Systematic Chemical Nomenclatures in the Computer Age

G. H. Kirby* and D. J. Polton

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

Received October 7, 1992

This paper draws upon experience in the development of computer software which generates and analyzes systematic organic chemical nomenclatures and highlights some general points where these nomenclatures cause difficulties in computer processing. It is suggested that the tackling of these points would, through consequent simplification, tightening, and improved consistency of the rules, help scientists and students in their use of nomenclature both by manual means and by computer. This might gradually lead to the development of a new systematic nomenclature, formally defined, which could co-exist for some time with current nomenclatures, supported by software to provide interconversion between it and them.

INTRODUCTION

Only quite recently has progress been made, other than for indexing purposes, in computer recognition and translation of general systematic organic chemical nomenclatures.[1-5] That it has taken so long to be able to recognize and generate systematic nomenclatures and to interconvert them with structural representations results from the many problems encountered in analyzing and generating nomenclatures by computer. This has led us to consider whether it is time for nomenclature to respond to the needs of computer processing through simplification, consistency, and tightening of the rules. This would have the added benefit that scientists could more readily form, understand, and use systematic nomenclature both with and without computer systems.

The principal use of chemical nomenclature is to give a compound a label that can be spoken, written, and used in printed indexes and from which the structure can be perceived by scientists. While trivial nomenclature has the benefit of conciseness, only systematic nomenclature, which to a certain extent gives pronunciation and semantics to a structure, is of use for unambiguously labeling a structure with a name that can safely be communicated worldwide. The compilation by Lees and Smith[6] of the papers presented at a symposium held in 1981 to discuss the use and the problems of nomenclature, and to help overcome the confusion and misunderstandings that existed, remains a valuable record of many of the issues. Typical of these is that IUPAC systematic organic nomenclature[7] has some similarities to a natural language[8] in that it has evolved. So it includes some commonly used trivial names, contains redundant information, allows a structure to have more than one name, and requires expertise in applying the rules to determine the name for a given structure.

As mechanization was applied to chemical information, stricter linear representations, or notations, of structures were developed. Such notations represent structures in coded form using symbols which consist to a greater or lesser extent of the ordinary characters for the chemical elements accompanied by other alphanumeric and special characters. Examples include the Wiswesser line notation (WLN)[9] and the Dyson-IUPAC notation.[10] These chemical notations were readily adapted to computer use: registration and substructure search systems based on them were developed many years ago.[11] More recently, systems have been developed which give rise to a unique line formula for each structure, sometimes also called notations, without any characters specific to that representation. These use only elemental symbols, numerals, and necessary characters such as parentheses to identify the full structure. Examples are SMILES[12] and ROSDAL.[13]

Notations, while more suited than nomenclatures for computer processing, have not succeeded in replacing the role of nomenclature in spoken and written communication and in education. Accepting then that systematic nomenclature has a future, we suggest that it is time for chemical nomenclatures to take account of the computer age. Possibilities might be the co-existence of a notation, for computer applications, and a nomenclature, for communication between scientists, as was the Dyson-IUPAC notation and IUPAC nomenclature, or the development of a new nomenclature of relatively simple construction with a condensed equivalent for computer use, as is the new nomenclature of Dyson.[14] Such a solution could co-exist with nomenclature in current use in the hope that the virtue of simplicity and the availability of computer software would eventually result in gradual acceptance into general use. It would be essential that any new nomenclature must be interconvertible by computer with existing nomenclatures.[5]

The object of this paper is to collect together, from our experience, the problems in computer processing of some current nomenclatures for organic compounds including, as well as the aforementioned IUPAC and Dyson systems, the nodal nomenclature originated by Lozac'h (LNN)[15] and part of the HIRN system.[16] There follows an outline of the general points where it is felt that chemical nomenclatures lack the strict formality desirable for computer use, leading to the necessity of repeatedly carrying out procedures which lengthen the time taken in processing. Recommendations are made for some components of a nomenclature that would ease communication between scientists and between scientist and computer.

PROBLEMS WITH CURRENT NOMENCLATURES

Cyclics. It is necessary to distinguish between acyclic and cyclic structures, and this is done by the use of the prefix "cyclo" in most nomenclatures. However, the number of individual rings in a polycycle is also indicated by the use of a prefix (bi, tri, etc.) This information is redundant, as the ring count is inherent in the name, and like group and section multipliers below, can also be regarded as confirmatory. However, complex polycycles may be very difficult to name manually, making computer analysis essential, in which case retention of the ring count in the name serves no useful purpose.

Perhaps the usage in chemical nomenclatures which is the most difficult to compute is the naming of a cycle, and then

0095-2338/93/1633-0560$04.00/0

reducing or enlarging its size by the use of the prefix nor- or homo-, such as exists in the naming of steroids in IUPAC nomenclature. In addition, seco- is used to cut a bond in a ring system, thereby altering entirely the configuration of the polycycle as originally described. Not only does computer derivation of names with such prefixes require complex procedures, manual allocation of the correct name requires the utmost care.

Ring fusion nomenclature methods require a polycycle to be correctly orientated before naming can begin. Perhaps in the simpler polycycles this makes naming easy, although time consuming, for the scientist, but it creates major problems in computing. As computer analysis is essential in determining the correct name for the more complex polycycles, the problems created are of significance. Software to orientate graphically input structures is straightforward, but if input is in the form of a character string, this must be broken down into its individual atoms and a spatial representation of the structure constructed. Such a process is complex and must be executed whenever a polycycle, even the simplest bicyclic structure, is handled.

Acyclics. There is a need here for consistent and logical ordering of chains, locants, noncarbon atoms, and bond types.

Chains and their locants are arranged in current nodal nomenclatures in opposing numerical order. Whereas the chain lengths are entered in decreasing order of length, the locants which precede or follow them are in increasing order. For example, a structure containing two each of 2-methyl and 3-ethyl substituents would have these in the order 2-ethyl 3-ethyl 2-methyl 3-methyl, the size decreasing and the location within the size increasing. Since these, expressed as pairs of numbers, alternate with one another (2–2 2–3 1–2 1–3 size first or 2–2 3–2 2–1 3–1 location first), when deriving names by selecting from alternative enumerations, the process of comparison of entire chain–locant sequences requires repeated switching between ascending order and descending order sorts. It is logical to arrange locants in increasing order, but the only reason for ordering chains in decreasing order of size is because the nomenclature system is based on longest chain substitution. If the reverse were true, priority being given to the shorter chains, the size–locant pairs would be in sequence. The new nodal numbering system[17] referred to later does use such a method, with both locant and size in ascending order, giving the sequence 2–1 2–2 3–1 3–2 for the above example, i.e. 2-methyl 2-ethyl 3-methyl 3-ethyl.

IUPAC nomenclature is alone in ordering chains alphabetically, which puts for example the first six, meth, eth, prop, but, pent, hex, in the length order 4–2–6–1–5–3. Note that whilst names for the alkyl series from five upward are regularly formed from a word that designates a number, the first four in IUPAC nomenclature are irregular. Dyson, in his new nomenclature, regularizes these to mon, di, tri, and tetr and does not use alphabetical order. However, this does not ease computer processing since some look-up table or rule is still needed to associate them with their numeric equivalents.

In the case of alkyl radicals, alphabetical ordering necessitates a sort or a check of the substituents and their associated locants after they have been generated or recognized. The lack of consistency in the IUPAC rules for alphabetic ordering has already been pointed out.[1] Thus, in computer generation of IUPAC names, the merging of two alkyl substituents of the same length must be done after sorting into alphabetic order. 2-Methyl and 4-methyl are sorted and then merged to 2,4-dimethyl because the group multiplier is ignored for purposes of alphabetic ordering. However, a parenthesized branched substituent which begins with a group multiplier, e.g., (2,4-dimethylpentyl), is sorted on the full group beginning with the multiplier (here dimeth...). How difficult this must be for the student learning to formulate correct IUPAC names.

Another case of difference in ordering, and hence of complexity when converting either manually or automatically between nomenclatures, is in the order given to the noncarbon elements. Some orderings are based on alphabetic order of symbol or name, others on the position of the elements in the periodic table. When generating a name by computer, the ideal way would be to arrange the elements in increasing atomic number order, though an element table would need to be represented in the software. This would however create some difficulty in manual name formation.

Finally, the apparently anomalous ordering of the triple bond in most, if not all, nomenclature systems between the single and double bond has been discussed at some length in a recent paper.[18]

Principal Component. Computer studies have indicated that it is advantageous to deal with degree of substitution and bridging rather than size when selecting the principal component of a structure for nomenclature purposes. For example, Davidson[19] has criticized the IUPAC tie-breaker rules for equal length candidate main chains in favor of side-chain complexity minimization. Very recently he has extended this to ring–chain assemblies.[20] The authors of the AUTONOM system[21] for generating IUPAC-compatible names from structural diagrams also draw attention to problems with the IUPAC rules for some highly branched acyclic hydrocarbons and to the complexity of ring system seniority. Furthermore, in our experience, it is better to handle shorter before longer branches and bridges.[5] These aspects of chemical nomenclature have been covered in a recent paper.[17]

Locants. An inconsistency in some nomenclatures, and in some cases within the same nomenclature, is the omission of the 1-locant. In IUPAC and nodal nomenclatures it should only be omitted if any other locant would be meaningless or incorrect. However, a few names where the 1-locant is understood are permitted, like cyclohexene. In LNN, the 1-locant is included in a monoheterocycle, such as 1-oxacyclo-[06]hexane, but it is omitted in the unsaturated monocycle cyclo[06]hexanene. Few extra characters are needed to include the 1-locant. The name achieves uniformity, and software does not need to allow for both possibilities. The early notation systems suffered from omission of the 1-locant, which had to be reinserted before efficient computer systems could be developed.[22]

Locants of suffixes should be in one fixed place in a name, unless that place is already occupied by locants of other endings when the alternative must be clearly defined. The options of the beginning of the name, if it does not cause confusion, or immediately before the suffix, that are permitted in IUPAC nomenclature, add unnecessary uncertainty. The way software has to deal with locants in different positions is seen in names such as 1,3-octadien-5,7-diyne. The "a" is lost from the triple bond ending -adiyne here, the whole becoming -adienyne (Rule A-3.3). Elision of the vowel is similarly applied in LNN where the -ane ending of the saturated name is retained, as in heptane-2,5-diene, but the "e" is lost in heptan-2-ene. The ideal situation for computer processing would be for consistency of name fragments and hence no elision.

Multipliers. Some modern nomenclatures follow the practice of IUPAC nomenclature in using a group multiplier in addition to the locants when describing a multisubstituted structure. Thus, the name 2,4-dimethylpentane shows disub-

stitution by virtue of the two locants. The advantage of the repetition implied by the multiplier is in spoken communication and for computer checking and correction of erroneous nomenclature.[23] The disadvantage is in education and computer generation of a name, where the determination and insertion of the group multiplier is an extra task which, with other similar ones, increases the length, complexity, and execution time for student and software. Thus the multiplier can be regarded as confirmatory or superfluous.

Section multipliers (e.g., bis-) are also used in these nomenclatures which complicate the issue still further and tend to repetitiveness. For instance, 2,4-bis(2,4-dimethylpentyl)cyclo... could just as well be 2,4-(2,4-methylpentyl)cyclo... without any loss of information.

Alternative Names. Historically, nomenclatures have been expected to accomplish many objectives.[6] One-to-one correspondence between a structure and a name has been but one of these. In our view, rules that allow alternative ways of naming a structure defeat the principal objective of a systematic chemical nomenclature, which should be to give a unique name to each structure. Yet this is a feature of many systems of systematic nomenclature. A good example from IUPAC nomenclature is that two rings of the same size joined by a single bond may be named with the hydrocarbon ending "ane" or the radical ending "yl". So both 1,2′-binaphthyl and 1,2′-binaphthalene are permitted names for the same structure. In LNN the unsaturation terms "axene" and "arene" are interchangeable, "arene" being permitted if the aromatic nature of the compound is to be emphasized.

Coupled with the permitted use of alternative names is the use of a dual system of naming polycycles. IUPAC nomenclature offers two methods for the naming of polycycles, namely, the extended von Baeyer system for bridged polycycles and the ring fusion system for ortho-fused polycycles. However, where an ortho-fused polycycle is bridged, the bridge is designated as a bivalent substituent such as "methano" (Rule A-34). This complication is avoidable by permitting only one naming method for polycycles, as in LNN.

Trivial Names. IUPAC nomenclature also includes a plethora of trivial names and radicals, which furthermore are not always consistent in their enumeration. This is for historical reasons, and whereas it does offer consistency with old terminology and avoids long names, it means that computer software must incorporate all the rules for their different syntax and enumeration. Trivials have a place in nomenclature as alternatives to long systematic names and should perhaps only be used in scientific publications after first being associated with the corresponding systematic name.

Superscripts and Subscripts. One of the major problems in inputting chemical nomenclature to computer systems, via keyboard or scanner, and of printing out names is the extensive use of superscripts and subscripts. Superscript numbers and other characters are used extensively in nodal nomenclatures. It is just as cumbersome for scientists to communicate, vocally, chemical names containing superscript and subscript character strings as it is to create a special means to communicate them to a computer system.

Although superscripts and subscripts may be unavoidable in, for example, the designation of charge and isotopes, it is perfectly possible to design a system of nomenclature in which they are deliberately avoided[17] in the name in general.

Character Set. Software, for example for scientific word processing and typesetting, that can handle nonstandard characters, such as Greek and Roman characters, and superscripts and subscripts is now readily available. Screens and printers for outputting these are also commonplace. Nevertheless, the universality of an international standard chemical nomenclature is such that the names it defines must be capable of being input, processed, displayed, and printed on all computer systems, from the simplest PC, word processor, and printer in a school to the graphics workstation, specialist typesetting software, and laser printer in a research environment. A truly linear set of printable ASCII characters, without control characters to designate special effects, would be beneficial to education, printed communication, and computer processing of systematic chemical nomenclatures.

DISCUSSION AND CONCLUSION

The above points all show the disadvantages found in one or more chemical nomenclature systems insofar as computer processing is concerned. It cannot be overemphasized that computer software to allocate correct chemical names will become an essential part of everyday science. Science is an exact subject, so chemical nomenclature should also be constructed in an exact manner, without the vagaries which have been allowed to develop over the years. IUPAC chemical nomenclature is under constant scrutiny by the IUPAC commissions, whose interests include computer processing. Of course common usage has become so ingrained that it is not possible to change overnight to a completely new system just for the sake of being computer-oriented. But this, we believe, has to occur gradually. The initial step taken by Lozac'h in the idea of a nodal nomenclature[15] has been followed up by the development of a computer-oriented nodal numbering system.[17] If there is any future in these ideas this can only be determined by their general use, when the simplicity, consistency, and logic of a new nomenclature for communication between scientists and in education would in time overcome the inertia against change.

In the future, the scientific literature that is not available in machine readable form may be made so by optical recognition techniques, and it will be necessary to isolate chemical names for translation into whatever nomenclature and/or notation is in use. There will be problems with the older literature, where less care was taken in allocating unique and unambiguous names, and with literature in different languages, where enumeration patterns in polycycles sometimes differ.

Any new nomenclature proposed should be formally defined in an appropriate meta-language as are computer programming languages and as was GENSAL, a language to describe generic chemical structures.[24] This approach has the benefit of forcing the syntax of the language to be concise and unambiguous, and is much better done at the definition stage than later, as was the case for parts of the IUPAC organic chemical nomenclature.[25]

ACKNOWLEDGMENT

The authors are grateful to a reviewer for comments that have led to an improved structure in this paper.

REFERENCES AND NOTES

(1) 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.

(2) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 4. Concise Connection Tables to Structure Diagrams. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 122–127.

(3) Wisniewski, J. L. AUTONOM: System for Computer Translation of Structural Diagrams into IUPAC-Compatible Names. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 324–332.

(4) Raymond, K. W. A LISP Program for the Generation of IUPAC Names from Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 270–274.
(5) Polton, D. J. Computer Studies in the Interconversion and Development of Linear Representations of Chemical Structures. Ph.D. Thesis, University of Hull, Hull, England, 1991. Polton, D. J. *Chemical Nomenclatures and the Computer*; Research Studies Press: Taunton, Somerset, England, in press.
(6) Lees, R.; Smith, A. *Chemical Nomenclature Usage*; Ellis-Horwood Ltd.: Chichester, England, 1983.
(7) IUPAC. *Nomenclature of Organic Chemistry, Sections A-F and H.* Pergamon: Oxford, England, 1979.
(8) 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.
(9) Smith, E. G. *The Wiswesser Line-Formula Chemical Notation (WLN)*; McGraw-Hill: New York, 1968.
(10) IUPAC. *Rules for the Notation for Organic Compounds*; Longmans Green and Co.: London, England, 1961.
(11) Polton, D. J., A computer process for substructure searches on compound structures ciphered in the IUPAC notation. *Inf. Storage Retr.* **1972**, *8*, 191–201.
(12) Weininger, D. SMILES: a Chemical Language Information System. 1. Introduction to Methodology ane Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(13) Barnard, J. M.; Jochum, C. J.; Welford, S. M. 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, 1988; pp 76–81.
(14) Dyson, G. M. *Some New Conepts in Organic Chemical Nomenclature*, unpublished, 1976.
(15) Lozac'h, N.; Goodson, A. L.; Powell, W. H. Nodal Nomenclature-General *Principles. Angew. Chem. Int. Ed. Engl.* **1979**, *18*, 887–899.
(16) Hirayama, K. *The HIRN System, Nomenclature of Organic Chemistry.* Maruzen: Tokyo, Japan; Springer-Verlag: Berlin, Germany, 1984.
(17) Polton, D. J. A new nodal numbering system for cyclic and acyclic structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 430–436.
(18) Polton, D. J. A Comment on Nomenclature and the Unsaturated Bond. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 82–83.
(19) Davidson, S. An Improved IUPAC-Based Method for Identifying Alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 151–155.
(20) Davidson, S. Algorithm for Selecting the Parent Structural Unit of a Ring-Chain Assembly. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 215–221.
(21) Goebels, L.; Lawson, A. J.; Wisniewski, J. L. AUTONOM: System for Computer Translation of Structural Diagrams into IUPAC-Compatible Names. 2. Nomenclature of Chains and Rings. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 216–225.
(22) Polton, D. J. Conversion of the IUPAC notation into a form for computer processing. *Inf. Storage Retr.* **1969**, *5*, 7–25.
(23) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 6. (Semi)-automatic Name Correction. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 153–160.
(24) Barndard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 2. GENSAL, a Formal Language for the Description of Generic Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 151–161.
(25) 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.