(10) Rice, C. N., "Toward a National Systems Resource in Toxicology," *J. Chem. Doc.*, **9**, 181 (1969).
(11) Chemical Abstracts Service, "Desk-Top Analysis Tool for the Common Data Base," ACS, Columbus, Ohio, 1968, Federal Clearing House for Scientific and Technical Publications, PB 179 900.
(12) Smith, E. J., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968.
(13) Patterson, A. M., L. T. Capell, and D. F. Walker, "The Ring Index," ACS, Chemical Abstracts Service, Columbus, Ohio, 2d ed., (1960), Suppl. I (1963), Suppl. II (1964), Suppl. III (1965).
(14) Chemical Abstracts Service, "Wiswesser Line-Notations Corresponding to Ring Index Structures," ACS, Columbus, Ohio, 1968, Federal Clearing House for Scientific and Technical Publications, PB 180 901.

Application of the MCC Topological Screen System to a Small File of Pesticides

SAMUEL T. MORNEWECK*
Esso Research and Engineering Co., Linden, N. J. 07036

BRUCE G. HAWTHORNE
Esso Mathematics and Systems, Inc., Florham Park, N. J. 07932

Received August 19, 1970

A file of 8600 compounds, which were tested for pesticide activity, was manually coded and processed by computer to give substructure indexes using the MCC Topological Screen System developed at the University of Pennsylvania. Information is presented on coding rates and coding errors, computer processing times, and substructure searching experience. The indexes are found to be relatively inexpensive to produce and to provide considerable substructure searching capabilities.

The nature of pesticides research requires the ability to search chemical files by substructure. It is frequently necessary to assemble all the compounds containing a particular substructure as a first step in preparing structure-activity relations. On the other hand, the need to recall a particular compound is rather infrequent and can usually be met by intersecting two or three substructure lists if the substructure strings are reasonably long or by maintaining a separate empirical formula index.

A number of manually generated fragment coding systems are known, but an extensive experiment with such a system was unsatisfactory. The system used was developed at Esso and was based on the A.P.I. Thesaurus of Chemical Aspects. The vocabulary was substantially expanded, and the definitions of roles and links were modified to indicate connections between groups of atoms. Problems of thesaurus maintenance and the expense of coding, both of which would be common to all manually generated fragment codes, required that another system be found.

The following criteria were established for a substructure search system:

Reasonably long, descriptive fragments
Minimal cost for coding and processing
Printed indexes so that computer searching was not required for each request.

These criteria seemed to be fulfilled in a system that was being developed by Lefkovitz and coworkers at the University of Pennsylvania, the MCC Topological Screen System (TSS).[1,2] This system uses a nonunique line notation for input, which promised even easier coding than the popular Wiswesser Line Notation, and generates printed indexes of reasonably long fragments, which appear to be easy to search.

With the agreement of the sponsors, the application of this system on a file of 8600 compounds was undertaken. The compounds were intermediates and final products, which were synthesized in a pesticide research program, and selected compounds from other company laboratories, which were tested for pesticidal acitivity. Thus, the file had very few small molecules (<10 non-H atoms) and very few large, polycyclic molecules.

MCC CODING

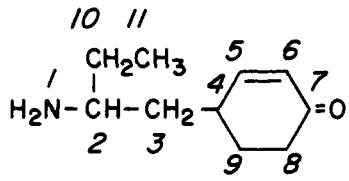
Compounds were coded from the data sheets submitted by the chemists using the Mechanical Chemical Code (MCC) described in reference 1. The MCC uses standard atomic symbols except for elements that normally have hydrogen attached. Thus

is a, $-CH_2-$ is b, $-CH_3$ is c, $-NH-$ is M, $-NH_2$ is Z, $-OH$ is Q. Certain other contractions are used for common groupings, i.e., $>C=O$ is L, $-NO_2$ is NX,

* To whom inquiries should be addressed; present address, Department of Chemistry, St. Peter's College, Jersey City, N. J. 07306.

—SO_2— is SX, benzene ring is R (not required—the ring can also be specified as a string of a's and C's). Since lower case letters or subscripts could not be used, certain of the symbols were changed, i.e., a = A, b = B, c = U, / means subscript follows (B/3 = B_3 = BBB). These changes also required that boron and uranium be coded as +BO and +UR, instead of the standard atomic symbols.

Briefly stated, the MCC is usually generated by starting at one end of the molecule and specifying all the atoms or groups that are present. The attachments between atoms do not have to be specified unless an atom is not attached to the immediately preceding atom in the code. For example:



is ZABAAALBB4B2U

Additional provision was made for materials that were made up of two discrete chemical compounds having no formal bonds between them, such as hydrates or mixtures of isomers. Such MCC's were separated by periods as shown in the following example:

$CH_3CH_2CH_2Cl + (CH_3)_2CHCl$

UBBG·UAGU

These compounds then appeared separately in the rotated screen indexes as decimals. Thus 1-chloropropane was assigned the number $n \cdot 0$ (where n is the compound number) and 2-chloropropane was assigned $n \cdot 1$. This change made it possible to make some helpful breakdowns in the file, e.g., amine salts of acids were treated as two compounds, thereby making them accessible through either the acid or the amine and avoiding any confusion between them and esters or quaternary salts. Thus trimethylammonium acetate was coded as ULQ·NUUU, rather than ULOJUUUH.

The simplicity of MCC, which allows the coder to start anywhere in the molecule and proceed in any direction using a very small basic set of symbols, permits it to be taught to anyone who can be taught to write an empirical formula for a structural formula. Lefkovitz[3] reports a training time of six hours with a high school graduate, which is in line with our experience.

Coding of the files was done mainly by three people: a clerk with high school education who had been previously trained to write empirical formulas from structural formulas, a chemical technician with some college training, and a college graduate in chemistry. Their rates of coding varied substantially as shown in Table I.

The MCC's were keypunched on an IBM 026 printing keypunch along with the compound number. Generally only one card was needed per compound, but as many continuation cards can be used as are needed.

For the initial file of 8582 compounds the error rate for coding was a very high 15% (Table II).

The following eight common errors in coding were found, with 1 and 2 contributing at least 2/3 of all the coding errors (the underlined symbols are the ones omitted or incorrectly used):

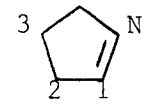
Table I. Rates of MCC Coding

Coder	Compounds Hr
Clerk	82
Chemical technician	100
College graduate	200

Table II. Errors in Preparation of File

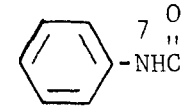
Total number of compounds	8582
Coding errors	1291 (15%)
Coding errors missed by program	370 (29%)
Keypunch errors	54 (<1%)
Keypunch errors missed by program	0

1. Failing to indicate the double bond before a locant, e.g.,



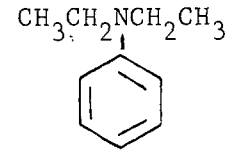
is AB/3N<u>D</u>1

2. Failing to locant all connections out of R, e.g.,



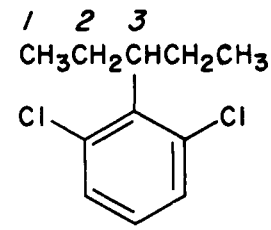
is RM1<u>L</u>

3. Treating R as a univalent substituent, e.g.,



is UBNRB<u>3</u>U

4. Misnumbering R when it is locanted to an earlier atom, e.g.,



UBABUR3G<u>6</u>G<u>10</u>

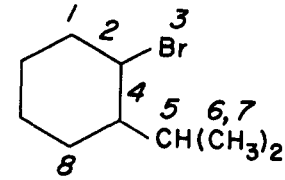
(The highest numbered atom in the ring is attached to the locant.)

5. Failing to specify valences for elements, e.g.,

$(CH_3)_3SnOCOCH_3$ is <u>=4</u>+SNUUUOLU

(Sn is assumed to be divalent unless specified.)

6. Failing to specify the locant when returning to the main chain, e.g.,



BAEAAUUB<u>4</u>BB1

7. Mixing related symbols, e.g., A,B,C or N,M.

Table III. Computer Processing Times for TSS

Procedure	Minutes/1000 Compounds CPU	I/O
Edit and screen assignment	4.39	0.31
Complete screen indexes	0.56	0.38
Acyclic subscreen indexes	1.65	0.59
SMF screen indexes	0.23	0.26
ERP screen indexes	0.18	0.26
Total	7.01	1.80

8. Failing to use a slash to denote a subscript, e.g., NH_3 is NH/3.

Although no breakdown of errors by coder was made, there was an indication that the error rate increased with the coding rate, which is not unlikely. It is expected that the error rate can be reduced substantially by emphasis of the above common errors during the training period.

COMPUTER PROCESSING

Computer processing was carried out on an IBM System 360 Model 65 requiring 150K bytes of core storage, four 9-track, 1600 BPI tape drives, and six 2314 disk work data sets. The operating system was release 15/16 of OS/MVT with HASP-II. All the TSS programs are written in FORTRAN IV and were compiled under the G level compiler. The Edit and Screen assignment procedure consists of one FORTRAN program while each index procedure requires three programs and two IBM sorts. Processing times are given in Table III.

Processing is carried out in two stages. In the first stage, the input is read, a connection table is generated and checked for satisfaction of all valences, and all the screens are generated and written on four tapes. The printout includes error messages and one or more of the following options: MCC, connection table, screens assigned to each compound. It was useful to have the MCC and the screens printed, but the connection table print required a lot of paper and was not very useful in detecting or diagnosing errors.

As noted in Table II, the error check in the program found 71% of the coding errors and all the keypunch errors. Unfortunately, the program was particularly likely to miss the most common coding errors, 1 and 2. The program checks the connection table that it generates from the MCC to see that all valences of all atoms are satisfied. Since locants and multiple bonds are only specified when an atom is not bonded to an immediately preceding atom in the code, the program can frequently generate a valid connection table by rearranging the bonds. This problem is substantially aggravated by the program's unchecked ability to create separate compounds even where they were not specified in the MCC. Thus if the rearrangement of bonds results in the termination of a chain, the next atom is assigned to a new chemical compound and the checking proceeds. Of the 370 coding errors missed by the program, 155 were the result of this creation of separate compounds. A change in the program that would prevent this option of creating more compounds than were specified would produce a substantial reduction in coding errors missed by the machine check. (Lefkovitz has indicated that several changes have now been made in the program, i.e., disconnected structures will not be permitted unless the MCC contains specific symbols to delineate them, and Type 2 errors will not be missed because it is now assumed that the next MCC symbol is connected to symbol 6 of the ring. These changes should catch most of the errors mentioned except in a very few cases where rearrangement of bonds can produce a valid formula.)

Even with the above changes in the program, 17% of the coding errors would still have gone undetected. The easiest way to check for other errors missed by the program is to compare the screens assigned with the structural formula. A check of the SMF screens that give the atom count of each ring in the compound is frequently sufficient to indicate any errors. If these are correct or there are no rings in the compound, the central atoms in the acyclic complete screens are the other checkpoint. These two checks are much easier to make than reconstructing the coding of the MCC, which could have started anywhere in the molecule. (Overall, it required an estimated 30 hours of professional time to find and correct the 370 errors that were missed by the program check.)

The corrected MCC's are again passed through the edit and screen generation phase, and the correct screens are added to the output tapes. These corrected tapes then pass individually into the second phase where the rotated and inverted indexes are produced and printed. The sizes of the indexes are given in Table IV.

THE TSS INDEXES

Each of the four index sets for TSS includes a rotated index and an inverted index. (This brief description of the TSS indexes has been included for the benefit of the reader. Further information and examples can be found in reference 3.) The following paragraphs and Figure 1 (taken from reference 3) summarize the screen assignment algorithms.

The screen index specifies the central atom of an acyclic branch point and the chains radiating from that atom. The MCC expressions for the chains go up to and include terminal atoms or ring attachments (indicated by asterisks) but do not include another central atom. If two central atoms are adjacent, they are included in the radiating chain of each other. The program rotates the chains radiating from the central atom and prepares an alphabetic index of all the rotated forms.

The subscreen index specifies the central atom of an acyclic branch point and one of the branches that radiates from it. The subscreens are rotated so that each MCC symbol appears in the principal column.

The Elementary Ring Population (ERP) screens list separately all the rings in the compound, giving their total atom count, molecular formula in standard symbols, and the number of double bonds. [Resonant bonds can

Table IV. Index Sizes for 8582 Compounds

Index	Number of Screens
Acyclic screens	3909
Acyclic subscreens	3357
Elementary Ring Population screens	208
Skeleton Molecular Formula screens	375

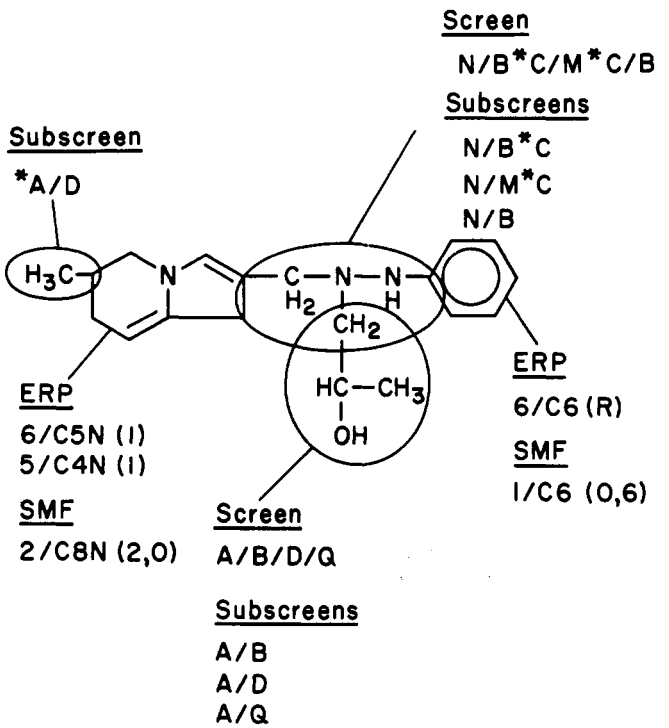


Figure 1. TSS screen assignment algorithms

be indicated, but this requires special coding. We chose not to use this special code, so our benzene rings became 6/C6 (3).] The program finds the smallest set of smallest rings for polycyclic structures and uses this set in the ERP indexes. The screens are rotated so that every symbol appears in the principal column.

The Skeleton Molecular Formula (SMF) screens indicate the number of rings in each cyclic fragment, the molecular formula of the fragment in standard symbols, and the number of double bonds in the fragment (resonant bonds are also listed if coded). As above, the screens are rotated so that every symbol appears in the principal column.

SUBSTRUCTURE SEARCHING

The four indexes from the TSS system for the original 8582 compounds have been in use in our laboratory for about one year. Although many searches were done, no comprehensive records were kept of the results. Such results are reported elsewhere.[3]

The indexes proved generally useful for the types of searches required in pesticide research. In most cases it was possible to furnish the chemist with a list of compounds containing the pertinent substructures within 30 minutes of the receipt of his request. It was not usually necessary to confirm the structures by review of the structure sheets, since the chemists usually wanted to go to the files to see the testing data on the compounds. Since these data sheets also contained the chemical structure, one was able to eliminate quickly any false drops. Significantly, there were no complaints about excessive false drops. The prompt response, which would not have been possible without the printed indexes, was very favorably received and led to increased use of these search facilities.

There are some problems with the use of the fragment indexes. If several screen lists have to be intersected, it sometimes requires much flipping back and forth between lists in the inverted indexes. This could be easily solved by putting the four inverted indexes on microfilm for use in a reader-printer. Then a quick copy could be made of one list that could be compared with the other lists to find the intersections.[4] An even better alternative would be to have the inverted lists loaded on magnetic disk files to be intersected by computer. Such a program has been written.[3]

There are other use problems that are not so readily corrected. One particularly troublesome case results from requests for groups separated by a certain length of carbon chain that can be substituted or unsubstituted. Since each carbon can possibly be a central atom for a screen, the number of possible screens increases exponentially with the number of carbons in the chain. In general, "don't care" options in open chain segments lead to problems in searching, but such options for ring substituents are no problem.

Another problem area is the lack of specificity of atom placement in rings containing two or more heteroatoms. This can lead to a lot of false drops, for example, in a search calling for pyridazines in a file containing many diazines.

Although it did not occur in our searching, questions that specify a particular orientation for substituents on rings, e.g., *p*-chloroanilines, cannot be handled with the screens alone. A structure look-up must follow the narrowing of the file to those rings containing the proper substituents.

The TSS system, therefore, works best with questions that require rather specific open-chain fragments and/or relatively general ring fragments. These kinds of queries were not uncommon in our experience with the system.

It is always easy to lose sight of the real advantages of an information system when the discussion turns to its limitations. It must be emphasized that the MCC Topological Screen System is a relatively inexpensive system to operate and provides substantial substructure search facility. Furthermore, this system has now been shown to operate outside the inventor's own facilities and has been made available to any company who would like to try it.[5] These are impressive credentials if substructure searching is required in an information program.

LITERATURE CITED

(1) Lefkovitz, David, "A Chemical Notation and Code for Computer Manipulation," *J. Chem. Doc.*, **7**, 186–91 (1967).
(2) Lefkovitz, David, "Substructure Search in the MCC System," *J. Chem. Doc.*, **8**, 166–73 (1968).
(3) Lefkovitz, David, and A. R. Gennaro, "A Utility Analysis for the MCC Topological Screen System," *J. Chem. Doc.*, **10**, 86–94 (1970).
(4) Kaback, S. M. "User Benefits from Secondary Journals on Microfilm," *J. Chem. Doc.*, **10**, 7–9 (1970).
(5) Lefkovitz, David, in "Tutorial on Available Computer Programs for Information Retrieval," ACS National Meeting, New York, September 11, 1969.