

A Structural Molecular Formula for Flexible and Efficient Substructure Searching of Large Databases

R. GEOFF. DROMEY*

Research School of Chemistry, Australian National University, P.O. Box 4, Canberra. A.C.T. 2600, Australia

Received September 21, 1977

The molecular formula provides a simple description of the nature of molecular structure at the atom composition level. The structural molecular formula introduced gives a description of molecules at the much more explicit atom structural identity level. This latter level of description is ideally suited for both substructure and complete structure screening. It also provides a simple yet powerful mechanism for establishing the upper bound of the maximal substructural commonality of a series of compounds. Unlike other fragment codes the structural formalism used requires a basis set of only 26 descriptors. The simple formalism makes it easy for a user to phrase powerful structural queries without needing to learn about a complex set of descriptors.

1. INTRODUCTION

The molecular formula is a primitive fragment code for indexing molecular topology. As such it is of limited use for structural and substructure screening of large databases because it contains virtually no information about the implicit structural relationships among atoms. In an effort to develop more suitable and effective screening systems, a number of different fragmentation codes have been introduced.¹⁻⁵ Inherent structural variability forces the subset of these codes that aim at generality of application to become rather unwieldy and involve a large number of descriptors.

One way out of this predicament is to adopt the simpler approach of working with an atom's structural identity. It is then found that the number of descriptor types can be made small (in this case 26) without conceding significant losses in differential descriptive power (structural resolution, cf. accurate mass resolution). In fact there are a number of other important advantages that accompany the proposed simpler approach. The formalism can be incorporated into a "pseudo-molecular formula" system to provide a powerful "complete-structure" screen. Alternatively the structural molecular formula can be readily adapted to a special type of inverted file suitable for interactive substructure searching. It can also be used very effectively to find the maximal substructural commonality for a series of compounds, a problem that is computationally prohibitive in many instances. Finally it would seem that the structural molecular formula represents a logical extension of the molecular weight and molecular formula descriptions of molecular structure. The molecular weight is a description at the atom mass level, the molecular composition is a more refined description at the atom composition level, and the structural molecular formula is a further refinement of molecular description at the atom structural identity level.

2. ATOM STRUCTURAL IDENTITY AND STRUCTURAL MOLECULAR FORMULAS

Before proceeding with the discussion on structural molecular formulas it is necessary to give an operational definition of an atom's structural identity. *The structural identity of an atom defines its structural environment with respect to its location in a ring or chain.* Table I lists the 26 structural identities and bond types (the alphabetic character set) that are employed. This set of descriptors is adequate for differentiating among the most common atom environments. Atoms in, and directly connected to, rings of various sizes are distinguished. Ring positions of structural significance in-

Table I

descriptor	structural identity
A	aromatic ring atom
B	ring atom at end of bridge
C	substituent atom directly connected to 3-membered ring
D	substituent atom directly connected to 4-membered ring
E	terminal atom—at end of chain
F	fused ring atom—only for two rings
G	substituent atom directly connected to 5-membered ring
H	substituent atom directly connected to ring of more than 6 atoms
I	atoms in a 3-membered ring
J	atoms in a 4-membered ring
K	one of a pair of substituent atoms attached to same ring atom
L	atoms in a 5-membered ring
M	atoms in a 7-membered ring
N	atoms in a ring of size greater than 7
O	coordination bond
P	peri-fused ring atom—involving 3 rings
Q	spiro ring atom—connected to 4 other ring atoms
R	six-membered carbocyclic ring atom
S	substituent atom directly connected to an aromatic ring
T	substituent atom directly connected to 6-membered carbocyclic ring
U	double bond (not in aromatic ring)
V	substituent atom attached to a fused atom
W	triple bond
X	chain atom with 4 nonhydrogens attached
Y	chain atom with 3 nonhydrogens attached
Z	atom in an aliphatic chain (nonterminal)

volving fused, perifused, bridging, and spiro atoms are identified along with terminal atoms and chain environments. Where possible the alphabetic structural identifiers have been chosen to correspond with the first character of simple structural mnemonics or pseudo-structural template characters (e.g., A for an aromatic ring atom, X for a tertiary connected atom).

A structural molecular formula is derived by identifying and making a count of the structural identities of all atoms in the molecule. The encoding is complete and nonoverlapping since all atoms are structurally identified and assigned. The procedures for identifying atoms are self evident in almost all cases. Several assignment rules and definitions given below resolve possible conflicts.

Structural Descriptor Format. The basic format for a structural descriptor is

<Structural Identity><Atom Symbol(s)><Atom Count>

For example, AC₆ indicates that there are six aromatic carbons

* Address correspondence to Department of Computing Science, University of Wollongong, Wollongong, N.S.W. 2500, Australia.

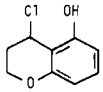
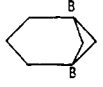
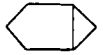
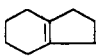
(a) $\text{CH}_3\text{CH}_2\text{CH}_2\text{OH}$	$\text{EC}_1\text{EH}_4\text{EO}_1\text{ZC}_2\text{ZH}_4$
(b) $\text{CH}_3\text{-O-CH}_2\text{CH}_3$	$\text{EC}_2\text{EH}_5\text{ZO}_1\text{ZC}_1\text{ZH}_2$
(c) $\text{CH}_3\text{CH}_2\text{CHO}$	$\text{EC}_1\text{EH}_3\text{ZC}_2\text{ZH}_3\text{EO}_1\text{EU}_1$
(d) $\text{CH}_3\text{-C(=O)-CH}_3$	$\text{EC}_2\text{EH}_6\text{YC}_1\text{EO}_1\text{EU}_1$
(e) 	$\text{AC}_4\text{AH}_3\text{FC}_2\text{RC}_3\text{RO}_1\text{TC}_1\text{RH}_5\text{SO}_1\text{SH}_1$
(f) 	$\text{RC}_3\text{RH}_6\text{BC}_2\text{BH}_2\text{JC}_2\text{JH}_4$
(g) 	$\text{IC}_1\text{IH}_2\text{FC}_2\text{FH}_2\text{LC}_3\text{LH}_6$
(h) 	$\text{LC}_3\text{LH}_6\text{FC}_2\text{FU}_1\text{RC}_4\text{RH}_8$

Figure 1. Structural molecular formulas for some simple compounds.

C_6H_8 Chain Molecules		
$\text{CH}_3\text{CH}_2\text{-C}\equiv\text{C-CH=CH}_2$	$\text{EC}_2\text{EH}_5\text{ZC}_4\text{ZH}_3\text{EU}_1\text{ZW}_1$	2174
$\text{CH}_3\text{CH}_2\text{-CH=CH-C}\equiv\text{CH}$	$\text{EC}_2\text{EH}_4\text{ZC}_4\text{ZH}_4\text{ZU}_1\text{EW}_1$	2153
$\text{CH}_3\text{CH}_2\text{C(=CH}_2\text{)-C}\equiv\text{CH}$	$\text{EC}_3\text{EH}_6\text{ZC}_2\text{ZH}_2\text{YC}_1\text{EU}_1\text{EW}_1$	1406
$\text{CH}_3\text{-CH=CH-C}\equiv\text{C-CH}_3$	$\text{EC}_2\text{EH}_4\text{ZC}_4\text{ZH}_2\text{ZU}_1\text{ZW}_1$	2594
$\text{CH}_3\text{-CH=CH-CH}_2\text{-C}\equiv\text{CH}$	$\text{EC}_2\text{EH}_4\text{ZC}_4\text{ZH}_3\text{ZU}_1\text{EW}_1$	2153
$\text{CH}_3\text{-C}\equiv\text{C-CH}_2\text{-CH}_3$	$\text{EC}_3\text{EH}_6\text{ZC}_2\text{YC}_1\text{EU}_1\text{ZW}_1$	1847
$\text{CH}_3\text{-C(=CH}_2\text{)-CH}_2\text{-C}\equiv\text{CH}$	$\text{EC}_3\text{EH}_6\text{ZC}_2\text{ZH}_2\text{EU}_1\text{EW}_1\text{YC}_1$	1406
$\text{CH}_2\text{=CH-CH}_2\text{-C}\equiv\text{C-CH}_3$	$\text{EC}_2\text{EH}_5\text{ZC}_4\text{ZH}_3\text{EU}_1\text{ZW}_1$	2174
$\text{CH}_2\text{=CH-CH}_2\text{-CH}_2\text{-C}\equiv\text{CH}$	$\text{EC}_2\text{EH}_5\text{ZC}_4\text{ZH}_4\text{EU}_1\text{EW}_1$	1733
$\text{CH}\equiv\text{C-CH=C(CH}_3\text{)-CH}_3$	$\text{EC}_3\text{EH}_7\text{ZC}_2\text{ZH}_1\text{YC}_1\text{EW}_1\text{ZU}_1$	1826
$\text{CH}\equiv\text{C-C(CH}_3\text{)=CH-CH}_3$	$\text{EC}_3\text{EH}_7\text{ZC}_2\text{ZH}_1\text{YC}_1\text{EW}_1\text{ZU}_1$	1826
$\text{CH}_2\text{=CH-CH(CH}_3\text{)-C}\equiv\text{CH}$	$\text{EC}_3\text{EH}_5\text{ZC}_2\text{ZH}_1\text{YH}_1\text{YC}_1\text{EU}_1\text{EW}_1$	1405
$\text{CH}_2\text{=CH-CH=CH-CH=CH}_2$	$\text{EC}_2\text{EH}_4\text{ZC}_4\text{ZH}_4\text{EU}_2\text{ZU}_1$	2245
$\text{CH}_2\text{=C(CH=CH}_2\text{)-CH=CH}_2$	$\text{EC}_3\text{EH}_6\text{ZC}_2\text{ZH}_2\text{YC}_1\text{EU}_3$	1501

Figure 2. Structural molecular formulas for some C_6H_8 acyclic isomers.

in a molecule while FC_2 identifies the presence of two fused carbons.

Terminal Atom. A terminal atom is an atom connected to only one other nonhydrogen atom.

Bridge Substructure. A bridge substructure is defined when a pair of rings of smallest size share more than one atom-pair bond (a detailed formalism for interpreting bridging systems is given elsewhere⁷).

ENCODING RULES

(a) Identification of an atom as a terminal atom takes precedence over chain atom assignment.

(b) An atom (or bond) is encoded as a ring "substituent" in preference to being encoded as a terminal atom or bond (Figure 1e).

(c) Hydrogen atoms take the structural identity of the type















C_6H_8 Ring Structures		
	$\text{IC}_2\text{IH}_2\text{IU}_1\text{QC}_1\text{JC}_3\text{JH}_6$	1047
	$\text{JC}_4\text{JH}_6\text{DU}_1\text{DC}_1\text{EC}_1\text{EH}_2\text{EU}_1$	847
	$\text{IC}_3\text{IH}_6\text{QC}_1\text{FC}_2\text{FH}_2$	738
	$\text{LC}_5\text{LH}_6\text{LU}_1\text{GU}_1\text{GC}_1\text{GH}_2$	1289
	$\text{IC}_1\text{FC}_2\text{FH}_2\text{JC}_2\text{JH}_4\text{CU}_1\text{CC}_1\text{CH}_2$	856
	$\text{RC}_6\text{RH}_8\text{RU}_2$	2196
	$\text{RC}_6\text{RH}_8\text{RU}_2$	2196
	$\text{JC}_2\text{JH}_4\text{BC}_2\text{BH}_2\text{LC}_2\text{LH}_2\text{LU}_1$	896
	$\text{IC}_1\text{IH}_2\text{FC}_2\text{FH}_2\text{LC}_3\text{LH}_4\text{LU}_1$	1014
	$\text{IC}_2\text{IH}_4\text{FC}_4\text{FH}_4$	564
	$\text{IC}_6\text{IH}_8\text{WU}_1$	1203
	$\text{IC}_4\text{QC}_1\text{IH}_6\text{CU}_1\text{CC}_1\text{CH}_2$	1002
	$\text{JC}_4\text{JH}_4\text{DU}_2\text{DC}_2\text{DH}_4$	800
	$\text{JC}_4\text{JH}_4\text{DU}_2\text{DC}_2\text{DH}_4$	800

Figure 3. Structural molecular formulas for some C_6H_8 cyclic isomers.

of atoms to which they are attached, (e.g., a hydrogen atom attached to a terminal atom is given a terminal atom (E) descriptor).

(d) Atoms at the ends of bridges are given a bridge structural identity (Figure 1f).

(e) In bridge systems the ring type identity of an atom is established by assigning it to the ring of *smallest* size of which it is a member (Figure 1f).

(f) Multiple bonds are included in the structural molecular formula. They take their structural identity descriptions from the atoms to which they are attached according to rules a-e. Where uncertainties are not resolved by the above rules, assignment is made to the identity occurring *earliest* in the alphabet.

Something of an assessment of the power and limitations of the present system for structural differentiation can be obtained from examination of Figures 1, 2, and 3. In Figure 1e it can be seen that the ring identity of the heteroatom is defined. The type of ring to which each of the substituents is attached is also assigned. At this level of description the relative positions of heteroatoms and substituents are not defined. In a later section it will be shown how a simple extension of the formalism can be used to take these characteristics into account. If necessary, structural molecular formulas can be written more concisely by grouping together different atoms of the same structural identity (e.g., the formula in Figure 1a could be written as $\text{EC}_1\text{H}_4\text{O}_1\text{ZC}_2\text{H}_4$). Ordering subelements of structural formulas first by the alphabetic value of structural descriptors and subordinately by the alphabetic value of element descriptors can provide additional standardization which is useful in some applications.

The imprecision at the level of description given is not altogether a negative aspect in that it allows flexibility in structure searching and guarantees that closely similar



Molecular Formula = C_6H_6

Integer representation = 80056

Structural Molecular formula = $IC_2IH_2IU_1QC_1JC_3JH_6$

Integer structural identity = 1047

∴ Structural Molecular formula integer representation = 800561047

Figure 4. Derivation of integer representation of a structural molecular formula.

compounds will be retrieved without formulating complicated search queries. In many instances a search at this level is essential.

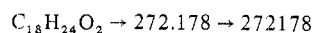
The results of the DENDRAL group in their exploration of the structural spaces of molecular compositions give practical confirmation of the fact that dramatic screening can be achieved with relatively small amounts of structural information (e.g., NMR data relating to the number of terminal methyl groups reduced the number of plausible structures for what was identified as *N,N*-dimethyl-*n*-octadecylamine from 1 284 792 to just one structure;⁸ obviously in many cases the pruning would not be this dramatic). These results together with the fact that the structural populations of most molecular compositions are sparse even in large databases suggest that the structural molecular formula can act as a powerful screening tool.

3. COMPLETE STRUCTURE SEARCHING OF A LARGE DATABASE USING THE STRUCTURAL MOLECULAR FORMULA

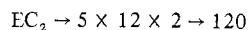
It has been recently suggested that standardized accurate molecular masses derived from molecular compositions provide a suitable basis for designing a molecular formula retrieval system.⁹ A small extension of this representation can lead to the construction of a compact and efficient structural molecular formula retrieval system.

The structural molecular formula contains two types of information: the molecular composition and the composite atomic structural identities. With respect to structural resolution the latter represents a considerable refinement.

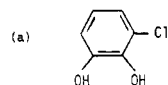
A six- or seven-digit representation of molecular formulas can be obtained using accurate atomic weight information. Such a representation provides a compact and yet almost unique numeric mapping for molecular compositions. A corresponding integer representation is obtained by making the appropriate magnitude change, e.g.,



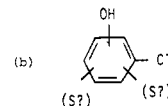
The structural resolution of integer molecular formulas (e.g., 272 178) can be extended an *extra* four digits with an integer representation that characterizes the structural identities conveyed in the structural molecular formula. The latter representation is derived by assigning the structural identity characters the numeral corresponding to their position in the alphabet (e.g., A = 1, E = 5, etc.) and taking the product of this with the rounded atomic weight and number of occurrences, e.g.,



The derivation for a C_6H_8 isomer is given in Figure 4. If the four most significant digits of the structural integer component (e.g., 1047 in Figure 4) are placed after the least significant digit of the integer molecular formula, an integer representation that can be represented by 36 bits is obtained. This new hierarchical molecular-formula-structural-identity nu-

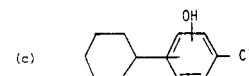


$AC_6AH_3SO_2SH_2SCl_1$

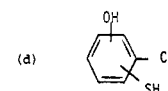


$AC_6AH_2SO_1SCl_1SH_1$

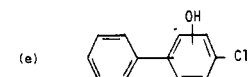
$AC_6.AND.AH_2.AND...$



(to be excluded) NOT RC_1



(to be excluded) NOT SS_1



(to be excluded) NOT AC_7

Figure 5. Boolean strategy for substructure searching with the structural molecular formula.

meric code can form the basis of a "complete-structure" screening system. Complete structure representations (molecular compositions plus structural identity) would require approximately 36 bits which is less than or equivalent to one computer word size on many large computer systems.

For very large files of structures the data could be most efficiently searched by either a binary search technique or a hashing technique.¹⁰ The performance of both the search systems suggested on files like that of structural molecular formulas is well established and clearly defined. The former file organization would demand that the file be ordered on the magnitude of the numeric code. Such an organization would enable the file to be indexed on molecular weight to localize the binary search.

The fact that the molecular composition and structural identity index are integrated into a single representation in a hierarchical fashion ensures that retrieval from the file proceeds in a very efficient manner. *That is, the search can take full advantage of the screening capability of the molecular composition before incorporating the structural screen.* Although it does not show particularly well in the restricted set of examples given, it is found that in general ring compounds have lower indices than acyclic structures with the same composition. Within a set of ring compounds those containing aromatic rings will have the lowest indices. In acyclic systems with a given composition the most highly branched structures will have the lowest indices.

If necessary the resolution of the structural formula could be extended even further at very small storage and almost no processing cost by incorporating the interaction formalism described in section 6.

4. AN INVERTED FILE STRUCTURE FOR SUBSTRUCTURE SCREENING WITH STRUCTURAL MOLECULAR FORMULAS

The construction of a mechanism for flexible substructure searching of a large database represents a far more challenging task than the complete structure problem. If the information in structural molecular formulas is encoded in an inverted bit

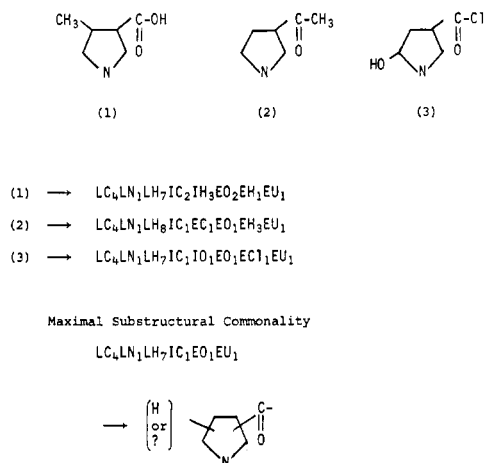
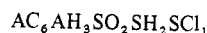


Figure 6. Mechanism for establishing the maximal substructural commonality for a series of compounds using structural molecular formulas.

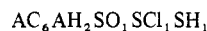
map in a special format to be described, it is possible to come close to the desired goal at least to the level of resolution of the structural molecular formula. The method employed can provide a sufficient reduction of the structure space to allow other more detailed substructure searching techniques to become practical. The only information needed in the proposed inverted file is that derived from structural molecular formulas. The inverted file must consist of a set of key descriptors that represent atom structural identities for a range of elements and for a range of occurrences (e.g., AC₁, AC₂, AC₃, . . . and RN₁, RN₂, RN₃, . . .).

Consider how the structure in Figure 5a with the structural formula



would be represented in the inverted file to facilitate substructure searching. There are six aromatic carbons (AC₆) and so bits corresponding to AC₁, AC₂, AC₃, AC₄, AC₅, and AC₆ must be set for the appropriate record number in the inverted file. Similarly there are two oxygens substituted on an aromatic ring and thus bits for SO₁ and SO₂ must be set and so on until *all* atoms in the molecule have been encoded.

As an example of a substructure search suppose it was necessary to find all chlorophenols that possessed no more than four substituents (Figure 5b); then the relevant query would be



If it were necessary to exclude all chlorophenols that included an additional six-membered, nonaromatic ring (Figure 5c), then a Boolean NOT with the descriptor RC₁ would accomplish the task. If it were also required that none of the chlorophenols should contain a substituted sulfur, then the descriptor SS₁ would be used with a Boolean NOT. The possibility of obtaining chlorophenols with a second aromatic ring could be eliminated by doing a NOT with AC₇. As this small scenario suggests, the inverted structural formula file can be used very easily, *but with care*, to do powerful constrained substructure searches. *What this method of screening does, in effect, is to establish a lower (existence) bound for substructural commonality.*

Construction of the search profiles is straightforward and does not involve a commitment to learning a complicated descriptor formalism. Substructure screening can be enhanced even further by implementing the structural interaction formalism described in section 6. With the latter level of description, it would be possible to restrict the search to o-chlorophenols if that level of precision were necessary.

5. ON FINDING THE MAXIMAL SUBSTRUCTURAL COMMONALITY FOR A SERIES OF COMPOUNDS

Many instances arise where it is desirable to find out the maximal structural commonality for a series of compounds. For complex structures this can prove to be a difficult and very time-consuming procedure when approached directly by graph-theoretical techniques.^{11,12} However, when the problem is approached first at the structural molecular formula level, it reduces to a much simpler and relatively efficient procedure. *The structural molecular formula provides a filter that establishes the upper bound for the substructure shared by a group of compounds.* In so doing it can reduce the substructure space to a much more tractable size. It may then be possible to employ advantageously more explicit and time-consuming techniques.^{11,12} An example of substructural commonality identification is given in Figure 6. In comparison with a graph-theoretical search of the structures, an analysis of a set of structural molecular formulas for a maximum bound is a trivial task.

The structural molecular formulas can also be used usefully to ask important questions of a database, like "which compound on a structural atom basis best matches a query compound?"

6. AN INTERACTION FORMALISM FOR ENHANCING THE POWER OF THE STRUCTURAL MOLECULAR FORMULA

In the discussion thus far, consideration of interatom relationships has been explicitly avoided because of the complexities that it can introduce. One way to simplify this problem to a large degree is to consider interatom interaction at the rather general level of structural identity. The resulting formalism, when taken in combination with the structural formula, can provide comprehensive structural descriptions.

There are two basic atom structural relationships that need to be considered, the interaction between atoms in the *same* structural environment (e.g., the relationship of a heteroatom to a substituent in a ring, or the relationship of two substituents in a ring, etc.) and the interaction between atoms in *different* structural environments (e.g., the relationship of a heteroatom in one ring to a substituent in another ring). The former are considered as *internal* relations while the latter are classified as structurally *external* interactions.

6.1 Internal Relationships between Atom Structural Identities.

Internal relations have the following general four component format

$$\left\{ \begin{array}{l} \text{ring or} \\ \text{chain} \\ \text{identifier} \end{array} \right\} \left\{ \begin{array}{l} \text{structural identity} \\ \text{of} \\ \text{first component} \end{array} \right\} \left\{ \begin{array}{l} \text{structural identity} \\ \text{of} \\ \text{second component} \end{array} \right\}$$

$$\left\{ \begin{array}{l} \text{numeric atom} \\ \text{displacement between} \\ \text{the two identities} \end{array} \right\}$$

The numeric atom displacement is defined as the number of atoms on the *shortest path* joining the two entities as measured from the start of the first entity to the start of the second entity. A summary of important internal relationships for rings and chains is given in Table II. The ring identifier for a five-membered ring is L, and so on. The character C is used as the chain identifier. Structure type identifiers are expressed in alphabetic order (e.g., HS rather than SH). Some examples are given in Figure 7 to illustrate how the relational concept is used.

The relational term LHH₃ in Figure 7a specifies that there are two heteroatoms in a five-membered ring that are on a three-atom chain. The term AHS₃ in Figure 7b indicates that there is a heteroatom in an aromatic ring separated by a path of length 3 from a substituent. CES₂ means that the end of

Table II. Internal Structural Identity Interactions

interaction	base descriptors
substituent-substituent	SS, TT, CC, DD, GG
heteroatom-heteroatom ^a	HH
fused atom-fused atom	FF
heteroatom-substituent	HS, HT, HC, HD, HG
fused atom-substituent	CF, DF, FG, FS, FT
fused atom-heteroatom	FH
double bond-double bond	UU

^a Note with a small change in the formalism it would be possible to represent heteroatoms at their atom identity level of description rather than the general H level.

a chain attached to an aromatic ring is on a path of length 2 from the atom directly attached to the ring. LAF₃ in Figure 7g indicates that on a five-membered ring there is a directly attached aromatic ring that is in position 3 relative to a fused atom.

The extension of the formalism to cover all types of internal relationships (e.g., FQ, BT, EX, etc.) is straightforward. The combination of internal structural identity interactions with the structural molecular formula description provides a sufficiently well-resolved structural description for most screening purposes. The duplicates in Figures 2 and 3 are resolved when internal interactions are taken into account. The level of description can be refined a stage further by taking into account the external relationships between atom structural identities.

6.2 External Relationships between Atom Structural Identities. The extension of the internal relationship formalism to the external case is straightforward. The only difference is that just one external identifier X is used. The generalized format is:

$$\left\{ \begin{array}{l} \text{external} \\ \text{identity} \\ X \end{array} \right\} \left\{ \begin{array}{l} \text{structural identity} \\ \text{of} \\ \text{first component} \end{array} \right\} \left\{ \begin{array}{l} \text{structural identity} \\ \text{of} \\ \text{second component} \end{array} \right\}$$

$$\left\{ \begin{array}{l} \text{numeric atom} \\ \text{displacement between} \\ \text{the two atom identities} \end{array} \right\}$$

Several examples are given in Figure 8. The term XST₄ in Figure 8b indicates that a substituent on an aromatic ring is four atoms away from a substituent on a six-membered carbocyclic ring.

The extension of the formalism to all other pairs of interactions is self evident. The level of implementation for a given database is necessarily determined by storage, cost limitations, and search requirements. This need not prevent the user from taking advantage of the flexibility of the system that incorporates all types of internal and external interactions. A table can be kept of all interaction descriptors that have been implemented. A query interpreter can then check the descriptor table and inform the user if a particular descriptor has not been implemented.

The same technique can be employed for identity descriptors. For example, suppose a query involved the search for a substructure containing AC₂₅ (25 aromatic carbons) and the actual database considered aromatic descriptors up to AC₂₀. The command interpreter could implement this search as AC₂₀ and inform the user. In this situation very little discriminating power would be lost since AC₂₀ is most probably almost as effective as AC₂₅ in eliminating unwanted structures.

CONCLUSIONS

A structural molecular formula formalism has been introduced as an alternative screening method to existing systems. The formalism provides a comprehensive and logical framework in which to address questions of molecular structure. It has been demonstrated that the formalism has

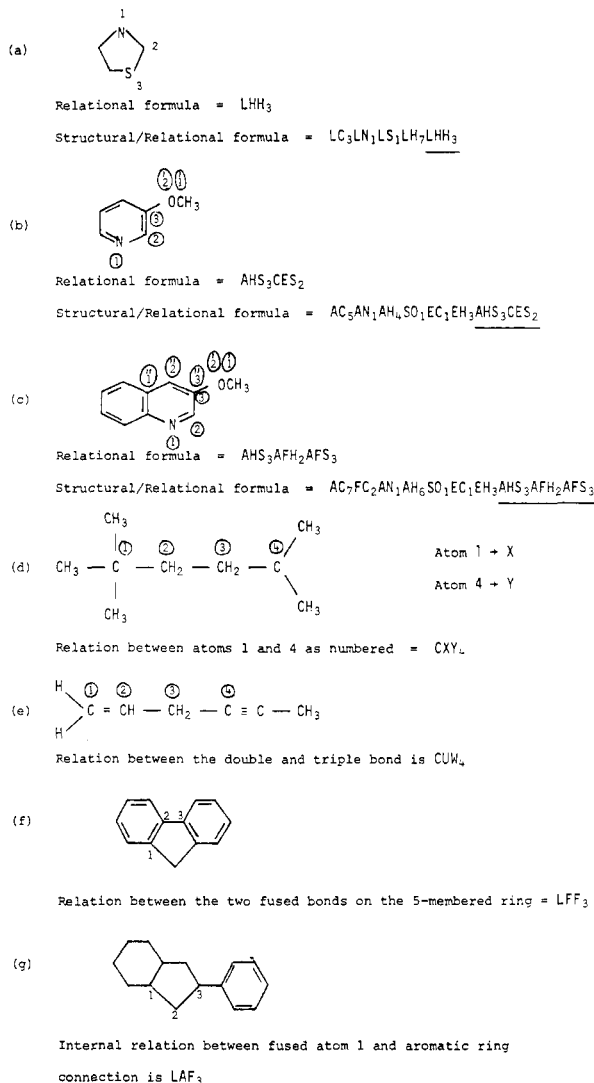


Figure 7. Examples of the extended internal interaction formalism.

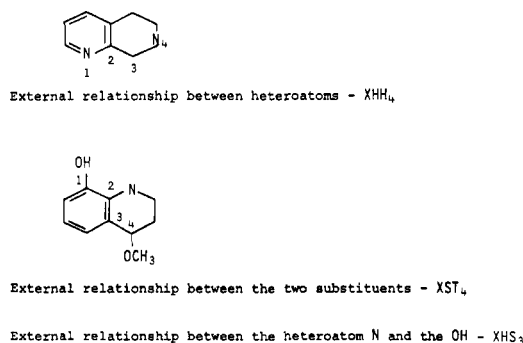


Figure 8. Examples of the extended external interaction formalism.

sufficient resolving power to be used for both substructure and complete structure searching. An important attribute of the system is that it is very simple to use because the description has been taken at the atom structural identity level. Although it has not been emphasized here, it is clear that connection table representation at the atom structural identity level could lead to much more efficient substructure searching of such tables. A somewhat related approach has been adopted in the CROSSBOW connection table system.⁶ CROSSBOW, however, works at a fragment level description. Automatic generation of structural molecular formulas from connection tables and linear notations is certainly a realizable task. It would provide a mechanism for incorporating structural

molecular formulas into existing systems without large manpower commitments. A subset of the techniques used for the much harder problems of structure drawing,¹³ and the generation of WLN formulas from connection tables,¹⁴ should meet the demands of structural molecular formula generation.

ACKNOWLEDGMENT

The author wishes to thank Professor F. W. McLafferty for stimulating his interest in the maximal substructural commonality problem. The author thanks Patrick Keogh for timing tests on the UNIVAC 1108. Thanks are also due to Carol Jacobs, Lorraine Scarr, and Glenda Gregor for their help in preparing the manuscript.

REFERENCES AND NOTES

- (1) P. N. Craig and H. M. Ebert, "Eleven Years of Structure Searching Using the SK&F Fragmentation Codes", *J. Chem. Doc.*, **9**, 141 (1969).
- (2) E. Meyer, "Superimposed Screens for the GREMAS System", Proceedings of the FID-IFIP Conference, Rome 1967, K. Samuelson, Ed., North-Holland, Amsterdam, 1968, p 280.
- (3) G. W. Adamson, J. Cowell, M. F. Lynch, A. H. McLure, W. G. Town, and A. M. Yapp, "Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files", *J. Chem. Doc.*, **13**, 153 (1973).
- (4) M. Milne, D. Lefkowitz, H. Hill, and R. Powers, "Search of CA Registry (1.25 Million Compounds) with the Topological Screens System", *J. Chem. Doc.*, **12**, 183 (1972).
- (5) C. E. Granito, G. T. Becker, S. Roberts, W. J. Wiswesser, and K. J. Windlinx, "Computer-Generated Substructure Codes (Bit Screens)", *J. Chem. Doc.*, **11**, 106 (1971).
- (6) E. Hyde in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke et al., Ed., Wiley, New York, N.Y., 1974.
- (7) R. G. Dromey, "A Simple Tree-Structured Line Formula Notation for Representing Molecular Topology", *J. Chem. Inf. Comput. Sci.*, submitted for publication.
- (8) B. G. Buchanan and J. Lederberg, "The Heuristic Dendral Program for Explaining Empirical Data", Stanford University Computer Science Dept., Report No. CS-203, 1971.
- (9) R. G. Dromey, "A Highly Compressed Inverted File for Molecular Formula and Homologous Series Searching of Large Data Bases", *Anal. Chem.*, **49**, 1982 (1977).
- (10) N. Wirth, "Algorithms + Data Structures = Programs", Prentice-Hall, Englewood Cliffs, N.J., 1976.
- (11) F. W. McLafferty, private communication.
- (12) D. G. Corneil and C. C. Gotlieb, "An Efficient Algorithm for Graph Isomorphism", *J. Assoc. Comput. Mach.*, **17**, 51 (1970).
- (13) E. J. Corey and W. T. Wipke, "Computer Assisted Design of Complex Organic Syntheses", *Science*, **166** 178-192 (Oct 10, 1969).
- (14) G. A. Miller, "Encoding and Decoding WLN", *J. Chem. Doc.*, **12**, 60 (1972).

Computer-Assisted Simulation of Chemical Reaction Sequences. Applications to Problems of Structure Elucidation^{1,2}

TOMAS H. VARKONY, RAYMOND E. CARHART, DENNIS H. SMITH,* and CARL DJERASSI

Departments of Chemistry and Computer Science and Genetics, Stanford University,
Stanford, California 94305

Received February 14, 1978

An interactive computer program (REACT) for simulation of chemical reaction sequences and its application to problems of structure elucidation are described. The program is supplied with information about laboratory operations such as chemical reactions, separations, and data about structural features of the products. The program applies this information to structures in the computer memory and allows a chemist to examine the results by displaying the reaction/separation sequence or drawings of structures. A sequence demonstrating the use of REACT is illustrated for the structure elucidation of palustrol (1). A detailed description of some of the algorithms is presented.

Chemical reactions are fundamental tools in the study of molecular structure far beyond obvious applications to synthetic organic chemistry. Other areas of research where chemical reactions play a crucial role include mechanistic studies, e.g., cyclizations and rearrangements, and a wide variety of reactions applied to unknown structures during the course of structure elucidation, e.g., functionalization, derivatization and degradation, or simplification reactions. Modern instrumentation, including X-ray crystallography, mass spectrometry, and ¹³C NMR spectroscopy have dramatically altered the methodology of structural studies in both mechanistic organic chemistry and structure elucidation. There remain, however, many structural problems where knowledge gained from the results of applications of chemical reactions is crucial to determining mechanistic pathways or structural identity.

We have developed an interactive computer program, called REACT,^{3,4,6} to help a chemist to explore certain aspects of chemical reactions applied to representations of molecular structure. REACT was designed to provide a general tool for studying the results and implications of chemical reactions. In structure elucidation problems such results can reduce further the number of possible structures for an unknown, thus helping the chemist to focus attention on the correct structure. In mechanistic studies it is particularly useful to determine

exhaustively all possible interconversion pathways and intermediates, or to decipher all structural implications of measurements on products subsequent to an extended sequence of reactions.³ The reactions typically utilized in such problems are relatively general (i.e., they apply in a wide variety of structural contexts) and well understood. Although many are relatively simple reactions (hydrogenation, hydrolysis, Wagner-Meerwein-type rearrangements), REACT provides the capability for definition of reactions of considerable complexity which can be constrained in several ways to express details of the structural context in which they apply.

In addition we can use REACT for structure generation. For example, if an unknown was obtained from a known starting material through a set of well-defined reactions, REACT can generate the products obeying the constraints of the reactions. These products then are candidate structures for the unknown. Thus REACT provides a "computer laboratory" in which representations of chemical reactions can be applied to structures in ways which parallel actual experimental studies or plans.

The REACT program is a logical extension³ of the CONGEN program for computer-assisted structure elucidation.⁵ CONGEN, by generating all possible structures consistent with given physical or chemical data, provides candidate structures for an unknown, or potential precursors