User's Guide:
IUPAC International Chemical Identifier (InChI) Program

Version 1, Software version 1.01

Last revision date May 11, 2006

CONTENTS

I. OVERVIEW

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.

This version of 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.

II. ABOUT THE InChI PROGRAM

This document is accompanied by the version 1 of the InChI generator. This program runs under 32-bit Microsoft Windows Operating Systems. The main program, wInChI-1.exe, is a conventional Windows application, although a 'command line' version is also included (cInChI-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 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 InChI PROGRAM

The InChI generation program is provided along with sample chemical structures in a 'zip' file – InChI-1.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 wInChI-1.exe that was extracted from the zip file. Figure 1 then appears on your monitor.

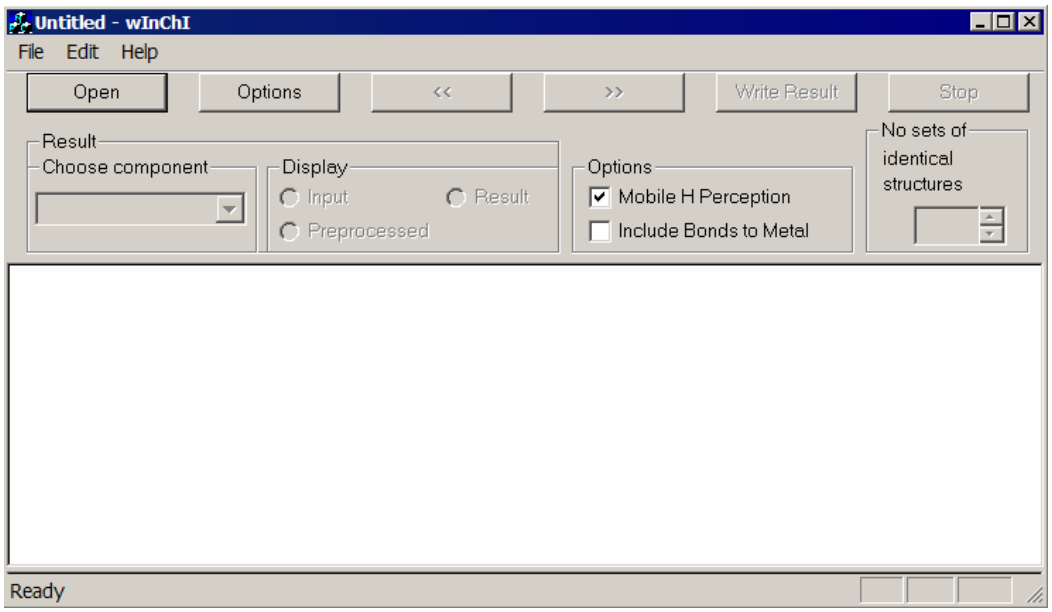


Figure 1

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),

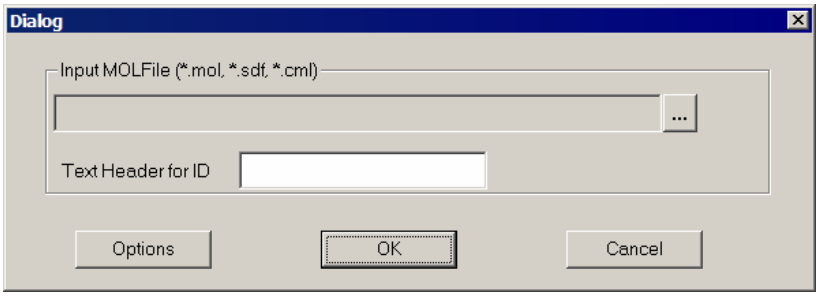


Figure 2

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.

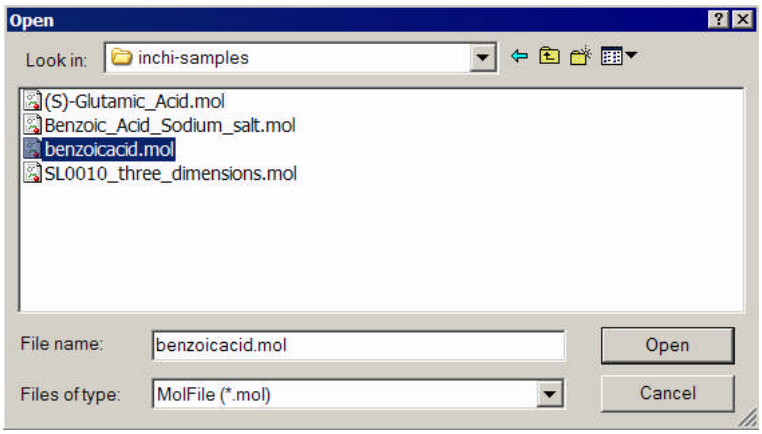


Figure 3

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.)

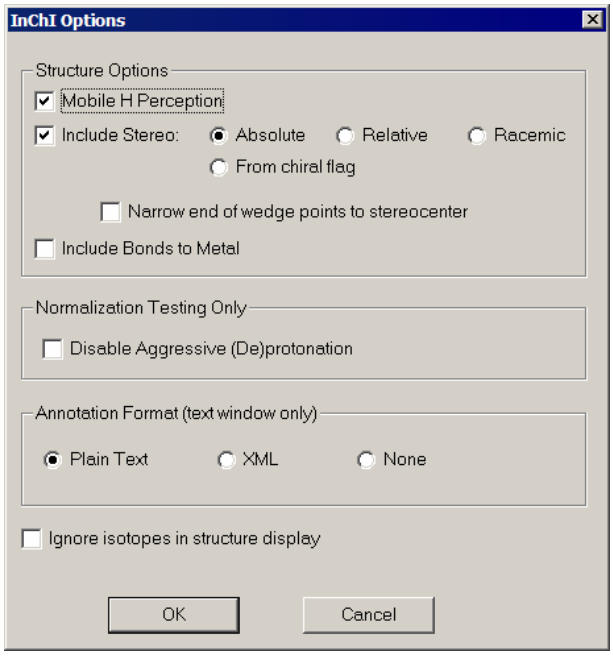


Figure 4

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

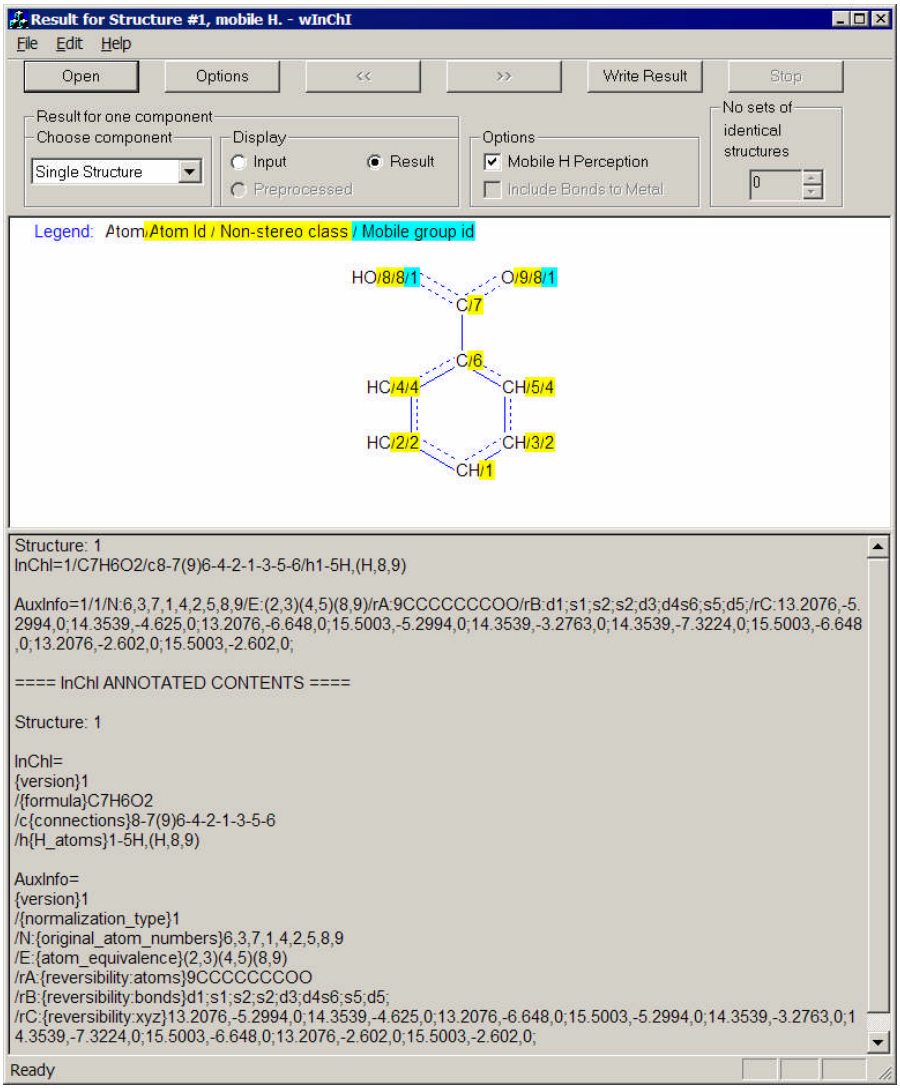


Figure 5

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.

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 the same as in the “Options” Dialog Box. “Mobile H Perception” removes “fixed-H” part of the identifier. Figure 9 shows the same structure with the option “Mobile H Perception” off.

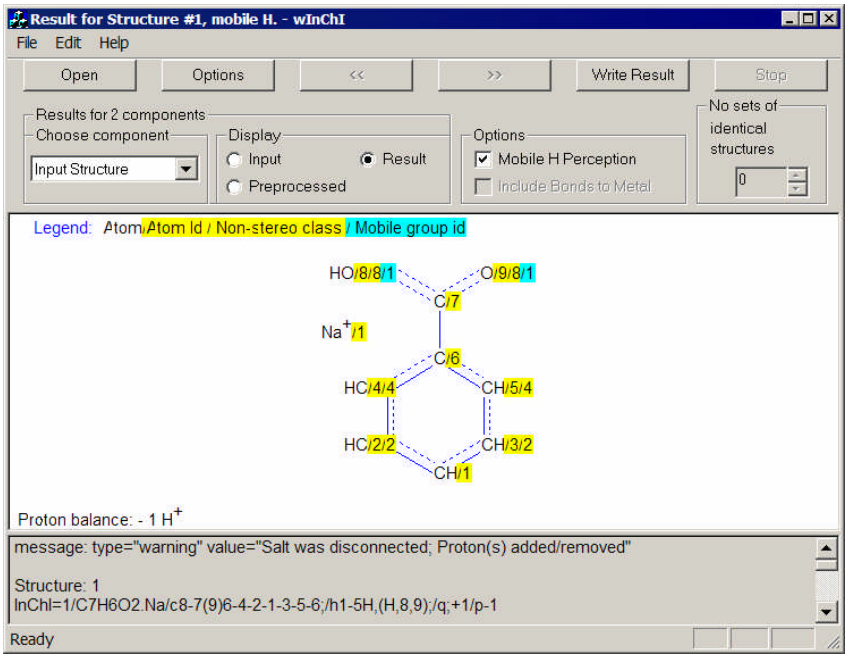


Figure 6

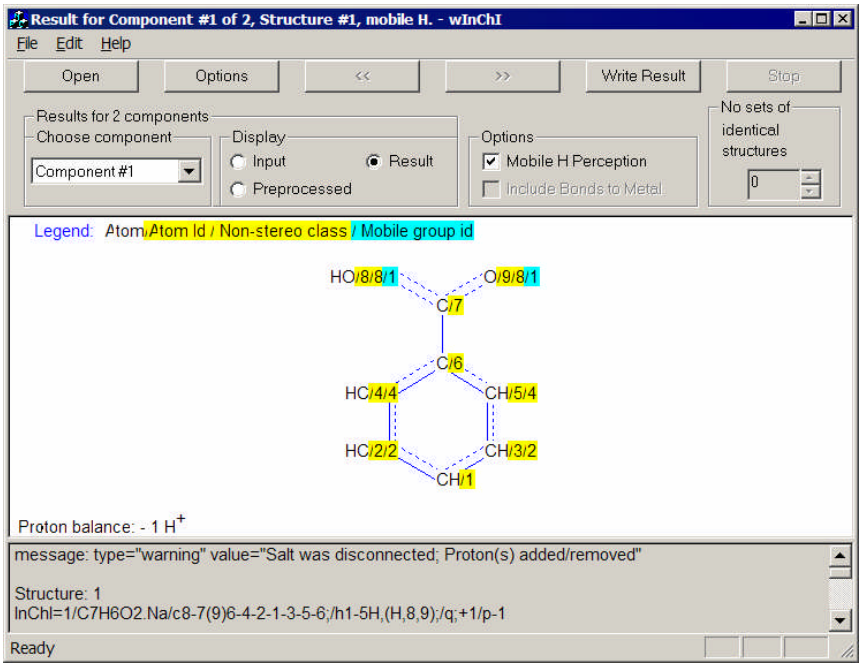


Figure 7

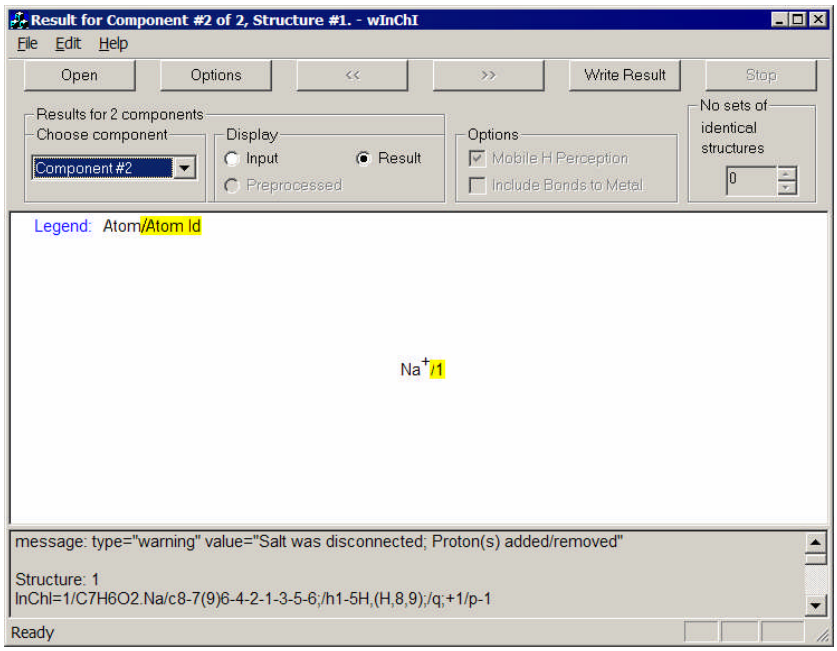


Figure 8

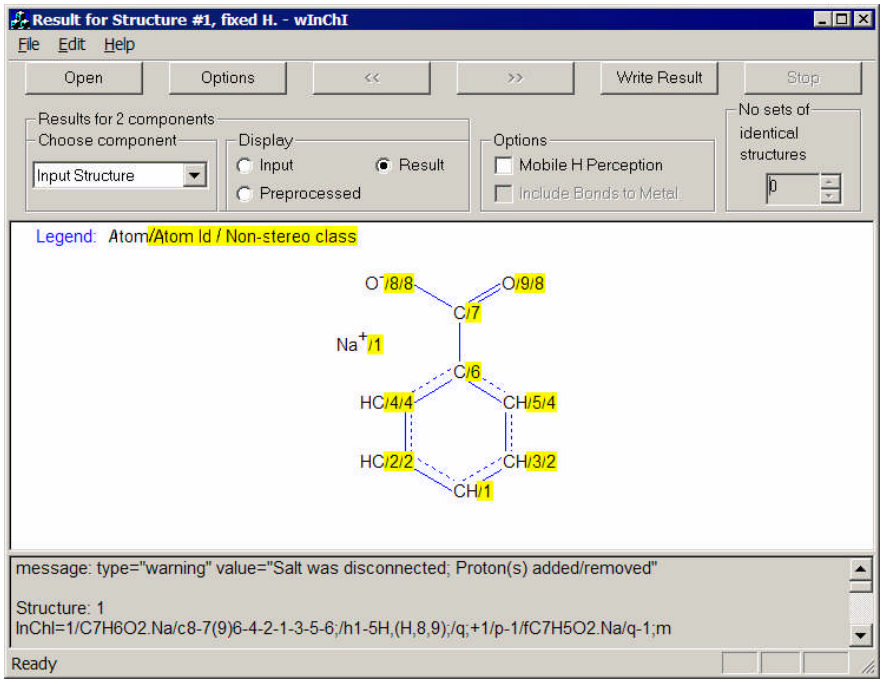


Figure 9.

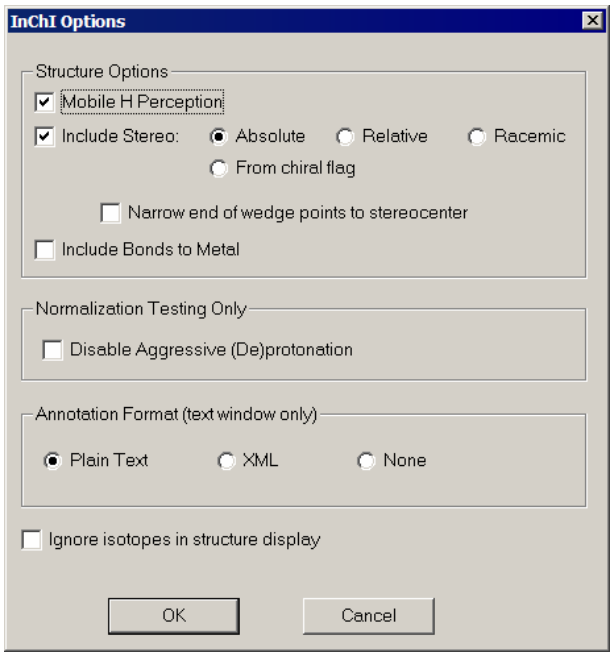
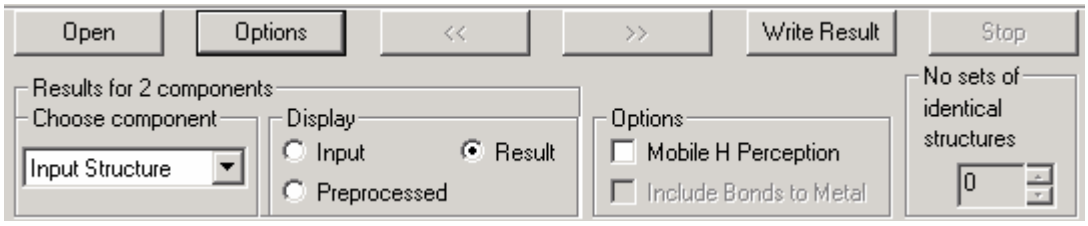


Figure 10

Figure 10 shows the options box with the various choices. "Include Bonds to Metal" includes the reconnected structure (but will not reconnect a metal atom if it is part of a salt.).



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 indicates that the stereochemistry has been explicitly entered as 'unknown'. 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, as can be seen in Figure 9. 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 10):

- Mobile H Perception – Turning Off will fix all H-atoms (disallow H-migration), this allows the generation of a fixed-H section of the Identifier;
- Include Stereo (Absolute, Relative, Racemic, From chiral flag) – Include stereo layer and choose its type or exclude all stereo information from the identifier. If the last option is selected then in presence of a chiral flag stereochemistry is considered absolute otherwise relative.
- Narrow end of wedge points to stereocenter – Turns on another 2D-specific convention for wedged and hatched stereo bond interpretation. Normally the InChI algorithm assumes that a stereobond affects both atoms it connects. With this option turned on, the bond affects the stereochemistry of only the atom 'pointed to' by the wedge.
- Include Bonds to Metal –Turning On will add a layer that includes specific bonding to metals (in case of salts bonds between a metal and an acid cannot be reconnected – as seen in Figures 6-9 where that choice is "grayed out" and can not be ticked or checked.)
- Disable 'Aggressive (De)protonation' – For building the protonation layer, a series of 'aggressive' rules, as described in the technical manual, is used. To find if this is the cause of changes made to the input structure, you may

disable this processing. Please let us know you feel this causes problems (different compounds being assigned to the same tautomer group).

- 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 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 11 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 12, 13, 14 show the content of the three (3) output files. The .prb file is, of course, empty, since there were no problems encountered in generating the InChI for Benzoic Acid.

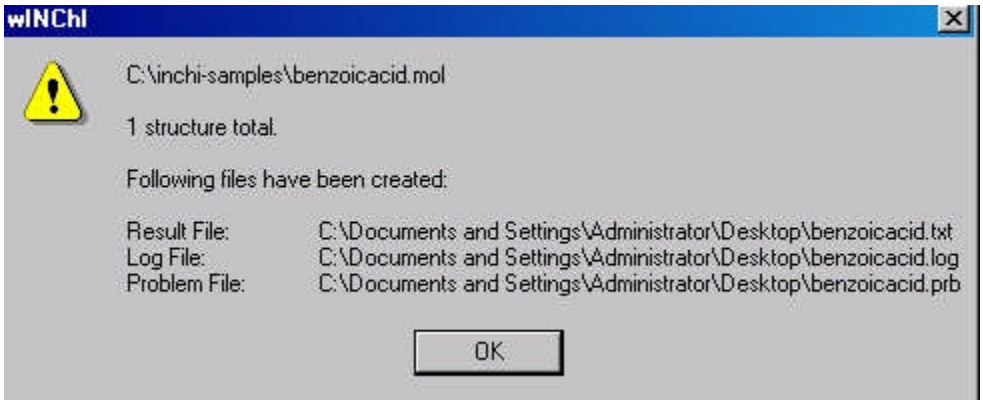


Figure 11

```
* Input_File: "C:\inchi-samples\benzoicacid.mol"
Structure: 1
InChI=1/C7H6O2/c8-7(9)6-4-2-1-3-5-6/h1-5H,(H,8,9)/f/h8H
AuxInfo=1/1/N:6,3,7,1,4,2,5,8,9/E:(2,3)(4,5)(8,9)/F:m/E:(2,3)(4,5)
/rA:9CCCCCCCOO/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;
```

Figure 12. File benzoicacid.txt

```
InChI version 1
Opened log file ‘C:\Docments and Settings\Administrator\Desktop\benzoicacid.log'.
Opened input file 'C:\inchi-samples\benzoicacid.mol'.
Opened output file ‘C:\Docments and Settings\Administrator\Desktop\benzoicacid.txt'.
Opened problem file ‘C:\Docments and Settings\Administrator\Desktop\benzoicacid.prb'.
Options: Mobile H Perception OFF
Isotopic ON, Absolute Stereo ON
Omit undefined/unknown stereogenic centers and bonds
Full Aux. info
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 13. File benzoicacid.log

Figure 14. File benzoicacid.prb

Command Line Program

For those familiar with the Windows ‘Command Prompt’, an executable program is also provided – cInChI-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.

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 cInChI-1.exe, ‘C’ source code for InChI Library that may be compiled into a Dynamic Link Library (DLL) under Windows or Shared Object (SO) under Linux or Unix, and ‘C’ and Microsoft Visual Basic 6.0 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 on gcc version 3.4.2 (MinGW) under Windows and under i386 Linux. The description of the InChI library interface is located in "inchi_api.h" 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; the other (SamplesTechMan.sdf) contains examples from the InChI Technical Manual. These SDfiles contain names of the structures. To display them enter word "name" (without quotes) in "Structure ID Header" field (Fig. 2).

InChI Creation Options

These tell InChI software which layers should be included in the Identifier, what kind of output is requested, and how to treat the input structure data. Not all options are available in all available InChI software.

Options Availability			Command line option (without / or – prefix)	Explanation
wInChI, Win-dows	cInChI, Win-dows	Lib. or other O.S.		
Yes	Yes	Yes	SAbs	Calculate absolute stereochemistry
Yes	Yes	Yes	SRel	Calculate relative stereochemistry
Yes	Yes	Yes	SRac	Calculate racemic stereochemistry
Yes	Yes	Yes	SUCF	Use chiral flag to calculate stereo-chemistry: On=SAbs, Off=SRel
Yes	Yes	Yes	SNon	Exclude stereochemical layer
-	Yes	Yes	SUU	Include omitted or undefined stereodescriptors
Yes	Yes	Yes	NEWPS	Narrow end of wedge points to stereocenter
Yes	Yes	Yes	RecMet	Include bonds to metal
-	Yes	Yes	DoNotAddH	Do not add H according to usual valences
Yes	Yes	Yes	FixedH	Turn off Mobile H perception
-	Yes	Yes	AuxNone	Do not produce Auxiliary Information
Yes	Yes	Yes	NoADP	Disable Aggressive (De)protonation
-	Yes	Yes	Compress	Output in compressed format
Always	Yes	-	D	Display the structure
Yes	Yes	-	Equ	Display sets of identical components
-	Yes	-	F*number*	Set display font size (points)
60 sec	Yes ***)	Yes***)	W*number*	Set time-out per structure in seconds

Yes	Yes	Yes	SDF:*name*	Read from the input SDfile the ID under the named data header
*)	Yes	-	CML	Input in CML format
-	Yes	-	NoLabels	Omit structure number, DataHeader and ID from InChI output
-	Yes	-	Tabbed	Separate structure number, InChI, and AuxIndo with tabs
-	Yes	Yes	OutputSDF	Convert InChI created with default auxiliary info to a SDfile
-	Yes	**)	STDIO	Use standard input/output streams
-	Yes	Yes	WarnOnEmptyStructure	Warn and produce empty InChI for empty structure

*) wInChI 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 cInChI the default is 60 seconds (W60). Time-out applies separately to mobile-H structure, fixed-H structure, and "Include bonds to metal" structure calculations.

Lib. or other O.S. column refers to

- cInChI compiled with "INCH_ANSI_ONLY" source code option, which usually means not MS Visual C++ or not under Microsoft Windows
- InChI Library software

Additional options included in InChI version 1 software version 1.01 are described in Appendix 2.

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-methyl anthracene’, this information would appear in the in SDfile as 3 lines (the last line is blank):

> <Name>
2-methyl anthracene

In this case, if the tag ‘Name’ is entered in the ‘Structure ID Header’ field in the input dialog box, ‘2-methyl anthracene’ 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. InChI LAYERS

This program the 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.

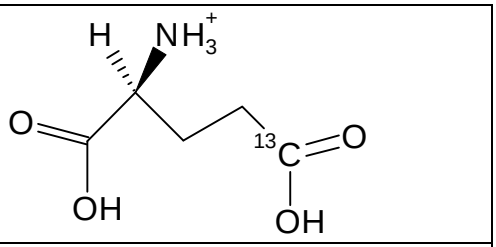


Figure 15. (S)-Glutamic acid

To provide an example of some of the InChI layers for a “real” molecule, we have chosen the structure of isotopically substituted (S)-Glutamic acid in Figure 15 above for illustrative purposes.

Figure 17 shows the input structure display. Figure 18 – “Preprocessed” – shows the result of the preprocessing – an attempt to eliminate charges with purpose to reduce different protonation forms to one. Figure 19 shows the result of the structure analysis by the InChI algorithm. Figure 20 shows the results with “Mobile H Perception” turned off. Notice that the contents of the text window has changed: a string that starts with “/f” has been appended to InChI. Fihure 16 shows the Identifier.

InChI=1/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
Figure 16. The Identifier for (S)-Glutamic acid with “Mobile H Perception” turned on.

Figure 21 shows full contents of the text output window in case of “Mobile H Perception” turned off. The “InChI ANNOTATED CONTENTS” provides annotations to each item of the Identifier and Auxiliary information. Note that the Auxiliary information is not a part of the Identifier.

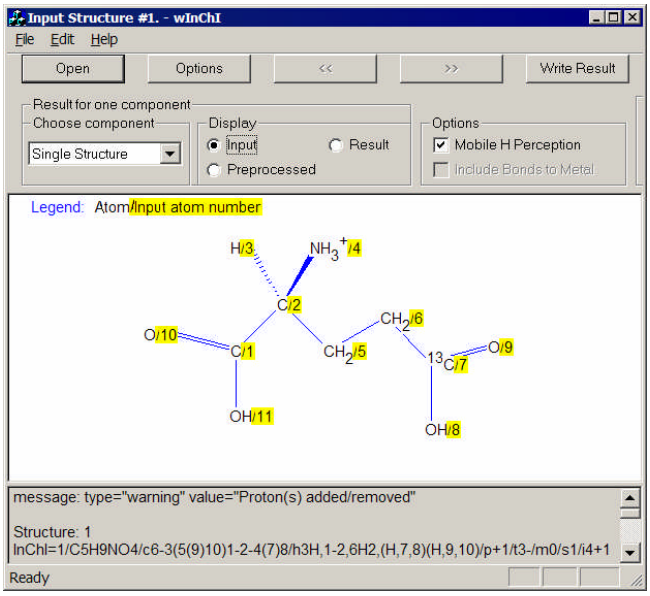


Figure 17.

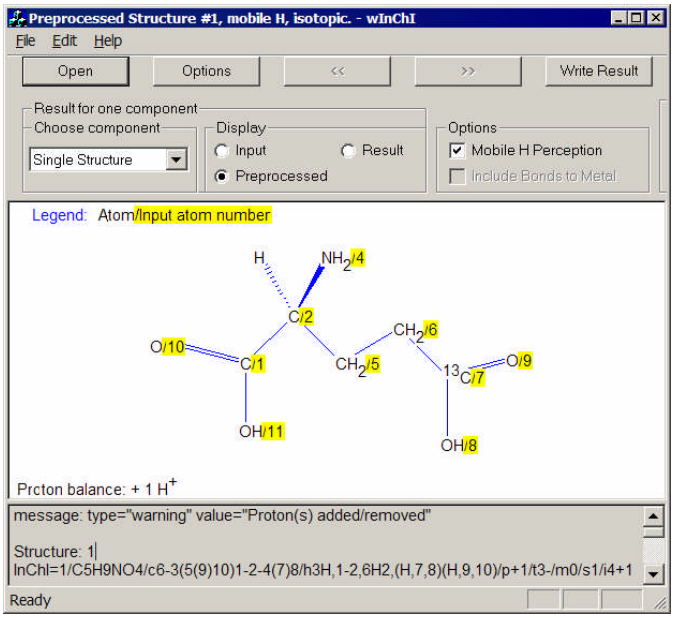


Figure 18.

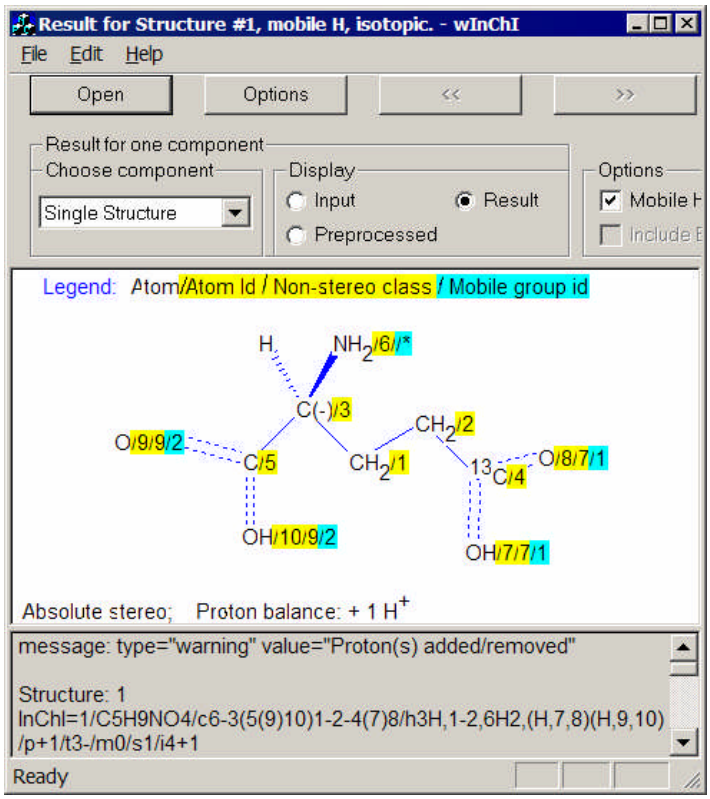


Figure 19.

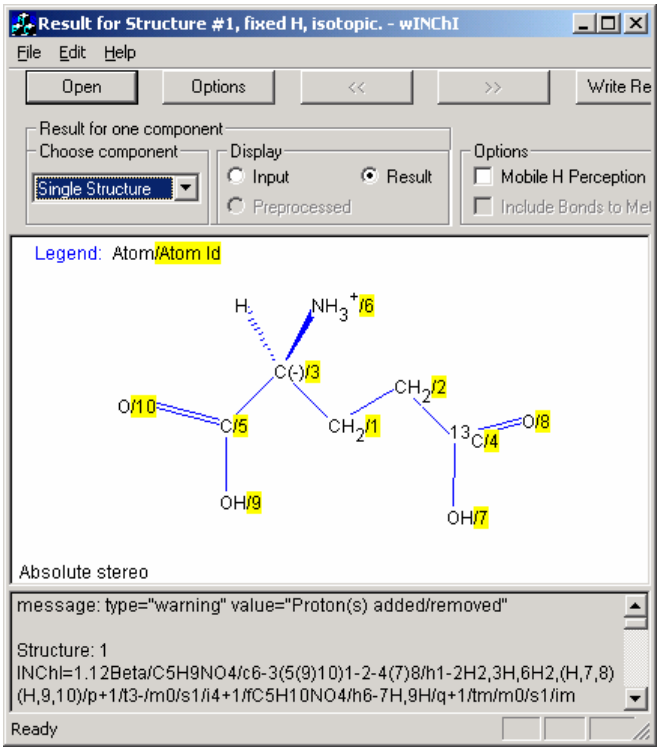


Figure 20.

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

Structure: 1
InChI=1/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/fC5H10NO4/h6-7,9H/q+1

AuxInfo=1/1/N:5,6,2,7,1,4,8,9,10,11/E:(7,8)(9,10)/it:im/l:/E:m/F:5,6,2,7,1,4,8,9,11,10/it:m/rA:11CCHN+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;

==== InChI ANNOTATED CONTENTS ====

Structure: 1

InChI=
{version}1
/{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
/f{fixed_H:formula}C5H10NO4
/h{fixed_H:H_fixed}6-7,9H
/q{fixed_H:charge}+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
/l:{isotopic:original_atom_numbers}
/E:{isotopic:atom_equivalence}m
/F:{fixed_H:original_atom_numbers}5,6,2,7,1,4,8,9,11,10
/it:{fixed_H:abs_stereo_inverted:sp3}m
/rA:{reversibility:atoms}11CCHN+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 21. The Identifier, Auxiliary information, Annotated Identifier and Auxiliary information for (S)-Glutamic acid with “Mobile H Perception” turned off.

The InChI represents the structure of a covalently bonded compound in five distinct ‘layers’:

1. Main Layer

1.1 Chemical Formula

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

In the example in Figure 21, the formula is:
/{formula}C5H9NO4

1.2 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 21, 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.

2.1 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.

2.2 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 21 the proton(s) are:
/p{protons}+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 an designation of whether absolute stereochemistry is required (and if this requires inversion of the relative stereochemistry).

In the example in Figure 21 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 21 the isotopic layer is:
   /i{isotopic:atoms}4+1

5. *Fixed-H Layer*
   When present, this layer provides the location of H-atoms considered mobile in earlier layers along with any needed changes to earlier layers.

   In the example in Figure 21 the Fixed-H layer is:

   /f{fixed_H:formula}C5H10NO4
   /h{fixed_H:H_fixed}6-7H,9H
   /q{fixed_H:charge}+1
   /t{fixed_H:stereo:sp3}m
   /m{fixed_H:stereo:sp3:inverted}0
   /s{fixed_H:stereo:type (1=abs, 2=rel, 3=rac)}1
   /i{fixed_H:isotopic:atoms}m

Note that these names of layers are used in 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. Also, if no mobile H atoms are perceived, there will be no indication of fixed-H layer or mobile H groups in the Identifer.

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 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 an 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 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 two or three lines:

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

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.

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)}
```
Fixed H layer
```
/f{fixed_H:formula}*
/h{fixed_H:H_fixed}
/q{fixed_H:charge}
/b{fixed_H:stereo:dbond}
/t{fixed_H:stereo:sp3}
/m{fixed_H:stereo:sp3:inverted}
/s{fixed_H:stereo:type (1=abs, 2=rel, 3=rac)}
```
Fixed H layer (isotopic part)
```
/i{fixed_H:isotopic:atoms}*
/b{fixed_H:isotopic:stereo:dbond}
/t{fixed_H:isotopic:stereo:sp3}
/m{fixed_H:isotopic:stereo:sp3:inverted}
/s{fixed_H:isotopic:stereo:type (1=abs, 2=rel, 3=rac)}
/o{transposition}
```

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.

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 u. 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 indicates that the stereochemistry has been explicitly entered as ‘unknown’.

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 requested type of stereo is shown in /s segment as /s1 (absolute), /s2 (relative), or /s3 (racemic).

Fixed-H layer

The fixed-H (/f) layer follows. It adds information required to fix the positions of all mobile H atoms. It is structured the same as the main layer, except for absent connections and H_fixed replacing H_atoms layer. For example, H_fixed for acetic acid is 3H: a single mobile H position is fixed at atom 3. Negative values should be subtracted from corresponding H_atoms; in this case lowercase h is used.

Layer transposition

The order of the components in the main section of the identifier may differ from the order in fixed-H section. This is shown in the /o (transposition) segment. The transposition usually occurs because in the sorted order of the components constitution has higher priority than stereo and isotopic layers.

Mobile-H Limitations

Not all possible forms of tautomerism are represented. Specifically, when there are no charged heteroatoms (normalization_type, the first segment in the Auxiliary Info, is 1) this version perceives 1,3 (and limited 1,4 and 1,5)-H-atom transfer as well as 1,2-H-atom migration in 5-membered rings. ; In case of charged heteroatoms the detection of mobile atoms H and removal of protons may be ‘aggressive’ (see InChI Technical Manual). Note that the mobility of H-atoms depends on the environment of a substance as well as its structure, hence can at best be a useful approximation. The definitions used by InChI were intended to represent, to the degree possible, common, current practice.

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}
```
Fixed H part
```
/F:{fixed_H:original_atom_numbers}
/E:{fixed_H:atom_equivalence}
/it:{fixed_H:abs_stereo_inverted:sp3}
/iN:{fixed_H:abs_stereo_inverted:original_atom_numbers}
```
Fixed H isotopic part
```
/I:{fixed_H:isotopic:original_atom_numbers}*
/E:{fixed_H:isotopic:atom_equivalence}
/it:{fixed_H:isotopic:abs_stereo_inverted:sp3}
/iN:{fixed_H: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 wInChI 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.

The Identifier of a reconnected structure (bonds to metals are disconnected by default), if requested, is separated by /r and may contain all layers describer earlier.

Note: The compressed format (available in command-line version) uses base 27 numbers (27 digits are @,a,b,…,z); each number starts with an uppercase letter (never starts with @=zero), following it “digits” (if present) are lowercase. For example,
A=1, A@ = 1x27+0 = 27, Abc = 1x27x27 + 2x27 + 3 = 786.

Error/Warning Output

If problems are encountered during the processing of a structure, they are shown in the first line of Winchi text window or in cInChI 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 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 InChI generation code is written in the 'C' language and the user interface code of wInChI 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 InChI. We still encourage questions and comments of all kinds. Some specific issues for consideration by the users are:

1) Does the software identify features of input structures as expected (is the "normalization" step correct)?
2) Is the InChI for a given compound the same regardless of how it was drawn and are differences appropriate for different input structures?
3) Is H-atom mobility handled in a reasonable manner? Is there a need for tautomer definition be expanded or modified?
4) How can the program or documentation be improved?
5) Is there a need for the format for expression of the InChI to be changed?
6) What capabilities should be added to the next version of InChI?

XII. CONTACT INFORMATION

Mass Spectrometry Data Center
Building 221, Room A111
100 Bureau Drive
National Institute of Standards and Technology
Gaithersburg, Maryland 20899-8380

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. InChI version 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(#)"

Appendix 2. What's New in InChI Software Version 1.01

Compatibility

Identifiers produced by the InChI version 1 software version 1.01 are same as those produced by released in April 2005 InChI version 1 provided that both were generated with the same InChI options.

AuxInfo could be different for some of the structures containing "Either" (wavy) single stereo bonds.

Ordinal numbers of __stdcall entry points exported from Win32 inchilib.dll with __declspec(dllexport) have changed. Ordinal numbers of __cdecl entry points remain unchanged.

New Features

New software features

The new software features are turned on by command line options summarized in the following table.

Command line option (without / or – prefix)	Explanation
InChI2InChI	Check input InChI string for significant syntax errors; may remove layers; the result is an InChI string. Meaningful additional options: FixedH, RecMet (keep these layers), Snon (remove stereo layer), NoLabels, Tabbed, STDIO, Compress.
InChI2Struct	Convert input InChI string into a 0D structure. cInChI-1 outputs structure in AuxInfo; InChI software library produces a binary connection table. A description of differences in comparing the input InChI and InChI generated out of the reconstructed structure are in a log file and (in case of InChI software library) in bitmaps related to InChI layers. If AuxInfo containing coordinates is available in the input then cInChI-1.exe would display the reconstructed structure in case of options /d and /dcr.
SPXYZ	Treat Phosphines as stereogenic
SAsXYZ	Treat Arsines as stereogenic
FixSp3bug	Activate fixes of two known bugs in stereochemical sp^3 (/t) segment

None of these options are available in wInChI-1.exe. All of them are available in cInChI-1 and in InChI software library, libinchi.dll/libinchi.so, where new entry points were added for the first two options.

InChI validation protocol

The purpose of InChI validation is to verify that InChI code included an application or InChI code ported to a particular compiler or operating system produces same InChI as the software distributed by IUPAC, cInChI-1, software version 1.01.

Bugs Fixed

Several bugs have been fixed. As a result, so far no structure that would cause InChI software version 1.01 to fail is known. Most of the bugs were discussed at InChI-discuss,

https://lists.sourceforge.net/lists/listinfo/inchi-discuss

The elimination of the following two bugs in creating stereochemical sp^3 (/t) segment may lead to identifiers different from those produced by InChI version 1 software version 1.00:

- A stereocenter connected by 3 bonds in a 2D structure could be undetected if an average bond length is greater than 20.
- A parity of a stereocenter connected by 4 bonds in a 2D structure such that a stereobond exactly overlaps with another single bond may become “undefined” depending on the order of the atoms in the structure.

To maintain compatibility with software version 1.00, the fix for these bugs is activated by a new command line option, FixSp3bug.

Processing Files Greater Than 2 Gigabytes

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:

cInChI-1 /NEWPS input.sdf output.txt logfile.log NUL

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

Running wInChI-1 under Linux

Reportedly, wInChI-1.exe could be run under WINE build 20050524 running on Mandrake 10.1