User's Guide:

IUPAC Standard International Chemical Identifier (InChI) Program

Version 1, Software version 1.02

Last revision date January 10, 2009

This is the release of the IUPAC standard International Chemical Identifier with InChIKey, version 1, software version 1.02. (http://www.iupac.org/projects/2000/2000-025-1-800.html http://www.iupac.org/inchi).

The release conforms to standard InChI and standard InChIKey definitions (see Standard InChI User's Guide) as established by IUPAC InChI Subcommittee at its September 15, 2008 meeting (for details, see http://sourceforge.net/mailarchive/message.php?msg_name=20081002155703.FXZB29597.aamtaout03-winn.ispmail.ntl.com%40ALAN).

CONTENTS

I. OVERVIEW

About InChI

The IUPAC International Chemical Identifier (InChI) provides unique labels for well-defined chemical substances. These labels are generated by converting an input chemical structure, in the form of a ‘connection table’, to a unique and predictable series of ASCII characters. They offer a means for representing chemical compounds in a manner that does not depend on how they were drawn. Note that they are re-expressions of chemical structures, they are not registry or registration numbers and do not require access to a database. They were developed primarily as a means of ‘naming’ a compound in digital media although it is expressed as simple text that may be manually interpreted. This document describes the operation and output of the final version of the program that generates this Identifier.

The Identifier is designed to process single, well-defined chemical compounds. These compounds may be composed of multiple components.

Technical details are given in a separate document, the InChI Technical Manual, and in an earlier version on the IUPAC Website (http://www.iupac.org/projects/2000/2000-025-1-800.html). The basic algorithms were taken from the literature, with selection, testing and implementation done at NIST. In the several years of its development, many individuals contributed to the development of the InChI at meetings and through correspondence. The chemical rules employed are intended to represent a consensus view of the concept of chemical identity. The computer program described in this document applies these algorithms to input structures and generates both the Identifier and an annotated depiction of the structure.

Derivation of the InChI from an input chemical structure proceeds through three steps: 1) normalization – all input information not needed for structure identification is discarded and structure information is divided into ‘layers’; 2) canonicalization – each atom is given a label that depends only on its position in the structure; 3) serialization – a string of characters, the Identifier, is generated from the canonical labels. All ‘chemical’ rules are applied in the first step.

About standard InChI

InChI has a layered structure which allows one to represent molecular structure with a desired level of details. Accordingly, InChI software (releases 1.01 and 1.02-beta) was able to generate different InChI strings for the same molecule, dependent on a multitude of options (e.g., distinguishing or not between tautomers).

This flexibility, however, may be considered a drawback with respect to standardization/interoperability. The standard InChI was defined by IUPAC InChI Subcommittee in response to these concerns.

As related to its internal (layered) structure, standard InChI v.1, introduced in this v. 1.02 release of InChI software, is a subset of IUPAC International Chemical Identifier v.1 implemented and described in previous software releases, v. 1.01 in 2006 and v. 1.02-beta in 2007.

The standard InChI was defined to reach following goals.

- Standard InChI is for the purposes of interoperability/compatibility between large databases/web searching and information exchange.
- Standard InChI and non-standard InChI are always distinguishable.
- Standard InChI is a stable identifier; however, periodic updates may be necessary; they are reflected in the identifier version designation, which is included in the InChI string.
- Any shortcomings in standard InChI may be addressed using non-standard InChI.

The layered structure of the standard InChI conforms to the following requirements.

- Standard InChI organometallic representation should not include bonds to metal for the time being.
- Standard InChI distinguishes between chemical substances at the level of ‘connectivity’, ‘stereochemistry’, and ‘isotopic composition’, where:
  - connectivity means tautomer-invariant valence-bond connectivity (different tautomers have the same connectivity/hydrogen layer);
  - stereochemistry means configuration of stereogenic atoms and bonds; unknown stereo designations are treated as undefined;
  - isotopic composition means mass number of isotopic atoms (when specified)

In the light of the above requirements, the following options are selected for generation of standard InChI

- include tautomerism (i.e., turn mobile H perception on, exclude “fixed hydrogen atoms layer);
- omit reconnection of bonds to metal atoms;
- only a narrow end of a wedge points to a stereocenter;
- include all bug fixes (previous command-line options “Fb”, “Fb2”, “Fnud”) without a possibility to turn them off;
- exclude unknown/undefined stereo if no other stereo is present;
- include stereochemistry of phosphines and arsines;
- treat stereochemistry as absolute (not relative or racemic).

The standard InChI is designated by prefix:

“InChI=1S/……….. “

(that is, letter ‘S’ immediately follows the version number; standard InChI version numbers should always be whole numbers).

Note that InChI v.1 Technical Manual still remains the main technical description of IUPAC International Chemical Identifier. Standard InChI is a subset of InChI described in the Technical Manual.

About standard InChIKey

The InChIKey (introduced in InChI v.1 software v. 1.,02-beta) is a character signature based on a hash code of the InChI string. A hash code is a fixed length condensed digital representation of a variable length character string. Providing a hash derived from an InChI string should be helpful in search applications, including Web searching and chemical structure database indexing; also, this hash may serve as a checksum for verifying InChI, for example, after a transmission over a network.

Standard InChIKey, introduced in this v. 1.02 release of standard InChI software, is computed only from the standard InChI. It serves for the principal purpose of a search-engine-style lookup of chemical information.

Standard InChIKey is a stable identifier; however, periodic updates may be necessary; this is reflected in the version included in the InChIKey string.

Note that the format of the standard InChIKey is different from that introduced in InChI software v. 1.02-beta.

Standard InChIKey has five distinct components.

(1) 14-character hash of the basic (Mobile-H) InChI layer;
(2) 8-character hash of the remaining layers (except for the "/p" segment, which accounts for added or removed protons: it is not hashed at all; the number of protons is encoded at the end of the standard InChIKey.)
(3) 1 flag character,
(4) 1 version character
(5) the last character is [de]protonation indicator.

The overall length of InChIKey is fixed at 27 characters, including separators (dashes):

AAAAAAAAAAAAAA-BBBBBBBBFV-P

Here
(1) **AAAAAAAAAAAAAA** is a 14-character hash.
(2) **BBBBBBBB** is an 8-character hash
(3) **F** is a flag indicating standard InChIKey (produced out of standard InChI): it always has the value 'S'.
(4) **V** is a flag for InChI version character: 'A' for version 1, 'B' for version 2, etc.
(5) **P** is an indicator for the number of protons; this number is not encoded in the hash but is indicated as a separate 2-character block at the end, where one character is a hyphen, as –N for neutral, -M for -1 hydrogen, -O for +1 hydrogen, etc.

More details about standard InChIKey may be found in Appendix 1.

II. ABOUT THE STANDARD INCHI PROGRAM

This document is accompanied by the version 1.02 of the standard InChI generator. This program runs under 32-bit Microsoft Windows Operating Systems. The main program, stdwinchi-1.exe, is a conventional Windows application, although a ‘command line’ version is also included (stdinchi-1.exe). A version recompiled under i386 Linux without any changes is also included. The program takes an input structure and generates both graphical and text output in a form designed to allow critical examination of the standard InChI. The Identifier and associated text output may be parsed and annotated in either a simple plain text or XML (eXtensible Markup Language) format.

As structure input, the program currently accepts standard SDfiles, Molfiles [see “Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited” by Arthur Dalby, James G. Nourse, W. Douglas Hounshell, Ann K. I. Gushurst, David L. Grier, Burton A. Leland, and John Laufer, Journal of Chemical Information and Computer Sciences, 1992; 32(3); pp. 244-255; a more recent description of V2000 format may be downloaded from http://www.mdli.com/downloads/public/ctfile/ctfile.jsp], CML files [http://www.xml-cml.org] or its own output produced when the “Full auxiliary information” option is selected. Input may originate from individual disk files or through the Windows clipboard. InChI may be also generated directly from an application programming interface (API). This is described later.

III. INSTALLING AND RUNNING THE STANDARD INCHI PROGRAM

The standard InChI generation program is provided along with sample chemical structures in a ‘zip’ file – STDINCHI-1-BIN .zip. To use this program, first extract the contents of the file to a directory of your choice. To start the program, run the file stdwinchi-1.exe that was extracted from the zip file. Figure 1 then appears on your monitor.

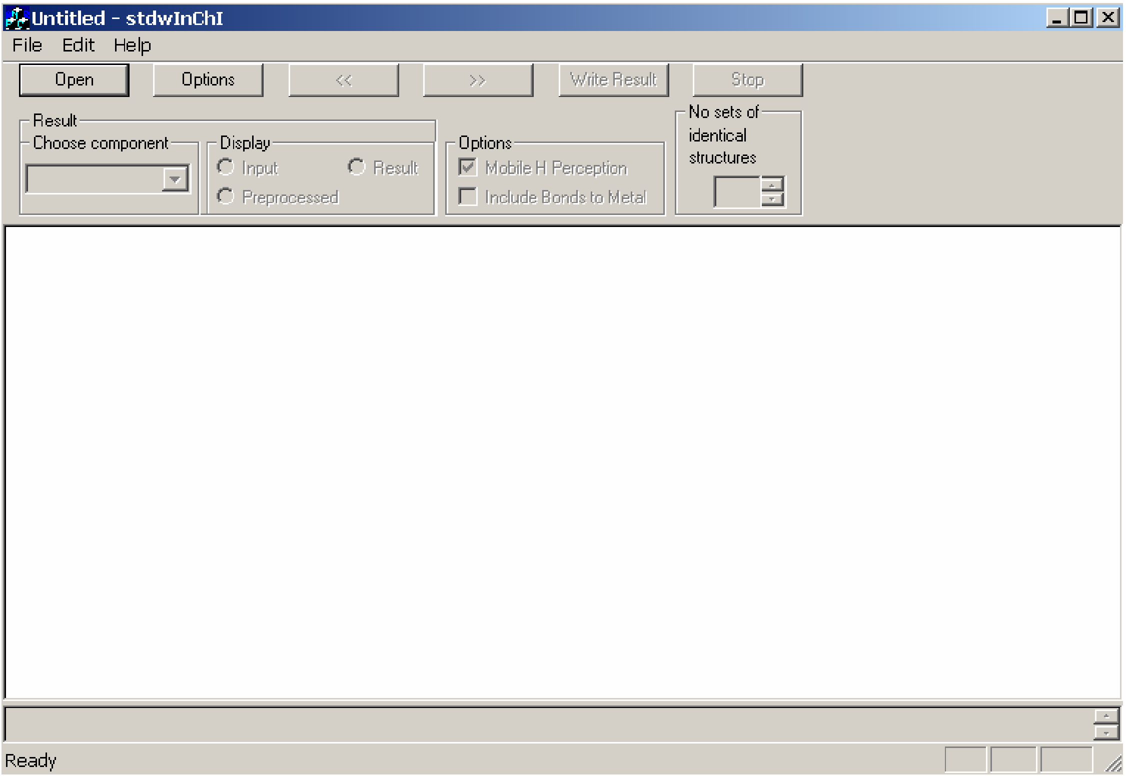


Figure 1

Generating an InChI begins with the selection of an input structure file. The simplest way is to drag the input structure file from Windows Explorer directory list into the InChI window. Structures also may be copied from certain chemical structure editors (ISIS/Draw with “Copy Mol/Rxnfile to the Clipboard” option or from ACD/ChemSketch) and pasted into the InChI window (Select Edit → Paste from InChI menu). Input structure file pathname may be provided as a command line option when you start wInChI. Selection of the input structure file may also be done by first clicking on the ‘Open’ button (top left corner of Figure 1) and then, in the Dialog box that appears (as shown in Figure 2),

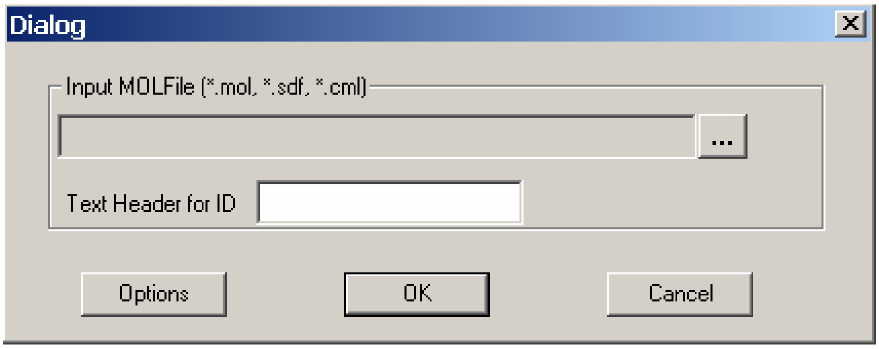
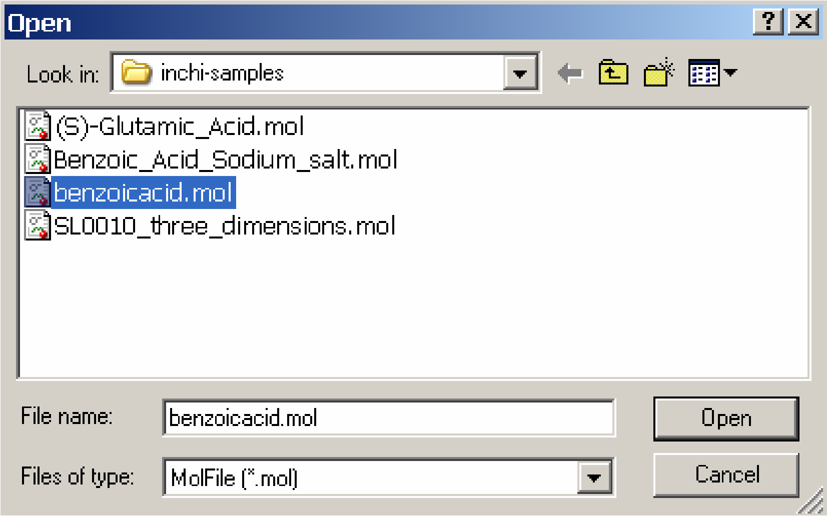


Figure 2

selecting a structure file using the '…' button on the right of the 'Input Structure File' field. You may select any of the sample .mol, .sdf or .cml files for initial testing. In this Dialog you may also enter "Text Header for ID"; this will simply add to the InChI header a structure ID if it is present in an input SDfile (from other input formats the header and ID are extracted automatically). Ignore this box for now.



Look in: inchi-samples

(S)-Glutamic_Acid.mol
Benzoic_Acid_Sodium_salt.mol
benzoicacid.mol
SL0010_three_dimensions.mol

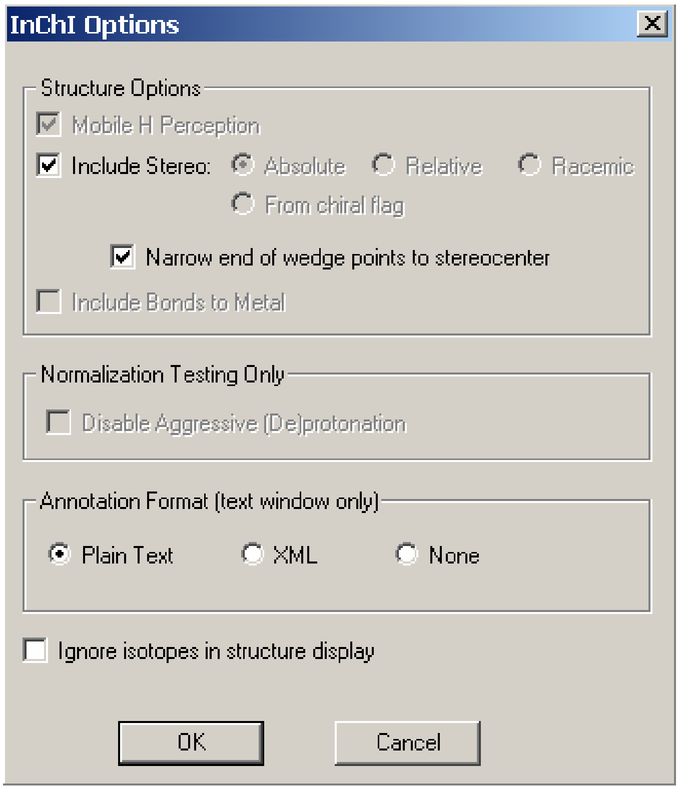
File name: benzoicacid.mol

Files of type: MolFile (*.mol)

Figure 3

Figure 3 shows the selection of a structure file. In this case it is entitled benzoicacid.mol, which was prepared by a separate structure-drawing program. Clicking the file name copies it into "File name:" line. After that click "Open" to close the dialog.

At this point you may also change InChI processing options. (The choices for the options that can be changed are shown in Figure 4, but no changes are made in this example.)


InChI Options

Structure Options

☑ Mobile H Perception

☑ Include Stereo: ● Absolute ○ Relative ○ Racemic
○ From chiral flag

☑ Narrow end of wedge points to stereocenter

□ Include Bonds to Metal

Normalization Testing Only

□ Disable Aggressive (De)protonation

Annotation Format (text window only)

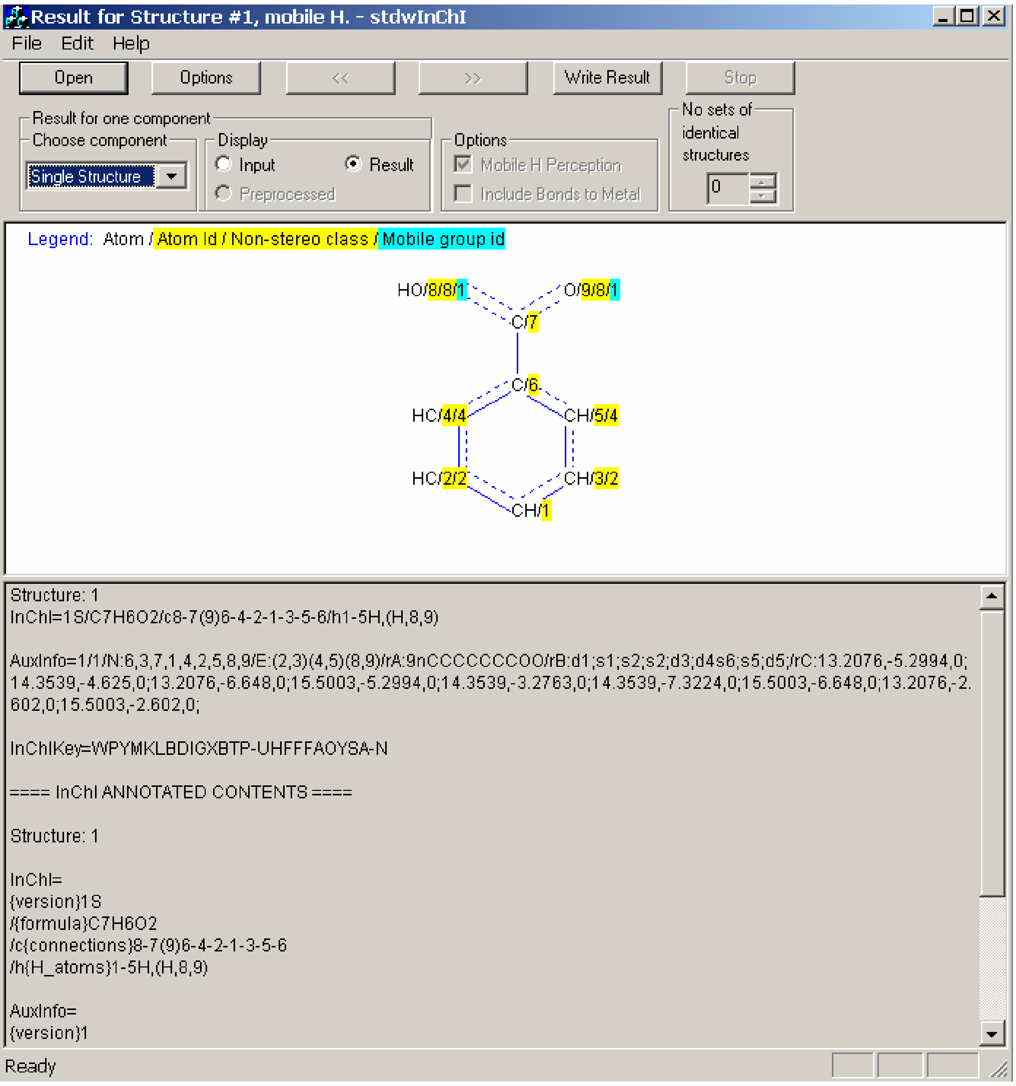
● Plain Text ○ XML ○ None

□ Ignore isotopes in structure display

Figure 4

Close InChI Options dialog if you opened it and select OK in the dialog (Fig. 2) when done; the result is Figure 5.

The main output window is composed of two sections, the upper section (shown in white in Figure 5) shows structural information graphically and the lower section (shown in gray in Figure 5) shows text output.


Result for Structure #1, mobile H. - stdwInChI
File
Edit
Help
Open
Options
<<
>>
Write Result
Stop
Legend: Atom / Atom Id / Non-stereo class / Mobile group id
HO/8/8/1
O/9/8/1
C/7
C/6
HC/4/4
CH/5/4
HC/2/2
CH/3/2
CH/1
Ready

Result for one component

Choose component: Single Structure

Display: ○ Input ● Result ○ Preprocessed

Options: ☑ Mobile H Perception □ Include Bonds to Metal

No sets of identical structures: 0

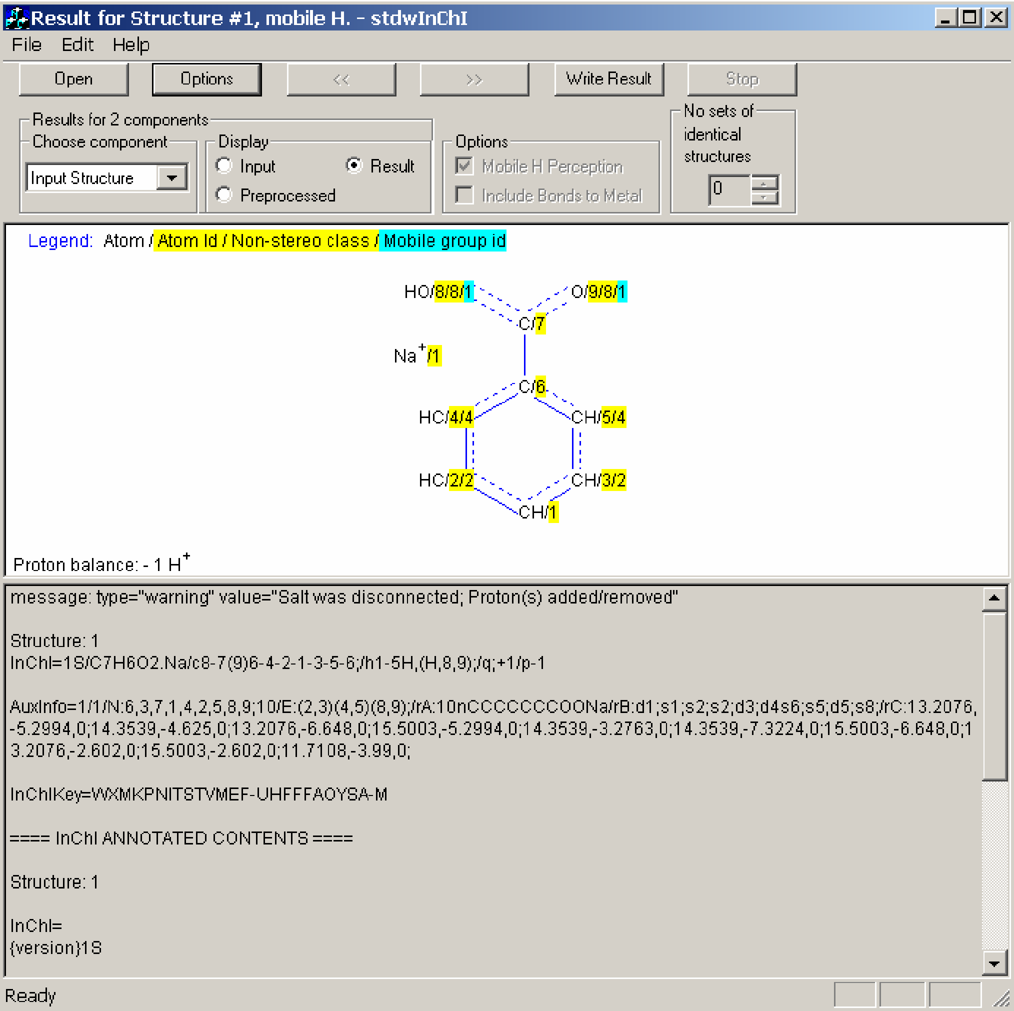
```
Structure: 1
InChI=1S/C7H6O2/c8-7(9)6-4-2-1-3-5-6/h1-5H,(H,8,9)

AuxInfo=1/1/N:6,3,7,1,4,2,5,8,9/E:(2,3)(4,5)(8,9)/rA:9nCCCCCCCOO/rB:d1;s1;s2;s2;d3;d4s6;s5;d5;/rC:13.2076,-5.2994,0;
14.3539,-4.625,0;13.2076,-6.648,0;15.5003,-5.2994,0;14.3539,-3.2763,0;14.3539,-7.3224,0;15.5003,-6.648,0;13.2076,-2.
602,0;15.5003,-2.602,0;

InChIKey=WPYMKLBDIGXBTP-UHFFFAOYSA-N

==== InChI ANNOTATED CONTENTS ====

Structure: 1

InChI=
{version}1S
/{formula}C7H6O2
/c{connections}8-7(9)6-4-2-1-3-5-6
/h{H_atoms}1-5H,(H,8,9)

AuxInfo=
{version}1
```

Figure 5

Upper section

The structure is displayed along with labels generated by InChI algorithms. In cases where an SDF or CML file is input, the first structure shown is the first entry in the input file. The example shown in Figure 5 is a single component example. If more than one component (independent structure) is found in the first structure file (such as Benzoic Acid, Sodium salt shown in Figure 6), each may be separately examined using the “Choose component” ‘combo box’ on the upper left of the screen, although they are treated as part of a single compound by InChI (Figures 7 and 8).

The buttons under “Display” permit viewing of the input structure and the preprocessed structure if it differs from the input structure. The buttons under “Options” are disabled (shown in gray; these options are not available for standard InChI).


Result for Structure #1, mobile H. - stdwInChI
File
Edit
Help
Open
Options
<<
>>
Write Result
Stop
Legend: Atom / Atom Id / Non-stereo class / Mobile group id
HO/8/8/1
O/9/8/1
C/7
Na+/1
C/6
HC/4/4
CH/5/4
HC/2/2
CH/3/2
CH/1
Ready

Results for 2 components

Choose component: Input Structure

Display: ○ Input ● Result ○ Preprocessed

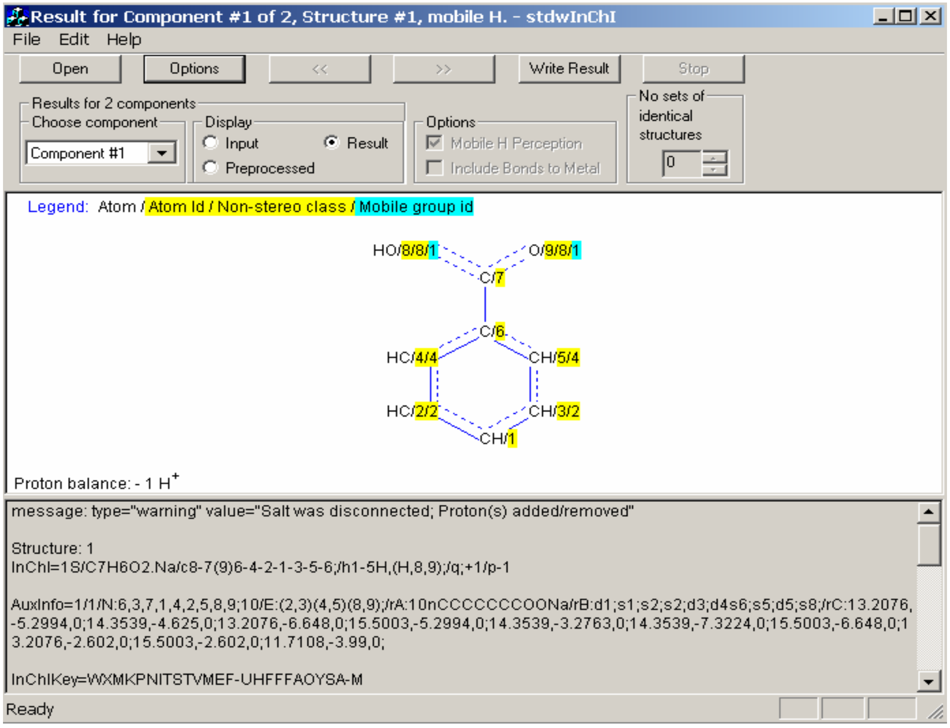
Options: ☑ Mobile H Perception ☐ Include Bonds to Metal

No sets of identical structures: 0

Proton balance: - 1 H+

```
message: type="warning" value="Salt was disconnected; Proton(s) added/removed"

Structure: 1
InChI=1S/C7H6O2.Na/c8-7(9)6-4-2-1-3-5-6;/h1-5H,(H,8,9);/q;+1/p-1

AuxInfo=1/1/N:6,3,7,1,4,2,5,8,9;10/E:(2,3)(4,5)(8,9);/rA:10nCCCCCCCOONa/rB:d1;s1;s2;s2;d3;d4s6;s5;d5;s8;/rC:13.2076,
-5.2994,0;14.3539,-4.625,0;13.2076,-6.648,0;15.5003,-5.2994,0;14.3539,-3.2763,0;14.3539,-7.3224,0;15.5003,-6.648,0;1
3.2076,-2.602,0;15.5003,-2.602,0;11.7108,-3.99,0;

InChIKey=WXMKPNITSTVMEF-UHFFFAOYSA-M

==== InChI ANNOTATED CONTENTS ====

Structure: 1

InChI=
{version}1S
```

Figure 6



Results for 2 components

Choose component: Component #1

Display: ○ Input ● Result ○ Preprocessed

Options: ☑ Mobile H Perception ☐ Include Bonds to Metal

No sets of identical structures: 0

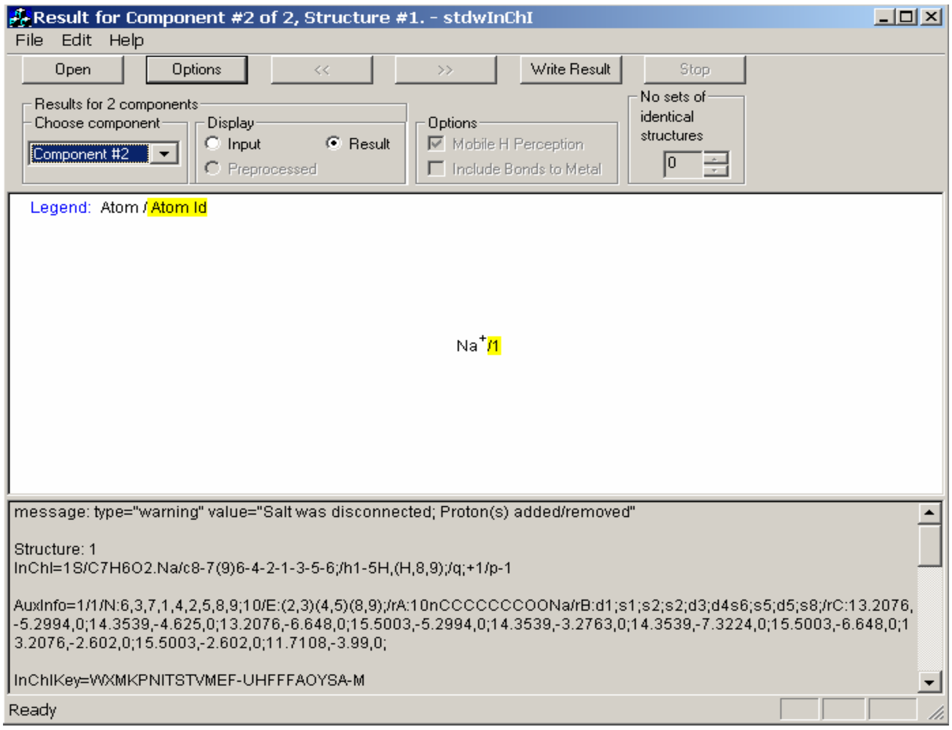
Legend: Atom / Atom Id / Non-stereo class / Mobile group id

Proton balance: - 1 H^+

```
message: type="warning" value="Salt was disconnected; Proton(s) added/removed"

Structure: 1
InChI=1S/C7H6O2.Na/c8-7(9)6-4-2-1-3-5-6;/h1-5H,(H,8,9);/q;+1/p-1

AuxInfo=1/1/N:6,3,7,1,4,2,5,8,9;10/E:(2,3)(4,5)(8,9);/rA:10nCCCCCCCOONa/rB:d1;s1;s2;s2;d3;d4s6;s5;d5;s8;/rC:13.2076,
-5.2994,0;14.3539,-4.625,0;13.2076,-6.648,0;15.5003,-5.2994,0;14.3539,-3.2763,0;14.3539,-7.3224,0;15.5003,-6.648,0;1
3.2076,-2.602,0;15.5003,-2.602,0;11.7108,-3.99,0;

InChIKey=WXMKPNITSTVMEF-UHFFFAOYSA-M
```

Ready

Figure 7



Results for 2 components

Choose component: Component #2

Display: ○ Input ● Result ○ Preprocessed

Options: ☑ Mobile H Perception ☐ Include Bonds to Metal

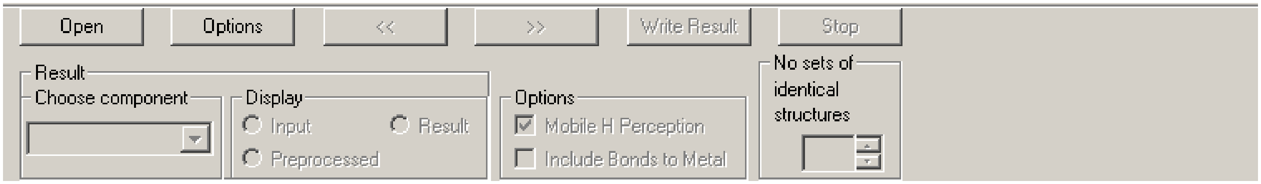
No sets of identical structures: 0

Legend: Atom / Atom Id

```
message: type="warning" value="Salt was disconnected; Proton(s) added/removed"

Structure: 1
InChI=1S/C7H6O2.Na/c8-7(9)6-4-2-1-3-5-6;/h1-5H,(H,8,9);/q;+1/p-1

AuxInfo=1/1/N:6,3,7,1,4,2,5,8,9;10/E:(2,3)(4,5)(8,9);/rA:10nCCCCCCCOONa/rB:d1;s1;s2;s2;d3;d4s6;s5;d5;s8;/rC:13.2076,
-5.2994,0;14.3539,-4.625,0;13.2076,-6.648,0;15.5003,-5.2994,0;14.3539,-3.2763,0;14.3539,-7.3224,0;15.5003,-6.648,0;1
3.2076,-2.602,0;15.5003,-2.602,0;11.7108,-3.99,0;

InChIKey=WXMKPNITSTVMEF-UHFFFAOYSA-M
```

Ready

Figure 8



○ Input ○ Result
○ Preprocessed

☑ Mobile H Perception
□ Include Bonds to Metal

InChI Toolbar

On the InChI Toolbar the rightmost box displays number of sets of equivalent components. When equivalent components are found, they may be highlighted by making a selection in the box. This provides a quick way to determine if two depictions of the same compound are considered to be the same by InChI algorithms, although the actual InChI generated will represent the collection of structures as a single compound.

The structure display shows the canonical identification number of each atom along with the non-stereo equivalence class number assigned to that atom. The canonical number is the unique number given to an atom and used for ‘serialization’ (creation of the actual InChI). The non-stereo equivalence class number is a number assigned to each set of equivalent atoms (all atoms having the same equivalence class number are indistinguishable, *ignoring stereochemistry*; the equivalence class number is the smallest canonical identification number in the class of equivalent atoms). This information is only intended to assist in the understanding of results of InChI processing and is not directly used in InChI generation except in the processing of stereochemistry.

Stereochemical parities of bonds and atoms are also displayed. A question mark symbol indicates that stereochemistry is possible, but has not been specified. A ‘u’ symbol displayed in non-standard InChI versions to indicate that the stereochemistry has been explicitly entered as ‘unknown’, is replaced with the question mark in standard InChI. Bonds that have been found to be variable by alternation or movement of mobile H-atoms or charges are shown by dotted lines. This information is used only for deciding which bonds may exhibit double bond (Z/E) isomerism. By design, the Identifier does not explicitly represent bond types.

Lower Section

The InChI along with auxiliary data and explanatory information is shown in the lower section of the output window, such as seen in Figure 6 (see Section VI). Unlike the graphical display, even if more than one disconnected component is found, all textual results for a single input structure file are shown together. This reflects the important point that all components of a submitted structure are considered by InChI to be part of a single compound. Results for different (disconnected) components of a single substance are separated by semicolons, except for chemical formulas, which, in keeping with common conventions, are separated by dots.

Options

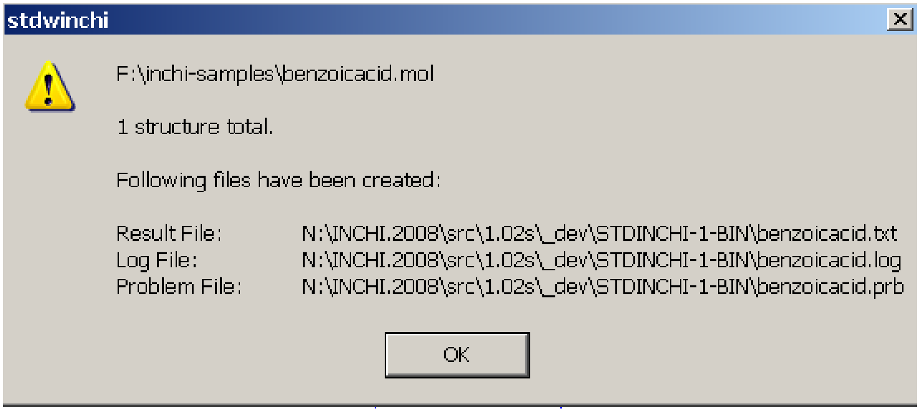
Pressing the Options Button opens InChI Options Dialog Box. The following options are then available (as seen in Figure 4):

- Include Stereo (Absolute) – Include stereo layer or exclude all stereo information from the identifier. Standard InChI default is Absolute stereochemistry. The only other option is to completely exclude stereo (which should be used if and only if the user is completely sure that stereo information should not be perceived).
- Narrow end of wedge points to stereocenter – for generation of standard InChI, this option should be turned on, i.e. the bond affects the stereochemistry of only the atom 'pointed to' by the wedge. However, it may be turned off if and only if the user is completely sure that a stereobond affects both atoms it connects (that is, for 2D structures complying to the legacy "perspective" stereochemistry drawing style).
- Annotation Format (Plain Text; XML, None) – Choose appropriate format for explanatory information.
- Ignore Isotopes in Structure Display – This does not change the identifier, it only affects the structure appearance and displaying sets of equivalent components.

Text File Output

At any time you may select 'Write Result' to analyze the input file and write all textual results to an output file located in the same directory as the program. The name of this file is derived from the name of input structure file and is displayed when it is created (the name has extension .txt). Figure 9 is an example of this for Benzoic Acid. It shows the directory/location on the computer as well as the file names given to the three (3) output files. Two other files to assist in diagnosing problems, should they occur, are created and their names displayed. One of them is a log file; it contains names of input and output files, a list of selected options, warning and error messages, number of processed structures, processing time, etc. The name of this file has extension .log. Another file – a problem file -- contains input structure file records that caused errors. This file (its name has extension .prb) may be important to determine reasons for the errors. A listing of errors and warnings is given in the Appendix 1.

Figures 10, 11, 12 show the content of the three output files. The .prb file is, of course, empty, since there were no problems encountered in generating the InChI for Benzoic Acid.



```
F:\inchi-samples\benzoicacid.mol

1 structure total.

Following files have been created:

Result File:     N:\INCHI.2008\src\1.02s\_dev\STDINCHI-1-BIN\benzoicacid.txt
Log File:        N:\INCHI.2008\src\1.02s\_dev\STDINCHI-1-BIN\benzoicacid.log
Problem File:    N:\INCHI.2008\src\1.02s\_dev\STDINCHI-1-BIN\benzoicacid.prb
```

Figure 9

```
* Input_File: "C:\inchi-samples\benzoicacid.mol"
Structure: 1
InChI=1S/C7H6O2/c8-7(9)6-4-2-1-3-5-6/h1-5H,(H,8,9)
AuxInfo=1/1/N:6,3,7,1,4,2,5,8,9/E:(2,3)(4,5)(8,9)/rA:9nCCCCCCCOO/rB:d1;s1;s2;s2;d3;d4s6;s5;d5;/rC:13.2076,-
5.2994,0;14.3539,-4.625,0;13.2076,-6.648,0;15.5003,-5.2994,0;14.3539,-3.2763,0;14.3539,-7.3224,0;15.5003,-
6.648,0;13.2076,-2.602,0;15.5003,-2.602,0;

InChIKey=WPYMKLBDIGXBTP-UHFFFAOYSA-N
```

Figure 10. File benzoicacid.txt

```
InChI version 1, Software version 1.02 release 01/10/2008
Opened log file 'C:\inchi-samples\benzoicacid.log'
Opened output file 'C:\inchi-samples\benzoicacid.txt'
Opened problem file 'C:\inchi-samples\benzoicacid.prb'
Generating standard InChI
Full Aux. info
Generating standard InChIKey
Input format: MOLfile
Output format: Plain text
Timeout per structure: 60.000 sec; Up to 1024 atoms per structure
End of file detected after structure #1.
Finished processing 1 structure: 0 errors, processing time 0:00:00.00
```

Figure 11. File benzoicacid.log

Figure 12. File benzoicacid.prb

Command Line Program

For those familiar with the Windows ‘Command Prompt’, an executable program is also provided – stdinchi-1.exe. This program uses ‘command line’ arguments that are shown by invoking the program without any arguments. The principal use of the program is to allow batch processing within other programs for the processing of multiple structure files. At present, this program is intended primarily for processing SDF files. This program may be recompiled and used under Linux without any changes. The Linux version does not display chemical structures.

Standard redirection may be used to suppress stdinchi-1 console output.
Under Windows:

```
stdinchi-1 /AuxNone input.sdf  output.txt logfile.log NUL  2>NUL
```

Under Linux:

```
stdinchi-1 -AuxNone input.sdf  output.txt logfile.log NUL  2>/dev/null
```

“>” or “1>” redirects standard output, “2>” redirects standard error output.

To process files greater than 2 GB with cInChI-1, the output of a problem file should be suppressed. To do that, the output and log file names should be included in the command line; the name of the problem file should be NUL, for example:

```
stdinchi-1 input.sdf  output.txt logfile.log NUL
```

stdwinchi-1.exe cannot process files greater than 2 GB.

Standard InChI Software Library

For advanced users who may want to create the Identifier in their own software the InChI Software Library is provided in a separate package. The package contains ‘C’ source code for stdinchi-1.exe, ‘C’ source code for standard InChI Library that may be compiled into a Dynamic Link Library (DLL) under Windows or Shared Object (SO) under Linux or Unix; also, there are ‘C’ and Python examples of simple applications that read input Molfile and use InChI Library to produce Identifiers. The InChI Library does not display structures and is not able to read chemical structural data from the input file. It uses specially formatted input binary data and produces out of them three strings: the Identifier, the Auxiliary information, and, if necessary, an error or warning message. The source code is accompanied with makefiles tested with gcc under Windows and Linux. The description of the standard InChI library interface is located in “Standard_InChI_API_Reference” document and in “inchi_api.h” header file included in the package.

Test Files

A number of Molfiles (*.mol), CML files (*.cml) and two SDfiles (*.sdf) are included with the program for illustrative purposes. Some Molfiles contain more than one fragment – each may be viewed separately using the ‘combo-box’ on the upper left of the screen. Multiple structures are given in the SDfiles, which may be viewed in order by pressing the ‘Next Structure’ (“>>”) and ‘Previous Structure’ (“<<”) buttons. File Samples.sdf contains all of the individual Molfiles from Samples.zip. These SDfiles contain names of the structures. To display them enter word “name” (without quotes) in “Structure ID Header” field (Fig. 2).

Standard InChI Software Options

According to definition of standard InChI, the number of options affecting the result is reduced to minimum. The available options are related to regulation of input, output, and perception of input structures rather than to InChI generation.

Options Availability			Command line option (without / or – prefix)	Explanation
stdwinchi-1	stdinchi-1	Library		
Input				
-	Yes	**)	STDIO	Use standard input/output streams
-	Yes	-	InpAux	Input structures in InChI default aux. info format (for use with STDIO)
Yes	Yes	Yes	SDF:*name*	Read from the input SDfile the ID under the named data header
*)	Yes	-	CML	Input in CML format
Output				
-	Yes	Yes	AuxNone	Do not produce Auxiliary Information
-	Yes	-	NoLabels	Omit structure number, DataHeader and ID from InChI output
-	Yes	-	Tabbed	Separate structure number, InChI, and AuxIndo with tabs
Always	Yes	-	D	Display the structure
Yes	Yes	-	Equ	Display sets of identical components
-	Yes	-	F*number*	Set display font size (points)

-	Yes	Yes	OutputSDF	Convert InChI created with default auxiliary info to a SDfile
-	Yes	Yes	SdfAtomsDT	Output Hydrogen Isotopes to SDfile as Atoms D and T
Structure perception				
Yes	Yes	Yes	newpsOFF	Both ends of wedge point to stereocenters
-	Yes	Yes	DoNotAddH	Do not add H according to usual valences
Yes	Yes	Yes	SNon	Ignore stereo information in input structures
Generation				
60 sec	Yes ***)	Yes***)	Wnumber	Set time-out per structure in seconds
-	Yes	Yes	WarnOnEmptyStructure	Warn and produce empty InChI for empty structure
Always	Yes	- ****)	Key	Generate InChIKey
Conversion				
-	Yes	-	InChI2Struct	Convert standard InChI string(s) into structure(s)

*) stdwinchi recognizes CML file file format if the file name has extension ".CML".
**) Yes for executable files, No for InChI Library.
***) W0 means unlimited time. In InChI Library the default is W0, in stdinchi the default is 60 seconds (W60).
****) In standard InChI Library, generation of standard InChIKey is performed via a separate function call.

IV. CHEMICAL STRUCTURE INPUT

Molfiles, CML or the program output produced with the "Full auxiliary information" option may be used for input. Molfile structures may be submitted either as a single Molfile or as a series of concatenated Molfiles (an SDfile). A number of programs, some of them freely available, may be used to create these Molfiles. Information on how to produce and convert CML files may be found at http://www.xml-cml.org. If an input structure contains more than one independent structure, each component is individually shown in the graphical output section of the program, though this has no effect on the InChI. Text results are given for all layers and all components (different components of a single substance are separated by semicolons in each layer, except for chemical formulas, which, by convention, are separated by dots.).

While structure normalization methods built into the program perceive a range of different structure drawing conventions, it is possible that other conventions may not be properly recognized. Examination of the graphical results of InChI processing, especially for equivalent atom classes and stereo labeling, should reveal such problems.

If a SDfile is ‘labeled’, the program can supply these labels in its output. If the tag name is ‘Name’ and the data field is ‘2-methylanthracene’, this information would appear in the in SDfile as 3 lines (the last line is blank):

```
>  <Name>
 2-methylanthracene

```

In this case, if the tag ‘Name’ is entered in the ‘Structure ID Header’ field in the input dialog box, ‘2-methylanthracene’ will appear in the output text.

A variety of structure files are provided for testing. Individual MOL files have extension .MOL, concatenated MOL files have extension .SDF, CML files have extension .CML.

V. STANDARD INCHI LAYERS

This program parses and annotates the InChI and associated auxiliary information and displays it in the textual output region. An understanding of this information requires an understanding of InChI ‘layering’, which is described in detail in the Technical document. A summary is presented here for understanding program output.

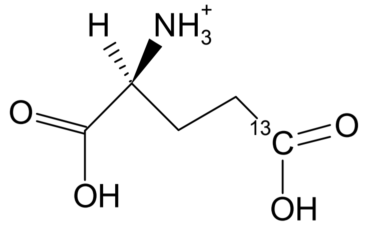


Figure 13. (*S*)-Glutamic acid

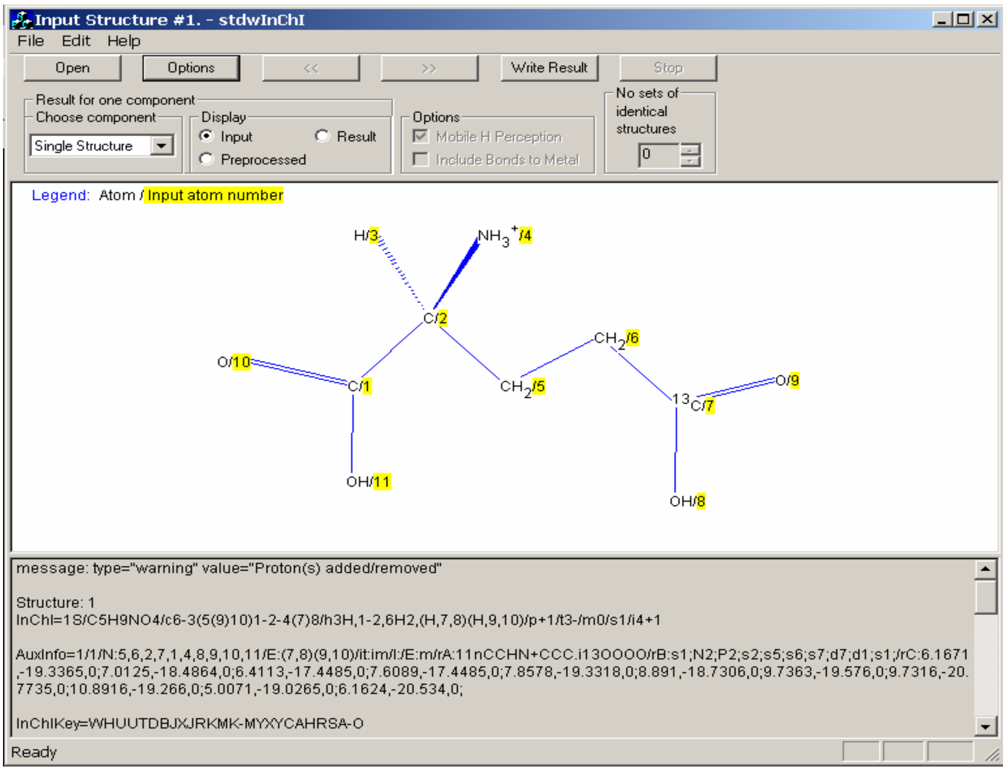
To provide an example of some of the standard InChI layers for a “real” molecule, we have chosen the structure of isotopically substituted (*S*)-Glutamic acid in Figure 13 above for illustrative purposes.

Figure 15 shows the input structure display. Figure 16 – “Preprocessed” – shows the result of the preprocessing – an attempt to eliminate charges with purpose to reduce different protonation forms to one. Figure 17 shows the result of the structure analysis by the InChI algorithm.

Figure 14 shows standard InChI, AuxInfo and standard InChIKey.

```
Structure: 1
InChI=1S/C5H9NO4/c6-3(5(9)10)1-2-4(7)8/h3H,1-2,6H2,(H,7,8)(H,9,10)/p+1/t3-/m0/s1/i4+1

AuxInfo=1/1/N:5,6,2,7,1,4,8,9,10,11/E:(7,8)(9,10)/it:im/I:/E:m/rA:11nCCHN+CCC.i13OOOO/rB:s1;N2;P2;s2;s5;s6;s7;d7;d1;s1
19.3365,0;7.0125,-18.4864,0;6.4113,-17.4485,0;7.6089,-17.4485,0;7.8578,-19.3318,0;8.891,-18.7306,0;9.7363,-19.576,0;9.7316,
20.7735,0;10.8916,-19.266,0;5.0071,-19.0265,0;6.1624,-20.534,0;

InChIKey=WHUUTDBJXJRKMK-MYXYCAHRSA-O
```

Figure 14. The standard InChI, AuxInfo and standard InChIKey for (*S*)-Glutamic acid.

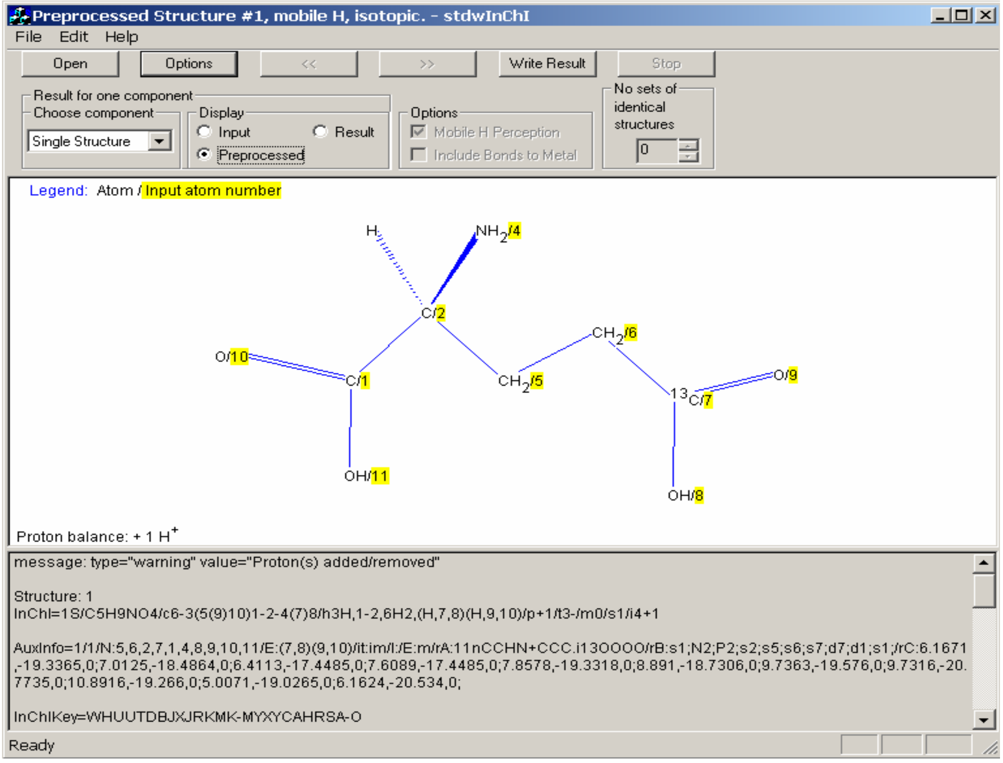


Result for one component
Choose component: Single Structure
Display: ● Input ○ Result ○ Preprocessed
Options: ☑ Mobile H Perception □ Include Bonds to Metal
No sets of identical structures: 0

```
message: type="warning" value="Proton(s) added/removed"

Structure: 1
InChI=1S/C5H9NO4/c6-3(5(9)10)1-2-4(7)8/h3H,1-2,6H2,(H,7,8)(H,9,10)/p+1/t3-/m0/s1/i4+1

AuxInfo=1/1/N:5,6,2,7,1,4,8,9,10,11/E:(7,8)(9,10)/it:im/I:/E:m/rA:11nCCHN+CCC.i13OOOO/rB:s1;N2;P2;s2;s5;s6;s7;d7;d1;s1;/rC:6.1671
,-19.3365,0;7.0125,-18.4864,0;6.4113,-17.4485,0;7.6089,-17.4485,0;7.8578,-19.3318,0;8.891,-18.7306,0;9.7363,-19.576,0;9.7316,-20.
7735,0;10.8916,-19.266,0;5.0071,-19.0265,0;6.1624,-20.534,0;

InChIKey=WHUUTDBJXJRKMK-MYXYCAHRSA-O
```

Ready

Figure 15.



Result for one component

Choose component: Single Structure

Display: ○ Input ○ Result ● Preprocessed

Options: ☑ Mobile H Perception □ Include Bonds to Metal

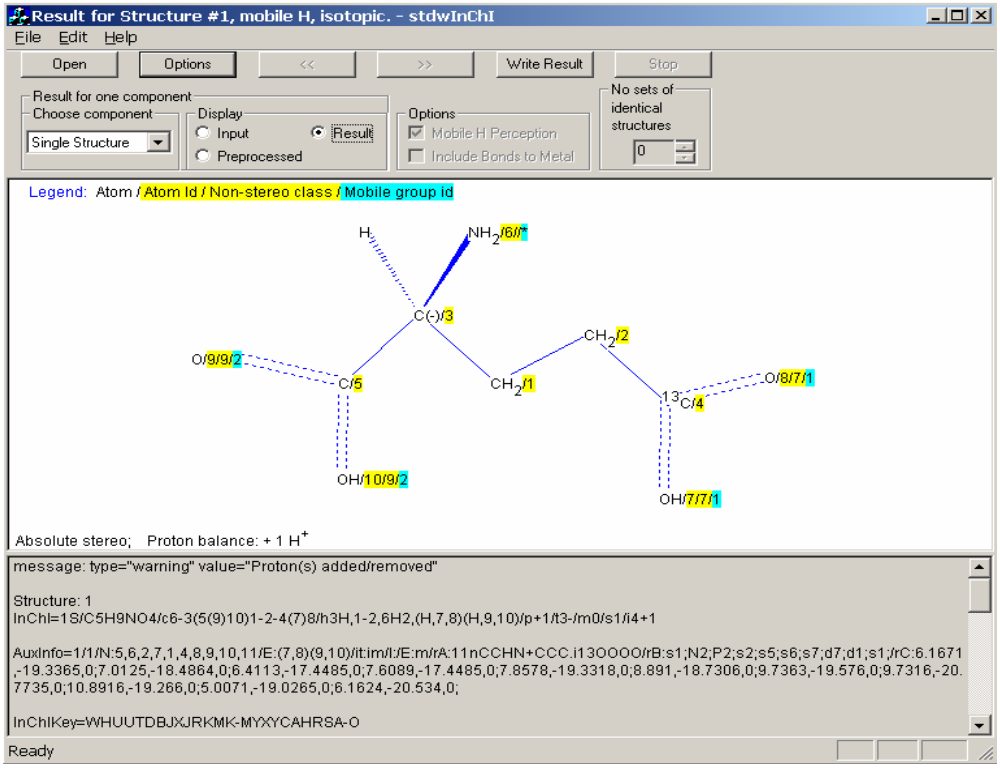
No sets of identical structures: 0

Proton balance: + 1 H^+

```
message: type="warning" value="Proton(s) added/removed"

Structure: 1
InChI=1S/C5H9NO4/c6-3(5(9)10)1-2-4(7)8/h3H,1-2,6H2,(H,7,8)(H,9,10)/p+1/t3-/m0/s1/i4+1

AuxInfo=1/1/N:5,6,2,7,1,4,8,9,10,11/E:(7,8)(9,10)/it:im/I:/E:m/rA:11nCCHN+CCC.i13OOOO/rB:s1;N2;P2;s2;s5;s6;s7;d7;d1;s1;/rC:6.1671
,-19.3365,0;7.0125,-18.4864,0;6.4113,-17.4485,0;7.6089,-17.4485,0;7.8578,-19.3318,0;8.891,-18.7306,0;9.7363,-19.576,0;9.7316,-20.
7735,0;10.8916,-19.266,0;5.0071,-19.0265,0;6.1624,-20.534,0;

InChIKey=WHUUTDBJXJRKMK-MYXYCAHRSA-O
```

Ready

Figure 16.



Result for one component

Choose component: Single Structure

Display: ○ Input ● Result ○ Preprocessed

Options: ☑ Mobile H Perception □ Include Bonds to Metal

No sets of identical structures: 0

Absolute stereo; Proton balance: + 1 H^+

```
message: type="warning" value="Proton(s) added/removed"

Structure: 1
InChI=1S/C5H9NO4/c6-3(5(9)10)1-2-4(7)8/h3H,1-2,6H2,(H,7,8)(H,9,10)/p+1/t3-/m0/s1/i4+1

AuxInfo=1/1/N:5,6,2,7,1,4,8,9,10,11/E:(7,8)(9,10)/it:im/I:/E:m/rA:11nCCHN+CCC.i13OOOO/rB:s1;N2;P2;s2;s5;s6;s7;d7;d1;s1;/rC:6.1671
,-19.3365,0;7.0125,-18.4864,0;6.4113,-17.4485,0;7.6089,-17.4485,0;7.8578,-19.3318,0;8.891,-18.7306,0;9.7363,-19.576,0;9.7316,-20.
7735,0;10.8916,-19.266,0;5.0071,-19.0265,0;6.1624,-20.534,0;

InChIKey=WHUUTDBJXJRKMK-MYXYCAHRSA-O
```

Ready

Figure 17.

Figure 18 shows the contents of the text output window.

Notes:

- Since the displayed InChI is the standard InChI, it is marked with "InChI=1S/" prefix.
- The Auxiliary information is not a part of the Identifier.
- The standard InChIKey contains protonation flag not equal to 'N' (the last character of the InChIKey is 'O') which indicates that the "p" segment of the standard InChI string is not empty.

The "InChI ANNOTATED CONTENTS" provides annotations to each item of the Identifier and Auxiliary information.

The standard InChI represents the structure of a covalently bonded compound in four distinct 'layers':

1. *Main Layer*

   Chemical Formula
   This is a conventional Hill-sorted formula with components separated by periods (dots).

   In the example in Figure 18, the formula is:
   /{formula}C5H9NO4

   Connections
   Defines the covalent bonds between atoms in the structure. It is partitioned into as many as three sublayers: H-atoms omitted, immobile H-atoms included and, mobile H-atoms included.

   In the example in Figure 18, the connections are:

   /c{connections}6-3(5(9)10)1-2-4(7)8
   /h{H_atoms}1-2H2,3H,6H2,(H,7,8)(H,9,10)

   where part (H,7,8)(H,9,10) is responsible for mobile H

2. *Charge Layer*
   This simply represents net charge, and may appear in two sublayers. Unlike other layers, this layer is independent of all others and when omitted indicates that the charge is not specified.

   Component charge
   The net charges of the components are represented in this layer as independent tags. By design, the InChI does not distinguish between

structures that differ only by the formal positions of their electrons.

```
message: type="warning" value="Proton(s) added/removed"

Structure: 1
InChI=1S/C5H9NO4/c6-3(5(9)10)1-2-4(7)8/h3H,1-2,6H2,(H,7,8)(H,9,10)/p+1/t3-/m0/s1/i4+1

AuxInfo=1/1/N:5,6,2,7,1,4,8,9,10,11/E:(7,8)(9,10)/it:im/I:/E:m/rA:11nCCHN+CCC.i13OOOO/rB:s1;N2;P2;s2;s5;s6;s7;d7;d1;s1;/rC:6.1671,-
19.3365,0;7.0125,-18.4864,0;6.4113,-17.4485,0;7.6089,-17.4485,0;7.8578,-19.3318,0;8.891,-18.7306,0;9.7363,-19.576,0;9.7316,-
20.7735,0;10.8916,-19.266,0;5.0071,-19.0265,0;6.1624,-20.534,0;

InChIKey=WHUUTDBJXJRKMK-MYXYCAHRSA-O

==== InChI ANNOTATED CONTENTS ====

Structure: 1

InChI=
{version}1S
/{formula}C5H9NO4
/c{connections}6-3(5(9)10)1-2-4(7)8
/h{H_atoms}3H,1-2,6H2,(H,7,8)(H,9,10)
/p{protons}+1
/t{stereo:sp3}3-
/m{stereo:sp3:inverted}0
/s{stereo:type (1=abs, 2=rel, 3=rac)}1
/i{isotopic:atoms}4+1

AuxInfo=
{version}1
/{normalization_type}1
/N:{original_atom_numbers}5,6,2,7,1,4,8,9,10,11
/E:{atom_equivalence}(7,8)(9,10)
/it:{abs_stereo_inverted:sp3}im
/I:{isotopic:original_atom_numbers}
/E:{isotopic:atom_equivalence}m
/rA:{reversibility:atoms}11nCCHN+CCC.i13OOOO
/rB:{reversibility:bonds}s1;N2;P2;s2;s5;s6;s7;d7;d1;s1;
/rC:{reversibility:xyz}6.1671,-19.3365,0;7.0125,-18.4864,0;6.4113,-17.4485,0;7.6089,-17.4485,0;7.8578,-19.3318,0;8.891,-
18.7306,0;9.7363,-19.576,0;9.7316,-20.7735,0;10.8916,-19.266,0;5.0071,-19.0265,0;6.1624,-20.534,0;
```

Figure 18. The standard Identifier, Auxiliary information, standard InChIKey and Annotated Identifier and Auxiliary information for (*S*)-Glutamic acid with "Mobile H Perception" turned off.

Protons

Number of protons removed from or added (if the number is negative) to the substance to make same components with variable protonation (e.g. amino acids) identical.

In the example in Figure 18 the proton(s) are:
/p{protons}+1

Note that standard InChIKey indicates this as protonation: the last character is 'O' which corresponds to "p+1".

3. *Stereochemical Layer*

   This layer is composed out of two sublayers. The first accounts for double bond, sp^2, and the second for sp^3 tetrahedral stereochemistry and allenes. The latter stereo descriptions are first given for relative stereochemistry only, followed by a designation of whether the absolute stereochemistry is required (and if this requires inversion of the relative stereochemistry).

   In the example in Figure 18 the stereo layer is:

   /t{stereo:sp3}3-
   /m{stereo:sp3:inverted}0
   /s{stereo:type (1=abs, 2=rel, 3=rac)}1

4. *Isotopic Layer*

   This is a layer in which different isotopically labeled atoms are distinguished from each other. Mobile isotopic hydrogen atoms are listed separately. The layer also holds any changes in stereochemistry created with the presence of isotopic atoms.

   In the example in Figure 18 the isotopic layer is:
   /i{isotopic:atoms}4+1

Note that these names of layers are used in the annotated InChI output. In the identifier itself the layers are preceded by two characters, a '/' followed by a letter.

For any input structure, the first layer will always be generated. Other layers will appear only when the input structure contains the associated information. For instance, if Z/E stereochemistry, but not sp^3 stereochemistry is entered, only the Z/E (sp^2) stereochemistry sublayer will be represented.

The contents of a layer may depend on prior layers. For instance, the stereochemical layer uses identification numbers of atoms defined in the formula layer.

The Charge layer is simply the overall charge of the component, hence is independent of the other layers. It is possible to extend this layer by adding other 'whole molecule' attributes, such as electronically excited state, vibrational/rotational state and state of aggregation (phase).

The Protons layer refers to the entire structure. The specific state of protonation (or deprotonation) may be ignored by omitting this layer.

VI. OUTPUT TEXT FORMAT

The text output from the standard InChI program is written in plain text format as described below. This text is visible in the lower region of the main window and in the text file generated by selecting ‘Write Result’ in the main window.

Note that a standard InChI for a substance is strictly defined as a string of characters composed of a series of text fields. The specific text format described here is meant only for those interested in the details of the representation and is not required for effective use of the standard InChI.

The actual fields present in a given representation will depend on the information present in the input structure and the intent of the structure author. If, for instance, it is desired to represent a structure with mobile H-atoms, a fixed H-atom layer is not generated. If a structure cannot have stereoisomers, no stereo layers will be present.

All text output originating from a single chemical substance input (structure file) is provided in up to four lines; all lines except the second one may be suppressed:

Structure NUMBER. STRUCTURE_ID_HEADER =VALUE
InChI=1S/…
AuxInfo=1/…
InChIKey=…

where NUMBER is the sequence number of the structure in the input file. When Molfiles or SDfiles are used and a “STRUCTURE_ID_HEADER” has been entered in the “Structure ID Header” field in the input dialog box (see Chemical Structure Input section above), VALUE represents the contents of that field. If the field was left blank then STRUCTURE_ID_HEADER=VALUE is omitted. The output AuxInfo line is optional. The Tabbed option produces output of the same items merged in one line with tab characters as separators.

InChI Output

Following the /? InChI delimited tags are individual layer values. Curly braces contain annotations. (the values for each layers follow the closing curly brace)

Main Layer (immediately follows the InChI version)

```
/{formula}
/c{connections}
/h{H_atoms}
```

Charge layer:

```
/q{charge}
/p{protons}
```

Stereo layer:

```
/b{stereo:dbond}
/t{stereo:sp3}
/m{stereo:sp3:inverted}
```

```
/s{stereo:type (1=abs, 2=rel, 3=rac)}
```

Isotopic Layer

```
/i{isotopic:atoms}*
/h{isotopic:exchangeable_H}
/b{isotopic:stereo:dbond}
/t{isotopic:stereo:sp3}
/m{isotopic:stereo:sp3:inverted}
/s{isotopic:stereo:type (1=abs, 2=rel, 3=rac)}
```

The Main Layer is divided into three layers:
Formula layer, Connections layer, and H atoms layer.

The stereo layer is divided into four layers:
Stereo double bond, sp^3 stereochemistry, sp^3 inversion flags, and type of sp^3 layer.

A description of the contents of each of these layers follows:

Main Layer.
This provides the elemental composition and connectivity of the structure. This layer, which is always present, is subdivided into several segments. The first segment is a conventional chemical formula, which also provides the InChI identification numbers used for each atom. These numbers are determined by the sequence numbers of elements in the chemical formula (excluding H). In the formula each element is represented by the form 'El#', where 'El' is the element symbol and '#' is the number of atoms. For example, in case of C2H6O two atoms C have identification numbers 1 and 2 and atom O (3rd non-H atom) has number 3. Atoms H are not given any identification numbers, except for bridging atoms H which, when present, are given the highest identification numbers. When a given component is present multiple times, this formula may be preceded by this number of occurrences. The case of H+ is special – it is represented simply as '1' in the Protons layer (the formula and connections segments are empty).

As noted above, the position of each element in the first (formula) segment is used as its identification number. These numbers are used in the second segment of the InChI, connections (/c), to indicate bonding partners. To illustrate, in this segment isobutane (C4H10) is represented as "1-4(2)3", which means that the 1st atom listed is bonded to the 4th, the 4th is bonded to the 2nd and the 3rd atoms. If the connections segment is empty (for example, in case of methane) it is omitted entirely.

The 3rd segment, the hydrogen layer (/h), describes positions of hydrogen atoms attached to the molecular skeleton described by formula and connections layers. For isobutene, "1-3H3,4H" means that each of atoms 1 to 3 has 3 H, and atom 4 has one H. A mobile H may be migrate between different atoms. For example, acetic acid (C2H4O2) has connections "1-2(3)4" and a hydrogen layer "1H3,(H,3,4)". Parentheses contain the number of mobile H (one in this case) and the identification numbers of the atoms that share these mobile H atoms (3 and 4).

Isotopic layer

Isotopic layers consist of a series of isotopic atoms with their identification numbers. Specific isotopes are represented by integers giving their atomic mass relative to the rounded average atomic mass of the element. For example, if atom number 6 is ^{37}Cl, it is represented as 6+2 (average atomic mass of Cl is 35.453, rounding to the nearest integer gives 35, 37 – 35 = +2).

Hydrogen isotopes are exceptions to this labeling rule. They are explicitly denoted as aDn (deuterium) and aTn (tritium), where a is the identification number of the atom to which they are attached and n is the number of these atoms attached to the ath atom; n=1 is omitted. Isotopic hydrogen atoms that are mobile or belong to atoms that are recognized as proton donors or acceptors are considered to belong to the whole substance and shown in the /h (exchangeable_H) segment of the isotopic layer.

Stereo layer

The stereo layer expresses the ‘parity’ of the atoms and bonds that define the stereochemistry. This layer is divided into four sub-layers, the first, dbond (/b), provides double bond (Z/E) stereochemistry, the second, sp3 (/t), represents sp3 tetrahedral stereochemistry and allenes, the third, sp3:inverted (/m), is present only in case of absolute configuration, the fourth (/s) describes whether the stereo representation is absolute, relative or racemic. Future versions of the InChI may add other forms of stereochemistry. Standard InChI supports only absolute stereochemistry or the absence of it in case of non-chiral chemical structures.

The stereo label of a double bond is represented in the format a-bX, where a and b are the identification numbers of the bonded atoms (a>b) and X is a parity label, with the possible values: +, -, or ?. The + and - labels indicate that the stereochemical configuration has been defined, however these values only have meaning relative to the atom identification numbers assigned in the labeling process. The numbers do not coincide with CIP priorities. If, for example, atoms in a similar structure were given different labels, the parity might change even if a chemist might consider the stereochemistry to be the equivalent. Additional rule-based processing would be needed to label a bond as ‘Z’ or ‘E’, for example. A question mark (‘?’) indicates that stereochemistry has not been specified. A ‘u’ symbol that in non-standard InChI indicates that the stereochemistry has been explicitly entered as ‘unknown’ is replaced with ‘?’.

Labels for sp^3 stereochemistry are expressed in the format nX where n is the identification number of the atom and X is the parity, as computed by InChI. The parity is allowed the same values as discussed for double bond stereochemistry. Also, as for double bond stereochemistry, parity values themselves depend on the particular labeling of the structure and are not readily converted to standard CIP notation. Currently the user may request absolute, relative, and racemic sp3 stereo. In case of absolute stereo the algorithm processes both the input structure and inverted structure; after that "the smallest" sp3 layer is chosen. Therefore enantiomers have an identical sp3 section. The fact of choosing the inverted configuration is shown as 1 in sp3:inverted (/m) segment, otherwise there is 0 or period if inversion does not bring a change. The type of stereo is shown in /s segment as /s1 (absolute). Used in non-standard InChI /s2 (relative), or /s3 (racemic) are not generated in standard InChI.

Auxiliary Information Output

A variety of additional information is optionally provided along with the Identifier. A mapping of canonical identification atom numbers on original atom numbers, constitutional equivalence, inverted sp3 stereo and its numbering, isotopic and fixed-H layer information, and 'reversibility' information which allows the redrawing of the original structure and recalculation of the identifier. This additional analysis information is shown in the line that starts with AuxInfo=

```
AuxInfo=
{version}1
/{normalization_type}
```

Main part

```
/N:{original_atom_numbers}
/E:{atom_equivalence}
/gE:{group_equivalence}
/it:{abs_stereo_inverted:sp3}
/iN:{abs_stereo_inverted:original_atom_numbers}
```

Isotopic part

```
/I:{isotopic:original_atom_numbers}*
/E:{isotopic:atom_equivalence}
/gE{isotopic:group_equivalence}
/it:{isotopic:abs_stereo_inverted:sp3}
/iN:{isotopic:abs_stereo_inverted:original_atom_numbers}
```

Reversibility part

```
/CRV:{charge_radical_valence}
/rA:{reversibility:atoms}
/rB:{reversibility:bonds>
/rC:{reversibility:xyz}
```

The original number of an atom with identification number of n is given as the n^{th} member of this list for a component; the lists are separated with ";".

Classes of equivalent atoms or groups are given as lists of identification numbers within parentheses.

Inverted absolute sp3 stereo provides the stereo layer of the inverted (reflected in a mirror) substance.

Unusual valences, atomic charges, and radical locations in the input data are shown in the charge-radical-valence (/CRV:) section. Together with the identifier this information allows to reconstruct a representative of a set of structures each of which produce same identifier. The examples are: 22+1, 22.3, 22+1.3, 22d, 22d3, where 22 is atom identification number, +1 is charge, 3 is valence, d is radical-doublet (t=triplet, s=singlet).

Upon requested “Full auxiliary information” (always ON in the stdwInChI program) the reversibility section is added; it includes all input information that allows the display of the input structure and the calculation of the identifier. This reversibility information is not used in InChI2Struct conversion.

Error/Warning Output

If problems are encountered during the processing of a structure, they are shown in the first line of stdwinchi-1 text window or in stdinchi-1 log file.
In the structure display, stereogenic atoms that caused warnings "Ambiguous stereo: center(s)" are displayed in red as well as atoms that caused warnings "Ambiguous stereo: bond(s)" concerning stereogenic bonds. Parities of these stereogenic elements are also displayed in red.

VII. PRINTING

The upper or lower sections of the output display may be printed by pressing the RIGHT mouse button with the cursor over the section and then selecting the print option. Text in the lower section may be copied using standard Windows controls.

VIII. OTHER OUTPUT FILES

In addition to the standard InChI output file discussed above (extension .txt), selection of the ‘Write Results’ option generates two other files that use the same base name as the input structure file. One is a .log file that records the progress of the program. The other is a .prb that records processing problems. We would appreciate being sent a copy of these files if problems are encountered with program operation.

IX. SOURCE CODE

The basic standard InChI generation code is written in the ‘C’ language and the user interface code of stdwinchi-1 is written in C++ using Microsoft Foundation Classes. All ‘C’ language source code, including Microsoft Visual C++ project files and gcc makefiles, is available at http://www.iupac.org/inchi

X. FEEDBACK

A key objective of the previous test versions was to facilitate a discussion of the implementation and scope of the standard InChI. We encourage questions and comments of all kinds.

XII. CONTACT INFORMATION

Mass Spectrometry Data Center
Building 221, Room A111
100 Bureau Drive
National Institute of Standards and Technology
Gaithersburg, Maryland 20899-8320

301-975-2670 (FAX)

steve.stein@nist.gov
301-975-2505

dmitrii.tchekhovskoi@nist.gov
301-975-4673

stephen.heller@nist.gov
301-975-3338

Appendix 1.Standard InChIKey details

The InChIKey is a character signature based on a hash code of the InChI string. A hash code is a fixed length condensed digital representation of a variable length character string.

The standard InChIKey is computed from (and only from) a standard InChI.

The overall length of standard InChIKey is fixed at 27 characters, including separators (dashes):

AAAAAAAAAAAAAA-BBBBBBBBFV-P

Here

AAAAAAAAAAAAAA is a 14-character hash which encodes molecular skeleton (connectivity). That is, it hashes part of standard InChI string

BBBBBBBB is an 8-character hash which encodes stereochemistry and isotopic substitution.

F is a flag indicating standard InChIKey (produced out of standard InChI): it always has the value ‘S’.

V is a flag for InChI version character: ‘A’ for version 1, ‘B’ for version 2, etc.

P is an indicator for the number of protons; this number is not encoded in the hash but is indicated as a separate 2-character block at the end, where one character is a hyphen, as –N for neutral, -M for -1 hydrogen, -O for +1 hydrogen, etc.

The exact layout is presented in the following table.

Char	Protons	Char	Protons
N	0		
M	-1	O	+1
L	-2	P	+2
K	-3	Q	+3
J	-4	R	+4
I	-5	S	+5
H	-6	T	+6
G	-7	U	+7
F	-8	V	+8
E	-9	W	+9
D	-10	X	+10
C	-11	Y	+11
B	-12	Z	+12
A	< -12 or > +12		

All symbols of InChIKey except the delimiter (a dash, that is, a minus) are uppercase English letters representing a "base-26" encoding.

The two hash blocks of InChIKey are based on truncated SHA-256 cryptographic hash function (http://en.wikipedia.org/wiki/SHA_hash_functions#SHA-2).

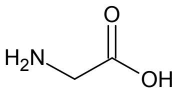
A theoretical – optimistic – estimate of collision resistance (i.e., the minimal size of a database at which a single collision is expected, that is, an event of two hashes of two different InChI strings being equal) is 6.1×10^{9} molecular skeletons × 3.7×10^{5} stereo/ isotopomers per each skeleton ≈ 2.2×10^{15}.
To exemplify: the probability of a single first block collision in a database of 1 billion compounds is 1.3%. In other words, a single first block collision is expected in 1 out of 100/1.3=75 databases of 10^{9} compounds each. For 10^{8} (100 million) compounds in a database this probability is 0.014%.

Alternative estimate of collision resistance is by a chance of an accidental collision upon adding a new entry to an existing collection. For a collection of 1 billion different InChIKey entries, the estimated probability of an accidental collision of the first layers for a newly added structure is 2.7×10^{-9} % and for both layers is $2.0*10^{-20}$ %.

Note, however, that by design of standard InChIKey different tautomers of the same compound (as far as their particular tautomerism is perceived by InChI) will have the same standard InChI and InChIKey strings.

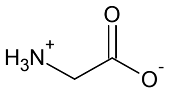
Note also that the different protonation states of the same compound will have standard InChIKey's which differ only by the last character, protonization flag (unless the both states have number of inserted/removed protons > 12; in this case the protonation flag will also be the same, 'A').

This is exemplified below by standard InChIKey's as well as standard InChI strings for neutral, zwitterionic, anionic and cationic states of the glycine (note that neutral and zwitterionic states do not differ by total number of protons so they have the same standard InChI/InChIKey):



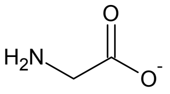
InChI=1S/C2H5NO2/c3-1-2(4)5/h1,3H2,(H,4,5)

InChIKey=DHMQDGOQFOQNFH-UHFFFAOYSA-N



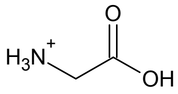
InChI=1S/C2H5NO2/c3-1-2(4)5/h1,3H2,(H,4,5)

InChIKey=DHMQDGOQFOQNFH-UHFFFAOYSA-N



InChI=1S/C2H5NO2/c3-1-2(4)5/h1,3H2,(H,4,5)/p-1

InChIKey=DHMQDGOQFOQNFH-UHFFFAOYSA-M



InChI=1S/C2H5NO2/c3-1-2(4)5/h1,3H2,(H,4,5)/p+1

InChIKey=DHMQDGOQFOQNFH-UHFFFAOYSA-O

Appendix 2. Standard InChI v. 1 Warning and Error Messages

Two varieties of problems detected during processing are reported. Warnings provide processing information that show any ambiguities in the input structure or special actions taken during processing. An InChI will be produced. When an error is generated, a valid InChI cannot be produced due to invalid input. It is expected that additional errors and warnings will reported in the final version.

Notes:

1. Messages ending with ":…" are followed by additional information
2. Symbol # represents an integer

Types of Warnings/Errors

- Input structure warnings
- Input structure errors
- InChI calculation errors
- Reading MOLfile warning messages
- Reading MOLfile error messages
- Reading pre-existing InChI output errors
- Internal errors (possible software error)

List of InChI warning and error messages

Input structure warnings

"Proton(s) added/removed"
"Charges neutralized"
"Omitted undefined stereo"
"Ambiguous stereo: [center(s)][bond(s)]"
"Unusual valence(s):…"
"Charges were rearranged"
"Salt was disconnected"
"Metal was disconnected"
"Not chiral"

Input structure errors

"Unknown element(s):…"
"Bond to nonexistent atom"
"Multiple bonds between two atoms"
"Atom has more than 3 aromatic bonds"
"Too many atoms"
"Empty structure"
"Atom 'X' has more than 20 bonds" (X is the chemical element symbol)

InChI calculation errors

"Output buffer overflow"
"Cannot process free radical center"
"Time limit exceeded"
"User requested termination"

Reading MOLfile warnings

"Too long counts line"
"Too long atom block line"
"Too long properties block line"
"Charge not recognized:…"
"Radical not recognized:…"
"Isotopic data not recognized:"
"Too long SData line truncated" (SData line was truncated to 200 characters)

Reading MOLfile errors

"Unrecognized bond type:#"
"Unrecognized bond stereo"
"Program error interpreting MOLfile"
"Unknown error"
"Cannot read counts line"
"Cannot interpret counts line:…"
"Cannot read atom block line"
"Cannot interpret atom block line:…"
"Cannot read bond block line"
"Cannot interpret bond block line:…"
"Cannot read STEXT block line"
"Cannot read properties block line"
"Unexpected SData header line
"Bypassing to next structure"

Reading pre-existing InChI output errors

"Missing atom data"
"Wrong atoms data"
"Wrong number of atoms"
"Wrong bonds data"
"Wrong bond type"
"Wrong number of bonds"
"Missing atom coordinates data"
"Wrong atom coordinates data"

"Wrong number of coordinates"
"Wrong version of auxiliary information"
"Cannot interpret reversibility information"
"Program error interpreting InChI aux"
"Unknown error"

Internal errors (possible software error)

"Out of RAM"
"Cannot disconnect metal error"
"Fatal undetermined program error"
"Cannot allocate output data. Terminating"
"Cannot distinguish components”
"Cannot extract Component"
"ARRAY OVERFLOW"
"LENGTH_MISMATCH"
"OUT_OF_RAM"
"RANKING_ERR"
"ISOCOUNT_ERR"
"TAUCOUNT_ERR"
"ISOTAUCOUNT_ERR"
"MAPCOUNT_ERR"
"ISO_H_ERR"
"STEREOCOUNT_ERR"
"ATOMCOUNT_ERR"
"STEREOBOND_ERR"
"REMOVE_STEREO_ERR"
"CALC_STEREO_ERR"
"STEREO_CANON_ERR"
"CANON_ERR"
"UNKNOWN_ERR(#)
"No description(#)"