# InChI version 1 validation protocol

## *1. Purpose*

The purpose of InChI validation is to verify that InChI code included in an application or InChI code ported to a particular compiler or operating system (such an application containing InChI code later in this document is called "software") produces the same InChI as the software distributed by IUPAC, cInChI-1, software version 1.01.

## *2. Method*

The method of validation is to compare InChI produced by the software out of a representative set of chemical structures and with various options to InChI produced by the official IUPAC software. Identical results are necessary for the software to pass the validation test.

## *3. Representative set of chemical structures*

This InChI validation suite includes a minimal set of chemical structures presented as a single SDfile, InChI_TestSet.sdf.

To obtain a representative set of structures, a greater collection of structures has to be included in validation procedure. Such a collection is NCI Open "September 2003 SD File of Combined DTP Releases, 2D/3D, with Canonical Properties Added," containing 260,071 structures. The compressed SDfile may be downloaded from
http://cactus.nci.nih.gov/ncidb3/download_ncidb3.html
The direct link to the file is "260,071 structures in SDF format",
http://cactus.nci.nih.gov/DownLoad/NCI-Open_09-03.sdf.gz

To decompress this file, a gzip, WinZip, or similar utility is needed.

To refer to a single chemical structure, the following SDF IDs should be used:

| Collection | InChI_TestSet | NCI Open |
|---|---|---|
| Sdf:ID | ID | NSC |

## *4. Sets of InChI options used for validation*

Below are 28 combinations (sets) of options used in validation procedure:

| Options set | FixedH | RecMet | newps | Stereo | InChI options |
|---|---|---|---|---|---|
| 01 | no | no | no | none | /Snon |
| 02 | no | no | no | absolute | /fb |
| 03 | no | no | no | relative | /Srel /fb |
| 04 | no | no | no | racemic | /Srac /fb |
| 05 | yes | no | no | none | /FixedH /Snon |
| 06 | yes | no | no | absolute | /FixedH /fb |
| 07 | yes | no | no | relative | /FixedH /Srel /fb |
| 08 | yes | no | no | racemic | /FixedH /Srac /fb |
| 09 | no | yes | no | none | /RecMet /Snon |
| 10 | no | yes | no | absolute | /RecMet /fb |
| 11 | no | yes | no | relative | /RecMet /Srel /fb |
| 12 | no | yes | no | racemic | /RecMet /Srac /fb |
| 13 | yes | yes | no | none | /FixedH /RecMet /Snon |
| 14 | yes | yes | no | absolute | /FixedH /RecMet /fb |
| 15 | yes | yes | no | relative | /FixedH /RecMet /Srel /fb |
| 16 | yes | yes | no | racemic | /FixedH /RecMet /Srac /fb |
| 17 | no | no | yes | absolute | /newps /fb |
| 18 | no | no | yes | relative | /Srel /newps /fb |
| 19 | no | no | yes | racemic | /Srac /newps /fb |
| 20 | yes | no | yes | absolute | /FixedH /newps /fb |
| 21 | yes | no | yes | relative | /FixedH /Srel /newps /fb |
| 22 | yes | no | yes | racemic | /FixedH /Srac /newps /fb |
| 23 | no | yes | yes | absolute | /RecMet /newps /fb |
| 24 | no | yes | yes | relative | /RecMet /Srel /newps /fb |
| 25 | no | yes | yes | racemic | /RecMet /Srac /newps /fb |
| 26 | yes | yes | yes | absolute | /FixedH /RecMet /newps /fb |
| 27 | yes | yes | yes | relative | /FixedH /RecMet /Srel /newps /fb |
| 28 | yes | yes | yes | racemic | /FixedH /RecMet /Srac /newps /fb |

To reduce the size of the output, two additional options should be added to each set: /AuxNone /NoLabels.
For example, the 01 *set of InChI options* is:
/Snon /AuxNone /NoLabels

A simple Win32 command file TestSet2InChI.bat was used to generate all 28 result files out of InChI_TestSet.sdf. The result files have names InChI_TestSet_01.txt, …, InChI_TestSet_28.txt. These files are included in the package.

## *5. The command line syntax to generate a result file*

cInChI-1 SDfile ResultFile LogFile NUL <*set of InChI options*>

For example, to generate the 01 result file:
cInChI-1 SDfile ResultFile01 LogFile01 NUL /Snon /AuxNone /NoLabels

For platforms other than Win32 the options should be preceded with a minus instead of the slash:
cInChI-1 SDfile ResultFile01 LogFile01 NUL -Snon -AuxNone -NoLabels

NUL instead of the problem file name is necessary if the SDfile size is greater than 2 gigabytes.

## *6. Validation procedure*

The following steps constitute the InChI validation procedure. The 28 sets of InChI options needed to generate the 28 result files are listed in Section 4; the command line syntax is explained in Section 5.

a) Assemble a representative set of chemical structures (InChI_TestSet.sdf and NCI Open collection described in Section 3)
b) Generate 28 result files for each SDfile using distributed by IUPAC cInChI-1 program, software version 1.01. This creates the standard set of InChI
c) Generate 28 result files for each SDfile using the software being validated. This creates the test set of InChI
d) Compare these two sets of InChI. If the standard set and the test set contain exactly same InChI in exactly same order then the software has passed the validation.

## *7. Notes*

1. The original InChI software released in 2005 will not pass this validation procedure. The 2006 InChI version 1 software version 1.01 is needed because it has several bugs fixed.

2. In general, this validation procedure establishes necessary but not sufficient conditions for the software compliance: it does not mathematically prove that there does not exist a structure such that the identifier produced by the software would be different from that produced by cInChI-1. A practical solution is to test the software on as great as possible variety of structures. Therefore, at the discretion of the tester, other structure collection(s) may be included in validation. One of such publicly available collections is the PubChem collection of Compound chemical structures. It may be found at
ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound/
The SDF ID in this collection is PUBCHEM_COMPOUND_CID.
The latest Compound structures are usually located in the directory
ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound/CURRENT-Full/SDF/
(643 compressed SDfiles containing 5,223,363 structures as of 4/15/2006)

## *8. An additional tool for checking InChI syntax*

InChI version 1 software version 1.01 has a built-in tool able to detect significant syntax errors in an identifier. This may be needed to quickly find syntax errors in an InChI of an unknown origin. The detection is a two-stage procedure. To check InChI in file input_inchi_file.txt:

1. Run cInChI-1:
cInChI-1 input_inchi_file.txt output_inchi_file.txt logfile.log NUL /inchi2inchi /FixedH /RecMet

2. Compare identifiers in input_inchi_file.txt to identifiers in newly created output_inchi_file.txt

If a difference has been found then there is a significant InChI syntax error in file input_inchi_file.txt. In most cases the error may be detected at stage 1; in this case it is described in the logfile.log file.

Note that there are InChI errors that cannot be detected with this tool.

## *9. List of files included in the InChI validation protocol package*

InChI_v1_validation_protocol.doc
InChI_v1_validation_protocol.pdf
InChI_TestSet.sdf
TestSet2InChI.bat
TestSet2InChI.sh
InChI_TestSet_01.txt
InChI_TestSet_02.txt
InChI_TestSet_03.txt
InChI_TestSet_04.txt
InChI_TestSet_05.txt
InChI_TestSet_06.txt
InChI_TestSet_07.txt
InChI_TestSet_08.txt
InChI_TestSet_09.txt
InChI_TestSet_10.txt
InChI_TestSet_11.txt
InChI_TestSet_12.txt
InChI_TestSet_13.txt
InChI_TestSet_14.txt
InChI_TestSet_15.txt
InChI_TestSet_16.txt
InChI_TestSet_17.txt
InChI_TestSet_18.txt
InChI_TestSet_19.txt
InChI_TestSet_20.txt
InChI_TestSet_21.txt
InChI_TestSet_22.txt
InChI_TestSet_23.txt
InChI_TestSet_24.txt
InChI_TestSet_25.txt
InChI_TestSet_26.txt
InChI_TestSet_27.txt
InChI_TestSet_28.txt
<end of the list>