An Application of Interactive Graphics— The Nested Retrieval of Chemical Structures

RICHARD J. FELDMANN* and STEPHEN R. HELLER
Computer Center Branch and Heuristics Laboratory, Division of Computer Research and Technology, National Institutes of Health, Department of Health, Education and Welfare, Bethesda, Md. 20014

Received November 2, 1971

A technique for structuring and searching a large file of chemical structures is presented. The technique involves generating a nested, structured tree based on the Wiswesser analysis of chemical structures. An important virtue of the tree is the rapid and inexpensive file updating. An example of this technique is shown for a 5001 compound subset of the *Chemical Abstracts* connection tables of the Common Data Base.

The availability of interactive graphics for the chemist to use has prompted a number of research projects in this facility.[1-5] Included in this work in the area of chemical retrieval has been a sequential sub-structure search (SSS)[5-8] and the rapid structure search (RSS).[1] The former is a time consuming and economically expensive computer program. Of course, it is noninteractive for all but small files—i.e., less than 3000 structures. In the DCRT SSS, the user draws on a Rand Tablet a chemical fragment or structural component which is automatically encoded and the file sequentially searched to see if the fragment is embedded in a complete structure that is in the file. If that is the case, the user inputs a complete chemical structure and, using the technique of hash coding, quickly tests the address of the structure in the file. The RSS permits the user to ask only one question of the file, but allows him to do so in a matter of seconds for a large file (say 1.7 million). Thus these two techniques represent the extremes of response *vs.* query breadth. The nested tree structure technique description which follows is an attempt to structure and search a large file of chemical structures. It allows the user to alter dynamically the response time *vs.* query breadth extremes in such a manner as to tailor the search to the individual user's own specific needs.

In analyzing a chemical structure into its various components for structuring the file, the concepts developed by Wiswesser and Smith for the Wiswesser Line Notation (WLN) have been used.[9] The WLN for the compound shown in Figure 1 is: T6 N COTJ B3 DR B2& E1.

The meaning or breakdown of the string is:

"T6" signifies a six member ring with a hetero-atom or stituent

"N" signifies a nitrogen stituent in the A position

"CO" signifies an oxygen stituent in the C position

"TJ" makes a statement about ring saturation

"B3" puts a three atom carbon chain in the B position of the ring

"DR" puts a benzene ring in the D position off the ring

"B2" puts a two carbon atom chain off the B position of the benzene ring

"&" signifies the end of the benzene substitution

"E1" puts a methyl group in the E position of the substituted ring

Thus, the Wiswesser analysis decomposes the structure into the parts:

1	Ring Nuclei
2 & 3	Hetero-atom or Stituent Pattern Sequence
4, 5 & 6	Substituent pattern, structure and sequence

The WLN has the effect of distributing the components of a chemical structure in a linear string. The linear string is an historic artifact of the computer technology some 20 years ago. The availability and ease of use of punched cards and line printers made this a virtual necessity. Today, with the advent of large computers with vast disk storage capabilities and techniques, the greatest asset of the WLN, namely, its decomposition analysis, has led to the concept of the nested, structured tree. The structure analysis follows the WLN quite closely, but, by changing to random access disk storage, it allows for a new use of this type of representation for chemical structures. Whereas WLN represents each chemical structure as an isolated entity (as does the Chemical Abstracts Service connection tables), the nested structured tree represents chemical structures as classes. In bringing together all structures of the same class, structural redundancies can be eliminated. Once the structural redundancy is eliminated, an algorithm can be provided for linking together related

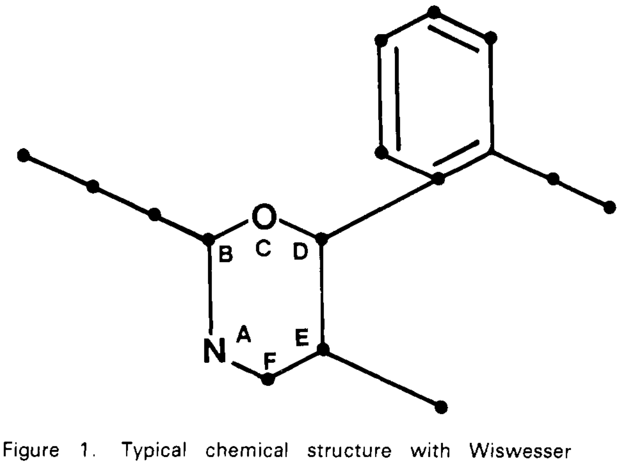


Figure 1. Typical chemical structure with Wiswesser locants of the central ring

*To whom correspondence should be addressed.

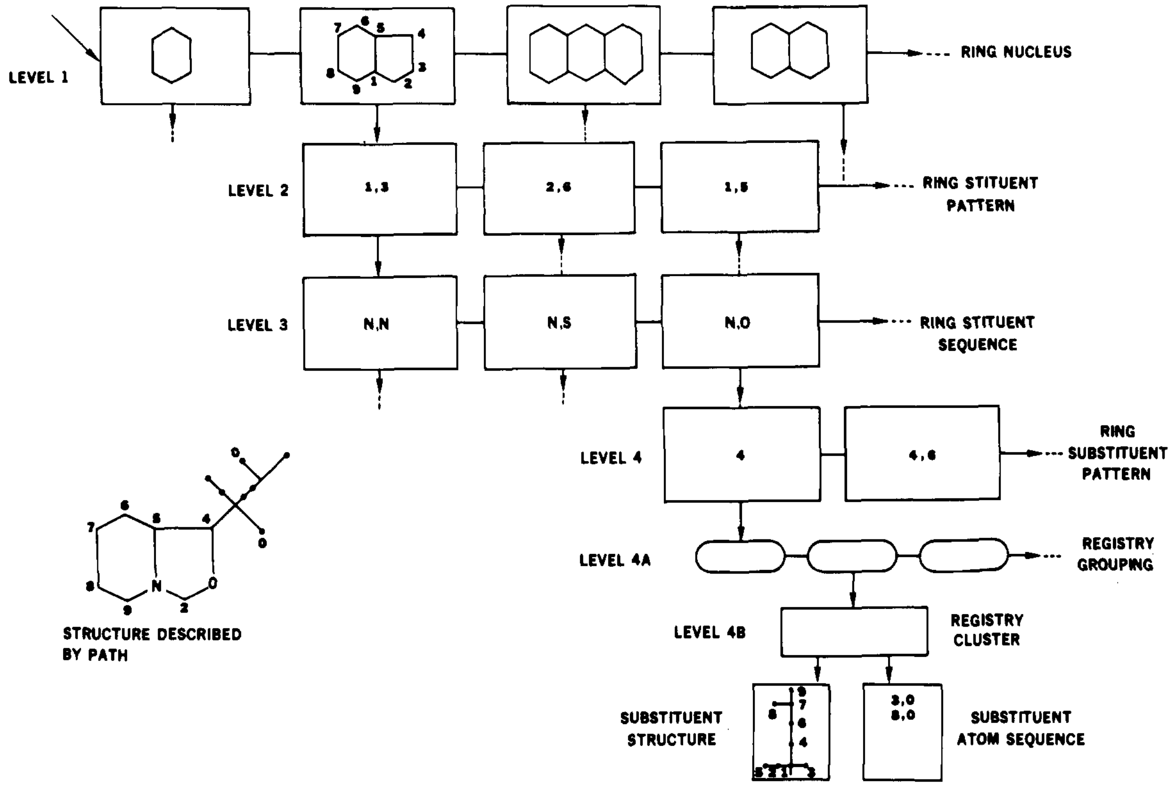


Figure 2. The nested, structured tree

classes. The result of this process is the nested structure shown in Figure 2.

DESCRIPTION OF THE TREE

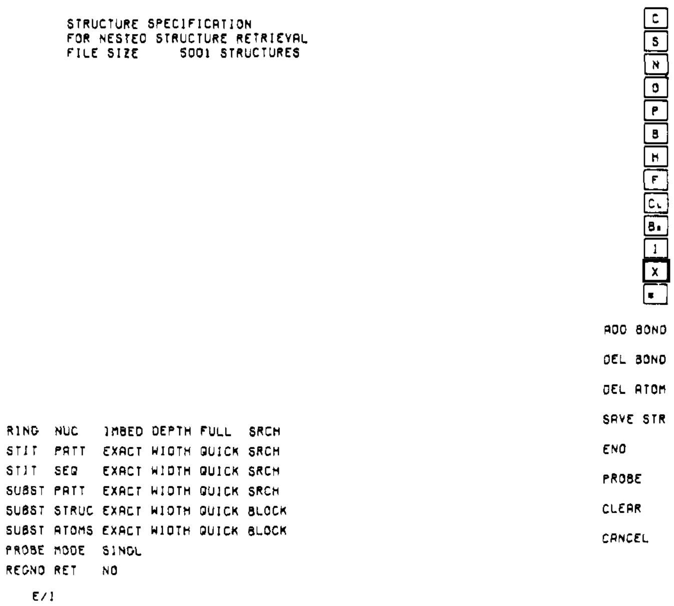
Each leaf or node of the tree in Figure 2 is represented by five computer words:

WORD	CONTENT
1	Pointer to antecedent level (father)
2	Pointer to consequent level (son)
3	Pointer to next node on this branch (brother)
4	Counter for structures in the tree below this node
5	Data

The topology of each chemical structure is stored once and only once. For example, the topology of the six-membered ring (which is estimated to occur perhaps 1.1 million times in the CAS structure file of 1.7 million) is represented only once. Even greater savings in topology occurs for the structure class of steroids. In all previous tape-based sequential SSS file and search systems, the topology is represented for each steroid structure. The nesting of chemical structures eliminates this topological redundancy. In so doing, structures of the same class are brought together. This is the heart of the structure tree, and the means by which one can devise algorithms to interactively search a large file of chemical structures.

GENERALIZED EXAMPLE

To understand better the structural tree, it is worthwhile to consider a hypothetical example. Figure 3 shows the display image the user sees when the program begins.



```
STRUCTURE SPECIFICATION
FOR NESTED STRUCTURE RETRIEVAL
FILE SIZE     5001 STRUCTURES
```

```
RING  NUC    IMBED DEPTH FULL  SRCH
STIT  PATT   EXACT WIDTH QUICK SRCH
STIT  SEQ    EXACT WIDTH QUICK SRCH
SUBST PATT   EXACT WIDTH QUICK SRCH
SUBST STRUC  EXACT WIDTH QUICK BLOCK
SUBST ATOMS  EXACT WIDTH QUICK BLOCK
PROBE MODE   SINGL
REGNO RET    NO
  E/I
```

Figure 3. The display image of the CRT that the user sees upon initialization of the Nest Program

DESCRIPTION OF KEYS:

Upper Right
  Atoms C, S, N, O, P, B, H, F, Cl, Br, I
  X Cause the 103 elements to appear on the screen, the user then chooses one atom.
  *Returns a structure node to the unspecified state.
Lower Right
  ADD BOND: Add a bond between the two atoms to be specified

DEL BOND: Delete a bond between the two bonds to be specified

DEL ATOM: Delete the atom to be specified on the Rand Tablet

SAVE STR: Put the structure now on the CRT screen out onto the disk, or if there is no structure on the CRT, bring in the one specified from the disk

END: (Exit from this program)

PROBE: (Search the nested structure tree)

CLEAR: (Re-initialize the program)

CANCEL: (Ignore the last action and restore structure to its last state)

Lower Left

At each level

EXACT or IMBED
WIDTH or DEPTH
QUICK or FULL
SEARCH or BLOCK

With the probe mode in single mode the user must depress the Rand Tablet pen over PROBE to search the structure tree, whereas in the AUTO mode each user action causes a search of the structure tree automatically

REGN Retrieval: YES or NO

The number of structures in the response from this example is only for illustrative purposes, a real (and more limited) example will be described later. There are six switches in the retrieval program, all of which are set to "EXACT" when the program is initialized. In the "EXACT" mode, the retrieval program looks for exactly the specified structure. The alternative mode is "IMBED," which means that the specified structural component can be imbedded in a larger chemical structure.

THE SCENERIO:

User: The user draws a six-membered ring

Program: There are 1.2 million structures with a six-membered ring

User: The user adds a six-membered ring as a fused ring

Program: That there are 105,000 structures with this ring nucleus

User: The user adds a third six-membered ring

Program: There are 11,503 structures with this ring nucleus

User: The user adds a five-membered ring to form a steroid nucleus (Figure 4)

Program: There are 50,506 structures with this ring nucleus.

User: The user specifies that at site X there is to be a ring substitution.

Program: There are 4562 structures with exactly this heteronuclear or stituent pattern

User: The user adds another ring substitution site specification Y (Figure 5)

Program: There are 9532 structures with exactly this pattern

User: The user changes the hetero-atom or stituent pattern switch from exact to imbed

Program: There are 9255 structures with exactly the given ring nucleus and at least the given hetero-atom or stituent pattern

User: The user specifies that sity *Y* is to be a nitrogen

Program: There are 0 structures which have exactly this stituent sequence. This response is due to the fact that the hetero-atom stituent switch is set to exact

User: The user changes the stituent sequence to imbed

Program: 5422 structures have at least the given stituent sequence

User: The user specifies that the site *X* can be either an oxygen, nitrogen or sulfur. (Figure 6)

Program: 255 structures have at least the given stituent sequence class

User: The user specifies that at site Z there is to be a substituent (Figure 7)

Program: There are 15 structures with exactly this substituent pattern

User: The user changes the substituent pattern switch to imbed

Program: 193 structures have at least this substituent pattern

User: The user specifies a substituent at site W (Figure 8)

Program: The program responds that 13 structures have at least this substituent pattern

User: The user asks to see the 13 structures

Program: The program presents in REGN sequence the 13 structures with the given specifications

User: The user changes site Y to be an oxygen

Etc.

The preceeding example would take about 1 minute of real time, given normal user facility with a display and Rand Tablet as well as normal time-sharing loading. The

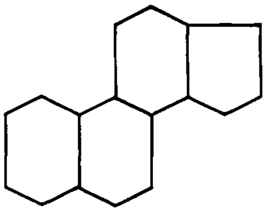


Figure 4. Steroid Ring nucleus

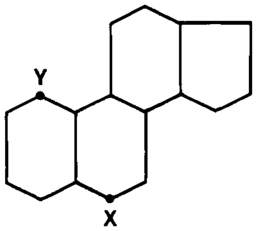


Figure 5. Steroid Ring system with two hetero atoms

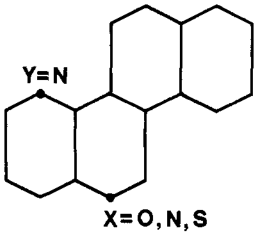


Figure 6. Steroid Ring system with hetero atom positions partially specified

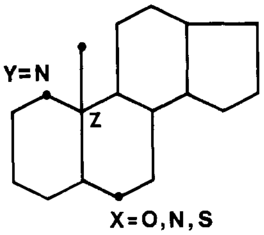


Figure 7. Steroid Ring system with three hetero-atom or stituent positions specified

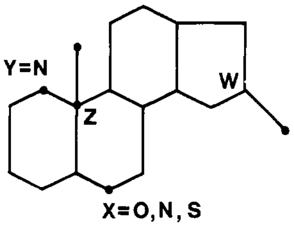


Figure 8. Steroid Ring system with four hetero atom or stituent positions specified

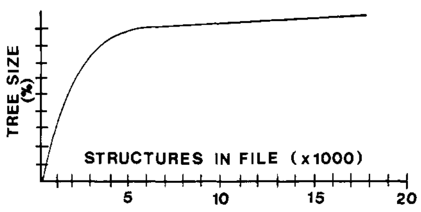
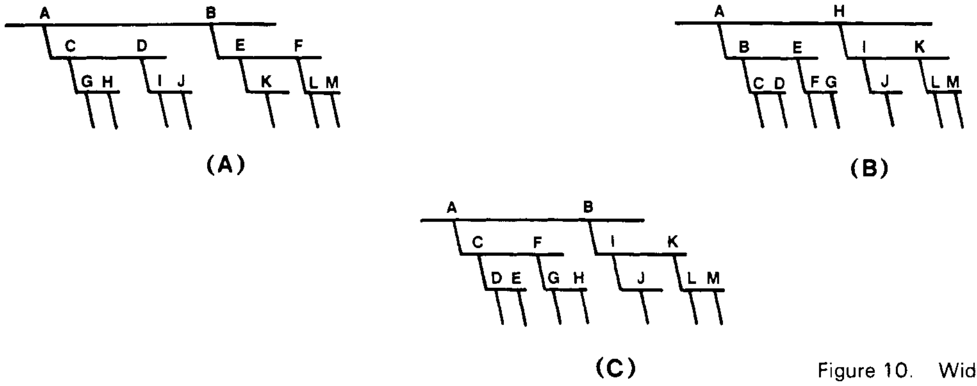


Figure 9. Growth of the nested, structured tree, levels 1-4



(A)

(B)

(C)

Figure 10. Width and depth searches of a structure tree

example indicates a high degree of interaction between the retrieval program and the user. The "EXACT-IMBED" switches at the six levels are:

1. Ring nucleus
2. Ring stituent pattern
3. Ring substituent sequence
4. Ring substituent pattern
5. Substituent structure
6. Substituent atom sequence

These permit the user to alter dynamically the search mode at a given level. The "EXACT" search makes use of the hash addressing technique to ask one question at a level. The "IMBED" search makes use of the SSS technique to ask a class question at a level. By properly arranging the file when it is constructed and by separating branches at each level, a tree is formed.

AN ACTUAL FILE SEARCH

While the previous example was mainly designed to illustrate the tree, work has been done on a subset of the CAS connection table file. The FDA/NLM Common Data Base file of about 20,000 structures has been processed several times during the testing of the file generation and search program. The growth of levels 1-4 of the nested, structured tree is shown in Figure 9. Note that by 5000 structures, the tree has essentially reached full development. Of course, large file scale trends, which can't be estimated, can distort this picture.

The file being used for the following retrieval example has 5001 structures from the Common Data Base. The user changes the state of the program by drawing a structure on the CRT screen, or if a structure has already been drawn, the user depresses the Rand Tablet pen over one the options from the menu shown in Figure 3. For example, if the user depresses the stylus over the "clear" key, the center of the screen is cleared of any structure, and the retrieval program is re-initialized.

The primary method for controlling the search of the nested structure tree is by altering the "EXACT"/"IMBED" switch at any or all levels. The preceeding example gave some idea of the utility of these switches. The "EXACT" search mode will always function more rapidly than an "IMBED" search since the exact search looks for only one node with the specified property. Aside from looking at all the nodes in a branch at A-level, the "IMBED" search is in itself more complicated. At level 1, the "IMBED" search is an atom—by—atom match of the query and the nucleus represented by the node. At the other levels, the imbedment search must filter out what is desired from what is present.

The search of the nested structure tree can be controlled by altering the sequence in which branches of the tree are examined. A width search at A-level in the tree finds all the nodes which satisfy the search. After all the nodes at one level are examined, the nodes of the next level are examined. The alphabetic sequence in Figure 10A illustrates a width search. A depth search at a level finds the first node in the tree which satisfies the search. The search then proceeds to the next level. When no nodes can satisfy the search at a level, the search backs-up to the previous level. The alphabetic sequence in Figure 10B illustrates a depth search. A width search gives a broad view of all of the potential search results at each level. The user, however, must wait until the search of the whole level is finished before the results are presented. A depth search gives the user detailed knowledge of the structures under a particular node in the tree. The six width/depth switches permit the user to search the structure tree in any combination of width and depth. Figure 10C illustrates a search which is a combination of width (level 1) and depth (levels 2 and 3).

The user can decide to obtain decision-making information at the expense of search time by asking the program to show the structure components as they are found in the search of the structure tree. For example, in width search of the first level of the tree (ring nucleus), the full switch would cause the pictures of the ring nuclei to be shown. The user can limit the depth of the search. Some queries may require only certain information. Since the lower levels of the tree are broader than the higher levels, the user can increase the rate of interaction by eliminating the search of unwanted levels.

SEARCH OF THE NESTED, STRUCTURED TREE

The following figures present a detailed description of the search of a nested, structured tree which represents 5001 structures. The ease and speed of performing these searches can only be conveyed by actually using the program or by a video tape or film. It will take you longer to look at the figures than it would to perform the searches actually.

It was clear from previous processing of the Common Data Base file that six-membered rings were very numerous. Figure 11 shows that 5001 structures produced 4443 six-membered rings. The total yield of rings was 7132. The six-membered rings account for 62% of the rings. Of the 4443 six-membered rings, 3620 have no ring substitutions. Most probably these rings are phenyl. Levels 2 and 3 are identical since the null ring stituent pattern can only have a null ring stituent sequence. The 78 rings which have no substituents are most probably multi-atom fragments. Note that the exact match at all four levels causes

```
FOUND   4443 STRUCTURES UNDER      1 NODES AT RING  NUC     EXACT MATCH
THERE ARE NO STITUENTS SPECIFIED
FOUND   3620 STRUCTURES UNDER      1 NODES AT STIT  PATT    EXACT MATCH
FOUND   3620 STRUCTURES UNDER      1 NODES AT STIT  SEQ     EXACT MATCH
FOUND    290 STRUCTURES UNDER      1 NODES AT SUBST PATT    EXACT MATCH
```

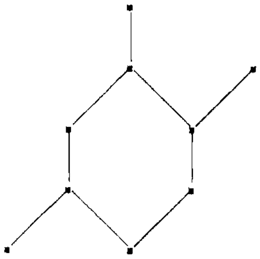


Figure 11. Single six-membered ring with three substituents

```
FOUND   4443 STRUCTURES UNDER      1 NODES AT RING  NUC     EXACT MATCH
FOUND    524 STRUCTURES UNDER      1 NODES AT STIT  PATT    EXACT MATCH
FOUND    340 STRUCTURES UNDER      1 NODES AT STIT  SEQ     EXACT MATCH
THERE ARE NO SUBSTITUENTS SPECIFIED
FOUND      7 STRUCTURES UNDER      1 NODES AT SUBST PATT    EXACT MATCH
```

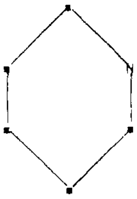


Figure 12. Single six-membered ring with one nitrogen stituent

```
FOUND   4443 STRUCTURES UNDER      1 NODES AT RING  NUC     EXACT MATCH
FOUND    823 STRUCTURES UNDER     12 NODES AT STIT  PATT    IMBED MATCH
FOUND    625 STRUCTURES UNDER     17 NODES AT STIT  SEQ     IMBED MATCH
THERE ARE NO SUBSTITUENTS SPECIFIED
FOUND     21 STRUCTURES UNDER      7 NODES AT SUBST PATT    EXACT MATCH
```

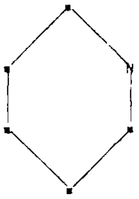


Figure 13. Single six-membered ring with one nitrogen stituent, imbedment match levels 2 and 3

```
FOUND    437 STRUCTURES UNDER      1 NODES AT RING  NUC     EXACT MATCH
THERE ARE NO STITUENTS SPECIFIED
FOUND    437 STRUCTURES UNDER     24 NODES AT STIT  PATT    IMBED MATCH
FOUND    437 STRUCTURES UNDER     32 NODES AT STIT  SEQ     IMBED MATCH
THERE ARE NO SUBSTITUENTS SPECIFIED
FOUND      2 STRUCTURES UNDER      2 NODES AT SUBST PATT    EXACT MATCH
```

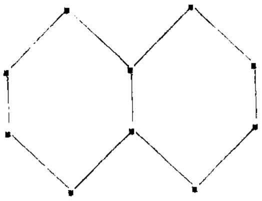


Figure 14. Two fused six-membered rings, imbedment match levels 2 and 3

```
FOUND    437 STRUCTURES UNDER      1 NODES AT RING  NUC     EXACT MATCH
THERE ARE NO STITUENTS SPECIFIED
FOUND    437 STRUCTURES UNDER     24 NODES AT STIT  PATT    IMBED MATCH
FOUND    437 STRUCTURES UNDER     32 NODES AT STIT  SEQ     IMBED MATCH
FOUND     35 STRUCTURES UNDER      5 NODES AT SUBST PATT    EXACT MATCH
```

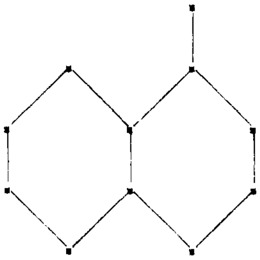


Figure 15. Two fused six-membered rings with one substituent, imbedment match levels 2 and 3

```
FOUND    261 STRUCTURES UNDER      1 NODES AT RING  NUC     EXACT MATCH
THERE ARE NO STITUENTS SPECIFIED
FOUND    261 STRUCTURES UNDER      3 NODES AT STIT  PATT    IMBED MATCH
FOUND    261 STRUCTURES UNDER      3 NODES AT STIT  SEQ     IMBED MATCH
FOUND    227 STRUCTURES UNDER     39 NODES AT SUBST PATT    IMBED MATCH
```

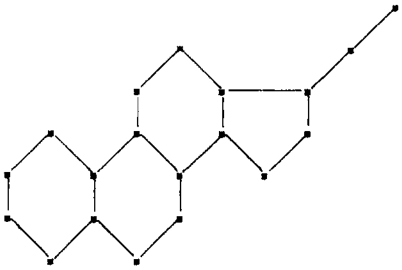


Figure 16. Steroid Ring system with one substituent, imbedment match levels 2, 3, and 4

only one node at each level to be found. The addition of one side chain changes the number of retrieved structures at level 4 to 1306 (29% of the six-membered rings and 18% of all rings). The presence of a second substituent reduces the number of retrieved structures to 330. Figure 11 shows the effect of adding a third substituent. The substituents shown in these figures are vestigal. The search is blocked at level 5. Therefore, all that is required at level 4 is an indication at which nodes on the perimeter of the nucleus the substituents occur. This is a distinct advantage since the user does not have to specify completely the substituents until they could affect the search at levels 5 and 6.

The presence of a hetero-atom reduces the number of retrieved structures at level 2. The 524 structures in Figure 12 represent 11% of the six member rings. Of the 524 structures with exactly one stituent 340 of these have nitrogen at that stituent. Only seven structures in Figure 14 have no substituents. These seven could be bond, isotope, or charge variations. Figure 13 shows the first use of the imbed switch. These are 823 structures with at least one stituent. Of these, 625 structures have one nitrogen. Note that the number of nodes being retrieved below level 1 is no longer one. The nodes in the tree at level 2 produce 17 nodes at level 3, but of these 17 nodes only 2 nodes have the exact substituent pattern. The 21 structures at level 4 of Figure 13 can be explained in any number of ways. Really, since level 4 and below are not specified, the user should block these searches (but the difference in time is so small in many cases that the "so what" effect sets in).

```
FOUND    261 STRUCTURES UNDER       1 NODES AT RING  NUC      EXACT MATCH
THERE ARE NO STITUENTS SPECIFIED
FOUND    261 STRUCTURES UNDER       3 NODES AT STIT  PATT     IMBED MATCH
FOUND    261 STRUCTURES UNDER       3 NODES AT STIT  SEQ      IMBED MATCH
FOUND    143 STRUCTURES UNDER      23 NODES AT SUBST PATT     IMBED MATCH
```

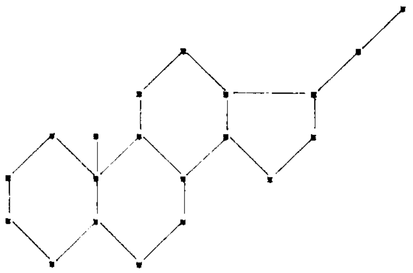


Figure 17. Steroid Ring system with two substituents, imbedment match levels 2, 3, and 4

```
FOUND    261 STRUCTURES UNDER       1 NODES AT RING  NUC      EXACT MATCH
THERE ARE NO STITUENTS SPECIFIED
FOUND    261 STRUCTURES UNDER       3 NODES AT STIT  PATT     IMBED MATCH
FOUND    261 STRUCTURES UNDER       3 NODES AT STIT  SEQ      IMBED MATCH
FOUND    134 STRUCTURES UNDER      22 NODES AT SUBST PATT     IMBED MATCH
```

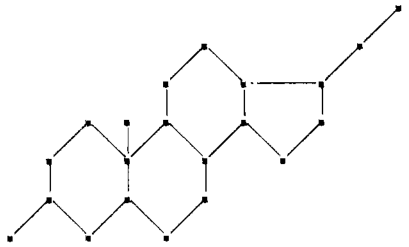


Figure 18. Steroid Ring system with three substituents, imbedment match levels 2, 3, and 4

Since the ring nucleus can be other than a single ring, Figure 14 is included to give some idea of the number of structures with two fused six-membered rings as the nucleus. Figure 15 retrieves 35 structures at level 4 with one substituent added to the nucleus. In the figure, note the number of nodes being searched at levels 2 and 3. By setting the imbedment switch at level 4 to "IMBED," 227 structures (not shown) give at least the given substituent which is six times the number of structures with exactly one substituent. The 32 nodes at level 3 produce 84 nodes at level 4. The user has a wealth of information at his disposal, and the access time is uniformly two seconds. The user is prompted to formulate strategies, make experiments, and then decisions.

Figures 16 and 19 show variations in the number of substituents with a steroid nucleus. These figures are included to give some support for the hypothetical example in the beginning of the paper. Note that while Figure 18 retrieves 134 structures at level 4 with an imbed search, Figure 17 retrieves no structures with exactly the given substituent pattern.

The imbedment search at levels 2 and 3 has been demonstrated in previous figures. Figure 20 shows an imbedment search at level 1 of a fused 5–6 ring nucleus. There are 134 nuclei which form level 1 for the file of 5001. Of the 135 nuclei, 55 have at least one five-membered ring and one six-membered ring. There are 46 nuclei with a fused 5–6 ring imbedded. Figure 21 shows the imbedment of a fused 6–6 ring nucleus. The imbedment search at level 1 is quite time consuming. The search of the 134 nuclei takes 10 seconds. The important thing to remember is that the user always has control over how his real time and the CPU time are to be spent.

Up until now only searches of levels 1 to 4 have been shown. Clearly the nuclei, the stituent pattern, the stituent sequence, and the substituent pattern provide powerful search capability. Figure 22 shows the retrieval of 1305 structures at level 4 for a six-membered ring with one substituent. In Figure 23, level 5 is not blocked. The 1305 structures at level 4 yield 54 structures at level 5 under exact match. Level 6 is not yet implemented. Clearly, the "EXACT" or "IMBED" matching of substituents is useful in reducing the number of structures.

```
FOUND    261 STRUCTURES UNDER       1 NODES AT RING  NUC      EXACT MATCH
THERE ARE NO STITUENTS SPECIFIED
FOUND    261 STRUCTURES UNDER       3 NODES AT STIT  PATT     IMBED MATCH
FOUND    261 STRUCTURES UNDER       3 NODES AT STIT  SEQ      IMBED MATCH
EXACT MATCH FOR SUBST PATT  NOT FOUND
```

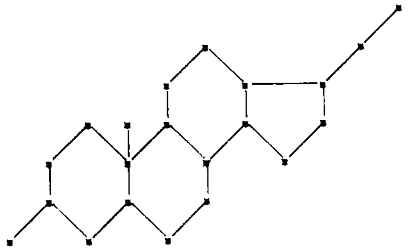


Figure 19. Steroid Ring system with three substituents, imbedment matches 2 and 3, exact match level 4

```
THERE ARE  134 NUCLEI        55 PASSED THE SCREEN
FOUND    774 STRUCTURES UNDER      46 NODES AT RING  NUC      IMBED MATCH
FOUND     12 STRUCTURES UNDER       1 NODES AT STIT  PATT     EXACT MATCH
EXACT MATCH FOR STIT  SEQ   NOT FOUND
```

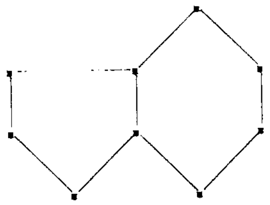


Figure 20. Two fused 5–6 membered rings, imbedment match level 1

```
THERE ARE  105 NUCLEI        61 PASSED THE SCREEN
FOUND   1355 STRUCTURES UNDER      46 NODES AT RING  NUC      IMBED MATCH
THERE ARE NO STITUENTS SPECIFIED
FOUND    737 STRUCTURES UNDER      27 NODES AT STIT  PATT     EXACT MATCH
FOUND    737 STRUCTURES UNDER      27 NODES AT STIT  SEQ      EXACT MATCH
THERE ARE NO SUBSTITUENTS SPECIFIED
FOUND     26 STRUCTURES UNDER      17 NODES AT SUBST PATT     EXACT MATCH
```

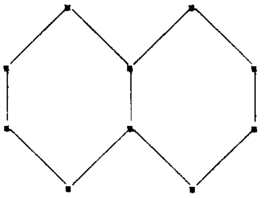
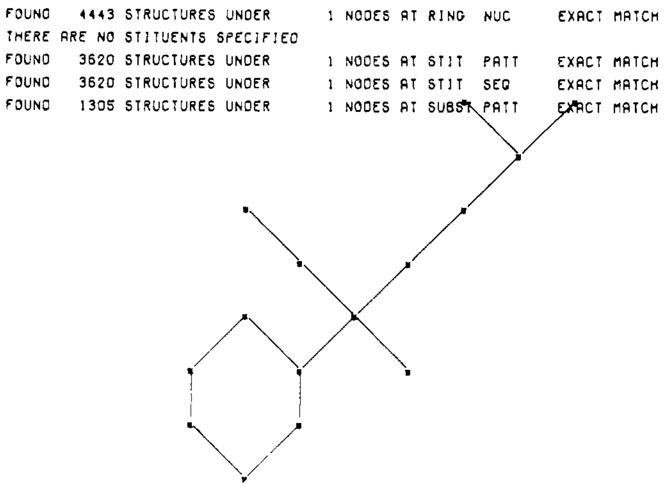
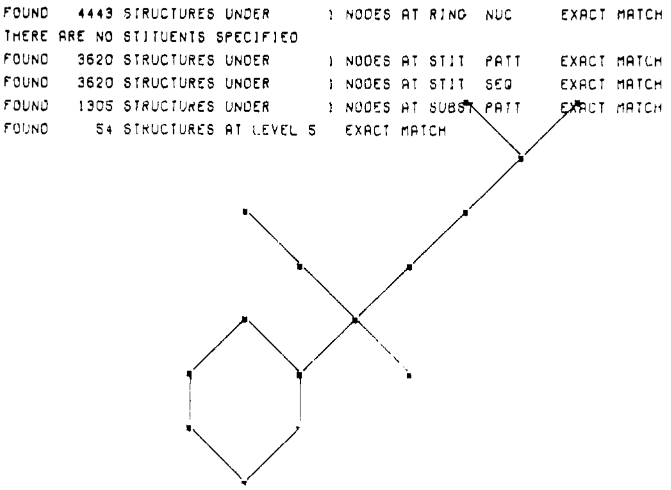


Figure 21. Two fused 6–6 membered rings, imbedment match level 1



```
FOUND   4443 STRUCTURES UNDER       1 NODES AT RING  NUC     EXACT MATCH
THERE ARE NO STITUENTS SPECIFIED
FOUND   3620 STRUCTURES UNDER       1 NODES AT STIT  PATT    EXACT MATCH
FOUND   3620 STRUCTURES UNDER       1 NODES AT STIT  SEQ     EXACT MATCH
FOUND   1305 STRUCTURES UNDER       1 NODES AT SUBST PATT    EXACT MATCH
```

Figure 22. One six-membered ring with complete substituent, level 5 blocked



```
FOUND   4443 STRUCTURES UNDER       1 NODES AT RING  NUC     EXACT MATCH
THERE ARE NO STITUENTS SPECIFIED
FOUND   3620 STRUCTURES UNDER       1 NODES AT STIT  PATT    EXACT MATCH
FOUND   3620 STRUCTURES UNDER       1 NODES AT STIT  SEQ     EXACT MATCH
FOUND   1305 STRUCTURES UNDER       1 NODES AT SUBST PATT    EXACT MATCH
FOUND     54 STRUCTURES AT LEVEL 5  EXACT MATCH
```

Figure 23. One six-membered ring with complete substituent research through level 5. Level 6 not yet implemented

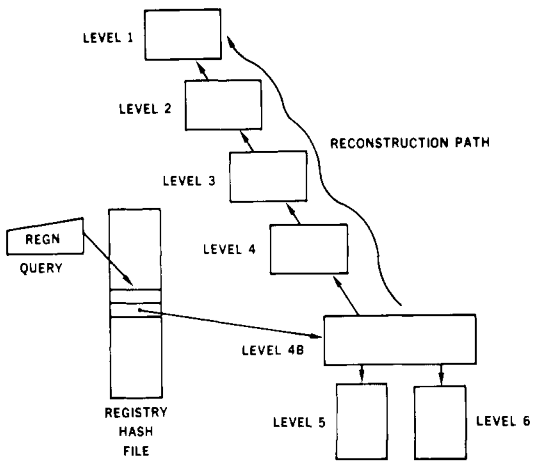


Figure 24. Retreival path for a structure given a REGN (Registry Number)

REGISTRY RETRIEVAL AND FILE UPDATE

The nested, structured tree can be used for registration of compounds. With the "EXACT" "IMBED" switches set to "EXACT," a probe into the tree will determine the presence or absence of a compound. Since the structure tree reaches virtually full development, the addition of a new compound will most probably affect only lower levels of the tree (4A–6). The nested, structured tree can be incrementally updated with ease since the basic mechanisms of linking together data elements are an essential part of the initial construction of the tree. Details of the file generation and update will be presented in a subsequent paper.

Also, the nested, structured tree can be used to retrieve a structure given the registry number. The registry number is contained in a hash file (Figure 24).[1] The hash file has a pointer to the registry cluster. The registry cluster contains all of the data concerning the ring substituents and the relationships between rings in a chain of rings situation. In addition, the registry cluster contains a pointer to level 4 in the structure tree. One word in each node of the structure tree is devoted to a pointer to the preceding level (Father). The retrieval of the structure given a REGN involves its reconstruction from the data in the registry cluster and in the structure tree.

CONCLUSIONS

SSS of large files is currently thought to be uneconomical. Using redundant-topology tape or disk-based sequential systems this is essentially true. It appears that the techniques shown which eliminate the topological redundancy and the formation of structure class representations reduces the search time as well as the amount of storage required to represent a file. The dynamically alterable search mode permits the user to formulate a query ranging from an exact structure match to a very broad class search. The inherent interactiveness of the retrieval program makes exploratory queries an integral part of the search process.

LITERATURE CITED

(1) Feldmann, R. J., Heller, S. R., Shapiro, K. P., and Heller, R. S., "An Application of Interactive Computing: A Chemical Information System," *J. Chem. Doc.* **12,** 41–7 (1972).

(2) Farrell, C. D., Chauvenet, A. R., and Koniver, D. A., "Computer Generation of Wiswesser Line Notation," *J. Chem. Doc.* **11,** 52–9 (1971).

(3) Heller, S. R., and Koniver, D. A., "Computer Generation Of Wiswesser Line Notation, II. Polyfused, Perifused and Chained Ring Systems," *J. Chem. Doc.* **12,** 55–9 (1972).

(4) Miller, G. A., "Encoding and Decoding WLN," *J. Chem. Doc.* **12,** 60–7 (1972).

(5) Feldmann, R. J., and Koniver, D. A., "Interactive Searching of Chemical Files and Structural Diagram Generation from Wiswesser Line Notation," *J. Chem. Doc.* **11,** 154–9 (1971).

(6) Meyer, E., "Versatile Computer Techniques for Searching by Structural Formulas, Partial Structures, and Classes of Compounds," *Angew. Chem. internat. Edit.* **9,** 583–9 (1970).

(7) Leiter, D. P., Morgan, H. L., and Stobough, R. E., "Installation and Operation of a Registry for Chemical Compounds," *J. Chem. Doc.* **5,** 238–42 (1965).

(8) Lefkowitz, D., "Substructure Search in the MCC System," *J. Chem. Doc.* **8,** 166–73 (1968).

(9) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, 1968.