The Concise Connection Table: Collected Definitions with Extensions for Stereochemistry and Saccharides

D. I. Cooke-Fox, J. F. Ewart, G. H. Kirby, M. R. Lord, and J. D. Rayner*

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

Received February 3, 1992

The Concise Connection Table is a representation of molecular structures based on an implicit hierarchy of substructural components. This paper introduces a number of enhancements and extensions to the original definition, particularly relating to representations of stereochemistry and saccharides, and summarizes the overall principles of the CCT in one definitive presentation.

BACKGROUND

The Concise Connection Table (CCT)[1] was first developed to provide a computer representation of a chemical structure corresponding to the semantic decomposition of systematic nomenclature. This was a requirement of the Hull University Nomenclature Translator (HUNT) project,[2-7] which has developed computer software for the automatic translation of IUPAC systematic organic nomenclature into CCT representation, and thence into graphical display of the corresponding molecular structure.[5] The CCT has also been converted into Standard Molecular Data (SMD) format[8] and may easily be expanded into fully explicit atom-by-atom connection table styles.

In developing the HUNT software to cover greater areas of the IUPAC nomenclature, various extensions to the original CCT definition have been introduced and a small number of refinements have also been achieved. Major extensions have accommodated the particular requirements of steroid nomenclature[6] and the representation of sugar structures according to the IUPAC/IUB joint nomenclature recommendations.[9] A technique for simple representation of common functional groups has been introduced to avoid the need to represent their sometimes uncertain conjugated bonding.

The purpose of this paper is to report the recent developments of the CCT for sugar structures and to draw together a coherent description of the CCT principles in a single definitive presentation.

GENERAL PRINCIPLES

The Concise Connection Table (CCT) consists of a table of positive integer values arranged into four columns which are referred to as LOCT, TIPE, SIZE, and SUBS, respectively. Each Main Entry row of the table describes a simple structural component, an atom or a chain for example, or will introduce a multirow description of a more complex substructure. The representation of ring structures (aromatic and alicyclic), steroids, carbohydrates, and similar complex molecules is achieved more efficiently by this means.

In general, a Main Entry introduces each substructure in the CCT, the LOCT field of which contains the locant of that particular substructure within its parent. This is a nonzero value for most main structures, and the first Main Entry LOCT field always contains the value 1.

The TIPE field is used to indicate the nature of the group/substructure being described. The third field, SIZE, can give more detailed information about the substructure referred to in the TIPE field. SUBS, the fourth field, is typically used to indicate the number of further substructures which are attached to the present one, to be described later on in the CCT.

The hierarchical nature of the table is illustrated by the fact that the parent structure of an organic molecule forms the initial entries in the CCT, while its substituent atoms and groups are added afterward. Each substituent may in turn give rise to more CCT entries, to describe its own further substituents, and so on. The overall linear sequence of CCT entries corresponds to a pre-order traversal of the substructure hierarchy. Figure 1 illustrates the overall appearance of the CCT and incorporates examples of the general entries described in the following sections.

MAIN ENTRIES

The Main Entry TIPE fields are coded to represent the major substructures that can occur in a molecule (Table I). Connectivities to the parent structure are indicated in the LOCT field, which specifies their point of attachment within the parent structure which will have been described earlier. Where a parent atom has more than one instance of the same group attached, this may be indicated concisely by a "repetition" entry in the table. The repetition entry has a positive integer value in the SUBS field of the CCT and is distinguished by zeros in the LOCT, TIPE, and SIZE fields. The SUBS value defines the number of occurrences of a substituent which is described by the immediately following sequence of CCT rows. A similar technique may be envisaged for representing polymers and copolymers with repeated sequences of monomeric units.

RING SEGMENT ENTRIES

Aromatic and alicyclic ring systems are described in the CCT by a group of entries, immediately following the appropriate main entry with a TIPE value of 0 or 3. The first ring segment entry (RSE) describes the ring containing locant 1 of the overall system, and subsequent RSEs add further components, ring or chain, in eventual locant order. The LOCT field is generally used to indicate the lowest locant in the partial ring system to which the new ring system is attached. The TIPE field is used to describe the ring segment in greater detail: whether it is an aromatic or an alicyclic ring, or a bridging chain. The SIZE field indicates the number of carbon atoms in the ring segment.

The fourth field, SUBS, is used to convey information about the attachment of the ring system to the parent structure. The SUBS field holds a value corresponding to the locant of the

0095-2338/92/1632-0556$03.00/0 © 1992 American Chemical Society

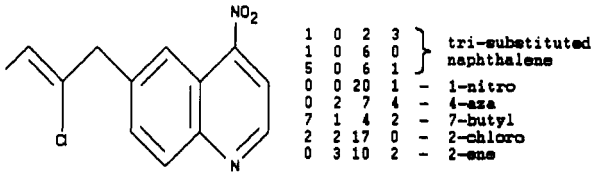


Figure 1. General example.

Table I. Main Entry Values (for a Nonzero LOCT Field, Indicating Attachment of Locant in Parent Structure)

TIPE value	meaning	interpretation of SIZE
0	an aromatic ring system	no. of further RSEs to describe ring system
1	an aliphatic chain	no. of atoms in chain
2	an atom	atomic number
3	an alicyclic ring system	no. of further RSEs to describe ring system
4	a carbohydrate	no. of further carbohydrate segment entries to describe atoms in sugar unit
5	a steroid	no. of further steroid segment entries to describe atoms in steroid nucleus

Table II. TIPE/SIZE Interpretation for Modification Entries Where LOCT Is Zero

TIPE	modification	interpretation of SIZE
0	SEFG or repetition	if SIZE = 0: denotes repetition entry if SIZE > 0: corresponds to SEFG code (Table III)
1	electronic charge	8 + value of charge, 8 may be an unpaired electron
2	heteroatom	atomic number of heteroatom
3	special bond	bond type relates to following code: 0, 1, 2 = ξ, α, or β bond 3, 4 = cyclo, seco (with locants in following entry) 5,6 = *R*, *S* stereochemical descriptors 7 = dative 8 = aromatic 9, 10, 11 = covalent single, double, triple bond

second point of attachment relative to the first point of attachment given by LOCT. For the first ring segment the SUBS field will hold a zero value, and for many aromatic situations it typically holds the value 1. Spiro attachments are indicated by a SUBS field of zero.

MODIFICATION ENTRIES

This is a special group of entries to represent structures that deviate in some way from the default carbon structure described in the preceding CCT entries. Modification Entries are given a LOCT value of zero which distinguishes them from all other entry types. The TIPE field indicates the nature of the modification with integer codes for modifications in electronic charge, heteroatoms, bond types, and special entries. The interpretation of the value held by the SIZE field depends on the value given in TIPE (i.e., the type of modification described). So, for example, the SIZE field can be used to describe a particular sort of bond, the atomic number of a heteroatom, a charge value, or specific information about a special entry group (Table II).

The SUBS field is used to denote the locant of the atom in the original structure at which the modification occurs, for all modification entries apart from the repetition entry, in which SUBS denotes the number of repetitions of the substructure to be described next.

Table III. Current SEFG Codes

size	nomenclature	line formula	size	nomenclature	line formula
1	amino	$-NH_2$	19	mercapto	–SH
2	azido	$-N_3$	20	nitro	$-NO_2$
3	carbaldehyde	–CHO	21	nitroso	–NO
4	carboxylic acid	–COOH	22	pentafluorothio	$-SF_5$
5	chlorosyl	–ClO	23	perchloryl	$-ClO_3$
6	chloryl	$-ClO_2$	24	seleneo	–SeOH
7	cyanato	–OCN	25	selenino	$-SeO_2H$
8	cyano	–CN	26	selenono	$-SeO_3H$
9	dihydroxyiodo	$-I(OH)_2$	27	sulfino	$-SO_2H$
10	dithiocarboxy	–CSSH	28	sulfo	$-SO_3H$
11	dithiosulfo	$-S_3OH$	29	thioaldehyde	–CHS
12	hydroperoxy	–OOH	30	thiocarboxy	–CSOH
13	hyroxy	–OH	31	thiol	–SH
14	iodosyl	–IO	32	selenol	–SeH
15	iodyl	$-IO_2$	41	'deoxy'	–H (from –OH)
16	isocyanato	–NCO	42	methyl	$-CH_3$
17	isocyano	–NC	43	hydroxymethyl	$-CH_2OH$
18	isothiocyanato	–NCS			

SPECIAL ENTRY FUNCTIONAL GROUPS

Certain common functional groups are represented in the CCT format by means of just one CCT entry each—the Special Entry Functional Group (SEFG). These entries are distinguished by the fact that both the LOCT and TIPE fields contain zeros. One of the main purposes of the SEFG is to allow the simple representation of functional groups without complex conjugated bonding details. A number of common line formulas have been given nonzero integer values which can be held in the SIZE field of the SEFG. The SUBS field contains the parent structure locant at which the functional group is attached. A list of the currently used SEFGs is given in Table III.

STEROID SEGMENTS

All steroids are based on a common tetracyclic ring structure with a special sequence of ring numbering. This is a good reason for developing a specialized CCT format to deal with this particular class of molecules in an efficient way. TIPE 5 in the main entry of a CCT indicates that the following table will represent a steroid. The SIZE field contains the number of CCT steroid segment entries needed to describe the steroid nucleus, and the SUBS field indicates the number of substituents and/or modifications that are made to the parent skeleton to give the required overall molecular structure.

Immediately following the main entry is the Steroid Nucleus Entry (SNE), not counted in the header SIZE value. This entry explicitly specifies the ring sizes that constitute the underlying structure of the molecule, typically 6 6 6 5 corresponding to rings A, B, C, D.

After the nucleus entry comes the given number of Steroid Segment Entries (SSE) to describe the steroid nucleus in greater detail. The description is based on differences from the stigmastane nucleus, which contains all possible components. The LOCT field of a steroid segment entry contains a value that corresponds to a specific locant on the steroid nucleus. The associated TIPE field can have one of two possible values: it contains a 1 if that particular entry specifies an aliphatic chain or a dummy entry and a 2 in this field indicates an atom entry. The SIZE and SUBS fields depend on the value held in the corresponding TIPE field (Table IV).

If TIPE contains a 1 and the SIZE field a 0 then a dummy entry is specified. In this case the SUBS field contains the

Table IV. TIPE, SIZE, and SUBS for SSEs

TIPE	SIZE	SUBS
1	0	count of dummy atoms
1	n	stereochemistry of attaching chain of length *n* (SIZE)
2	atomic no.	stereochemistry (as above)

number of atoms missing in this steroid, relative to the nonsystematic numbering. A nonzero SIZE field and a TIPE value of 1 indicates an aliphatic chain—the length of which is specified by the SIZE field. The SUBS field then contains a value of 0, 1, or 2 to indicate the stereochemistry at that locant. A SUBS field value of 1 indicates an α bond, a 2 corresponds to a β bond, and a zero value indicates unknown stereochemistry.

If the TIPE field contains a 2 (indicating an atom entry) the SIZE field will contain the atomic number of the substituent and the SUBS field will indicate the bond stereochemistry as described above.

CARBOHYDRATE SEGMENTS

The most recent enhancement of the CCT has been the development of a new format that will provide an efficient means of representing monosaccharide and polysaccharide carbohydrates. The "carbohydrate CCT" is distinguished from other CCT formats by the Carbohydrate Main Entry (CME) TIPE value of 4.

The rest of the CME contains details about the size of the carbohydrate and information about subunits within a sugar molecule. The SUBS field of the first CME in a polysaccharide specifies the number of monosaccharide subunits present, so a CCT describing just a monosaccharide will contain a value of 1 in this field. The LOCT field will also contain a 1 for a monosaccharide, but for polysaccharides, the LOCT field of the CME for each subsequent sugar unit gives the position of that particular monosaccharide subunit in the polysaccharide chain. The SIZE field specifies the number of carbon atoms, and hence the number of following Carbohydrate Segment Entries (CSE), that constitute the monosaccharide subunit indicated by the LOCT field.

The SUBS field of each intermediate CME in a polysaccharide contains the local branch count for that particular monosaccharide subunit which will typically be zero in a linear polysaccharide. The representation of branched polysaccharides is described below. The final subunit in the parent or a branch chain must have a SUBS value of zero (as it cannot have a substituent branch and be a terminal chain subunit). The LOCT value of this terminal subunit CME will have the same value as that in the SUBS field of the first subunit CME in the chain.

As with the steroid main entry, a carbohydrate main entry is immediately followed by one further CCT record now called the Carbohydrate Ring Entry (CRE). This entry specifies the nature of the carbohydrate ring system and the inter-subunit bonds that will exist in a polysaccharide. The TIPE and SIZE fields contain locants on the current monosaccharide (sub)unit that are connected via a bond to form the ring closure. The difference between these two values implies the shape of the carbohydrate: furan, pyran, or septan. The LOCT and SUBS fields also contain subunit locants—these refer to the local attachment points of inter-subunit bonds to the preceding and following units in a polysaccharide. In a monosaccharide these fields are zero.

The CSEs follow the CRE to specify each group attached to the sugar unit ring and the stereochemistry of each connecting bond. The LOCT field of a CSE specifies a locant on the monosaccharide structure. The TIPE field alludes to the chirality of the locant by means of three possible values. A 1 in this field indicates a group placed above the plane of a sugar ring drawn in the Haworth projection. It therefore follows that a 2 will specify a bond connecting the locant and a group below the plane of the ring. A zero value in this field indicates an achiral locant or a locant of unknown chirality. The SIZE field is used to specify a particular SEFG attached to the ring at the locant indicated by the LOCT field. The SUBS field is not used in the CSE of linear polysaccharides and therefore contains a zero in these entries. An example of a linear polysaccharide CCT and its associated structural formula is given in Figure 2.

1	4	6	2
0	1	5	1
1	1	13	0
2	2	13	0
3	1	13	0
4	2	13	0
5	2	13	0
6	0	43	0
2	4	6	0
4	1	5	0
1	1	13	0
2	2	13	0
3	1	13	0
4	1	13	0
5	2	13	0
6	0	43	0

O-β-D-galactopyranosyl-(1->4)-α-D-glucopyranose

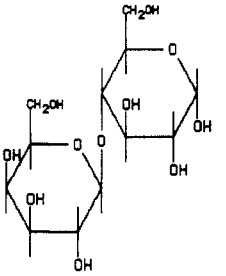
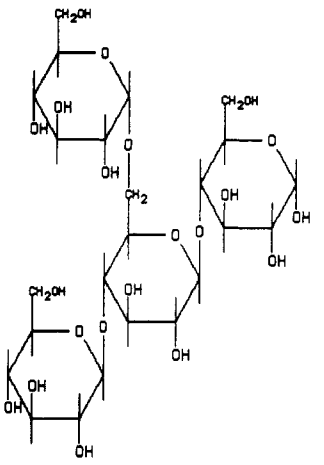


Figure 2. A disaccharide with its associated CCT.



O-β-D-glucopyranosyl-(1->4)-O-[α-D-glucopyranosyl-(1->6)]-β-D-glucopyranosyl-(1->4)-α-D-glucopyranose

1	4	6	3
0	1	5	1
1	2	13	0
2	2	13	0
3	1	13	0
4	2	13	0
5	2	13	0
6	0	43	0
2	4	6	1
4	1	5	1
1	1	13	0
2	2	13	0
3	1	13	0
4	2	13	0
5	2	13	0
6	0	43	0
3	4	6	0
4	1	5	0
1	1	13	0
2	2	13	0
3	1	13	0
4	2	13	0
5	2	13	0
6	0	43	0
6	4	6	1
1	1	5	0
1	1	13	0
2	2	13	0
3	1	13	0
4	2	13	0
5	2	13	0
6	0	43	0

Figure 3. A branched polysaccharide and its associated CCT.

BRANCHED POLYSACCHARIDES

The development of the carbohydrate CCT has had to provide for both linear and branched polysaccharide molecules. The representation of a linear polysaccharide is a straightforward case of joining monosaccharide subunit CCTs to form a larger, polysaccharide CCT. The order in which the subunit CCTs occur in the overall CCT reflects the ordering of the monosaccharides in the polysaccharide structure.

A branched polysaccharide can be thought of as one or more linear oligosaccharides joined to each other in a particular

way (Figure 3). The joining is always between one end of an oligosaccharide and any part of another one. This fact is exploited in the branched polysaccharide CCT in that the bulk of the branching information is held at the head of a linear CCT. The TIPE field of the CME starting a branch contains a 4 as in all other carbohydrate CCT main entries. The SIZE field also retains its usual value to indicate the size of the monosaccharide subunit following the main entry heading.

The SUBS field of the CME for a polysaccharide branch contains a nonzero value corresponding as usual to the chain length of the following chain of monosaccharide subunits. The LOCT field, however, contains information that differs from its usual main entry meaning.

The LOCT field of the branching subunit CME contains the locant in the monosaccharide subunit of the parent oligosaccharide chain to which the branch is attached. The LOCT field on the following, Carbohydrate Ring Entry, as usual holds the locant on the first branching unit that attaches the branch to its parent oligosaccharide chain. The rest of the monosaccharide subunit CCT entries in the branch follow the conventional format of linear polysaccharide representation.

Polysaccharide branches appear within the overall CCT in the sequence implied by the SUBS field local branch counts of each parental subunit CME. This is in line with the overall approach of the CME format, to describe first a parent structure, followed by its substituents in attachment locant order.

CONCLUSION

While the recent extensions to the CCT for steroids and saccharides have been introduced mainly for those structures as parent main entries, the overall hierarchic CCT format is capable of representing combinations of these specialist forms together with the other substructures. The format can support further types of main and modification entry being added in the future, and thus the CCT remains a flexible mechanism for coding structural information at a higher functional level than the classical atom/bond approach.

ACKNOWLEDGMENT

We are grateful for the continued support provided by the U.K. Laboratory of the Government Chemist and for the helpful advice of Mr. I. Cohen of its Chemical Nomenclature Advisory Service.

REFERENCES AND NOTES

(1) Rayner, J. D. A Concise Connection Table Based on Systematic Nomenclatural Terms. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 108–111.
(2) 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*, 101–105.
(3) 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.
(4) 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.
(5) Cooke-Fox, D. I.; Kirby, G. H.; Lord, M. R.; 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.
(6) Cooke-Fox, D. I.; Kirby, G. H.; Lord, M. R.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 5. Steroid Nomenclature. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 128–132.
(7) Kirby, G. H.; Lord, M. R.; 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.
(8) 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.
(9) IUPAC/IUB. Tentative Rules for Carbohydrate Nomenclature. *J. Biol. Chem.* **1972**, *247* (3), 613–635.