Techniques for Perception by a Computer of Synthetically Significant Structural Features in Complex Molecules

E. J. Corey,* W. Todd Wipke, Richard D. Cramer III, and W. Jeffrey Howe

Contribution from the Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138. *Received January 30, 1971*

Abstract: A computer program has been developed which starts with an internal machine representation of a chemical structure composed of atom and bond tables and effectively perceives a variety of structural units and relationships to be used for synthetic analysis by machine. Procedures are described for the recognition and storage of such synthetically significant information as functional groups, rings, and collections of certain types of atoms or bonds. A variety of data organizations, for example, *arrays*, binary *sets*, and linked *lists*, are utilized in the program.

The inspection of an organic structural formula by an experienced chemist involves perceptual processes which are remarkably direct and efficient. Although the detailed mechanisms of such perception are uncertain, it is clear that complex forms of symbolic and graphical recognition and memory correlation are involved. As a result the chemist obtains with a minimum of abstract or ordered analysis a complex mix of structural information which forms the basis of problem solving and creative thought. For synthetic analysis this information includes a knowledge of structural units such as functional groups, rings, steric screening groups, stereocenters, and stereorelationships, collections of chemically unstable groups, or groups under steric compression. An effective counterpart of this human method of examination which can be executed by a digital computer requires initially a simple internal representation of structure based on bonds and atoms and, additionally, data processing programs to derive complex information concerning polyatomic bond collections and relationships. This type of perceptual processing is the first step in synthetic analysis by computer and is required for each structure treated or generated in the analysis.

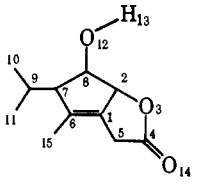
The present paper is concerned with the techniques and processes by which perceptual information can be generated by computer from the table of bonds and atoms which represent the structure. Application of this information to the selection of specific structural interconversions which lead to paths of synthesis linking specific intermediates is described in the following paper in this series.[1] The subject matter to be treated here may be divided as follows: (1) basic reorganization of structural data—construction of atom and bond sets; (2) ring perception; distinction between synthetically significant and nonsignificant rings; detection of aromaticity; (3) identification of functional groups; recognition of chemical sensitivity and instability common to several types of groups; the problem of unusual functional groups; (4) strategic disconnections—their use in planning syntheses; topologically important bonds and their perception.

An earlier paper[2] which briefly describes the machine representation of chemical structure in terms of a particular type of atom and bond (connection) table and the perception of rings and functional groups from these tables is intended to provide an introduction to the treatment which follows.

Data Organizations. For a given synthetic problem, structural perception is initiated by the computer from the atom and bond tables created as the chemist draws the structure. This first stage of machine perception involves an unambiguously ordered process of "sifting through" the data in the connection tables to uncover those arrangements of atoms and bonds that correspond to the major fundamental synthetic units. In order to understand how the computer perceives such arrangements, the data structures available to a computer for storing and organizing intermediate and final results of its perception must be examined.

The perception module of LHASA produces data which is organized in three ways differing in format:[3] one-dimensional *arrays*, binary *sets*, and linked *lists*. Briefly described, "*arrays*" are the familiar data organizations of FORTRAN, in which the *i*th *word* of a series of words contains a piece of information about the *i*th atom, bond, or other structural feature in the target molecule. A "*set*" organization of data is referenced by the name of a property, such as "nitrogen" or "secondary," and consists of a series of two or three words; the *i*th *bit* in the series of words will be 1 if the *i*th atom, bond, or other structural feature has the named property. "*Lists*" are a type of data organization used when the interrelationships among data are so complex that they must be presented explicitly along with the data. (To take a simple example of such a relationship, complete representation of a ring demands knowledge not only of the "names" of the atoms and bonds involved but also of the order in which they are encountered when proceeding around the ring.) Such interrelationships can be portrayed in a "*list*" structure by requiring each element in a list to contain not only data but also instruction codes for finding any elements that are significantly related to the present element. (By contrast, in an *array* there is no explicit information about relationships among the elements. Its interpretation depends on the implicit assumption that elements are ordered with the $i + 1$th following the *i*th.) An additional useful property of list storage derives from the fact that the memory spaces used need not be contiguous.

(1) E. J. Corey, R. D. Cramer III, and W. J. Howe, *J. Amer. Chem. Soc.*, **94**, 440 (1972).
(2) E. J. Corey and W. T. Wipke, *Science*, **166**, 178 (1969).
(3) A. T. Berztiss, "Data Structures, Theory and Practice," Academic Press, New York, N. Y., 1971.



Examples:						
OXYGEN	001	000	000	001	010	....
BOND2SET	100	101	000	000	010	....
PRIMARY	000	000	000	110	111	....
Elementary Set Operation:						
OXYGEN and BOND2SET → OXO						
OXO	000	000	000	000	010	....

Figure 1. Simple atom sets.

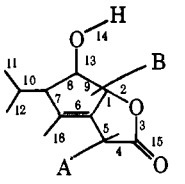
Some specific illustrations of each of the last two unfamiliar data organizations will now be considered. As the target molecule is being entered by the chemist, the atoms and bonds are numbered sequentially by the computer. Figure 1 shows a bicyclic lactone, with atoms given reasonable input sequence numbers. Immediately under the structure are shown parts of some of the sets which would be created by LHASA during perception of this molecule. For example, the oxygen atoms shown in the structure are numbered 3, 12, and 14. Therefore, the third, twelfth, and fourteenth bits in the atom set OXYGEN are ones and all remaining bits are zeroes. BOND2SET, which contains all atoms that are at one end of a double bond, has bits 1, 4, 6, and 14 as ones.

Sets can be manipulated and combined by the computer with remarkable facility, using basic instructions such as the logical *and*[4a] or the (inclusive) *or*.[4b] For example, the set OXO can be defined chemically as "those atoms which are oxygen and at one end of a double bond." This set of atoms may be constructed by performing a logical *and* between OXYGEN and BOND2SET and storing the result in OXO. A program statement which commands this process and an illustration of the resulting set OXO are shown in Figure 1. This method of processing sets is particularly powerful in that it represents *parallel* rather than *serial* handling of information; *i.e.*, a single *and* operation manipulates 18 pieces of information simultaneously.

A more advanced example of set manipulation programming is shown in Figure 2, where the problem of finding all appendage bonds attached to ring B is considered for the same structure. In this structure the bonds are labeled with the appropriate sequence numbers. Under each reference to a set by the program in Part I appears a Roman numeral, which is used in Part II to label the current contents of the set named.

At the start of this code sequence, RINGSETB (I) contains the bonds that constitute ring B (the nonlactone ring). Each bond in RINGSETB is extracted in turn, using the subroutine FRE(*i*, set), which on repeated calls returns as the value of *i* the name of the first item in the set beyond the input value of *i*. (For example, FRE(1, RINGSETB) would return 6, the sequence number of the next bit with value 1 beyond bit 1.) When no more items remain in the set, FRE returns a zero; this situation is tested for by the third instruction.

(4) (a) The result of the *and* operation between two sets is the intersection of the sets; (b) the result of the *inclusive or* operation on two sets is the union of the two sets.



Part I: Sample Program

```
                0 → i
newrgbond:  fre(i, ringset B → i
                      I
            if i = 0 then goto continue
            gab(i) AND notringset tempset
              II          III        IV
            tempset OR appendage appendage
                          V          VI
            goto newrgbond
continue:   . .
```

Part II: Contents of Sets during Program Execution

I:	100	001	111	000	000	000	···	
II:	000	000	101	100	100	000	···	(when i = 8)
III:	000	000	000	111	111	100	···	
IV:	000	000	000	100	100	000	···	(when i = 8)
V:	000	000	000	100	000	100	···	(when i = 8)
VI:	000	000	000	100	100	100	···	(when i = 8)

Figure 2. Example of programming using sets.

Then the subroutine GAB(*i*) is called to return the set of bonds attached to the *i*th bond, as illustrated for i = 8 (II). The resulting set of attached bonds is ANDed with NOTRINGSET (III), a set containing all bonds not included in any ring, to yield as TEMPSET (IV) only those bonds attached to the *i*th bond which are not members of rings. TEMPSET is then included in the accumulated results for the other bonds in ring B, APPENDAGE (VI). When this procedure is carried out for every bond in RINGSETB,[5] APPENDAGE will contain exactly those bonds which are roots of appendages to ring B.

The major lists produced during perception contain information about RINGS and functional GROUPS. The list handling programs written for LHASA produce a relatively simple list structure. Two consecutive memory words are required for each element in a list; the second contains the address of the next element in the list, and the first contains either data or the address of another list, said to be a "sublist" of the original list. The last element in a list or sublist contains a zero address.

As an illustration, the structure of the RINGS list for the bicyclic lactone is presented diagrammatically in Figure 3. Each box represents a single memory word; a pair of boxes thus represents a single list element. The addressing of one element by another is symbolized by an arrow pointing to the addressed element. Data about individual rings in the target structure are shown vertically as sublists of the main horizontal RINGS list. The first element in each sublist specifies

(5) An attentive reader may observe that examination of all bonds in ring B is unnecessary. All appendages to ring B may be picked up by examining only the attachments to alternate bonds (triplet 1, 3, 9 or triplet 2, 6, 7). Unfortunately, there is no way of determining the *order* of the bonds in ring B from the structure of RINGSETB, and hence no way of identifying alternate bonds. (The sequence of items within a set is quite arbitrary, reflecting at most perhaps the order in which bonds and atoms were drawn by the chemist.) It is true that the order of bonds in ring B is contained in the appropriate list. However, it turns out that extracting the appendage bonds by "wastefully" performing parallel operations on all the elements in a set is both faster and more economical of computer memory than extracting this information by performing sequential operations on alternate elements in a list.

the "name" and the size of the ring; subsequent elements name in order the bonds and atoms involved.

Each of these three data organizations has advantages and disadvantages. Some sort of list-like organization is irreplaceable for representing relationships among data. However, list structures consume memory rapidly and require indirect and relatively slow procedures for the storage, manipulation, and retrieval of data. Sets are peculiarly easy to manipulate and rather economical of memory. However, the range of data that can be represented in a set is limited to binary "true-or-false" descriptions about the properties of items that can be "named" by consecutive numbers. Furthermore, most computer languages and instruction sets do not permit direct access to individual bits in memory. Arrays present data in a very accessible organization, which, however, does not offer the unique advantages of list and set structures. The availability of diverse data structures has allowed the selection of the particular data structure which is most effective for a given application.

Basic Perception. The first step in perception is the creation of "basic sets" from the atom and bond tables. "Basic sets" are defined as those sets which are derived by simple reformulation of data already explicit in the tables. This category includes sets such as NITROGEN, PRIMARY, and CATION, which convey information such as "atom 2 is nitrogen" or "atom 6 is attached to just one atom." Table I contains a list of the basic sets. Reorganization of input data into basic sets facilitates subsequent handling in perception and other operations, both because of the parallel processing thereby possible (see above) and because a frequent operation, that of accessing only those atoms or bonds having a specified property, becomes more direct. (As an illustration of this operation, see the use of FRE() in Part I of Figure 2.)

As the basic sets are being filled from the atom and bond tables, certain "secondary sets" are also created. These sets contain information that, although explicit in the tables, cannot be obtained from a single atom or bond entry. A major example is the array of secondary sets GAM(i), in which the *i*th set names the atoms attached to the *i*th atom. Inspection of the information contained in individual atom and bond entries will show that a determination of the n atoms attached to some atom will require $n + 1$ look-up operations (one atom entry and n bond entries). Use of the sets GAM(*i*) greatly speeds subsequent structural searching operations.

Once these sets have been filled, the attempts to perceive functional groups and rings begin. One result of these operations will be the creation of additional useful sets, which may be classified into "topological" and "reaction site" sets, as seen in Table I.

Functional Group Perception. Recognition of simple functional groups is a prerequisite for even the most naive synthetic analysis. Both computer and human recognize functional groups by perceiving an identity between some combination of atoms and bonds in a molecule and a combination that they remember and have a name for. Within the computer this "recognition" involves atom-by-atom matching between some part of a current target structure and a table giving the structural requirements for functional groups.[2] Data

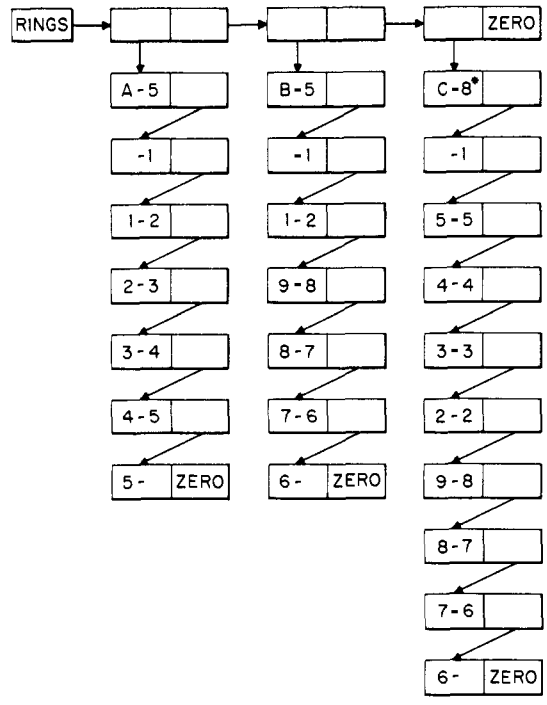


Figure 3. A list structure depicting a ring network. In a number pair, the first refers to the bond, while the second is the atom number. (The * designates a "pseudo" ring. See text.)

in this table include phrases that control the sequence of the matching processes, making group recognition a highly efficient procedure.

Details of the functional group recognizer will now be considered. First, LHASA does not try to match all of the combinations of atoms and bonds appearing in the current target against the data table. Only certain classes of atoms can initiate attempts to recognize a group. For example, the atom class first scanned by the recognizer is the set of atoms which are both "primary" (having only one attached atom appearing in the connection table) and "hetero" (O, S, N, or P).

Assume that the first atom in this set is an oxygen. The following portion of the data table then describes attachments to this primary oxygen atom which might be parts of recognizable functional groups:

```
DOUB N NUG A1D
DOUB C NUG A1C LAST
```

These lines instruct the program that if the primary oxygen atom is attached by a double bond to a nitrogen or carbon atom, a nitroso or carbonyl part structure has been detected. These substructures both may be parts of functional groups that LHASA recognizes, so NUG directs the recognizer next to examine the attachments to the next atom, nitrogen or carbon. These attachments will be matched against portions of the table addressed by either A1D or A1C, respectively. For example, the portion addressed by A1D appears as:

```
SING O MATCH NITRO LAST
```

If the attachments to the nitrogen of a nitroso part structure include an oxygen joined by a single bond, the word MATCH will signal the recognizer that a group whose name is NITRO has been found in the current target. The "names" of the atoms and bonds encountered during the matching process are strung

Table I. Atom, Bond, and Functional Group Sets Used in LHASA

Name	Set contents	Name	Set contents
1. Basic sets		ALPHST	Atoms attached to a member of CONNST
OCCUP	Atoms in the current structure	CJBD	Multiple bonds conjugated with other multiple bonds
CARBON	Carbon atoms	TERMINAL	Multiple bonds at the end of a conjugated series
HYDROGEN	Hydrogen atoms (explicit)	PRIM1	Atoms attached to just one carbon atom
NITROGEN	Nitrogen atoms	SEC1	Atoms attached to just two carbon atoms
OXYGEN	Oxygen atoms	TERT1	Atoms attached to just three carbon atoms
PHOSPHORUS	Phosphorus atoms	QUAT1	Atoms attached to just four carbon atoms
SULFUR	Sulfur atoms		
HALIDE	Halogen atoms		
HETERO	Nitrogen, oxygen, sulfur, or phosphorus atoms	3. Topological sets	
BOND1SET	Atoms to which at least one single bond is attached	RINGSET	Atoms belonging to any ring
BOND2SET	Atoms to which at least one double bond is attached	RINGBSET	Bonds belonging to any ring
BOND3SET	Atoms to which at least one triple bond is attached	NOTRINGB	Bonds not belonging to any ring
PRIMARY	Atoms attached to just one other nonhydrogen atom	AROMAT	Atoms belonging to an aromatic ring
SECONDARY	Atoms attached to just two other nonhydrogen atoms	RESON	Bonds belonging to an aromatic ring
TERTIARY	Atoms attached to just three other nonhydrogen atoms	JUNCTSET	Atoms belonging to more than one ring
QUATERNARY	Atoms attached to just four other nonhydrogen atoms	BRGHD	Bridgehead atoms
NEUTRAL	Atoms bearing no charge	SMALLR	Bonds belonging to a three- or four-membered *real*[a] ring
RADICAL	Atoms bearing an unpaired electron	RING5B	Bonds belonging to a five-membered *real*[a] ring
ANION	Atoms bearing a unit negative charge	RING6B	Bonds belonging to a six-membered *real*[a] ring
CATION	Atoms bearing a unit positive charge	LARGER	Bonds belonging to a *real*[a] ring larger than six members
OCCUPB	Bonds in the current structure	APPENDAGE	Bonds attached to a *real*[a] ring and in NOTRINGB
BOND1	Single bonds	EXOINR	Bonds attached to a *real*[a] ring and in RINGBSET
BOND2	Double bonds	FRAG(*i*)	Atoms in the *i*th fragment of the molecule
BOND3	Triple bonds		
CHEMSTB	Bonds specified by the chemist as "strategic" (see paper I)	4. Reaction site sets	
2. Secondary sets		FGORG	Atoms which are "points of attachment" of some functional group
GAM(*i*)	Atoms attached to the *i*th atom	WORG	Atoms which are "points of attachment" of those functional groups which are electron withdrawing
HETBOND	Bonds attached to an atom in the HETERO set	MBONDSET	Bonds which constitute functional "double" or "triple" bonds; *i.e.*, not aromatic or carbonyl
HBOND	Bonds attached to a hydrogen	GIF(*i*)	Functional groups in the *i*th fragment
HSET	Atoms having attached hydrogen (explicit or not)	INSTAB(*i*)	Functional groups unstable toward the *i*th type of synthetic reagent (see text, below)
XHSET	Atoms having explicit attached hydrogen		
ALLYLIC	Atoms attached to a member of BOND2SET or BOND3SET and themselves not members of those sets		

[a] "*Real*" rings are a subset of the rings present, defined below in the text.

onto a list, which is headed by a code number or "name" denoting the group type and a unique identification. The resulting list is made a sublist of the GROUPS list. Finally the matching process will be restarted with a different primary hetero atom.

On the other hand, if a primary oxygen atom is selected which is not attached by a double bond to carbon or nitrogen, a LAST in the table warns the recognizer that any other attachments to primary oxygen atoms could not be part of recognizable groups, *i.e.*, no further table need be searched. Thus, the present version of LHASA will not perceive groups such as sulfone or phosphate, since these groups contain primary oxygen attached to sulfur or phosphorus. However, since the group attributes are set up in a table format rather than imbedded in executable program code, extending the scope of the group recognizer is simply a matter of adding new data to the table. The "programmed" part of the recognizer—the section that interprets the table—does not need to be touched.

The substructures which LHASA currently recognizes as constituting functional groups are shown in Table II along with general electronic descriptors which are recognized for each group as defined in Table III.[2] The use of generalized descriptors which depend on the electronic properties of functional groups allows certain economies in the tabulation and processing of chemical data (see following paper[1]). However, the assignment of a particular group to a general family is warranted only if the general descriptor can be used as a valid synonym for *each member* of the group. Wherever a general descriptor is inappropriate, specific group names must be employed. As is discussed in detail in the following paper, the general descriptors and the specific functional group names are used both to *select* and *evaluate* synthetic processes.

Carbonium ions are ordinarily considered to be transient species, since they are seldom isolated and dealt with as stable synthetic intermediates. However, such structures undergo synthetically valuable changes which are important and sufficiently complex to require the special treatment accorded functional groups. A special provision then made is that targets containing a carbonium ion may undergo only those

Table II. Functional Groups. Target Substructures Defining Functional Groups to LHASA

No.	Name	Substructure	Generalized class[b]
		Standard Groups	
1	Acid	COH (O)	Oxo
2	Acid halide	CHal[a] (O)	Oxo
3	Alcohol	COH	D, X
4	Aldehyde	CH (O) or: CC (O)	Oxo, W
5	Amide	CN (O)	Oxo, W
6	Amine	CN	D, Z
7	Cyano	C≡N	W, Z
8	Double bond ("dbond")	C=C (nonaromatic)	
9	Epoxide	C—O—C	X
10	Ester	COC (O)	Oxo, W
11	Ether	COC	
12	Halide	CHal	X
13	Imine	C=N	W
14	Ketone	CCC (O)	Oxo, W
15	Nitro	O=N—O	W
16	Triple bond ("tbond")	C≡C	
		Special Groups	
17	Carbonium	C^+	
18	Vinylw	(W)C=C	W
19	Esterx	*COC* (O)	D, X
20	Amidz	*CNC* (O)	Z

[a] LHASA understands the symbol "X" to represent "halide," but to avoid confusion between this representation and that of the generalized "XGROUP" defined later in this paper, we will refer to a halide as "Hal." [b] Defined in Table III.

Table III. Reactive "Classes" for Functional Groups[a]

Name	Distinguishing feature
dgroup	Group which is electron donating because it contains an atom having an unshared pair (would be ortho, para directing and activating toward electrophilic aromatic substitution)
oxo	Group containing a carbonyl substructure
xgroup	A weakly nucleophilic group that could be introduced synthetically by nucleophilic displacement of a halide
zgroup	A strongly nucleophilic group (not a good anionic leaving group)
wgroup	Group sufficiently electron withdrawing to allow nucleophilic activation of an attached carbon

[a] The types of groups belonging to each "class" are shown in Table II.

retrosynthetic reactions (equivalent to the term "transforms" as defined in the following article of this series) in which the carbonium ion participates.

The need for a "vinylw" group arises from the ability of a double bond to "transmit" the electronic properties of an attached functional group to atoms at the other end of that bond. This well-known effect is particularly important for synthesis when the attached group is electron-withdrawing, because the power which electron-withdrawing groups have in promoting carbon–carbon bond formation is thereby extended two atoms. Therefore, LHASA has been instructed to create an additional special "vinylw" group whenever it encounters a double bond attached to an electron-withdrawing group.

Alternate names for the ester and amide groups are necessary to reflect adequately the variety of electronic effects that these groups have on attached carbons. The carbonyl portion of these groups weakly activates its carbon attachment toward electrophilic attack, whereas the nitrogen or oxygen atom weakly activates its attachments toward nucleophilic attack (by virtue of its potentiality as a leaving group). In subsequent synthetic analysis of an ester or amide, bonds α,β to the carbonyl group can be broken quite differently from bonds α,β to the other hetero atom. Therefore, for each ester group that is recognized, an additional "esterx" (ESTER, XGROUP) entry in the list of groups is created to recognize the possibility that this function allows the retrosynthetic disconnection to a carboxylate ion and an electrophilic carbon species. Similarly, "amidz" is created for an amide function RCONHR′ which could undergo the retrosynthetic disconnection to $RCONH^-$ and electrophilic R. (Strong nucleophiles are designated as "z" groups, whereas weak nucleophiles are indicated by the descriptor "x." One important difference between these is that whereas an "x" group β to carbonyl is readily eliminated by base, a "z" group is not.)

Perception of Molecular Properties Associated with Simple Functional Groups. The concept of a functional group makes useful generalizations about organic synthesis possible. (For example, any "alcohol" may be synthesized by Grignard addition to a "ketone.") A useful higher order generalization about groups is to label all those groups which have similar electronic effects on adjacent atoms by a single "class" name. For example, the observation that nitro, cyano, aldehyde, ketone, and ester groups are all sufficiently electron withdrawing to allow removal of a proton on an adjacent carbon can be expressed in general form by labeling all such groups "wgroups." This type of convention allows the requirements for such reactions as the Michael addition, ozonolysis, or nucleophilic epoxide opening to be expressed in a more compact way. The "classes" of groups recognized by LHASA and a description of their distinctive properties are given in Table III. The class or classes associated with a particular functional group type are checked in Table II.

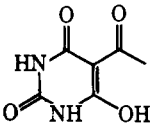
Information regarding the sensitivity of a target molecule toward the synthetic reagents required for a particular transform is necessary for the evaluation of that transform. For the most part, the overall sensitivity is a combination of the sensitivities of the individual groups present. LHASA therefore creates a set for each of the common reaction conditions involved in synthesis. The *i*th bit in such a set will be "on" if the *i*th group in the target molecule is unstable to the reaction conditions associated with that set.

A simple but useful procedure for the specification of reaction conditions has been devised on the basis of the following factors: (1) whether the medium is protic or aprotic, (2) whether the medium is acidic, neutral, or basic, and (3) reagent selectivity with regard to functional groups. A seven-level scale of acidity–basicity is used which would correspond in a *protic* medium to (1) concentrated strong acid, (2) pH 1–3, (3) pH 3–5, (4) pH 5–9 (effective neutrality), (5) pH 9–11, (6) pH 11–14, and (7) concentrated base. The table entry for a given transform usually includes a description of the medium in terms of the seven-level protic or aprotic scale. In addition, when a specific reagent is indicated for a transform, data can be made available from a table with regard to those functional groups which are sensitive to that reagent, and these data then can be utilized in the assignment of a rating to the transform. Clearly, this approach can in principle also be used by a program (to be developed later) for the application of protecting groups in a synthetic sequence. At a later stage of program development it will be advantageous to include two types of reagent tables. First, as mentioned above, for each reagent a set of sensitive functional groups, and second (and inversely) for each functional group the set of reagents (or reagent types) to which it is sensitive.

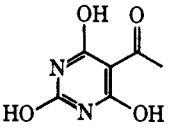
A few examples involving reducing agents will serve to illustrate the technique outlined above. Birch reduction conditions ($Na–NH_3–ROH$) would be classed as protic-strongly basic (P-level 7) with sensitivity of the following units: aromatic ring, wgroup (excluding COOH, CONH–), vinylw, tbond, dbond (conj), and halide. Chromous acetate or aluminum amalgam would be classed as protic–neutral (P-level 4) and reactive towards nitro, vinylw, halide, and W–C–D. Tributyltin hydride would be classed as aprotic–neutral (A-level 4), reactive toward acid halide, aldehyde, ketone, nitro, halogen. Formic acid would be classed as protic–acidic (P-level 2) and reactive toward C=N. Lithium aluminum hydride would be classed as aprotic–strongly basic (A-level 6), reactive toward w, vinylw, oxo (including COOH), C≡CCOH, halide.

Reactivity of Functional Groups. The effectiveness of organic synthetic reactions often depends on the selectivity which can be realized in molecules containing two or more functional groups of the same kind, as examples 1–5 will show: (1) primary OH, tertiary OH → primary OAc, tertiary OH; (2) primary Br, allylic secondary Br → primary Br, allylic OH; (3) β-lactam, amide → β-amino acid, amide; (4) $RCH{=}CH_2$, $R_2C{=}CR_2$ → RCH_2CH_3, $R_2C{=}CR_2$; (5) ether, ketal → ether, ketone. In cases where the application of a transform to some target molecule would generate a structure containing two or more functional groups of the same kind, the evaluation of that transform requires data on the relative reactivities of these groups, which in turn necessitates additional perception with regard to the environment of the functional groups. Provision for this sort of analysis has been made in LHASA and is discussed in the following paper. A more extensive implementation which will be used in later versions of the program will necessitate an expansion of the perception module to include perception of functional group environment and a data table which summarizes the effect of structure on reactivity. The perception of all highly reactive functional groups is crucial to the use of strategies for the effective and optimum ordering of a sequence of transforms.

Complex Functional Groups. Recognition of all the functional groups is not necessarily as straightforward a process as the preceding discussion implies. The elementary concept of a functional group, "groups of atoms which undergo characteristic reactions," is usually applicable only when "functional groups" are separated by insulating (*e.g.*, saturated carbon) atoms. Consider the problem of defining the functional groups in the following structure A.



A



B

One might begin by enumerating "fundamental functional units" containing no more than two atoms. These are found to be two amino groups, three oxo groups, a double bond, and a hydroxyl group. Of course, an amino group attached to an oxo group constitutes an amide group; there are at least three such attachments, and a fourth if the keto tautomer of the vinyl alcohol is considered. Also present are two vinyl ketone groups which share the same double bond. Combinations among three fundamental functional units yield two imide groups, a urea, and two vinylogous carboxylic acids. In addition, structure B, the aromatic tautomer of A, must be taken into account. In this structure the three hydroxyl groups are phenolic in type, and the two imino groups and endocyclic C=C in the context of an aromatic ring define a pyrimidine nucleus.

The functional group recognizer currently in use would identify only about half the groups present in this situation, and the particular groups detected would depend on the order in which the structure was drawn by the chemist. This is primarily because the present recognizer attempts to assign each hetero atom to just one group. That is, a particular nitrogen atom might be part of either an amine group, or an amide, or an imide, etc., but not part of more than one of these.

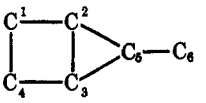
Perception of Rings and Networks. The perception of rings of atoms in chemical structures and the relationships between atoms in or on rings are central elements in synthetic analysis. The recognition of rings and ring sizes is fundamental to the application in synthetic design of the many important chemical reactions which generate or manipulate cyclic structures (*e.g.*, Diels–Alder, aromatic substitution). Thus the *selection* of the appropriate chemical manipulations for generating a synthetic tree of precursors to some target molecule requires detailed topological information. The *evaluation* of each intermediate and the chemical step which converts one intermediate into another also rests on such data. For example, a Grignard–carbonyl addition is generally inappropriate for the closure of a ring, although often of great value for the creation of a noncyclic link. On the other hand, the aldol reaction which is effective for the formation of certain types of rings is inoperable for other cyclic units (*e.g.*, three- or four-membered rings). Higher order topological information, such as the locations of appendage bonds

and common atoms, and, eventually, the perception of stereorelationships, depends on the perception of rings. Such topological information is presently used for strategic decisions and for avoiding the formation of intermediates containing highly strained ring systems.

The first step in ring perception by LHASA is the detection of every ring in the target molecule. Starting with the first atom in the connection table, the ring detector generates a random connected path through the target structure. Those atoms where an arbitrary choice among attached atoms had to be made are labeled "choicepoints."

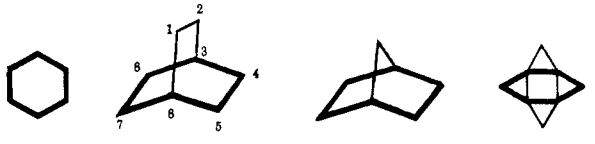
Whenever an atom which is already a member of the path is encountered, a ring has been detected. The atoms and bonds in this ring are stored in order as a sublist of the RINGS list. Then the path is shortened to the last choice point, and generation of a different path is begun. Eventually all choices will have been tried. If there are then atoms in the current target structure which have not been visited, the target structure consists of several separate structures, or "fragments."[6] The ring detector is then restarted at one of the unvisited atoms and path generation resumes.

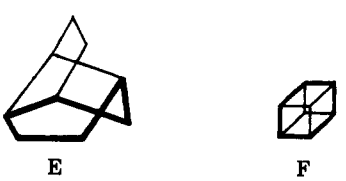
This ring detection process will now be illustrated for a methyl-substituted bicyclo[2.1.0]pentane. Each stage of path generation is shown; those atoms in a path which are choice points are starred.



1,2*,3*,4,1 . . . found the four-membered ring (1,2,3,4)
. . path trimmed to (1,2*,3*)
1,2*,3,5*,2 . . . found the three-membered ring (2,3,5)
. . path trimmed to (1,2*,3,5*)
1,2*,3,5,6,? . . . came to end of branch
. . backtrack to last choice point, leaving the path (1,2*)
1,2,5*,3*,2 . . . found three-membered ring again
. . backtrack to path (1,2,5*,3*)
1,2,5*,3,4,1 . . . found five-membered ring (1,2,5,3,4)
. . backtrack to path (1,2,5*)
1,2,5,6,? . . . came to end of branch
. . no more choice points in path so all rings have been detected

The RINGS list generated by this procedure is exhaustive and includes many rings of little chemical significance. For example, in the following series of alicyclic hydrocarbons, each of the darkened six-membered rings would be detected.





Both the original program OCSS[2] and the currently used program divide all rings into two discrete classes, "real," or significant rings, and "pseudo," or insignificant rings. The set of "real" rings for a given molecule has been defined previously in terms of set theory; the concepts involved will be restated here in relatively imprecise but perhaps more visually evocative language.

1. The set of real rings must contain all the bonds which are members of any ring.

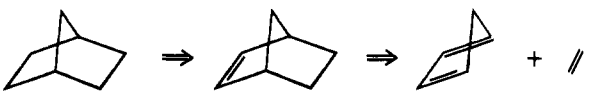
2. This set must be chosen so that the sizes of the individual rings are as small as possible.

3. The set of rings must include exactly the number of rings predicted from the molecular formula. **Exception.** If there are several equivalent sets of rings meeting these criteria, all the rings in all the sets must be "real." For example, the molecular formula of bicyclo[2.2.2]octane (B), C_8H_{14}, requires only two rings, but the structure actually contains three identical rings—(1,2,3,4,5,6), (1,2,3,6,7,8), and (3,4,5,6,7,8). Since any selection of a particular pair of rings would be arbitrary, all three rings must be considered "real."

The algorithm used for selecting the set of rings meeting these criteria has previously been described.[2] In the hydrocarbons shown above, the darkened rings in structures A and B would be real, and those in structures C through F would be pseudo. All the rings, "real" or "pseudo," remain on the RINGS list.

As will be seen in the following discussion, these criteria are not entirely satisfactory for testing ring significance. To devise an appropriate ring significance test, we must recall the contributions that ring detection make to synthetic analysis. Briefly, these are as follows: (1) Rings are "key synthetic units" for certain reactions. (2) Success or failure of most bond-forming synthetic reactions is strongly affected by the size of any ring being formed. (3) Rings are the basis for more sophisticated topological perceptions.

The rings designated as "real" in the originally used algorithm[2] do not include all the rings which might be "key synthetic units." A prominent example is the norbornane system, to which this algorithm assigns as real rings only the two five-membered rings. The neglect of the six-membered ring is serious here, since the perception of this ring can lead to the selection of synthetic precursors which combine by the Diels–Alder reaction.



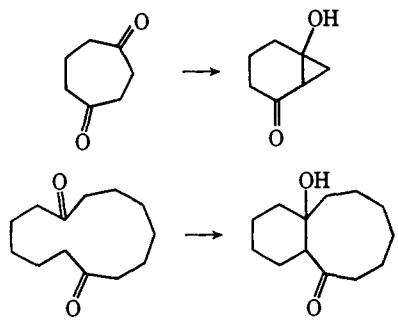
On the other hand, there are few synthetic reactions now known that have rings larger than six members as key synthetic units. Those reactions which do involve larger rings are successful only when the larger ring would be regarded as "real" now. Therefore, if all rings which may be key synthetic units are to be significant, the definition of significant can be "all rings of size 6 or smaller and all rings larger than 6 which are presently regarded as real." An algorithm to generate such a set of rings, starting with a connection table, has been devised[7] and coded in DECAL and FORTRAN and is available in LHASA.

The problem of determining whether or not a bond is in a ring of specified size, required for evaluation of many reactions, is rather simple. Ordinarily, when a bond in a target molecule is a member of several rings

(6) Current targets consisting of several "fragments" are often encountered as an analysis session proceeds, since a parent structure may be disconnected to give a pair of precursors.

(7) E. J. Corey and G. A. Petersson, *J. Amer. Chem. Soc.*, 94, 460 (1972).

of different sizes, the ease of forming that bond is determined primarily by the size of the *smallest* ring. For example, consider the plausibility of using an aldol reaction to form a six-membered ring that is fused to (a) a three-membered ring or (b) a nine-membered ring. In

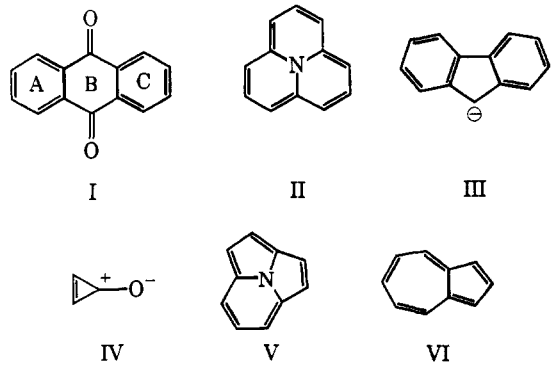


each case, the bond being formed is in two rings in the product, one of which has six members. The reaction goes well only when the six-membered ring is the smaller. Evidently for the purposes of evaluating reactions, each "ring" bond should be labelled by the size of the smallest ring it is in.

Aromaticity. Since the chemist enters aromatic structures using ordinary single and double bonds, LHASA has the responsibility for determining which, if any, atoms and bonds in the structure are aromatic. To be aromatic, a bond or atom must be a member of a ring, real or pseudo, that contains $(4n + 2)\pi$ electrons and whose constituent atoms will sustain a ring current. To sustain a ring current, an atom must have one or more of the following properties:

1. Be a member of a double or triple bond contained *wholly* within the ring; such an atom contributes one π electron.
2. Be an anion (two π electrons) or cation (no π electrons).
3. Be a hetero atom (oxygen, nitrogen, or sulfur); contributes two π electrons. (An atom having two of these properties contributes π electrons consistent with its first listed property. For instance, the nitrogen in pyridine meets criteria 1 and 3; 1 is listed first, so the nitrogen contributes one π electron.)

Results from this algorithm follow.



I II III

IV V VI

In anthraquinone (I), atoms and bonds in rings A and C are aromatic, whereas those contained wholly in ring B are not, since the carbonyl double bonds are not wholly in ring B. Of the nitrogen heterocycles II and V, II is wholly nonaromatic, whereas the perimeter of V is aromatic. All of the fluorene anion (III) is aromatic, no matter how it is drawn. All atoms and bonds in azulene (VI) are aromatic, since the perimeter contains 10 π electrons. Cyclopropanone (IV) will be aromatic if drawn in its polar form.

Relationships between Fundamental Perceptual Units. Many of the observations made by an experienced chemist about a molecular structure when analyzed turn out to be perceptions of relationships between two or more functional units. To cite some trivial examples, a ring system containing two five-membered rings which have three atoms in common is a norbornyl system. A six-membered ring with two double bonds in a 1,4 relationship is a substructure that suggests synthesis *via* Birch reduction. A cis relationship between two hydroxyl groups on adjacent alicyclic carbons suggests oxidation of a double bond.

LHASA examines relationships between rings for only one purpose at present, the identification of bridgehead atoms. A bridge exists whenever a pair of "real" rings has more than two atoms in common. The atoms within the bridge which are adjacent to only one other bridge atom are the bridgeheads.

Perception of relationships between units prior to the stage in which chemical disconnections and manipulations are selected is not feasible primarily because so many possible relationships are involved. A structure having n functional groups and r rings may contain up to $(r + 1)n(n - 1)/2$ pairwise relationships between groups, considering multiple paths between groups. Each of these relationships may then form a basis of other relationships. Furthermore, the quantity of information necessary to define a single relationship between units is usually greater than that necessary to define the units themselves. Therefore, only during the process of choosing chemical manipulations[1] is a relationship between functional groups defined, by requesting that the subroutine GTPATH() find a connecting path between the two groups. When two new groups are supplied to this subroutine, an "origin" atom from each group is chosen, defining the starting and end points of the path, and a maximum path length determined (relationships between groups connected by more than six atoms are not synthetically useful). Then, using the same subroutines as are used in detecting rings, a path is randomly traced through the structure from one of the origin atoms until either the origin atom is encountered, the maximum path length is reached, or the end of a chain is reached. In the latter two cases, the path is shortened to the last choice point and the search for the other origin atom continued. When it is found, the atoms and bonds in the completed path relationship are listed and returned to the manipulation selecting program.

Subsequent calls to GTPATH() yield alternate path relationships between the two groups, which will exist whenever the target molecule contains a ring.

A related perceptual problem occurring during the evaluation of possible chemical manipulations is that of finding the collection of all atoms or bonds at a small specified atomic separation from an initial atom or bond. This problem can be solved efficiently using set operations to move stepwise away from the starting atom or bond. In each step, the set of atoms or bonds attached to any member of the present set of atoms or bonds is generated. Then any atoms or bonds in the generated set that were present in the preceding set of

atoms and bonds are removed. This procedure is repeated the specified number of times.

Perception of Strategic Bond Disconnections. The synthesis of polycyclic bridged or fused ring systems is generally a more difficult matter than the synthesis of molecules of similar size and composition which possess no rings or at most only a few rings. The effective analysis of such problems requires a careful study of certain topological aspects of structure related both to the system under scrutiny and the methods available to chemists for the elaboration of cyclic systems. One objective of such topological considerations surely should be the discovery of those sequences of bond disconnective operations which lead in the fewest steps to the generation of acyclic or simple cyclic structures, starting from a target molecule. Those bond disconnections which are especially effective in achieving this topological simplification are designated herein as strategic bond disconnections, with the bonds involved termed "strategic bonds." In the present version of LHASA provision has been made to allow the chemist who inputs a structure to designate one or more bonds in that structure as strategic. This is done during graphical input of the structure as described in the preceding paper. The program then gives highest priority to those disconnective processes which lead to synthetic precursors formed by strategic bond disconnection. This highly useful and instructive feature takes advantage of the interactiveness of the program and at the same time increases the benefits which derive from that interactiveness.

However, a capability for automatic designation of strategic bond disconnections by the program is also highly desirable. Consequently, a number of provisions have been made in LHASA which constitute the first step in this direction. The following types of disconnections are preferred by the program: (1) endocyclic bonds to a bridgehead or fusion atom or, in other words, endocyclic bonds which are exo to another ring; (2) bonds between a ring and an appendage; (3) bonds which are in the "center" of the structure. A more extensive program for perception of strategic bond disconnections is planned for the next version of the program which will include indispensable information on stereochemistry (not available in the current program) and data tables with regard to the "chemical availability" of certain fundamental ring systems.

General View of Levels of Perception. It follows directly from the preceding discussion that in computerized as well as in human problem solving, the perception process must be carried out on a large number of levels, from primitive to highly complex and sophisticated. In going up this scale of levels, the units to be perceived become more complex, they are grouped into more complex collections, and the interrelationships between the collections themselves become more complex. A point is reached at which sufficient information has been collected by perception to allow the problem solving operation to efficiently employ the available "data" or "strategy" files. Further perception before access is made to these files would be inefficient to the extent to which the information to be perceived is not required for such access. In effect, at some point the perception process requires direction partly because the sheer mass of available information necessitates selectivity and because the most critical information is of a "specialized" type which is especially relevant to the specific problem. The effective utilization of the information available in the problem-independent data or strategy files (made possible by the first round of perception) may in itself require additional perception of information from the specific problem, and thus a second round of perception results. In the second round the perception process is driven by the data or strategy files and is highly selective. With regard to the operation of LHASA, a third stage of perception is operable because of the interactive nature of the program. In this stage the chemist assists the machine analysis by directing the processing toward new intermediates which appear more promising, or by selecting new strategies to be applied to the problem. The types of perception which are required for the selection of higher level strategies are currently under study and will be treated in due course.

Acknowledgment. We are indebted to the National Institutes of Health for financial assistance to this project and to the Advanced Research Projects Agency for their support of the Harvard Center for Research in Computer Sciences.