Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 4. Concise Connection Tables to Structure Diagrams

D. I. COOKE-FOX, G. H. KIRBY,* M. R. LORD, and J. D. RAYNER

Department of Computer Science, University of Hull, Hull HU6 7RX, England

Received October 12, 1989

The concise connection table (CCT) is the structure representation into which IUPAC systematic organic chemical nomenclature is translated by the methods described in previous papers in this series. Here, first, some enhancements to the original specification of the CCT are presented, resulting from experience of its use and extension of the functional groups and classes of names handled by the nomenclature translator. Then, software for the expansion of the CCT into an atom table is described which enables the structure to be displayed by character graphics. The conversion of the CCT into the Standard Molecular Data (SMD) format is reported, and this has enabled the molecular structure corresponding to input names to be displayed by a variety of graphics packages. The problems of linking the nomenclature translator and commercially available molecular graphics and database packages into one multiprocess software system under MSDOS on IBM PCs and compatibles are discussed.

INTRODUCTION

Previous papers in this series[1-3] have described the development of a formal grammar for IUPAC systematic organic chemical nomenclature and the production of a parser program to recognize valid names. The parser program also does semantic processing and generates a concise connection table (CCT) as a representation of the molecular structure of the corresponding name. The CCT was created explicitly for representing the structure of nomenclatural fragments.[4]

The current state of development of the CCT is outlined in the following section since a knowledge of the extensions is necessary in understanding how a structure diagram is produced from a CCT. The CCT can be converted into full atom-by-atom connection tables which provide a link not just into molecular graphics programs but also into a wide range of software requiring a description of chemical structure as input.

The production and display of molecular structure diagrams from CCTs completes the translation path from names to diagrams first discussed over 25 years ago by Garfield and republished in 1985.[5] Alternative approaches to this path were addressed in paper 1,[1] and quite recently a further development has resulted in software that recognizes Beilstein nomenclature and generates corresponding structure diagrams.[6]

ENHANCEMENTS TO THE CONCISE CONNECTION TABLE

Since the publication of the original specification of the CCT[4] a number of extensions and enhancements have been made as a result of several years' experience in its use. The chief motivation behind the development of the CCT was to provide an internal representation of chemical structure corresponding to the semantic decomposition of systematic nomenclature. The CCT is a linear table which represents implicitly the hierarchic structure of a molecule, as determined from the IUPAC nomenclature. Each CCT entry has four nonnegative integer fields, LOCT, TIPE, SIZE, and SUBS.

Essentially a nonzero LOCT field shows where a given substructure, described by this CCT main entry, is attached to its parent (sub)structure, described earlier in the CCT. TIPE 0 denotes an aromatic ring, 1 a chain, 2 a single atom, and 3 an alicyclic system, and SIZE gives the ring system or chain size or atomic number, respectively. Ring systems are fully described by additional ring segment entries for each ring component. SUBS indicates the number of substituents on the (sub)structure, these being described in subsequent CCT entries arranged in an implicit hierarchic order. A CCT entry with zero in the LOCT field is either a special entry having TIPE 0 or a modification entry where TIPE 1 indicates an electronic charge of value SIZE minus eight, TIPE 2 indicates a replacement heteroatom determined by SIZE, TIPE 3 represents a nondefault bond coded by convention in SIZE, and SUBS gives the locant of modification to the parent structure.

The principle underlying the CCT has been maintained in the introduction of additional CCT entry types for functional groups and steroid nomenclature. Further CCT modification entry types have been defined to represent certain stereochemical bond orientations and to improve the representation of multiple bonds in the bridges of polycyclic ring systems.

Special Entry for Functional Groups (SEFGs). The original CCT definition contained a number of points where extensions to the scheme could be introduced. One such was the area of special entries, characterized by LOCT and TIPE fields both having the value zero. To allow the simplified display of functional groups as conventional line formula segments with explicitly stated hydrogens [e.g., -COOH rather than -C(OH)=O or similar alternatives], the common groups are associated with a nonzero integer value held in the SIZE field of the SEFG (Table I). The locant at which the functional group is attached to its parent structure is then recorded in the SUBS (substitution) field, following the CCT convention of using this field as an alternative to LOCT in appropriate specific circumstances.

For example, from Table I it can be seen that decane-1,4-diol should now be represented as

LOCT	TIPE	SIZE	SUBS
1	1	10	2
0	0	13	1
0	0	13	4

rather than

LOCT	TIPE	SIZE	SUBS
1	1	10	2
1	2	8	0
4	2	8	0

Similarly, *m*-nitrotoluene now has the CCT

LOCT	TIPE	SIZE	SUBS
1	0	1	2
1	0	6	0
1	1	1	0
0	0	20	3

0095-2338/90/1630-0122$02.50/0

rather than

LOCT	TIPE	SIZE	SUBS
1	0	1	2
1	0	6	0
1	1	1	0
3	2	7	2
1	2	8	1
0	3	10	0
1	2	8	1
0	3	10	0

This latter example illustrates the increased brevity provided in the concise connection table through the use of SEFGs, including the avoidance of explicit description of the sometimes uncertain, conjugated bonding arrangements within functional groups.

Steroids and Stereochemical Bond Orientation. In extending our grammar of IUPAC organic nomenclature[2] to include aspects of steroid nomenclature, one consideration was the common feature of the steroid nucleus ring system and its particular locant-numbering arrangement. To preserve this high-level complexity within the CCT scheme, a new TIPE 5 main entry has been introduced which leads into a substantial area of extension.

A detailed discussion covering all aspects of steroid nomenclature handling from grammar development through structure diagram display will be presented in the next paper.[7] In brief, however, the TIPE 5 main entry (or steroid header entry; the digit 5 is mnemonically reminiscent of S for steroid) introduces a number of further entries (counted in the steroid header SIZE field) that describe details of the stereochemistry of the ring system and its standard chain substituents. The sizes of the four rings A, B, C, D of the steroid nucleus (typically 6 6 6 5) are given in the four fields of a partner CCT entry which always follows immediately the TIPE 5 main (steroid header) entry.

The introduction of facilities for handling steroids has led to the definition of new modification entry codes for bond stereochemistry. The original CCT definition provided for bonds other than those assumed by default according to the TIPE of a substructure, by means of the TIPE 3 modification entry. This used the SIZE field to denote a bond type by conventional coding, which has now been extended (Table II) to include designations of α, β, and ξ bonds as interpreted in the steroid context. Further codings are being considered to handle more general stereochemical features, by use of presently unallocated code values.

Polycycle Bridge Unsaturations. Ring systems are described in detail by ring segment entries (RSE) which follow a TIPE 0 or TIPE 3 main (ring header) entry. A bridge in a polycycle is represented by an RSE of TIPE 1, whose SIZE field contains the number of atoms in the bridge, which may be zero. The two points of attachment of the bridge segment to its parent system are given by the LOCT and SUBS fields, the LOCT field containing the lower valued locant and SUBS containing the offset (numeric difference) between that and the higher valued locant. (RSEs appear in the linear sequence of the CCT in order of ascending final locants, so that each new segment of the structure may be placed relative to atoms already introduced, and the locants ascribed to the atoms of the new segment are correct in the context of the final overall structure.)

As with aliphatic chains, the default bond type in a TIPE 3 ring system is covalent single, and variations from this are represented through TIPE 3 modification entries. Usually, the lower of two adjacent-valued locants of a bond modification is given in the SUBS field of the modification entry, and a SUBS locant value of zero is used to indicate a bond modification in the attachment of a substructure to its parent.

Table I. SIZE Codes for Functional Group Special Entries (SEFGs)

SEFG SIZE Code	nomenclature	line formula
1	amino	$-NH_2$
2	azido	$-N_3$
3	carbaldehyde	-CHO
4	carboxylic acid	-COOH
5	chlorosyl	-ClO
6	chloryl	$-ClO_2$
7	cyanato	-OCN
8	cyano	-CN
9	dihydroxyiodo	$-I(OH)_2$
10	dithiocarboxy	-CSSH
11	dithiosulfo	$-S_3OH$
12	hydroperoxy	-OOH
13	hydroxy	-OH
14	iodosyl	-IO
15	iodyl	$-IO_2$
16	isocyanato	-NCO
17	isocyano	-NC
18	isothiocyanato	-NCS
19	mercapto	-SH
20	nitro	$-NO_2$
21	nitroso	-NO
22	pentafluorothio	$-SF_5$
23	perchloryl	$-ClO_3$
24	seleneno	-SeOH
25	selenino	$-SeO_2H$
26	selenono	$-SeO_3H$
27	sulfino	$-SO_2H$
28	sulfo	$-SO_3H$
29	thioaldehyde	-CHS
30	thiocarboxy	-CSOH
31	thiol	-SH
32	selenol	-SeH

Table II. TIPE 3 Modification Entry Bond Codes

SIZE code	bond type	SIZE code	bond type
0	ξ	8	aromatic
1	α	9	single
2	β	10	double
6	polycycle bridgehead	11	triple
7	dative		

However, in the case of polycycle bridges, the locants of the bridgehead and the bridge end atoms are generally nonadjacent numerically and an alternative mechanism is necessary for these (relatively rare) situations.

The mechanism is indicated by the use of bond modification code 6 (see Table II) in the case of a double bond from bridge to bridgehead, and the SUBS field contains zero. An additional CCT entry then follows to give *both* locants involved, in the LOCT and SUBS fields, with the TIPE and SIZE fields unused and set to zero. This method represents a further instance of the CCT philosophy whereby common situations are represented in a few fields, with default assumptions where necessary. Relatively rarer or more complex situations use progressively more space and are introduced by specific "flag values" in designated fields of the simpler formats.

CONVERSION TO ATOM-BY-ATOM TABLES

In contrast with the CCT, most other connection table schemes are fully explicit in listing all the (non-hydrogen) atoms and bonds in an organic compound, with connectivity information for each atom.[8] Such connection tables are much used in chemical information systems for structure storage and retrieval.[8,9] Another use is as the starting point for the display of the molecule, in which case the coordinates of each atom may also be held in the table.[10] However, there are other situations in which a fully explicit connection table is not necessary and alternative concise schemes and notations have

been proposed.[11] Thus, it is the use for which a particular connection table is required that determines its complexity.[8]

Software for the expansion of the CCT into a full atom-by-atom connection table was therefore desirable, in particular to aid the eventual display of structure diagrams but also to permit the interfacing of the nomenclature translation software to a variety of chemical information systems that take connection table input.

Atom Table. A linked-list data structure has been implemented to store an atom table containing information about the non-hydrogen atoms within a molecule in the locant order resulting from the sequential processing of the CCT.

The topology of an organic molecule is predictable from the TIPE field of the parent structure in the CCT, namely, aliphatic chain, aromatic or alicyclic ring system, or individual atom. Procedures have been written to expand each of these parent structures. Thus, for aliphatic chains the SIZE field gives the length of the chain, enabling entry of the appropriate number of carbon atoms into the atom table. Display coordinates for each atom can also be generated, if required, from the positions of the preceding atoms according to the context inferred from the CCT. Initially, it is assumed that all bonds are single bonds.

For aromatic ring systems, each ring fusion, detected when a further ring segment entry has to be processed, requires an amendment to connectivity information for two atoms already in the atom table. The LOCT field for the additional RSE identifies the first of these, relative to the first atom of the whole ring system (i.e., the current parent). New atoms are inserted for the additional ring segment, taking account of the orientation of the fusion position, until the second fusion atom is reached, as indicated through the SUBS field. From the information given in the CCT, it is not possible to decipher directly the double-bond coordination in an aromatic system, so all the bonds are identified as aromatic.

Alicyclic ring system entries are similarly processed by using the SIZE field for overall size and LOCT and SUBS fields to detect any spiro attachments or bridges, as described previously.

The nature of an individual atom parent is determined from its SIZE field, which gives the atomic number. Coordinates can be computed by assumptions of standard bond lengths and angles.

The sequential processing of the CCT means that substituents and/or modifications to a parent structure are dealt with later, requiring changes to connectivities, bond, and atom types for atoms already inserted in the atom table. Before terminating, each parent structure procedure makes calls to a further, general procedure to process any substituents and modifications on that parent structure as indicated in the SUBS field of the parent entry. Parameters passed at each call include the locant of the substituent or modification, known from the LOCT field of the corresponding CCT entry and the pointer to the first atom of the current parent in the partially constructed atom table.

The substituents themselves are handled by a general substituents and modifications procedure with calls of each appropriate parent structure procedure to enter further atoms in the table, according to the TIPE field of the substituent CCT entry. These procedures are therefore mutually recursive through the general substituent procedure, which enables the substituents-on-substituents situation to be handled correctly.

Modifications to bonds or atoms (i.e., heteroatoms) and special entry functional groups, whose structure is known and cannot be substituted or modified, are dealt with by further separate procedures.

Standard Molecular Data (SMD). The SMD format has resulted from the Computer Assisted Synthesis Planning (CASP) project run by a consortium of seven German and Swiss chemical companies.[12] It was developed to provide a common interface for the exchange of molecular data between various chemically orientated systems rather than to provide a means of permanent storage.

An SMD file is a sequential ASCII text file containing one or more SMD structures. Each SMD structure contains all the relevant information for a specific compound or reaction and is divided into information (or main) blocks. Each block contains the relevant data for a particular property associated with the compound or reaction. Blocks are designated for structure name, connection table (CT), Cartesian coordinates (CO) and lables (LB), among others. It is the responsibility of the computer program, taking an SMD file as input, to select only those information blocks that are necessary for its operation.

Hierarchic structuring of certain information blocks is provided by the use of subblocks (in the CT, CO, and LB blocks) and superatoms. A superatom can be used, for example, to represent a fragment that occurs more than once within a molecule, thus eliminating the need for repetition. Data records are used to provide explicit information and are located at the lowest level in the information hierarchy.

We have developed software to produce the SMD file format directly from the CCT. The atom table described previously was designed for the display of structures by use of our own character graphics technique (see next section). Hence, it contains information specific to that task and is not easily transformed to the SMD format.

Once a valid CCT is produced by nomenclature translation, the SMD procedures can be called to create an SMD file containing the following blocks:

DTCR	creation date and time of the structure file
CT	connection table of the structure
CO	corresponding coordinates
FORM	empirical formula
NAME	name of compound

It is the CT block that is generated from the CCT and is described here. At present, the CO block contains character-based coordinates extracted from the atom table, although, in time, these will be replaced with display coordinates computed by other procedures.

Two linked-list data structures, namely, the atom and bond lists, were designed to allow construction and modification of the atom and bond entries during the sequential processing of the CCT. Each atom and bond entry consists of records that contain a separate field for every SMD CT field within that entry, plus pointers to the previous and next records to allow bottom-up and top-down data structure manipulation, respectively. In addition, an atom record contains two extra fields, atom position and locant position, which contain the atom number and its corresponding locant number and are only used in the construction of the CT block. As with the atom table, the TIPE field of the parent structure in the CCT results in an appropriate procedure being executed to handle the relevant parent.

For an aliphatic chain, the SIZE field of the first CCT entry is interpreted as the length of the chain. The numbers of atom and bond entries are equal to the chain length and one less than this, respectively.

With an aromatic ring system, the SIZE field of the first CCT entry is interpreted as the number of rings. The first CCT ring segment entry (RSE) is examined and the content of the SIZE field used to construct the atom and bond entries for a ring of that size. This initial ring is handled differently from any further rings on the parent since it has the same number of atom and bond entries, which are equal to its size.

A procedure to deal with fused rings is executed for each of the additional rings on the parent in turn. The LOCT field of a fused ring's RSE contains the fusion position to the previous ring system. The atom entry that contains this first fusion position is located, and the implied hydrogen field is set to zero for this entry and the following one. The linked-list of atoms is then broken between these two entries, the extra entries are inserted for the new ring and the links restored. The way in which this is done depends upon whether the aromatic ring system is linear or nonlinear.

The bond entries for the new ring are then added to the end of the bond list, after slight modification to the last entry. Once the last fused ring has been handled in this way, the fusion bonds are added to the bond list by searching through the atom list for those entries with zero implied hydrogens. Finally, the atom position and locant position fields in the atom list entries are updated, resulting in all fusion positions having a zero in the latter field.

For alicyclic ring systems, the SIZE field of the first CCT entry again gives the number of rings. The SIZE field of the next CCT RSE is used to construct atom and bond list entries for a ring of that size, the number of implied hydrogens per atom being two. The TIPE field of the third and subsequent CCT entries indicates the nature of additional rings as spiro or bridged. A spiro ring is indicated by a TIPE 3 RSE, whereas bridges are represented by TIPE 2 chain ring segments (CRS). For spiro compounds, the atom and bond list entries are derived from the remaining RSEs as a whole, rather than ring by ring as in aromatic systems. For bridged systems, the number of extra carbon atoms to be added to the atom list is obtained by summing the number of atoms in each of the bridges, as given by the SIZE fields in each CCT CRS. Extra bonds are added to the bond list to describe the connectivity between the atoms within a bridge and with their corresponding bridgeheads.

Steroids are handled differently, and this will be presented in the next paper of this series.[7]

As with the atom table, each parent structure procedure makes a call to another procedure to process any substituents and/or modifications on the particular parent before terminating. Each substituent or modification is handled in turn, and a relevant procedure is called to add to and/or modify the existing atom and bond lists. Complex substituents are dealt with by recursive calls to the appropriate procedure.

The molecular fragments, known as SEFGs, are represented in SMD format by a novel use of superatom entries. We assign a unique superatom symbol to each SEFG comprising a lower case s, e, f, or g plus a single digit. Thus, the 32 SEFGs in Table I are represented by the symbols

s1,s2,—,s9, e0,e1,—,e9, f0,f1,—,f9, g0,g1,g2

Alongside each superatom symbol in the atom list is given a four-character subblock name, which identifies the later explicit description of the superatom structure. The superatom subblock follows the atom list and should normally contain the resolved connection table of the superatom. However, for SEFGs, the subblock is in fact empty; that is, it comprises only a heading record. The reason for this is that SEFGs are represented as text strings on the structure diagram, rather than by their full molecular structure. Consequently, the SEFG subblock name, with the extension string field if necessary, can itself be the relevant fragment string, e.g., COOH, thus eliminating the need for an LB block entry to represent the label of the superatom.

Once the CCT has been fully processed, the atom and bond CT lists are traversed in a top-down manner with the relevant entry details being output to a file in SMD format. This file can then be picked up by any other program requiring SMD input.

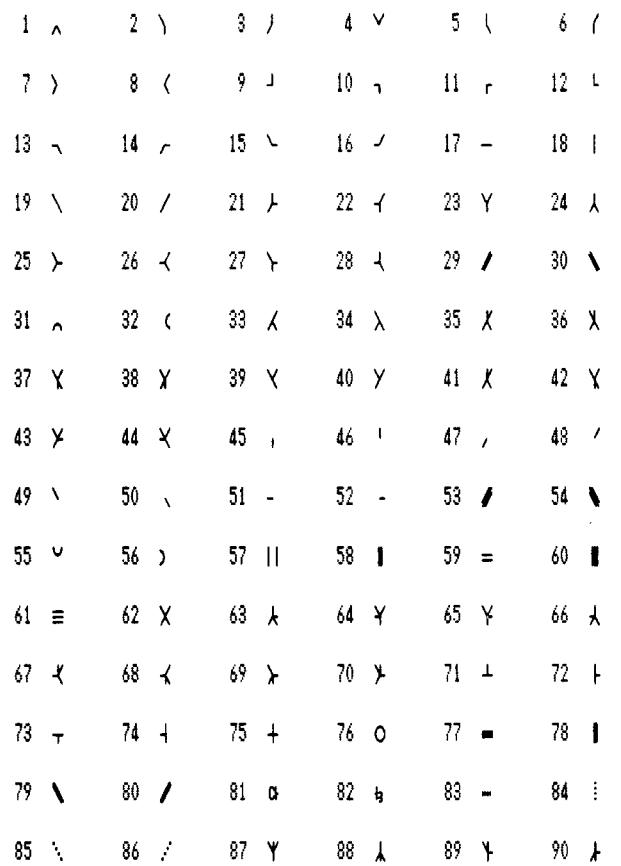

Figure 1. Graphic character set used for the display of structure diagrams.

STRUCTURE DIAGRAM DISPLAY

To confirm that a name corresponds to an expected structure, the display must be clear and unambiguous but need not have the sophistication provided by many current commercially available interactive molecular graphics programs. The two-dimensional unscaled structure diagrams used in organic chemistry textbooks are adequate, and we originally developed our own character graphics software to produce such displays. However, now that the IBM PC has become a chemical industry standard, there are many software packages available for the display, storage, manipulation, and searching of chemical structures.[13] As Warr[14] points out, the future for these various products lies in integration. However, the full integration of software processes in the PC environment remains constrained at present by inadequate systems support for multiple operation of such relatively large packages. These operational problems are discussed in the next section. The development of the CCT-to-SMD conversion software provides a standard data transfer medium that allows the nomenclature translator to provide structural data input to other chemically oriented software packages. This approach has been demonstrated for alternative molecular graphics packages to our own.

Character Graphics. For ring sizes from 4 to 8 carbon atoms and with aliphatic chains drawn as zigzag lines, short straight single or multiple lines of a few different lengths in a limited number of orientations are all that is needed to display the carbon skeleton of a conventional structure diagram with hydrogens assumed. A simple way of providing this on a wide range of microcomputers and printers is by use of a special chemical graphics character set or font. Other atoms and SEFGs can be shown explicitly by atomic symbols using the normal ASCII character set.

A minimal set of some 36 graphic characters initially used in this project for simple rings and chains was published[15] in 1983. This set has been considerably extended into a font of

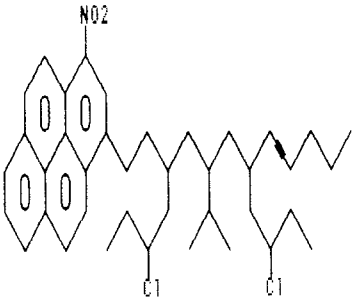

Figure 2. Structure diagram drawn by character graphics.

some 90 characters, shown in Figure 1, that permits a wide range of complex structure diagrams to be displayed. On the IBM PC, video displays are produced by two fundamently different modes, called text and graphics by IBM. Graphics mode is mainly used for complex drawings, but it can be used to display characters as well.[16] In graphics mode the extended ASCII character set (codes 128–255) is located by the PC via the 1FH interrupt vector. By creation of an array of bytes containing the definition of our chemical graphics character set and placement of the start address of this array in the 1FH interrupt vector, the extended set is replaced by the chemical one.

Entries in the atom table are associated by software with corresponding characters from either the graphic or normal character set. The diagram is constructed first in a screen array, an internal representation of the screen. The screen coordinates for each character are computed and stored in the atom table, from knowledge of the nature and position of the preceding character drawn and the context (i.e., chain, ring, atom).

The coordinates of the first character to be drawn are chosen to allow the diagram to expand in all directions. If, however, the structure diagram attempts to pass a boundary of the screen array, then the starting position is recalculated to allow for the greatest expansion in this direction and the diagram reentered into the screen array. A maximum of three reentries is allowed before the user is told that the structure diagram is too large for the screen.

The screen resolution is limited to 80 × 25 characters, and overlapping of characters cannot be allowed. The zigzag aliphatic chain substituents cause most problems with overlap of existing screen characters. Hence, a table of alternative character codes is used by the character graphics software to attempt to redraw the offending substituent in different directions or orientations. Alternatives are tried from the table until the current extent of the structure can be drawn without character overlap and without exceeding the screen size or until there are no more alternatives.

This redrawing is quite successful in enabling the display of structure diagrams that otherwise could not be represented, but there are cases where it is awkward to draw structures without overlap. Figure 2 is the structure diagram finally drawn by the character graphics software after the name 3-[3,7-bis(2-chlorobutyl)-5-isopropyldodec-8-ynyl]-1-nitropyrene has been parsed and processed semantically. Initially, its parent rings and main chain are constructed with the chlorobutyl substituent on locant 3 then drawn down and to the right of its attachment atom. However, this blocks some of the character positions required next for the isopropyl group on locant 5. In this case, the software finds that the chlorobutyl group can be redrawn to the left, as seen in Figure 2, which resolves the conflict.

It is hard to resolve problems of character overlap involving ring systems as substituents, since their rigid internal structure makes reorientation difficult within the partially constructed 2D display. SEFGs with several characters can be redrawn by reversing the order of the characters, e.g., from -COOH to HOOC-.

MOLIDEA. MOLIDEA[17] is a program for the IBM PC that automatically calculates atomic Cartesian coordinates from a molecular description and displays the 3D structure on the color graphics screen. The molecular structure can be displayed either as a wire-frame or a space-filling model, which can subsequently be rotated around the x, y, and z axes or around a selected bond. This package is intended for molecular modeling applications rather than structure diagram display. As such, it complements the other packages we are using.

The main advantage of this system is that coordinates do not need to be provided by the nomenclature translator. The molecular description required by MOLIDEA has been derived from the SMD CT block produced by the CCT-to-SMD conversion. Through collaboration with CompuDrug Ltd. a two-process system has been developed that can repeatedly read in a name and produce a MOLIDEA display of the corresponding structure, the molecular description being passed from the nomenclature translation process to the MOLIDEA process in a file. All the display facilities of the MOLIDEA system are available for manipulation of such displays.

PCMODEL and MOLGRAF. Both of these packages[18,19] allow the drawing and manipulation of 3D structures in graphics form on IBM PCs and compatibles. They are similar to MOLIDEA except that they require either Cartesian or X-ray coordinates. Neither package actually reads an SMD file, but their corresponding connection table inputs are very similar to SMD file format and an interface to perform the necessary conversion has been implemented.

PSIDOM. PSIDOM[20] (Professional Structure Image Database on Microcomputers) is a package for IBM PCs and compatibles providing connection table based input, storage, retrieval, and display of chemical structures. With assistance from Hampden Data Services Ltd., we developed another two-process system that translates names into SMD format and drops this into a file for reading by the PSIDOM package. All the facilities of PSIDOM are available once the name has been translated, including display of the corresponding 2D structure diagram, storage within a PSIDOM database, and modification of the structure followed by restorage in SMD format if required.

SYSTEMS CONSIDERATIONS

The nomenclature translation software forms a large program (14 000 lines of source code) which is structured as several modules for TurboPascal version 5.0 on an IBM PC compatible. Problems exist in linking this program with a variety of graphics, or other, packages to produce one multiprocess software system. These result from the size of the programs concerned, the use of different compilers for each package, limitations imposed by the MSDOS operating system used on IBM PCs and compatibles, and being unable to change the point of entry to packages for which only executable code is available.

It is possible under MSDOS to combine several processes by loading the first, dropping output into a predetermined file, halting and loading, or reloading, the next process, and so on. However, reloading the nomenclature translation software, in particular, is a lengthy process since the fragment dictionary and parse tables have to be copied back into main memory each time. This problem has been overcome with the DESQview multitasking extension to MSDOS.[21]

DESQview is essentially a resource manager for a PC. The nomenclature translation and other required software packages are loaded into memory. If there is insufficient memory to load a new process, DESQview has the capability to switch a previous process out to either expanded memory, a disk, or

RAM disk cache, by memory paging. Once the nomenclature translation software has successfully processed a particular name, a desired display package can be chosen by the user via a key which activates an operating system script, or macro, set up by the nomenclature translation program. This ensures that the requested process is (re)activated. The user can similarly return to the nomenclature translation software when appropriate.

The use of expanded memory or disk cache greatly reduces the transfer time between selected processes. However, each process usually still has to be restarted from the first initialization phase of that process, requiring the user repeatedly to negotiate menu structures, for example, before activating the required package feature. There is typically no means of selecting a particular display and maintaining that choice between activations of the process. Thus, it is not generally possible to construct a composite system based on the integration of separate proprietary packages. This will remain the case until such packages permit multiple entry and exit points for use in conjunction with multitasking operating systems.

DISCUSSION AND CONCLUSION

The concise connection table was originally devised to allow representation of structural fragments within a dictionary of systematic nomenclature morphemes. The CCT has been expanded in line with increasing scope of the Hull nomenclature translator, and the implicit hierarchic structure of the CCT has proved to be easily enhanced to support a variety of unforeseen extensions. The use of the CCT remains an important pivotal technique in the conversion of molecular structural information from systematic nomenclatural representations into alternative connection tables, for storage, retrieval, and graphical display.

A range of techniques and software systems exist for the graphical display of molecular structures, with a variety of data interface mechanisms. The CCT as output from the translator is potentially a further interface, but the need for a commonly agreed standard for communicating structural representations between different software modules is apparent. The Standard Molecular Data (SMD) proposal is welcome as a potential answer, and conversion of structures from CCT to SMD representations now gives a route from nomenclature into an increasing number of other software packages. In particular, a variety of alternative structure display modules may be entered, affording a choice of display styles for graphic output.

Problems have been encountered in the systems linkage between different proprietary software modules in the PC environment. The recent availability of resource manager operating systems for PCs now enables the prototyping of a multiprocess software system using sequences of processes of disparate origin, without any requirement to access protected source code. This is particularly necessary to enable the nomenclature translator to interface with other chemically oriented software packages, but further developments in systems support facilities will be necessary before an acceptable user interaction can be achieved.

ACKNOWLEDGMENT

We are pleased to acknowledge the funding provided by the U.K. Laboratory of the Government Chemist for this project. We are grateful to CompuDrug Ltd., Budapest, Hungary, and Hampden Data Services, Oxford, England, for their assistance in linking their software products with our nomenclature translator.

REFERENCES AND NOTES

(1) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 1. Introduction and Background to a Grammar-Based Approach. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 101–106.
(2) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 2. Development of a Formal Grammar. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 106–112.
(3) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 3. Syntax Analysis and Semantic Processing. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 112–118.
(4) Rayner, J. D. A Concise Connection Table Based on Systematic Nomenclatural Terms. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 108–111.
(5) Garfield, E. An Algorithm for Translating Chemical Names to Molecular Formulas. In *The Awards of Science and Other Essays*; ISI Press: Philadelphia, 1985; p 453.
(6) Goebels, L. The Role of Beilstein Today. Presented at the Conference on Chemical Nomenclature into the Next Millenium—Has It a Role?, London, Nov 12/13, 1987.
(7) Cooke-Fox, D. I.; Kirby, G. H.; Lord, M. R.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 5. Steroids. *J. Chem. Inf. Comput. Sci.* (following paper in this issue).
(8) Ash, J. E. Connection Tables and Their Role in a System. In *Chemical Information Systems*; Ash, J. E., Hyde, E., Eds.; Ellis Horwood: Chichester, U.K., 1974; Chapter 11.
(9) Ash, J. E.; Chubb, P. A.; Ward, S. E.; Welford, S. M.; Willett, P. *Communication, Storage and Retrieval of Chemical Information*; Ellis Horwood: Chichester, U.K., 1985; p 141.
(10) Dittmar, P. G.; Mockus, J.; Couvreur, K. M. An Algorithmic Computer Graphics Program for Generating Chemical Structure Diagrams. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 186–192.
(11) Barnard, J. M.; Jochum, C. J.; Welford, S. M. ROSDAL: A Universal Structure/Substructure Representation for PC–Host Communication. In *Chemical Structure Information Systems. Interfaces, Communications and Standards*; Warr, W. A., Ed.; ACS Symposium Series 400; American Chemical Society: Washington, DC, 1989; pp 76–81.
(12) Bebak, H., et al. The Standard Molecular Data Format (SMD Format) as an Integration Tool in Computer Chemistry. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 1–5.
(13) Meyer, D. E.; Warr, W. A.; Love, R. A. *Chemical Structure Software for Personal Computers*; ACS Professional Reference Book; American Chemical Society: Washington, DC, 1988.
(14) Warr, W. A. *Graphics for Chemical Structures. Integration with Text and Data*; ACS Symposium Series 341; American Chemical Society: Washington DC, 1987; p ix.
(15) Rayner, J. D.; Milward, S.; Kirby, G. H. A Character Set for Molecular Structure Display. *J. Mol. Graphics* **1983**, *1*, 107–110.
(16) Norton, P. *The Programmers Guide to the IBM PC*; Microsoft Press: Washington, DC, 1985; p 68.
(17) Lopata, S.; Gabanyi, Z.; Bencze, A. MOLIDEA Version 2.0, 1988, available from CompuDrug Ltd., Furst Sandor Utca 5, H-1136 Budapest, Hungary.
(18) Henkel, J. G.; Clarke, F. H. *PCMODEL Molecular Graphics on the IBM PC Microcomputer, Enhanced Version 2.0*; Academic Press: London, 1986.
(19) Barlow, R. B.; O'Donnell, C. H. *MOLGRAF Molecular Graphics for Microcomputers*; Elsevier-BIOSOFT: Cambridge, U.K., 1986.
(20) Hampden Data Services Ltd., PSIDOM Release 5.0, Oxford, U.K., 1987.
(21) Quarterdeck Office Systems, DESQview Version 2, Santa Monica, CA, 1987.