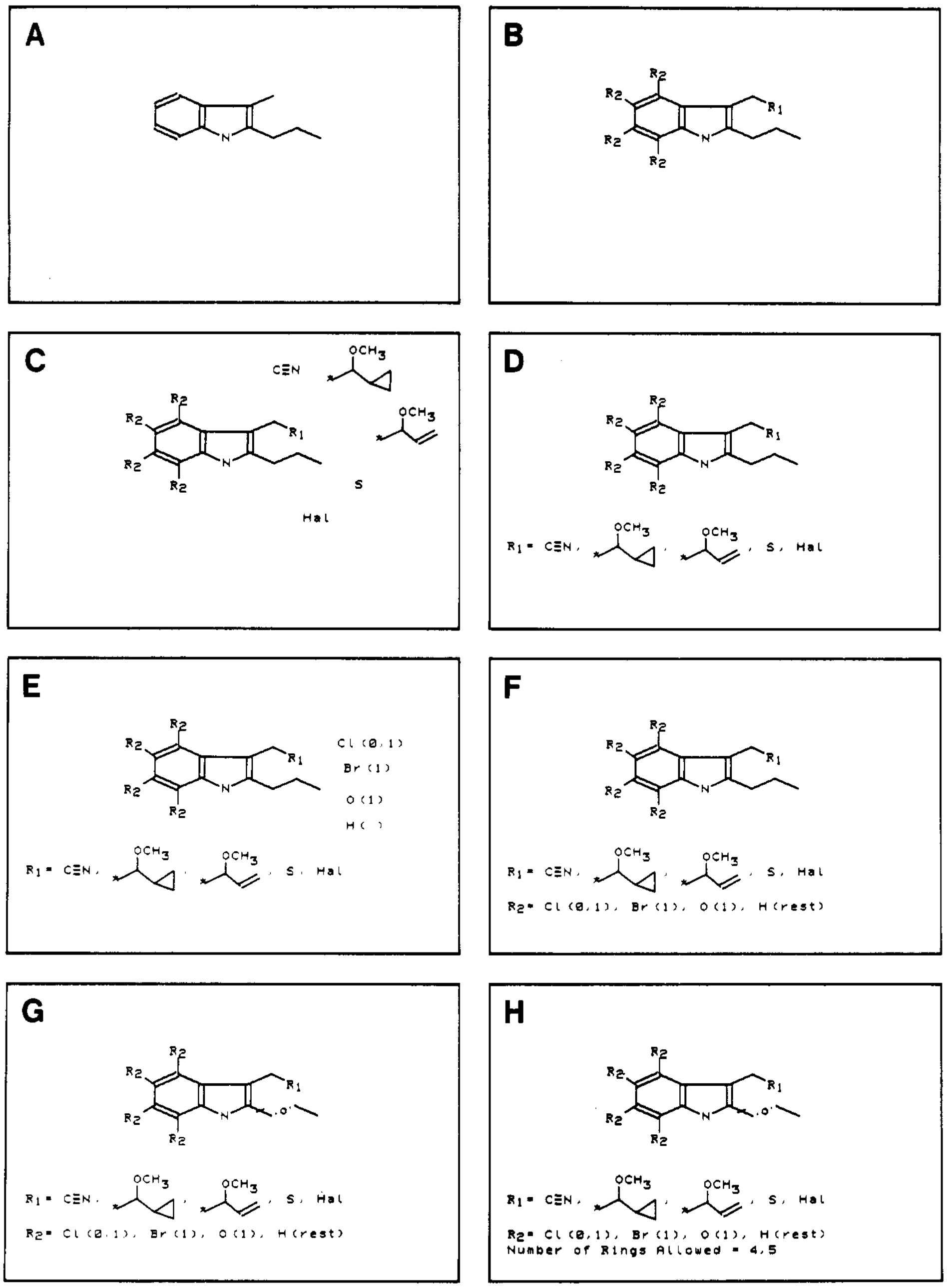

A
B
R2
R1
N
C
C≡N
OCH3
S
Hal
D
R1 = C≡N, OCH3, OCH3, S, Hal
E
Cl (0,1)
Br (1)
O (1)
H ( )
F
R2 = Cl (0,1), Br (1), O (1), H (rest)
G
O
H
Number of Rings Allowed = 4,5

Figure 5. Sample query construction sequence. Each of the frames was photographed directly from the display. Frames A and B were entered in the "Draw Substructure" and "Insert R Groups" steps, respectively. Frames C–F were entered in the "Define R Groups" step and illustrate the use of blank regions on the screen for R-group definition and the specification of restricting occurrence counts. Frames G and H were entered in the "Bond Types" and "Ring Counts" steps.

this button (a more detailed description appears in ref 1f). Other buttons enable atoms to be relocated or removed (MOVE and DELETE buttons).

Unlabeled ends of bonds are considered to be carbon atoms. To insert an atomic symbol the user may select from a cluster of atoms which appears at the left side of the drawing controls (Figure 3A). Also available are three other clusters containing additional atom symbols (Figure 3B), predrawn commonly occurring functional groups (Figure 3C; groups not represented here can be drawn by the user with the LINE button and the atom symbol buttons), and amino acid symbols (Figure 3D). These clusters appear in place of the normal atom cluster in Figure 3A when ATOMS, GROUPS, or PEPTIDES buttons are selected respectively. When an atom or group symbol is selected, it appears in place of the pen tracking cross and may be moved down into the substructure drawing. If an atom symbol is not represented in either of the atom clusters, then it does not exist in any data base compound.

A collection of predrawn rings appears when the RINGS button is selected. The substructure drawing is temporarily replaced by the rings in Figure 2, any of which can be selected by pointing to the number inside the ring. The selected ring then follows the pen motion as the other rings disappear and the substructure diagram reappears. A copy of the ring is inserted in the substructure diagram wherever the pen is depressed.

In the sample query construction sequence, Figure 5A has been entered in the Draw Substructure step. At this point it consists only of an indole ring system with two chains. Any attachments to the aromatic ring are allowed (including hydrogens), and the side chains may be true acyclic chains or part of another ring. The bonds in the side chains will match only single bonds.

(*b*) *Enter R-Groups Step.* This and the following step enable the user to indicate points in the query where choices of atoms, groups, or larger structural fragments are allowed.

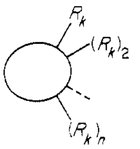
Before describing COUSIN's controls for inserting R groups into the fixed fragment and defining what the R groups are allowed to be, we will first discuss the notational format that was developed to support COUSIN's use of R groups.

R_k Method: It is characteristic of complex SS queries that somewhere within the query is an indication of a set of two or more atoms, groups, or larger fragments which represent "allowed" attachments at one or more positions in the query substructure. This is the basis of the commonly used generic or Markush[6] notation. It is a tool which permits the specification of a potentially large number of structures or substructures in a small space.

This section describes a flexible, yet easily learned method (termed the R_k method) for defining allowed variability in SS queries. Although it was originally developed for inclusion in COUSIN's SS query module, the notation is just as useful outside of the context of a computerized retrieval system. In this section we will be dealing strictly with the chemical meaning of the notation as it applies to SS query definition. The technique chosen for incorporating the R_k methodology into COUSIN and the graphical R-group definition tools available to the user will be discussed in the following sections.

The R_k method is based on the standard generic notation convention. In the simplest cases, where two or more atoms or groups are allowed to be attached at a particular position, it appears identical with the corresponding Markush form of notation. It was also designed to incorporate the descriptive power of the Merck "ZX" method[7] for handling indefinite positioning of variable groups but at the same time be simpler to use.

In formal terms the notation appears as follows, using generalized query **1**.



1

$$R_k = G_1(a_1, \ldots, a_x), G_2(b_1, \ldots, b_y), \ldots, G_r(f_1, \ldots f_z)$$

(1) An R_k symbol is attached to the fixed fragment at one or more positions (n = number of positions, k is an integer subscript).

(2) The R_k is then defined in a definition line which consists of the R_k followed by an equals sign followed by a definition expression.

(3) The definition expression consists of two or more structural fragments (G_1–G_r).

(4) Commas separating the fragments imply an AND relationship.

(5) Following each fragment is an optional parenthesized expression containing one or more occurrence count integers (e.g., a_1–a_x).

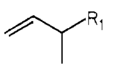
(6) Commas separating count values within a parenthesized expression imply an OR relationship.

(7) If a count expression is left out, the expression (0–n) is implied.

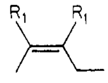
(8) A valid fragment combination is one in which the sum of count values, one selected from each parenthesized expression, equals n. For example, if $(a_1 + b_4 + \cdots + f_2) = n$, then a_1G_1's, b_4G_2's, and ... f_2G_r's taken together form a valid combination.

(9) The substitution patterns that correspond to each valid combination are those that result from permutation of the selected number of fragments over the n attachment points.

While the formal description of the R_k notation has been included here for completeness, the user can get by with a much less detailed view of the method since it is based on a familiar form of notation. In the simplest form of the notation, in cases where the R group is attached at only one position (variable group query) or where it is attached at more than one position (indefinite positioning query) but no restrictions on occurrences are desired, the occurrence counts mentioned in point 5 above can be left out. Example **2** therefore states that at the position marked by the R_1 in the fixed fragment, OH, Br, or N may be attached; nothing else is acceptable. Query **3** states that either a methyl or a bromine may be attached at either position on the double bond. There are three valid combinations (zero methyl, two Br; one methyl, one Br; two methyl, zero Br) which when permuted over the two attachment points give rise to four substitution patterns, since the monobromo–monomethyl case defines two separate patterns.

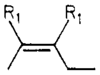


2, R_1 = OH, Br, N

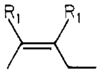


3, $R_1 = CH_3$, Br

Even though no occurrence counts have been specified in the two examples, there are *implied* occurrence counts. Example **2** could be rewritten as R_1 = OH(0,1), Br(0,1), N(0,1), and **3** could be rewritten as $R_1 = CH_3(0,1,2)$, Br(0,1,2). Example **3** is referred to as an *unrestricted* definition, since by not specifying occurrence counts the user is allowing all permutations of the specified groups (in this case four possible patterns). To restrict a definition to include a particular subset of all the possible patterns, occurrence counts are indicated on at least one of the fragments in the definition line. In **4**, for example, the definition line states that there must be exactly one methyl and one Br on the two positions. This combination selects two of the four possible patterns. The other two patterns are selected in example **5** which states that zero or two methyls and zero or two bromines are allowed. According to rule 8 of the formal statement of the R_k method, this means that if there are zero methyls there must be two bromines and if there are two methyls there must be zero bromines. Two methyls taken together with two bromines do not form a valid combination. Example **5**, therefore, selects the dibromo and dimethyl patterns. Thus, the inclusion of a count value does not *require* that number of the group in a pattern (unless the count value is the only one specified for a particular group as in example **4**), but rather it *allows* that number of occurrences of the group in a pattern. When that number is combined with allowed counts of the other groups *in such a way that the total of the counts equals the number of R-group attachment points*, then a valid combination has been defined.

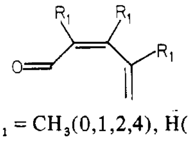


4, $R_1 = CH_3(1)$, Br(1)



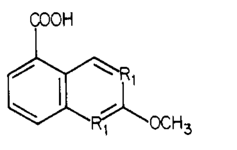
5, $R_1 = CH_3(0,2)$, Br(0,2)

In **6**, for example, two problems are evident. There is no way for the "2" in the CH_3 expression to take part in a valid combination unless there is also a "1" in the H expression. And there is no way for the "4" in the CH_3 expression to be part of any combination, regardless of what is in the H expression, since there are only three R_1 points of attachment. Errors of this type are caught by COUSIN and brought to the user's attention before the search is started.



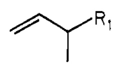
6, $R_1 = CH_3(0,1,2,4)$, H(2,3)

An R group may be embedded in the fixed fragment as in 7 where, according to the definition line, a nitrogen must be adjacent to the methoxy attachment point.

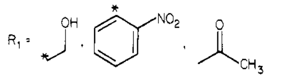


7, $R_1 = C(1), N(1)$

When dealing with larger structural fragments in a definition some indication of the attaching atom or "origin" of the fragment is usually needed. COUSIN uses asterisks for this purpose. In 8, for example, COUSIN will ask the user

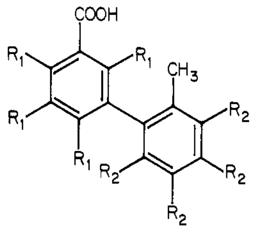


8



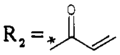
to specify origins on the first two fragments, while the third fragment has only one possible origin atom so an asterisk is not needed.

There are two devices that can be used to simplify a R_f notation without changing the meaning of a definition. The



9

$R_1 = Cl(2–4), H(rest)$

$R_2 =$  (1), N(0,1), H(rest)

first is simply the use of a dash to shorten count expressions in which three or more digits appear in sequence. In 9, the Cl count expression (2,3,4) has been shortened this way. The second is the use of "rest" in a count expression to remove from the user the burden of having to calculate valid counts.

In 9, for example, the user can state that on the R_1 positions there must be two, three, or four chlorines and the rest are hydrogens. The implied hydrogen count is H(0–2). In the R_2 definition the implied count is H(2,3).

Our use of the R_k method during the last 2 years has revealed several advantages in addition to its flexibility. The first is that the use of the notation need not be restricted to a computer search system. It can be used any time it is necessary to specify a class of compounds in which variability at one or more sites is involved. The notation also enables a user to specify allowed patterns in a way that is quite similar to the verbal statement of the same problem. Importantly, as query complexity increases it becomes difficult to build an R_k definition no faster than it becomes difficult to verbally state the nature of the query. And finally, the method permits a class of compounds to be defined in a rigorous fashion; a given definition will mean the same thing to different people. From examination of the R_k definition it is usually quite easy to distinguish between compounds that are covered and those that lie outside the class definition.

Inserting the R Symbols: We now return to the query entry display. When the "Enter R-Groups" step button is selected, the drawing controls that appear above the drawing box convert to those shown in Figure 4A. An R symbol may be picked up on the pen tip and inserted in the drawing in exactly the same way that atom and functional group symbols are manipulated in the first step. In Figure 5B of the sample query sequence, the user has attached to the substructure a single R_1 and four R_2 symbols. The user then goes to the next step to define R_1 and R_2 according to the R_k method.

(*c*) *Define R-Groups Step.* The graphical control buttons available to the user in this step are pictured in Figure 4B. Most of the buttons are the same as those available in the Draw Substructure step and enable the user to draw the allowed structural fragments for an R-group definition or modify the fixed fragment. There are some additional buttons which control the R-group definition and occurrence count insertion process.

To define an R group the user points to the DEFINE R button and then selects one of the R symbols in the fixed fragment. That symbol and any others with the same subscript begin to blink to indicate that all further drawing actions will be associated with the definition of that R group (i.e., the R group is "open" for definition). The user may then employ any of the drawing controls, including the predrawn rings, to define the allowed fragments in any vacant area of the drawing box. Referring to the sample sequence in Figure 5C, the user has selected R_1 (which is blinking) and has drawn five allowed fragments. The ORIGIN button (Figure 4B) was used to insert the asterisks. To terminate the definition, the DONE R button is hit. The fragments are automatically aligned below the fixed fragment and the R group is now closed (Figure 5D).

In the following frame (Figure 5E) the user has opened R_2 for definition and selected four atoms. To insert occurrence counts the user hits the INSERT # button and points to the desired fragment. A pair of blinking parentheses appears beside the fragment, and the digits can then be inserted simply by pointing to the desired digit buttons (See Figure 4B). In the example pictured in Figure 5E, the user is just about to hit the "rest" button to insert a rest count into the blinking parentheses beside the hydrogen. When DONE R is hit, the R_2 definition is closed, the R_2 symbols stop blinking, and the definition line is automatically formatted as in Figure 5F.

During R-group definition guidance messages appear frequently, and the level of error monitoring is increased. Also, at any time during query entry an R-group definition may be reopened for modification. Each of the R-group definition lines is subjected to a final detailed test of chemical reasonableness when the user eventually hits the SEARCH button to start the search.

(*d*) *Bond Types Step.* The buttons available in this step (Figure 4C) enable the user to relax or tighten constraints on individual bonds. Pointing to one of the bond symbol buttons and then pointing to a bond in the query causes a copy of the symbol to appear on the bond. In this manner, bonds can be forced to match only bonds with a particular cyclic character (the "cyclic", "cyclic-aromatic", "cyclic-nonaromatic", and "acyclic" buttons), or their bond orders can be relaxed to match more than one bond type (the "any", "single-double", and "double-triple" buttons). In Figure 5G the user has indicated that the bond connecting the lower side chain to the ring must be acyclic (slash symbol) and that the next bond must be in a ring (circle symbol) and may have any bond order (dotted bond symbol). The dotted "any" bond is used frequently for retrieving tautomers.

(*e*) *Ring Counts Step.* In this step (Figure 4D) the user can indicate the number of rings that must appear in retrieved compounds by selecting one or more of the digit buttons. A corresponding phrase appears automatically below the query. In the sample sequence (Figure 5H) the user has selected the tetra- and pentacyclic buttons. If the user had selected the monocyclic "1" button, COUSIN would detect the inconsistency (as there are already two rings in the query) and inform the user.

SEARCH EXECUTION AND SYSTEM HARDWARE

Once the query has been entered the user starts the search by hitting the SEARCH button. The query is then subjected to a detailed error analysis before the actual searching begins. There are approximately 20 conditions which will cause the system to ignore the search request and to inform the user of an adjustment that needs to be made. In particular, the R-group definitions are examined to ensure that all of the occurrence counts specified by the user can take part in legitimate variable group combinations.

The completed query is sent by the graphics terminal to the mainframe computer (an IBM 4341) in the form of an atom-bond connection table. There, it is subjected to a variety of chemical perception operations which extract from the query a list of search screens and which reformat the connection table. The "search formatted" query is then passed from the mainframe to a dedicted substructure search processor (a PDP11/55 minicomputer) where the actual search is performed.[5] The user is completely unaware of the distributed nature of the COUSIN system during SS searching. All other types of searching, including full-structure searching, are handled directly by the mainframe.

The results of the substructure search are sent back to the mainframe and placed in a temporary holding file which can then be manipulated in a variety of ways by user commands. The user also has the option of overlapping searching with the display of results. That is, matching compounds can be displayed as they are found while the search continues on the PDP11/55. However, since the time taken to search the entire file is less than 30 s in almost all cases (this includes an initial screening step followed directly by an atom-by-atom matching step)[5], the normal mode of operation is to wait until the search is complete before viewing the results. One of the options available to the user after completion of a search is to recall the most recent query to the query input display for modification. There are typically two reasons for doing this. Upon viewing the results of a search the user may see certain undesirable compound classes that had not been anticipated in the original query. If there is a large number of search results, the user will probably want to make the query more restrictive and rerun the search. Or if a search generates no matching compounds, users often begin to modify parts of the original query, making the structural constraints gradually less restrictive and rerunning the search to determine the structural level at which hits start to occur. Other command options enable both queries and result files to be assigned names and stored for future access or, in the case of result files, to permit Boolean operations on groups of compounds.

The front end of the COUSIN system consists so far of three graphics terminals, all using Digital Equipment Corporation's VT-11 graphics processors attached to PDP-11 minicomputers. Each of the graphics systems has two interactive devices, a Talos graphics tablet and a TI745 keyboard/printer. Two of the units also have Versatec 200 dot/inch printer-plotters attached for hard copy structure output. In addition, the COUSIN system can be accessed for nongraphical searches or information display via a large number of alphanumeric display terminals located throughout the company.

Although the PDP-11 graphics units are referred to as "terminals", each contains a 2.5 Mbyte disk drive which can act as system device, and each contains its own operating system. The graphics "terminals" are thus capable of standalone operation. When COUSIN is running, the mainframe generally assumes the role of master and the graphics units are slaves. However, when the user wishes to enter a full structure or substructure query, the mainframe turns over to the graphics terminal full responsibility for query construction. All graphical interaction with the user, connection table construction, button servicing, error feedback, and so on, are handled locally. The mainframe program only becomes active again when it receives the completed query connection table when the user hits the SEARCH button. This is a departure from the traditional approach which normally involves mainframe servicing of each button hit and each atom or bond insertion. By placing all the structure or substructure drawing software in the minicomputer, the interaction with the user is immediate, predictable, and unaffected by the level of mainframe system loading. The small disk drive attached to the minicomputer takes care of program segment overlays. These occur automatically during a query building session without user awareness and enable a rather large piece of graphics software to be run in a relatively small amount of memory.

DISCUSSION

Just as the substructure search module is only one part of the COUSIN system, the techniques described here for graphical entry of a SS query comprise only one part of the full SS module. In a subsequent article[5] we will examine in more detail what happens during the search itself. This will include a description of the search algorithms used and hardware and software considerations involved in optimizing search performance.

In the year and a half that COUSIN's substructure search facility has been in place it has been employed in a large number of searches not only by members of Upjohn's research population but also by those involved in research support functions. Our observations of the usage patterns during that period have enabled us to draw some conclusions about the effectiveness of graphics in an end-user oriented SS system. The first observation is that almost without exception users become very enthusiastic about being able to query a data base graphically. As a result, many scientists who had earlier chosen not to make use of a batch, intermediary-controlled SS system now use COUSIN routinely as a research tool. This is an especially important consideration for organizations that may now be in the process of deciding whether or not to upgrade their current system to a graphical, end-user controlled system. The current usage level of an older system is often used as a basis for inferring a research population's need for that capability. That is a mistake; the true need and true utility of SS in a research environment do not become evident until all the barriers which stand in the way of facile hands-on access to the data are removed.

The second observation concerns the amount of time spent at the terminal. Since COUSIN's R-group method enables the user to develop some very complex structural definitions in a single query, and since search times are relatively fast, the search needs of most users can be satisfied in a fairly short terminal session. Some users take advantage of that by leaving when they have what they want. Others take advantage of it by staying and exploring the data base much more deeply than they would otherwise be able to do in a reasonable time period. The intrinsic value of the data in the data base is thereby increased as the exploration can often generate ideas that were not originally sought.

The final observation concerns the frequency-of-use spectrum. As expected, there is a distribution of users which ranges from the heavy searchers who have integrated the system so thoroughly into their work that they seem not to be able to get along without it to those who use COUSIN once and then not again. Most fall somewhere in between. There are still a few people, however, who recognize a need to use the system but, for reasons of their own, choose to have a search intermediary do it for them or at least work with them on a graphical search session. At this point, therefore, it appears that the role of the search specialist has evolved away from searches originated by individual users (except in a few cases) who now can and should run their own searches, toward a function which satisfies information needs at project team and management levels.

REFERENCES AND NOTES

(1) For a representative sample of applications of interactive graphics for structure entry, see the following articles and references cited therein: (a) Corey, E. J.; Wipke, W. T.; Cramer, R. D.; Howe, W. J. "Computer-Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics". *J. Am. Chem. Soc.* **1972**, *94*, 421. (b) "Computer Assisted Organic Synthesis"; Wipke, W. T., Howe, W. J., Eds.; American Chemical Society: Washington, DC, 1977; ACS Symp. Ser. No. 61, Chapters 1, 5, 8. (c) Salatin, T. D.; Jorgensen, W. L. "Computer-Assisted Mechanistic Evaluation of Organic Reactions. 1. Overview". *J. Org. Chem.* **1980**, *45*, 2043. (d) "Computer Representation and Manipulation of Chemical Information"; Wipke, W. T., Heller, S. R., Feldmann, R. J., Hyde, E., Eds.; Wiley: New York, 1974; Chapters 3, 7. (e) Howe, W. J.; Hagadone, T. R. "Progress Toward an On-Line Chemical and Biological Information System at The Upjohn Company"; American Chemical Society: Washington, DC, 1978; ACS Symp. Ser. No. 84, Chapter 8. (f) Howe, W. J.; Hagadone, T. R. "Chemical Substructure Searching"; Proceedings of the Manufacturing Chemists Association Meeting, Arlington VA, Aug 1977; Manufacturing Chemists Association: Washington, DC, 1977. (g) Saxberg, B. E. H.; Blom, D. S.; Kowalski, B. R. "A Flexible Interactive Graphics System for Searching Atom Connectivity Matrices". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 233. (h) Blake, J. E.; Farmer, N. A.; Haines, R. C. "An Interactive Computer Graphics System for Processing Chemical Structure Diagrams". *Ibid.* **1977**, *17*, 223. (i) Dyott, T. M.; Stuper, A. J.; Zander, G. S. "Moly—An Interactive System for Molecular Analysis". *Ibid.* **1980**, *20*, 28.

(2) For a representative sample of the relatively few applications of interactive graphics for substructure query entry, see ref 3 and the following sources and references cited therein. (a) Dyott, T. M., et al. "An Integrated System for Conducting Chemical and Biological Searches"; American Chemical Society: Washington, DC, 1978; ACS Symp. Ser. No. 84, Chapter 11. (b) Blower, P. E.; Peercy, R. R.; Wade, L. G. "Design Your Own Information Service"; Chemical Abstracts Service, Columbus, OH, 1980; CAS Report 9. (c) The MACCS System, Molecular Design Ltd., Hayward, CA.

(3) "Substructure Searching of Large Chemical Files"; McNulty, P. J., Smith, R. B., Eds.; Proceedings of the Manufacturing Chemists Association Meeting, Arlington, VA, Aug 1977; Manufacturing Chemists Association: Washington, DC, 1977.

(4) COUSIN stands for "CompOUnd Search INformation system" and is a trademark of the Upjohn Company.

(5) Hagadone, T. R.; Howe, W. J., manuscript in preparation.

(6) Examples of Markush notations are seen in the following references: Kaback, S. M. "Chemical Structure Searching in Derwent's World Patent index". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 1. Sneed, H. M. S.; Turnipseed, J. H.; Turpin, R. A., Jr. "A Line Formula Notation System for Markush Structures". *J. Chem. Doc.* **1968**, *8*, 173.

(7) Brown, H. D., et al. "The Computer-Based Chemical Structure Information System of Merck Sharp and Dohme Research Laboratories". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 5.

A Periodic Table for Polycyclic Aromatic Hydrocarbons. Isomer Enumeration of Fused Polycyclic Aromatic Hydrocarbons.[†] 1

JERRY RAY DIAS

Department of Chemistry, University of Missouri, Kansas City, Missouri 64110

Received June 26, 1981

A comprehensive framework for the enumeration of all the isomers of even-carbon, fused polycyclic aromatic hydrocarbons (PAH) containing only hexagonal rings is defined. The basis for this framework is the molecular formula in contrast to the number of hexagonal rings. From the molecular formula, one can compute the number of rings from $r = (1/2)(N_c + 2 - N_H)$ and the total number of edges (σ C–C bonds) from $q = (1/2)(3N_c - N_H)$. These and other relationships are derived for the first time. Criteria for stipulating whether some specific molecular formula can be represented by a PAH containing only hexagonal rings is presented. An approach for possible computer enumeration of PAH isomers is proposed. The maximum number of five-membered rings that a structure corresponding to a PAH6 formula can contain in addition to hexagonal rings is given by $r_{5\max} \leq N_c - 2N_H + 6$.

Very few attempts have been made to systematically enumerate all possible polycyclic aromatic hydrocarbons (PAH; PNA for polynuclear aromatic compounds has been synonymously used).[1-4] In this paper, the scope and framework for achieving this goal is defined. The basis for this framework is the molecular formula in contrast to the number of hexagonal rings.[2] It is believed that this approach will be applicable to computer methodology. Also, a number of graph theoretical observations and theorems relevant to PAH's are espoused for the first time, and criteria for determining whether some specific formula can be represented by a PAH within the specified constraints are presented.

RESULTS AND DISCUSSION

Totally fused (condensed) PAH's having even-carbon nonradical systems possessing exclusively hexagonal rings will be discussed first (designated PAH6). Then an examination of analogue PAH systems containing pentagonal rings will be briefly pursued. Aromatic hydrocarbon systems that are not

Table I. Glossary of Terms

d_i	degree of vertex i of a graph
d_s	tree disconnections (of internal graph edges)
N_c	total number of carbon atoms in a PAH
N_H	total number of hydrogen atoms in a PAH
N_{Ic}	number of internal carbon atoms in a PAH having a degree of 3
N_{Pc}	number of peripheral carbon atoms in a PAH having a degree of 3
PAH6	fused polycyclic aromatic hydrocarbons possessing exclusively hexagonal rings
p_3	number of graph points (vertexes) having a degree of 3
q	number of graph edges (lines or C–C bonds)
q_I	number of internal graph edges
q_P	number of peripheral graph edges
r	number of rings
$r_{5\max}$	maximum number of pentagonal rings

[†] This work was presented as a special seminar to the research group of Professor Nenad Trinajstić at the Rudger Bošković Institute in Zagreb, Yugoslavia, on May 21, 1981.

0095-2338/82/1622-0015$01.25/0 © 1982 American Chemical Society