The IUPAC NIST Chemical Identifier (INChI)

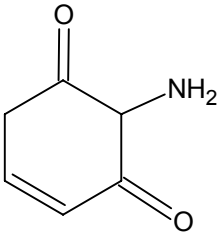
D. Tchekhovskoi, S. Stein (838), S. Heller (Department of Agriculture, retired)

The question of clearly identifying a chemical has been present almost since the beginnings of modern chemistry. As the number of chemicals grew, the need for systematic naming produced a number of results. The most widely adopted of these is that of the International Union of Pure and Applied Chemistry (IUPAC). But for many chemicals the resulting names are complex and so common names are still widely used. For most chemists, the graphical structure is the best method for identifying a chemical. The structure provides graphic information that can allow a rapid understanding of the properties of the chemical that a long text name can never provide.

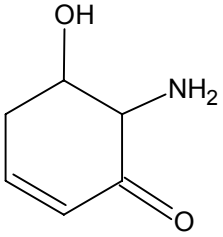
IUPAC

The goal of this research is to create a chemical naming system that would allow computers to uniquely identify a chemical, regardless of how it is drawn based entirely on the connectivity of the molecule – i.e., what atoms are connected to what other atoms.

The complications of names even for very simple molecules are illustrated below:



2-amino-4-cyclohexene-1,3-dione



6-amino-5-hydroxy-2-cyclohexen-1-one

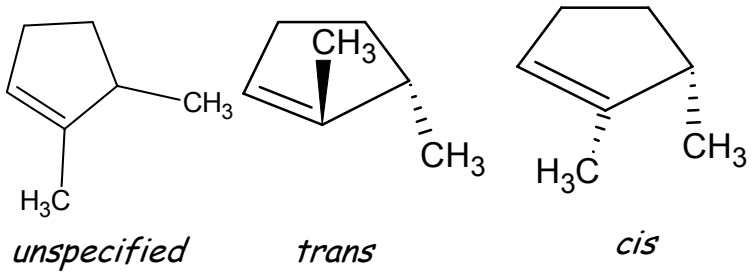
The numbering of both the amino group and the double bond changes in these names because of a set of rules for establishing precedence in functional groups. The structures allow the chemist to see the direct connection between the two molecules.

The complication with structures is that they are not easy to order in the sense that names can be ordered by simple alphabetical rules and they do not have an obvious storage method in the computer. Thus there has been a need to develop methods of describing a structure with a simple string that can be used both to order sets and to find a specific compound. Computer storage methods and search algorithms have been able for some time to process structural data but these have not been standardized (there is no agreed upon alphabet for structure representation in computers) and so the various systems for processing the data have not been able to easily communicate.

To accomplish this, much of what is normally viewed as "chemical information" was discarded and the molecules were trimmed to the minimum information needed to differentiate one from the other. In addition, a layered approach was developed to deal with some of the more complex issues of chemical structure.

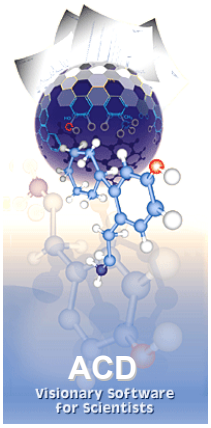
The IUPAC-NIST Chemical Identifier (INChI)

The need for a uniform and open standard that could be adopted by the entire chemical community prompted the NIST/IUPAC project to develop a chemical identifier. The aim of the project was not to create another naming system for normal communication, but rather to create a naming system that would allow computers to uniquely identify a chemical.



For example the two molecules in the illustration differ only in that one has the two methyl groups on the same side relative to the plane of the ring (on the right – called cis) and the other has the two methyl groups on opposite sides of the ring (in the center – called trans). On the left is a diagram that can be used to represent either of the molecules. The left diagram shows only the connectivity and does not specify if the molecule is cis or trans. The problem encountered prior to INChI is that the data retrieval was often dependent upon the way the molecule had been drawn. There is often a need to distinguish between the cis and trans form, and often a need to search for all possible forms, including cases where the configuration of the molecule was not known or it was known that a mixture was present.

The approach taken in developing INChI is a layered approach. This allowed as much information as was known to be specified, the search could be performed only on the information known, and the search could be stopped with less than full information. Thus, in the case above, a search for the cis isomer could be allowed to stop when it matched the connectivity or continued to find only the molecules that matched the geometric isomer. The methodology of INChI also conforms with the XML standards and the output of the method can be done in XML or in simple text.



The IUPAC NIST Chemical Identifier has been released for beta testing. The current version has been adopted by PubChem at NIH, is an integral part of the Chemical Markup Language (CML) standard, and has been integrated by ACD Labs in their widely used commercial drawing program, ChemSketch. In addition INChI will be integrated into the next version of the Chemistry WebBook so that anyone with access to the Internet can make use of this technology.