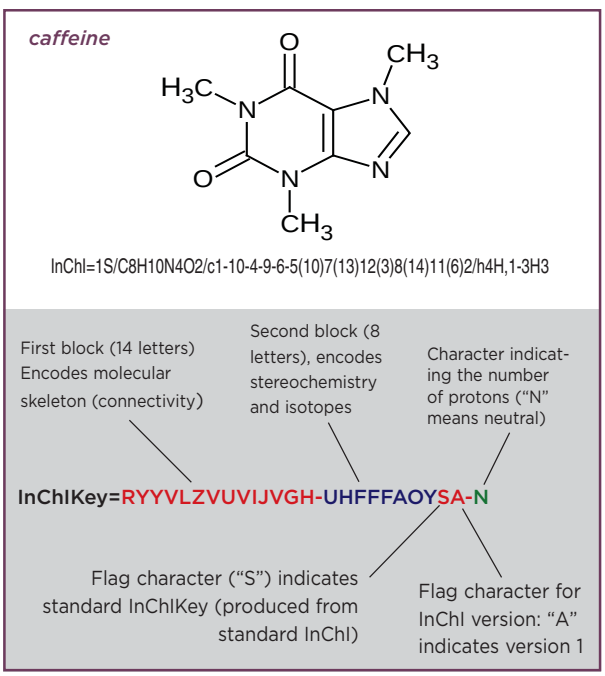
InChIs and Registry Numbers

by Jeffery Leigh

Constructing a systematic name of a chemical compound of known structure means that it is necessary for the reader to know the detailed nomenclature rules required to do this. Such a person must work within a particular system, of which IUPAC and the Chemical Abstracts Service (CAS) provide possibly the two most complete. These are both designed in the English language, but a person whose mother tongue is not English may face a further barrier to developing a name if another language, such as Russian or Japanese, should be the language of primary use. However, all trained chemists should be able to recognize a chemical structure displayed using atomic symbols and bond connections, etc., as these are independent of language, even if the basic chemical symbols, recognizable by all chemists, use the roman alphabet. IUPAC has recently developed a computer methodology for recognizing and codifying chemical structures, and, the converse, for reproducing the chemical structure from such a code. This code is called an InChI (IUPAC International Chemical Identifier), and there is a related, more abbreviated, version called an InChIKey.

Principles provides a brief introductory guide to InChIs, and InChIKeys, along with CAS Registry Numbers. These are quite distinct from each other. Registry Numbers are unique numbers used to identify a specific compound in the CAS database. The numbers are assigned to a compound the first time that it appears in Chemical Abstracts (CA) and can be used thereafter to find all references to this compound when it appears again in CA. However, it has no further significance, and does not contain structural information. The user of CA must be familiar with CA nomenclature. In contrast, the recently developed InChI and its related InChIKey are strings of numbers, letters, and other symbols that provide a complete description of the structure of a compound (see www.iupac.org/inchi).



The strings are not comprehensible to a casual reader, though they are to a computer that is equipped with the necessary programs. The InChI system can now deal with many, but as yet not all, compounds that appear in the literature, but it is still under development. In theory, InChI software will eventually provide an InChI character string from structural data for any compound. InChIs are already being used by most of the major chemistry publishers and databases. The InChITrust website (www.inchi-trust.org) lists some of the many organizations now using it.

The InChI software represents features of the compound structure as a sequence of levels and in strict order, starting with the formula and then dealing with various features, such as atom connectivity and stereochemistry. Production of an InChI is reversible, in that with the appropriate computer program it can be derived from a drawing of the structure and it can then be used to regenerate the structure from which it is derived. This is not true of the InChIKey. An InChI may contain many tens of characters. An InChIKey is an abbreviated form which contains only 27 characters. It cannot be used to regenerate its parent structure, but it is still unique and is designed primarily for searching databases. In that sense it is like a Registry Number, but unlike the Number, it derives ultimately from a compound's structure.

Of course, a primary requirement for someone to use both InChIs and InChIKeys is that they possess the programs that can construct and interpret them. Like all IUPAC products, these are freely available to the chemistry community. *Principles* contains enough detail and references for a beginner to obtain the programs and to start to use them.

Jeffery Leigh is the editor and contributing author of *Principles of Chemical Nomenclature—A Guide to IUPAC Recommendations*, 2011 Edition (RSC 2011, ISBN 978-1-84973-007-5). Leigh is emeritus professor at the University of Sussex and has been active in IUPAC nomenclature since 1973.

www.iupac.org/publications/ci/indexes/nomenclature-notes.html

Next in this series will be a review of Deciphering and Constructing Names.