(28) Hickey, J. P.; Butler, I. S.; Pouskouleli, G. "Carbon-13 NMR Spectra of Some Representative Hormonal Steroids". J. Magn. Reson. 1980, 38, 501-506.
(29) Gray, N. A. B.; Nourse, J. G.; Crandell, C. W.; Smith, D. H.; Djerassi, C. "Stereochemical Substructure Codes for ^{13}C Spectral Analysis". Org. Magn. Reson. 1981, 15, 375-389.
(30) Beierbeck, H. "Simple Stereochemical Structure Code for Organic Chemistry". J. Chem. Inf. Comput. Sci. 1982, 22, 215-222.
(31) Broadbent, T. A.; Paul, E. G. "Carbon-13 Nuclear Magnetic Resonance in Alkaloid Chemistry". Heterocycles 1983, 20, 863-980.

A Concise Connection Table Based on Systematic Nomenclatural Terms

J. D. RAYNER

Department of Computer Studies, The University of Hull, Hull HU6 7RX, U.K.

Received July 6, 1984

A connection table schema is introduced that can be used to describe concisely both molecular structures and mixtures, organic and potentially also inorganic, within a single uniform hierarchy. The schema can explicitly represent heteroatoms, charges, and different bond types but is nonexplicit in its coding of the most common individual carbon atoms in the molecule, being based on systematic nomenclatural terms for substructural units such as rings and chains.

BACKGROUND

During the course of an investigation[1] into the computer translation of IUPAC systematic nomenclature,[2,3] a target representation of molecular structure was required that could be formed as output from the developing nomenclature translation system. The immediate need was not for a molecular structural representation that could be fully explicit at the atomic level but rather a scheme that could include information at the level of detail found in the systematic name.

IUPAC names are formed from small fragments that often relate individually to polyatomic structural units—ring systems, carbon chains, etc.—in which the nature of bonding and the atom types are assumed to be generally uniform. Other fragments deal with deviations from these assumptions or denote smaller structural units such as individual atoms. Thus, in the concise connection table (CCT) described below, the bulk of the carbon skeleton in an organic structure is represented implicity, as it is in the nomenclature, but all non-carbon atoms and all unusual bonds (that is multiple or nonaromatic according to context) and all electronic charges that are described by a systematic name can be represented explicity in the table.

In contrast, other connection table schemes are generally fully explicit in their description of every (non-hydrogen) atom in a structure, and some contain redundancy in their duplication of bond information.[4] Such a redundant connection table may list every non-hydrogen atom in a structure, and by means of index numerals and bond multiplicity values, it can represent all the interatom connections. However, since every recorded atom has all its associated bonding represented within its own record, information on any one bond is duplicated, in the records of the two atoms that it joins. Such redundancy can be eliminated by recording only one bond in each atom record, for example, that to the adjacent atom with lowest index, but further records of a different type are then necessary to specify ring closure in cyclic compounds.

One advantage of the CCT described below is its implicit retention of ring closure information, while avoiding the use of any variant records. Further compactness, beyond the avoidance of such variants, is achieved in the CCT schema by dispensing with the bulk of bond information entirely, on the basis that for organic compounds (and therefore the bulk of those actively dealt with) the great majority of bonding is entirely predictable from the nature and configuration of the atoms themselves, in their four common environments—aromatic, aliphatic, alicyclic, and individual. The relatively few variations from these default states may still be represented with a net reduction in table size, as for instance in examples 8 and 9 of Figure 3 below.

GENERAL DESCRIPTION

A concise connection table (CCT) will comprise a sequence of table entries that are each of identical form but whose meaning will vary according to both content and context within the abstract hierarchy of the table. Since the CCT schema has been developed for use as the output of a nomenclature translation process, it is not unnatural for the table structure to be hierarchic, in resemblance of the approach of the IUPAC nomenclature. However, this is not to say that the IUPAC nomenclature is a necessary precursor to the use of the CCT schema for representing any molecular structure: any system that deals in substructural units may be potentially applicable, for example, the coding of chain fragments and ring system skeletons in the Wiswesser line notation (WLN).[5]

In the present application, the nomenclatural term that represents the parent structure of a molecule gives rise to the initial entries of the corresponding table. Substituents on the parent are represented by further entries, which are both hierarchically inferior and physically subordinate in the table. The relative ordering of individual table entries is further discussed below. The terms "parent" and "substituent" derive from a consideration of organic structures, but the CCT is also potentially capable of representing inorganic compounds.

The contractions possible for organic chains and rings of carbon cannot generally be echoed in the coding of the more varied assemblages of atoms and bonds that characterize inorganics. Bonding arrangements in the latter may be more accurately represented by further codings in the schema beyond those listed below, given appropriate extensions to the present definition.

The greater variety of structure found in an inorganic substance may thus cause a proliferation of table entries, with the result that a CCT for an inorganic substance can closely approximate to a full atom-by-atom representation. Nevertheless, since inorganic "parental" structures are typically single atoms, repeated substituents (ligands etc.) will generally have the unit locant, and their representation can then be condensed by means of the special entry for repetition (*qv*). Assemblages

0095-2338/85/1625-0108$01.50/0

of ions, and mixtures of general, can be represented as a collection of individual CCTs, preceded by a similar special entry.

TABLE ENTRIES

Each entry in a CCT is of the same fixed length and comprises four fields denoted in turn LOCT, TIPE,[9] SIZE, and SUBS. These fields may contain only positive integral numeric values. In general, the purpose of the fields is as follows: LOCT contains the locant (atom number) in the immediately parental (sub)structure to which is attached the substructure currently being defined. TIPE contains a code number to indicate the general nature of the (sub)structure being defined, e.g., ring, chain, heteroatom, bond, charge, etc. SIZE contains a value that gives more detailed information about the (sub)structure, interpreted in light of the TIPE code, for example, the number of atoms in a chain or the multiplicity of a bond. SUBS generally contains a count of the number of subordinate structures, yet to be described, that are attached to the current (sub)structure. However, in certain specific circumstances this field can contain a locant value that is either additional to or alternative to the contents of LOCT, depending on the context.

Table entries are categorized in three classes, as main entries, ring segment entries (RSE), and modification entries. In addition, the class of main entries contains the subclass of ring header entries (RHE). Most structures are handled through main entries, but ring systems use in addition a set of ring segment entries for their complete description, preceded by a main entry of the ring header subclass. Bonds and atoms within a structure are generally assumed to be of a standard, common type; where differences from these defaults occur, they are described by modification entries, which may also be used to describe other unusual features.

ORDERING OF CCT ENTRIES

The abstract hierarchy of the molecular structure being defined by a CCT is implicit in the physical ordering of the entries in the table. Thus, by inversion the broad physical ordering of CCT entries must be defined by a consideration of the underlying molecular structure. However, at any given level in the CCT, a set of hierarchically equivalent table entries will have to be placed in a physically linear order.

In the IUPAC systematic name, which has hitherto formed the source of structural information for coding in the CCT, substituents at any given level are generally ordered alphabetically, according to the text of the nomenclatural terms involved,[6] and for each substituent any locants will be quoted in numerical order as appropriate. Although the same ordering may be preserved in the CCT by default, a physical ordering based primarily on locant values, but with all modification entries preceding those for any substituents at each level, is considered to be potentially more beneficial. However, this question of detailed ordering for subordinate table entries at any one level does not affect the initial definition of the CCT as presented here.

DETAILED DESCRIPTION

The following description is arranged in sections according to the primary classification of entries as main and ring header, ring segment, or modification and defines the currently allowed field values for each of these entry classes. Each section concludes with appropriate illustrative examples of the CCT, including the corresponding name and structure in each case.

Main Entries. These entries introduce the parent and major substructures required for the description of a molecule. Substituents on a given structure are represented as full structures in their own right, by means of main entries in which

		LOCT	TIPE	SIZE	SUBS
(1)	nonane	1	1	9	0
(2)	2-chlorobutane	1	1	4	1
		2	2	17	0
(3)	6-bromo-5-(1,1-dimethylethyl)dodecane				
		1	1	12	2
		5	1	2	2
		0	0	0	2
		1	1	1	0
		6	2	35	0

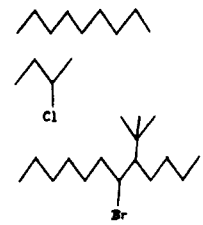

Figure 1.

the LOCT field gives the point of attachment to the parental structure. Where a substituent occurs more than once on the same parental atom, a special (repetition) entry of the form 0 0 0 n may precede a single instance of the substituent description, as detailed below. The interpretation of fields in a main entry will now be described.

The LOCT field must be nonzero, to indicate the atom of the immediate parental structure to which the currently described object is attached. If the entry is the first of the whole structure, then the field is set to 1 by convention.

The TIPE field may currently have one of four values, with the following meanings:

value	meaning
0	entry is a ring header entry for an aromatic ring system
1	entry describes an aliphatic chain
2	entry describes an atom alone
3	entry is a ring header entry for an alicyclic ring system

Further code values for the TIPE field must be defined if necessary in future extensions of the schema. The four codings above have been allocated essentially for their mnemonic value: the digits 0 and 3 are ringlike; 1 is chain-like.

The SIZE field is interpreted according to the value of TIPE, as follows:

TIPE	interpretation of SIZE
0	number of ring segment entries that follow
1	length of chain (number of atoms)
2	type of atom (atomic number)
3	number of ring segment entries that follow

The SUBS field contains a count of the number of substituents on, plus modifications to, the present structure, to be described by further main and modification entries, which follow immediately after any ring segment entries. Examples are shown in Figure 1.

Ring Segment Entries. These entries follow immediately after an associated main (ring header) entry of TIPE 0 or 3, which determines the basic nature of the ring system. The number of ring segment entries present is given in the SIZE field of the ring header entry. The interpretation of fields in a ring segment entry will now be described.

In general, the LOCT field denotes the first atom (lowest locant) in the current ring system, as so far described by the immediately preceding RSEs, and to which the new ring segment is fused or otherwise attached. The addition of each further ring segment after the first one implies an alteration of the locants in their assignment to the particular atoms of the ring system, to accord with the normal locant-assignment conventions.[7] Thus, for example, the atom specified by LOCT in the current ring segment entry will usually, after fusion, be assigned a notionally lettered locant whose numeric part is one less than the original LOCT value, and that wholly numeric locant will be moved to the first new atom in the new segment. This can result in the LOCT fields of adjacent RSEs having identical values, as in example 5 of Figure 2 below.

For the first RSE of a system, a LOCT field determines the initial source of locants, by reference to one particular atom of the first ring segment. By convention, the actual locant is given of the topmost atom in the first ring segment or of the

		LOCT	TIPE	SIZE	SUBS
(4)	naphthalene	1	0	2	0
		1	0	6	0
		5	0	6	1
(5)	triphenylene	1	0	4	0
		2	0	6	0
		5	0	6	1
		5	0	6	1
		9	0	6	1
(6)	tricyclo[5.4.0.$0^{4,8}$]undecane				
		1	3	3	0
		1	3	11	0
		1	1	0	6
		4	1	0	4
(7)	cubane	1	3	5	0
		1	3	8	0
		1	1	0	5
		2	1	0	3
		3	1	0	5
		4	1	0	3

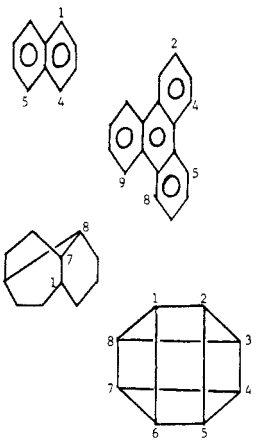

Figure 2.

atom that is attached to a parental structure. For this purpose, topmost refers to the single uppermost atom in the conventional diagrammatic representation of a six-membered ring, with vertical sides and an apex above, and for other ring sizes and orientations, it may be taken to indicate the equivalent uppermost atom or the atom to the right of the uppermost bond, as appropriate.

The TIPE field determines the nature of the ring segment as follows:

TIPE	nature of ring segment
0	segment is an aromatic ring
1	segment is an aliphatic chain
2	unused
3	segment is an alicyclic ring

The SIZE field indicates the number of atoms involved in the ring segment.

The SUBS field will be 0 for the first ring segment, and subsequently for later ring segments, it gives information about the second point of attachment, relative to initial locants given in each case by LOCT. Thus a zero in SUBS denotes spiro attachment, unity indicates fusion up to the next available unfused ring atom, and higher values will occur for polycyclic nonaromatic structures, to denote bridging. Hence in example 7 of Figure 2, the four bridging segments connect the atom pairs (1, 6), (2, 5), (3, 8), and (4, 7), respectively, the second locant being the sum of LOCT and SUBS values in each case.

The relative ordering of a group of RSEs will be determined by the requirement that, for any given ring segment, the points of attachment must have been defined earlier and the first ring segment to be defined must be that containing the lowest local locant. This effectively means that RSEs will appear generally in order of increasing first fusion locant (LOCT field value). Examples are shown in Figure 2.

Modification Entries. These entries describe deviations from default conditions in substructures previously introduced. They are characterized by a LOCT field of 0, with the actual locant information being held in the SUBS field. Current interpretations for the remaining fields in a modification entry are described below. The list should not be regarded as complete, since this is an area of the CCT schema where room is deliberately left available for possible future expansion.

(The LOCT field MUST be 0.)

The TIPE field determines the nature of the modification, as follows:

TIPE	nature of modification
0	denotes special entry (see below)
1	modification is an electronic charge
2	modification is a heteroatom
3	modification is a special bond

		LOCT	TIPE	SIZE	SUBS
(8)	pyridine	1	0	1	1
		1	0	6	0
		0	2	7	1
(9)	butadiene	1	1	4	2
		0	3	10	1
		0	3	10	3

Figure 3.

		LOCT	TIPE	SIZE	SUBS
(10)	dichromate ion	1	2	24	4
		0	0	0	2
		1	2	8	1
		0	3	10	0
		1	2	8	1
		1	2	24	3
		0	0	0	2
		1	2	8	1
		0	3	10	0
		1	2	8	1
		0	1	7	1
		1	2	8	1
		0	1	7	1
(11)	diammonium sulphate	0	0	0	3
		0	0	0	2
		1	2	7	5
		0	0	0	4
		1	2	1	0
		0	1	9	1
		1	2	16	4
		0	0	0	2
		1	2	8	1
		0	3	10	0
		0	0	0	2
		1	2	8	1
		0	1	7	1

Figure 4.

The SIZE field is interpreted according to the value of TIPE, as follows:

TIPE	interpretation of SIZE
0	see description of special entry
1	value of charge, plus 8, e.g., 7 → 1–, 10 → 2+; 8 may denote an unpaired electron
2	atomic number of heteroatom
3	bond type, according to the following code: 7 → dative, 8 → aromatic, 9, 10, 11 → covalent single, double, triple

Further values of modification TIPE and related SIZE fields could be defined to allow the description of rarer bonding arrangements such as, for example, the single-plus-dative combinations in borazine or, if it were felt useful, to represent accurately the conjugated bonding in carboxylic acid anions. The representation of absolute stereochemical configurations might also be handled similarly.

The SUBS field, for modification entries other than of TIPE 0, gives the locant of the atom in the current structure at which the modification occurs. Examples are shown in Figure 3.

Special Entries. These are a subclass of modification entries in which both LOCT and TIPE fields are 0. They provide a mechanism for condensing tables that contain multiple occurrences of a given structure or for describing mixtures and collections of structures that are not explicitly connected, e.g., salts.

The LOCT field MUST be 0. The TIPE field MUST be 0. The SIZE field is currently unused and should be assumed to be 0. The SUBS field contains a numerical count, either of the objects involved in a mixture or salt, if the special entry is the very first of the whole table, or else of the number of times that a substructure recurs, that substructure being described in the immediately following entry or entries. Examples are shown in Figure 4.

SUMMARY AND CONCLUSIONS

The concise connection table (CCT) is, physically, a linear arrangement of fixed-length table entries that consist of four

nonnegative integers each. These table entries describe the structure of a molecule or compound in gradually increasing detail, such that the complete CCT is interpreted as a hierarchy of subtables, linked together implicity by substituent counts and locant values.

For organic substances, most carbon atoms and their common bonds are represented implicitly, by the coding of complete rings and chains in single table entries, thus making the CCT particularly concise. Other atoms and bonds may be represented explicity, but provided that for some cases, particularly inorganics, further appropriate values are defined for bond modification entries.

It is widely recognized that the IUPAC nomenclature is nonunique, and this quality is often decried in relation to the needs of information storage and retrieval activities. However, it should not be overlooked that chemical nomenclatures are used in a wide range of human circumstances other than (indeed, perhaps more often than) in pure research and its related fields. For example, in general trade and commerce, together with its regulatory legislation, uniqueness is of a lower priority than the ever-essential nonambiguity, and particularly familiarity and, hence, relative ease of use over other structural representations are most important. It is here that the versatility of the IUPAC nomenclature is of great benefit and equally that the ability to verify and convert IUPAC names to other forms is of potential use. In this respect the CCT acts as an intermediary stage in a computer-transformation process just as do all connection tables in their own systems.

The CCT as so far defined has been used successfully to code the structural information contained within a variety of IUPAC nomenclatural terms, and complete tables have been generated algorithmically from complete names. Such tables, as exemplified throughout the text, are self-evidently concise in comparison with conventional atom-by-atom connectivity tables for structures of any great size, through the CCT representation of many atoms and bonds by a single table entry.

The expansion of the CCT to a full atom-by-atom connection table has been briefly investigated and shown to be feasible, thus indicating a potential route from the IUPAC nomenclature to many existing information systems based on such connection tables. The CCT is also suitable for further processing for the display of structural diagrams. Algorithms have been developed on the basis of a character-cell technique[8] to display unscaled line segment representations of rings and chains, given the corresponding CCT entries.

REFERENCES AND NOTES

(1) Rayner, J. D. "Grammar Based Translation by Computer of the IUPAC Systematic Chemical Nomenclature". Ph.D. Thesis, University of Hull, Hull, U.K., 1983.
(2) International Union of Pure and Applied Chemistry "Nomenclature of Organic Chemistry," 1979 ed.; Pergamon Press: Oxford, U.K., 1979.
(3) International Union of Pure and Applied Chemistry "Nomenclature of Inorganic Chemistry," 2nd ed.; Butterworths: London, U.K., 1971.
(4) Ash, J. E. In "Chemical Information Systems"; Ash, J. E.; Hyde, E., Eds.; Ellis Horwood: Chichester, U.K., 1975; Chapter 11, pp 156-176.
(5) Smith, E. G.; Baker, P. A. "The Wiswesser Line Formula Chemical Notation," 3rd ed.; Chemical Information Management Inc.: Cherry Hill, NJ, 1975.
(6) See pp 7 and 380 of reference 2.
(7) See pp 20-27 of reference 2.
(8) Rayner, J. D.; Milward, S.; Kirby, G. H. "A Character Set for Molecular Structure Display" *J. Mol. Graphics* **1983**, *1* (4), 107-110.
(9) The unusual spelling of "TIPE" arises historically through the use of the Pascal programming language for the coding of algorithms. The more preferable term "TYPE" is a reserved word in that language and is therefore unavailable as a field identifier.

Comparative Efficiency of Searching Abstract Text in the Chemical Abstracts Service Database†

M. HERZ, H. K. KAINDL, A. A. SALIB, and R. WARSZAWSKI

BASIC, Basel Information Center for Chemistry (Documentation Center of Ciba-Geigy Ltd., F. Hoffmann-La Roche & Co. Ltd., and Sandoz Ltd.), CH-4002 Basel, Switzerland

Received September 5, 1984

Text retrievals were carried out by utilization of Chemical-Biological Activities (CBAC) and Polymer Science & Technology (POST) tapes, which contained in computer-readable form all *Chemical Abstracts* data element fields. The subjects of the queries were randomly selected, and a natural language vocabulary was used for the text profiles. An evaluation of the contributions of all data elements to the retrievals showed that when searchable abstract text was added to a combination of the remaining fields, the recall was greatly improved and the probability of retrieving concepts not contained in the indexes was increased.

INTRODUCTION

Our investigation dates back to 1976 and was based on searches of Chemical-Biological Activities (CBAC) and Polymer Science & Technology (POST) tapes, two services available at that time containing all *Chemical Abstracts* (CA) data element fields in computer-readable form, including the abstract text.[1] As a result of improvements in its editorial process, Chemical Abstracts Service (CAS) began producing the full abstract text from all 80 CA sections in computer-readable form in 1975.[2] Prior to that time, the abstracts were only available in computer-readable form in CBAC and POST. CAS subsequently included abstract text in searches for some topics in its *CA Selects* current awarenesss service[3] and optionally for all searches in the *Individual Search Service* (ISS) batch current awareness service.[4] Although abstract text is the most voluminous data element in the CA text file (e.g., in 1976 CA text comprised about 21% of the complete database), we thought having the capability to search it would greatly improve the recall for certain queries. In an earlier study, Barker et al.[5] compared the search performance of the abstract text with that of the other free-text data elements (keywords and titles), which at that time were accessible in computer-readable services from CAS, and found that abstract text greatly improved recall. In addition, several studies have been devoted to searching the abstract text in other data bases.[6]

† Address correspondence to Dr. M. Schellenbaum, BASIC, Ciba-Geigy Ltd., CH-4002 Basel, Switzerland.

0095-2338/85/1625-0111$01.50/0