Computer Translation of Systematic Chemical Nomenclature to Structural Formulas—Steroids*

R. N. STILLWELL
Institute for Lipid Research, Baylor College of Medicine, Houston, Texas 77025

Received April 3, 1973

A program (FORMULA) is described which translates the systematic chemical names of steroids to structural formulas which with a suitable graphic output device could be used for direct typesetting.

Of the many possible representations of a chemical compound the most widely used are the structural formulas, trivial names, and systematic names. To the chemist, the most useful is the structural formula; unfortunately, this representation is difficult to typeset and impossible to index. For these purposes, the systematic name is the most useful representation in that it is also fairly readily understood by the chemist, although other representations such as the Wiswesser Line Notation (WLN) may be more efficient for information retrieval because of their compactness and because a given compound yields only a single WLN.

Translation between one representation and another by means of computers would be of obviously great assistance to the chemist, and a number of such programs have been reported.[1-10] Most of them deal with WLN as either the source or the target representation. This paper is a description of a program (FORMULA) which translates a systematic chemical name of a compound of a limited class (the steroids) to a structural formula, as nearly as possible in the format in which the chemist is accustomed to seeing it, and which with a suitable graphic output device could be used for direct typesetting.

The program was written in PL/I for an IBM 360/50 computer, and is highly modular and table-driven in order to be modified easily. In its present version, it is an interactive program running under the Baylor Executive System for Teleprocessing (BEST). Interaction with the user is through any of several terminals, and the graphic output can be directed either to a storage CRT terminal (a Tektronix 4010 with a hard copy unit), which may also be the interactive terminal, or to an incremental plotter. If he wishes, the user may enter no more than a systematic name (Figure 1), or he may specify other parameters such as the size of the structure, the relative size of the lettering, the position on the plot, etc. (Figure 2; the operator has asked for a debugging trace). The program is based on the IUPAC-IUB 1971 Definitive Rules for Steroid Nomenclature. Not all of the names which can be generated under the Rules can as yet be translated; the important exceptions are: configuration by the sequence rule, the prefix *ent-*, radical nomenclature (as 3-cholestanyl methyl ether), and complex substituents.

The four principal parts of the program are parsing, analysis, structure generation, and plotting. The main procedure (STEROTP) handles initialization and termination of the program and calls the other subroutines, while smaller subroutines handle most of the input-output. The program is heavily overlaid in order to run in a 75K partition of core, but because each principal overlay is called only once per structure this does not slow down the program appreciably. Information is passed between overlays by means of "external" tables in core.

The parsing section of the program is called STEROID and is essentially quite simple. The steroid nucleus (the stem of the parent name) is first identified. The entire name is then scanned from the beginning for positions and substituents, alternately, between hyphens. These are extracted first as character strings. The substituent is then identified as a member of a table (GROUP) and thereafter referred to by its index in the table. GROUP currently contains only 50 entries and includes, besides commonly occurring functional groups ("ol," "hydroxy," "en," etc.) the modifiers "nor," "homo," "seco," and "abeo." The position string is then analyzed according to rules which depend on the type of substituent (compare "3,5(10)-diene" with "3,4:5,10-diepoxy"), and an entry is made for each occurrence of the substituent in a table (SUBSTIT). If the position string immediately precedes the nucleus, it may be either a stereochemical designation (5β-estrane) or the position of unsaturation (5-estrene); if the former, a dummy substituent identification is assigned.

The analysis phase (ANALYZE and GENER) constitutes the largest part of the program. The entries in the position list are first sorted into pure position numbers (POSNO) and bond types (BONDTYP). Files describing the steroid nucleus are read in, along with a list of the normal angular substituents for that nucleus. This list is then checked against the SUBSTIT list for alterations at the angular positions. For example, the "substituents" 19-nor, 5β-(dummy), 7-ene, and 18-ol all affect angular positions. The remaining angular substituents are then added to SUBSTIT and redundant ones, principally "nor's" and "dummies," are removed. Remaining "nor's" represent atoms lost from the nucleus, so the nucleus "table" (*vide infra*) is edited accordingly. Homo-, seco-, *abeo-*, and hetero modifications are then processed and SUBSTIT is corrected again by removing hydrogen from new secondary or quaternary positions, adding $-CH_3$ to ends of bonds broken by seco-, etc.

The next task is to generate the drawing to be plotted. To simplify programming at this state, use is made of a 60° coordinate system (Figure 3). Eighty points of this plane are numbered, points 1–27 corresponding to the atoms of the cholestane nucleus with the same numbers. A table (NEXT) lists four positions "next" to each point, but not lying on the steroid ring system. Another table (IOCC) is used to keep tally of which points have been filled. This system avoids the necessity of calculating an-

* Presented in part at the Southwestern Regional Meeting, ACS, Baton Rouge, La., Dec. 6–8, 1972.

```
STANDARD SPECS?
YES
X, Y COORDINATES OR END?

NAME OF STEROID
17A,21-DIHYDROXY-4-PREGNENE-3,11,20-TRIONE
```

Figure 1. Minimum user interaction

```
STANDARD SPECS?
NO
 XLIM      20.000
YLIM       10.000
DOUBLE     0.150
DHEAVY     0.020
IDASH      6.000
LETSIZE    0.300
WIGGLE     0.060
BONDLEN    0.700
FTRACE     0.000
TITLE      1.000
ORGINC     0.000
(DATA FORMAT INPUT)
FTRACE=1;
```

Figure 2. Specifiable options

$$x' = x \cdot \sqrt{3}/2$$

$$y' = y - x/2$$

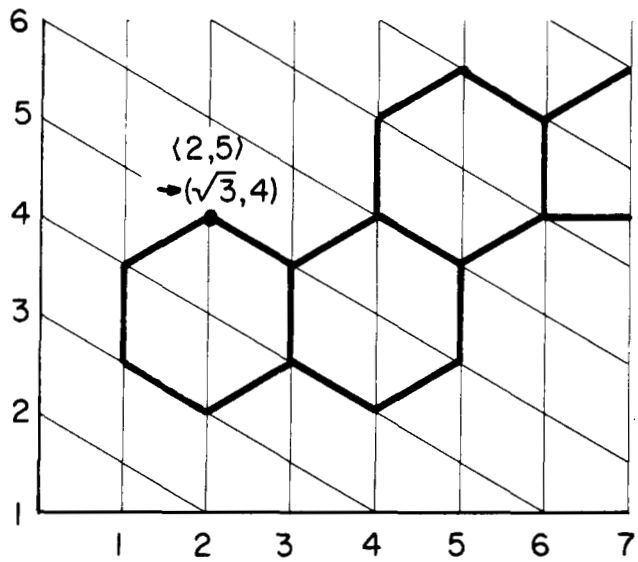

Figure 3. 60° Coordinate system and relationship to Cartesian coordinates

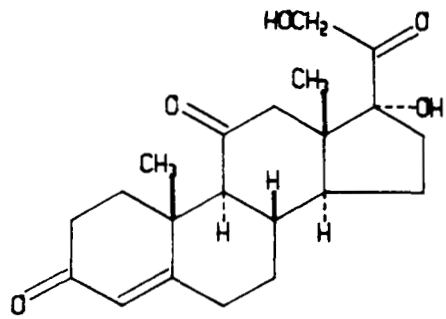

Figure 4. Formula generated from dialog of Figure 1

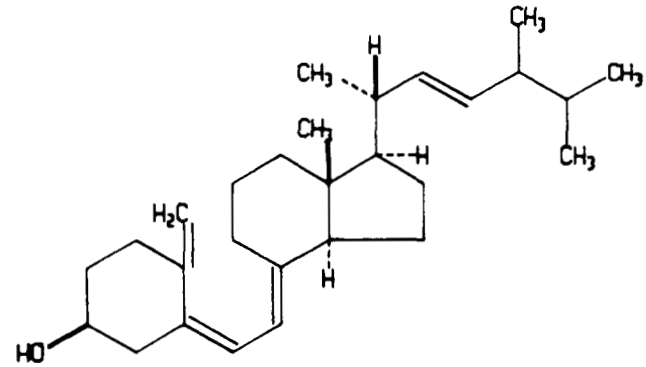

Figure 5. Formula generated for "9,10-SECO-5,7,10(19),22-ERGOSTATETRAEN-3B-OL"

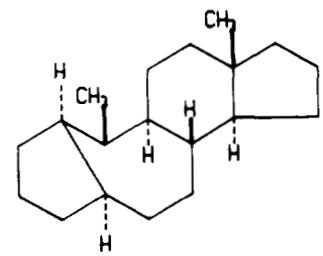

Figure 6. Formula generated for "5(10 − >1AH)-ABEO-5A ANDROSTANE"

gles and distances for each added substituent. The first step in generating the plot, then, is to fill in the positions of the "outer" ends of the substituents where this has not already been done. The nucleus table mentioned above is actually a character string separated by slashes into fields within which the entries are: x and y 60° coordinates to plot to, type of bond to plot (none, single, heavy, double, etc.), and (optionally) a character string to place at the end of the bond. A second such table is then made up containing the same information for the substituents. By this point, of course, all modifying substituents such as "nor" and "dummy" have been eliminated, and those which are left fall into three classes: substituents bonded to a single atom, substituents bonded to two atoms, and bonds (single, double, or triple) joining two atoms. The last (optional) entry in the substituent table is the name as originally given, to be plotted under the figure. Finally, the two buffers are passed to the plotting routine (PLOTTER), where the 60° coordinates are converted to Cartesian coordinates, and the various lines and letters are calculated and plotted via standard PLOT and LETTER calls. Examples of the resulting output are shown in Figures 4–6.

FORMULA provides the chemist with an easily used method of producing a structural formula for illustrations or slides, and at the same time permits him to check the accuracy of the systematic names he has constructed for his compounds. FORMULA also provides a demonstration that standard chemical nomenclature can be handled by computers if it is sufficiently well defined.

ACKNOWLEDGMENT

This work was supported by NIH grants GM-13901 and RR-00259.

LITERATURE CITED

(1) Farrell, C. D., Chauvenet, A. R., and Koniver, D. A., "Computer Generation of Wiswesser Line Notation," *J. Chem. Doc.* **11**, 52 (1971).
(2) Feldmann, R. J., and Koniver, D. A., "Interactive Searching of Chemical Files and Structural Diagram Generation from Wiswesser Line Notation," *Ibid.*, **11**, 154 (1971).
(3) Granito, C. E., Roberts, S., and Gibson, G. W., "The Conversion of Wiswesser Line Notations to Ring Codes. I. The Conversion of Ring Systems," *Ibid.*, **12**, 190 (1972).
(4) Hyde, E., Matthews, F. W., Thomson, L. D., and Wiswesser, W. J., "Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds," *Ibid.*, **7**, 200 (1967).
(5) Hyde, E., and Thompson, L., "Structure Display," *Ibid.*, 8, 138 (1968).
(6) Lynch, M. F., "Conversion of Connection Table Descriptions of Chemical Compounds into a Form of Wiswesser Notation," *Ibid.*, **8**, 130 (1968).
(7) Miller, G. A., "Encoding and Decoding WLN," *Ibid.*, **12**, 60 (1972).
(8) Petrarca, A. E., Lynch, M. F., Rush, J. E., "A Method for Generating Unique Computer Structural Representations of Stereoisomers," *Ibid.*, **7**, 154 (1967).
(9) Thomson, L. H., Hyde, E., and Matthews, F. W., "Organic Search and Display Using a Connectivity Matrix Derived from Wiswesser Notation," *Ibid.*, **7**, 204 (1967).
(10) Vander Stouw, G. G., Naznitsky, I., and Rush, J. E., "Procedures for Converting Systematic Names of Organic Compounds into Atom-Bond Connection Tables," *Ibid.*, **7**, 165 (1967).

A Microfilm Index to *Chemical Abstracts*

F. ROBINSON
CIU—Development, Imperial Chemical Industries Ltd., Imperial Chemical House, Millbank, London SW1P 4QG, England

Received November 2, 1971; Resubmitted April 9, 1973

To improve access to the recent *Chemical Abstracts*, a cumulative quarterly index, based on the keyword phrases, has been produced in microfilm form. By the use of the *CA Condensates* tapes and Computer Output Microfilm (COM) the index is available soon after the end of each quarter. Abstract titles are included in the index, thus increasing its value as a working tool.

Chemical Abstracts plays a fundamental part in the information services of most chemical companies. The printed volumes are scanned by most information units for current awareness and are also searched retrospectively on specific topics. The continuing growth in the number of items covered[1]—although it emphasizes the importance of CA as an information source—makes its use more difficult. The possibility of handling a reasonable proportion of the relevant information in original form is increasingly beyond the resources of any single company.

In common with many companies, ICI has considered the use of the various tape services offered by Chemical Abstracts Service and, in particular, *CA Condensates*. The first study was of the use of Condensates for retrospective search, and the problems of developing a computer system for in-house use were considered. These are complex, mainly because of the size of the data base, and it was felt that experience with Condensates for SDI using existing programs was necessary before implementing a system capable of retrospective search. The programs written by the National Research Council of Canada were chosen for this service; the NRC SDI system has been described elsewhere.[2]

However, the *CA Condensates* tapes could be processed relatively simply to provide, every few months, a cumulative index to ease the problems of retrospective search. The development of this tool and its use are described here.

The problems of searching *Chemical Abstracts* itself are well-known. The CA subject indexes are admirable for those volumes for which they have been issued, but real difficulties arise with the more recent volumes.[3] These are usually only approachable through the keyword index which was introduced in 1963.[4] The index is printed at the back of each issue and is an alphabetic list of keyword phrases, with each abstract being indexed by, on average, five phrases. The phrases generally consist of three or four words, all significant, which are permuted so that they all appear as indexing points. For example: the entries for Abstract 41444Q [Vol. 75 (6)], of which the title is "Transitions in Phases II-III-IV in high purity ammonium nitrate," are as follows:

Ammonium Nitrate Phase Transition
Nitrate Ammonium Phase Transition
Phase Transition Ammonium Nitrate
Transition Phase Ammonium Nitrate

Standardization in the keyword phrases is not achieved; the printing is small and constant use tends to strain the eyes. The delay in the issue of subject indexes at the time of the study mentioned above was averaging 18 months, which meant that a full search involved scanning the keyword indexes in some 78 issues. This task is so daunting that many people, particularly practising scientists, tend not to search the recent issues, thus missing the most up-to-date information.

For these reasons, therefore, ICI decided to provide cumulative indexes. The various alternatives considered during the development are discussed later. The result, however, is an index, based on the keyword phrases, arranged alphabetically, produced for 13 consecutive issues and stored on microfilm.

Each entry consists of:
Keyword Phrase (the first 80 characters)
Abstract Number
An Indicator for Patents
Title (the first 100 characters)
Specimen entries are shown in Figure 1.
The index is in upper/lower case format.

In addition to the keyword phrase index, a similarly formatted index based on patent assignees is produced. The subject index is produced on four spools of microfilm (equivalent to some 14,000 pages of computer printout), with a fifth spool holding the patent assignee index.

The *CA Condensates* records contain all the information needed to create the entries for each CA issue, and the index is produced by converting the weekly tapes into the index format. At the end of the quarter, these are