International Chemical Identifier for Reactions (RInChI)

The key to effectively managing reaction databases

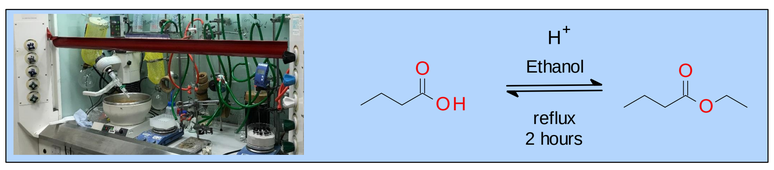
Guenter Grethe[1], Gerd Blanke[2], Jonathan M. Goodman[3], Hans Kraut[4]

Background

Since its inception, the IUPAC International Chemical Identifier (InChI) has found wide acceptance as a standard in the chemical community. In order to widen the applicability of the identifier, the IUPAC Division VIII Subcommittee and the InChI Trust have initiated several projects to extend the usage of the identifier. Among these is the development of a non-proprietary, international identifier for reactions (RInChI) to describe chemical reactions in a unique, machine-readable, character string based on the InChI algorithm suitable for data storage and indexing. Prototype versions of the RInChI supported by the IUPAC and the University of Cambridge have been available since 2011. The first official release (RInChI-V1.00), funded by the InChI Trust is now available for download.

Reactions are more difficult to codify than molecules. A molecule can usually be uniquely and precisely defined by a structural drawing, which shows the atom types, the bond connectivity, and the stereochemistry. A reaction includes molecules, but also times, temperatures, concentrations, rates of mixing, yields, and numerous other quantities that are recorded with varying levels of precision.

The goal of the development of RInChI is to define a reaction identifier that has enough detail to be useful and yet not so much that practically identical processes would be given different labels as described in the following diagram:



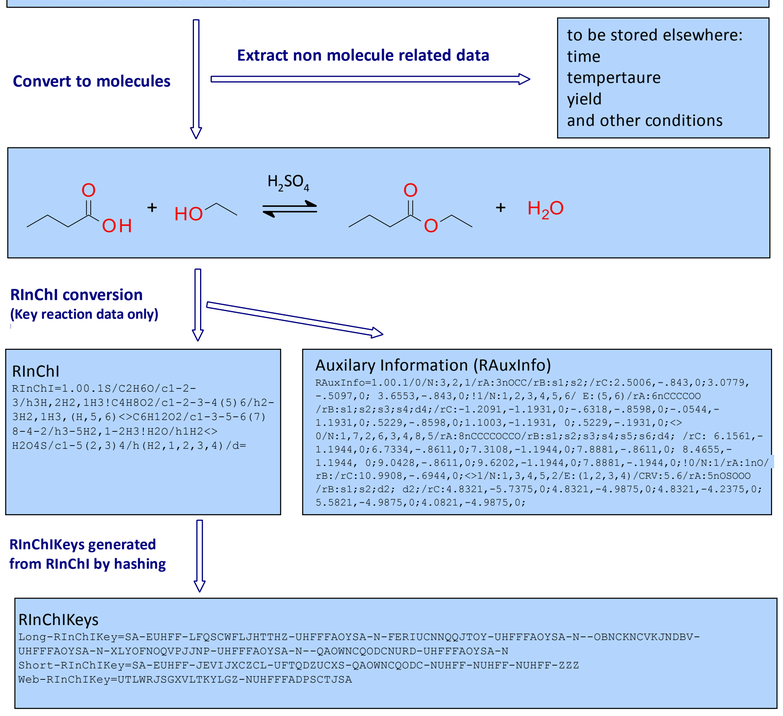
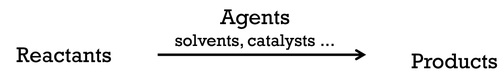


Fig 1: RInChI creation diagramm

RInChI as new type of identifier is extremely useful in the creation and analysis of reaction databases as well as in Web searches where reactions must be found in searches over multiple databases with different data models.

While identifiers for molecule exist for a long time (e.g. CAS-Number, MFCDNumber, InChI) there is no public identifier for reactions available up to now. In fact, the Web-RInChIKey introduced here is the first reaction identifier that is suitable for the special conditions of Web searches.

Types of RInChIs



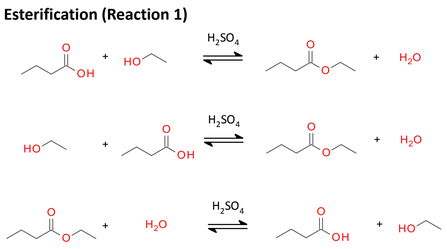
- RInChI
  - The RInChI is calculated from the InChIs of each reactant, product and agent
  - RAuxInfo consists of the AuxInfo of each reaction component
  - Recalculation of RXN/RD files from RInChI and RAuxInfo
- Long-RInChIKey
  - Calculated from InChIKeys of each reactant, product and agent.
- Short-RInChIKey
  - Fixed length hash over all reagents, products and agents
- Web-RInChIKey
  - Fixed length hash developed from the reaction components but ignoring the specific role within the reaction.

Format of RInChI

RInChI=1.00.1S/*layer2*<>*layer3*<>*layer4*/*d(+,-,=)*/*u#2-#3-#4*

- Layer 1 identifying the RInChI version 1.00 using Standard InChI version1
- Layer 2 / 3: InChIs of starting materials and products
- Layer 4: InChIs of all agents (catalysts, solvents, etc.)
- Layer 5: Directional identifier with d+ ≡ forward, d- ≡ backward and d= ≡ equi-librium reaction
- Layer 6: No-structure flag layer identified by /*u* with *#2* as number of no-structures in layer 2, *#3* number of no-structures in layer 3 and *#4* number of no-structures in layer 4
- Additional rules for building the RInChI (string)
  - The "InChI=1S/" is omitted from InChIs listed in the RInChI as it is already specified in layer 1.
  - Separators
    - Use "/" as divider between main layers
    - Use "<>" as divider between the reactant, product and agent layers
    - Use "!" as divider between molecules
    - Use "-" as divider for the occurrence number of no-structures
    - Separators are omitted if they have nothing to separate
  - Make RInChI representation unique by alphabetical ordering
    - Order InChIs within each layer alphabetically
    - Order layer 2 and 3 alphabetically.
      - In case that layer 2 and 3 are re-ordered the directional identifiers d+, d- must be exchanged to keep the reaction direction, i.e. d+ -> d- or vice versa d- -> d+. (d= implies an equilibrium, and we intend to introduce dx for reactions which do not work.)
  - The same molecule may not appear in both layer two and layer three, and no molecule should be repeated within each layer. Anything that is in both layers should be removed from both and added to layer four.
  - Empty layers are permitted.

Example Esterification (Reaction 1)

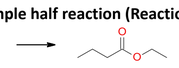


RInChI=1.00.1S/C2H6O/c1-2-3/h3H,2H2,1H3!C4H8O2/c1-2-3-4(5)6/h2-3H2,1H3,(H,5,6)<>C6H12O2/c1-3-5-6(7)8-4-2/h3-5H2,1-2H3!H2O/h1H2<>H2O4S/c1-5(2,3)4/h(H2,1,2,3,4)/d=

Special cases: "half" reactions and no structures

Half reactions where only the starting materials or the products are described are represented using an empty layer for the unknown part of the reaction; empty strings are not added to the RInChI.

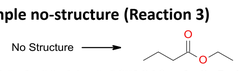
Example half reaction (Reaction 2)



RInChI=1.00.1S/<>C6H12O2/c1-3-5-6(7)8-4-2/h3-5H2,1-2H3/d+

The InChI for no-structures is an empty string as well. To distinguish no-structures in reactions from half reactions, no-structures are ignored within their group but are added as an additional flag at the right end of the RInChI string using the format u#1-#2-#A with # as integer representing the total number of No-Structures in group1, group2 and the agent group as 3rd group

Example no-structure (Reaction 3)



RInChI=1.00.1S/<>C6H12O2/c1-3-5-6(7)8-4-2/h3-5H2,1-2H3/d+/*u1-0-0*

RAuxInfo

The RAuxInfo contains the AuxInfos of the InChIs participating in the reaction, i.e. it contains the atom coordinates, the information the original connectivity including stereochemistry, tautomeric characteristics and other molecule properties that are not directly handled by InChIs.

The RAuxInfo is necessary to re-build the original RXN or RD file. Without RAuxInfo the standard InChI routines return molfiles with all coordinates set to 0, so that the molecule coordinates must be recalculated. The recalculation may end in other atom locations of the same molecule. Although the stereocenters are kept by InChI the re-calculation may lead into another distribution of the bound ligands around the stereocenter so that the stereo parity may flip.

Example Esterification (Reaction 1)

RAuxInfo=1.00.1/0/N:3,2,1/rA:3nOCC/rB:s1;s2;/rC:2.5006,-.843,0;3.0779,-.5097,0; 3.6553,-.843,0;!1/N:1,2,3,4,5,6/ E:(5,6)/rA:6nCCCCOO/rB:s1;s2;s3;s4;d4;/rC:-1.2091,-1.1931,0;-.6318,-.8598,0;-.0544,-1.1931,0;.5229,-.8598,0;1.1003,-1.1931, 0;.5229,-.1931,0;<>0/N:1,7,2,6,3,4,8,5/rA:8nCCCCOCCO/rB:s1;s2;s3;s4;s5;s6;d4; /rC: 6.1561,-1.1944,0;6.7334,-.8611,0;7.3108,-1.1944,0;7.8881,-.8611,0; 8.4655,-1.1944, 0;9.0428,-.8611,0;9.6202,-1.1944,0;7.8881,-.1944,0;!0/N:1/rA:1nO/rB:/rC:10.9908,-.6944,0;<>1/N:1,3,4,5,2/E:(1,2,3,4)/CRV:5.6/rA:5nOSOOO/rB:s1;s2;d2; d2;/rC:4.8321,-5.7375,0;4.8321,-4.9875,0;4.8321,-4.2375,0;5.5821,-4.9875,0;4.0821,-4.9875,0;

Example Half Reaction (Reaction 2)

RAuxInfo=1.00.1/<>0/N:1,7,2,6,3,4,8,5/rA:8nCCCCOCCO/rB:s1;s2;s3;s4;s5;s6;d4;/rC:5.7129,-7.3792,0;6.2903,-7.0459,0;6.8676,-7.3792,0;7.445,-7.0459,0;8.0223,-7.3792,0;8.5997,-7.0459,0;9.177,-7.3792,0;7.445,-6.3792,0;

Example No-Structure (Reaction 3)

RAuxInfo=1.00.1/<>0/N:1,7,2,6,3,4,8,5/rA:8nCCCCOCCO/ rB:s1;s2;s3;s4;s5;s6;d4;/rC:6.8243,-5.2486,0;7.3225,-4.961,0;7.8165,-5.2486,0;8.3105,-4.961,0;8.8046,-5.2486,0;9.3027,-4.961,0;9.7968,-5.2486,0;8.3105,-4.3903,0

Reaction InChIKeys (RInChIKey)

Like the InChIKey the RInChIKeys are created by hashing the InChIs ordered in RInChI format. Depending on the purpose three different types of RInChIKeys have ben developed: Long-RInChIKeys, Short-RInChIKeys and Web-RInChIKeys

Long-RInChIKeys

Long-RInChIKeys are created by building the InChIKeys for each of the components participating in the reaction based on the order of RInChI

Long-RInChIKey=SA-(F,B,E,U)UHFF-layer3--layer4--layer5

- Layer 1 identifying the InChI version SA as Standard InChI 1 (=A)
- Layer 2:
  - The first letter describes the direction of the reaction with **F**(orward) = 'd+', **B**(ackward) = 'd-', **E**(quilibrium) = 'd=', and **U**(nspecified)
  - The remaining 4 characters 'UHFF' are currently not used and will handle reaction conditions in a later version of RInChI
- Layer 3 to 5: InChIKeys related to the InChIs in the RInChI layers 2 to 4 using the order of InChIs in the corresponding RInChI layers
  - Multiple InChIKeys in the same layer are separated by a single hyphen "-".
  - Layers 3, 4, and 5 are separated by a double hash "--".
  - Each No-structure or one of the related pseudoatoms (R, X, A and *-atom) is repre-sented by the hash value for an empty string ("MOSFIJXAXDLOML-UHFFFAOYSA-N")
  - Layers are only displayed if they contain at least one InChIKey. If the last layer of the string is empty the separator "--" is omitted as well.

Long-RInChIKeys are a valuable tool for the database storage of reactions. Beside uniqueness checks, they allow the identification of each reaction component by simple text searches based on Standard InChIKeys.

Example Esterification (Reaction 1)

Long-RInChIKey=SA-EUHFF-LFQSCWFLJHTTHZ-UHFFFAOYSA-N-FERIUCNNQQJTOY-UHFFFAOYSA-N--OBNCKNCVKJNDBV-UHFFFAOYSA-N-XLYOFNOQVPJJNP-UHFFFAOYSA-N--QAOWNCQODCNURD-UHFFFAOYSA-N

Example Half Reaction (Reaction 2)

Long-RInChIKey=SA-FUHFF---OBNCKNCVKJNDBV-UHFFFAOYSA-N

Example No-Structure (Reaction 3)

Long-RInChIKey=SA-FUHFF-MOSFIJXAXDLOML-UHFFFAOYSA-N--OBNCKNCVKJNDBV-UHFFFAOYSA-N

Because any molecule within a reaction can be identified by a text search by its InChIKey the Long-RInChIKeys are a good tool to provide reaction component searches and can be used to build synthesis trees where e.g. an starting material is identified being a product within another reaction.

On the other hand the size of Long-RInChIKeys depends on the number of reaction components and is not fixed. Therefore, it is not the preferred tool for database actions like reaction indexing.

Short-RInChIKeys

Short-RInChIKeys have a fixed length of 55 letters plus 8 hyphens as separators resulting in a total of 63 characters. They are built by hashing each of the layers of the RInChI to a fixed length using the following format

*Short-RInChIKey=SA-(F,B,E,U)UHFF - **hash over all major layers in RInChI 2 – hash over all major layers in RInChI 3 – hash over all major layer in RInChI 4 – hash over all minor layers plus sum of protonation states in RInChI 2 – hash over all minor layers plus sum of protonation states in RInChI 3 – hash over all minor layers plus sum of protonation states in RInChI 4 – ### with # = Z, A, B…***

- Layer 1 and 2 correspond to layer 1 and 2 of the Long-RInChIKey (see above)
- Layer 3, 4, and 5 are hashes over all major layers of the InChIs from the layers 2, 3, and 4 of the RInChI keeping the first 10 letters.
- Layer 6, 7, and 8 are hashes over all minor layers of the InChIs from the layers 2, 3, 4 of the RInChI keeping the first 4 letters plus one letter representing the sum over all protonation states of the related InChIs
- Layer 9 consists of three letters, each of them representing the count of no-structures found in the sixth layer of RInChI. To simplify reading, "Z" corresponds to a count of zero no-structures (i.e. no no-structure), while "A" stands for one no-structure, B for two and so forth. The format is given by ### with # = Z, A, B,
- All layers are separated by an hyphen "-".

Example Esterification (Reaction 1)

Short-RInChIKey=SA-EUHFF-JEVIJXCZCL-UFTQDZUCXS-QAOWNCQODC-NUHFF-NUHFF-NUHFF-ZZZ

Example Half Reaction (Reaction 2)

Short-RInChIKey=SA-FUHFF-UHFFFADPSC-OBNCKNCVKJ-UHFFFADPSC-NUHFF-NUHFF-NUHFF-ZZZ

Example No-Structure (Reaction 3)

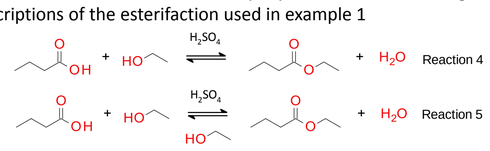
Short-RInChIKey=SA-FUHFF-UHFFFADPSC-OBNCKNCVKJ-UHFFFADPSC-NUHFF-NUHFF-NUHFF-AZZ

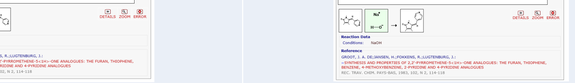
The fixed length of the Short-RInChIKey makes it especially suitable for

- exact searches of reactions in databases (and in the WEB)
- indexing reactions in databases
- Identifying identical reactions in different databases.

Web-RInChIKeys

The depiction of a chemical reaction is not uniquely defined. The following 3 depictions are alternative descriptions of the esterifaction used in example 1





The role of the components may differ depending on the rules used by the authors. If you want to compare reactions from different sources (e.g. for searches in the Web) you need a more flexible RInChIKey that ignores the roles. Therefore the more flexible Web-RInChIKey has been defined

Web-RInChIKey=major layer-minor layer

Building steps

1. All InChIs of the reaction components are ordered alphabetically. The role of the components are ignored.
2. Any duplicates are deleted from the list.
3. All major InChI layers are concatenated and hashed using the InChI hash functionality. The first 17 characters of the hash string define the "major" layer of the Web-RInChIKey
4. The protonation states of all participating InChIs are summed up with 0 as protonation state for those that are not explicitly defined in the InChI string. L = -2, M = -1, N = 0, O = +1, P = +2 … as 1-letter representation for the protonation state become the first character of the "minor" layer of the Web-RInChIKey.

   The minor layers of all InChIs are combined and the resulting string is hashed by the sha2 method provided by the InChI library. The first 12 characters of the mi-nor layers hash are joined with the one letter for the protonation state

With these definitions the reaction 4, 5, and 6 get following RInChIKeys

Example Reaction 4

Short-RInChIKey=SA-EUHFF-JEVIJXCZCL-UFTQDZUCXS-QAOWNCQODC-NUHFF-NUHFF-NUHFF-ZZZ
Web-RInChIKey=UTLWRJSGXVLTKYLGZ-NUHFFFADPSCTJSA

Example Reaction 5

Short-RInChIKey=SA-EUHFF-JEVIJXCZCL-UFTQDZUCXS-UAUFKIWNBD-NUHFF-NUHFF-NUHFF-ZZZ
Web-RInChIKey=UTLWRJSGXVLTKYLGZ-NUHFFFADPSCTJSA

Example Reaction 6

Short-RInChIKey=SA-EUHFF-FERIUCNNQQ-OBNCKNCVKJ-DNBJJWMYJT-NUHFF-NUHFF-NUHFF-ZZZ
Web-RInChIKey=UTLWRJSGXVLTKYLGZ-NUHFFFADPSCTJSA

While the Short-RInChIKeys (and Long-RInChIKeys and RInChI itself) differ and represent the reaction as drawn, the Web-RInChIKeys of all 3 reactions are identical.

The Web-RInChIKey has a fixed length of 47 characters with 17 letters in the major and 15 letters in the minor layer. Web-RInChIKey are especially suitable to compare reactions databases with different drawing rules and/or data models which is a typical situation for Web searches. Web searches are especially supported by the longer string sets for the major and minor layers which on its own already build unique identifier for a reaction.

Input formats

- RXN/RD files (hierarchical and flat)
- Molfiles for component roles (starting material, product, agent) as application example
- For the RInChI calculation an additional parameter must be set to make the reaction a backward or equilibrium reaction.
  - The standard input formats interpret reactions as forward reactions only.

RInChI for SPRESI

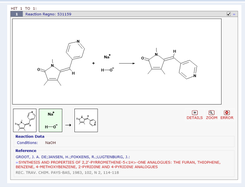
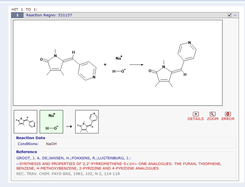
SPRESI by InfoChem is a chemical structure and reaction database that includes 5.52 million structures, 4.26 million reactions and 32 million factual data entries extracted from 675,000 references and 164,000 patents with the major focus to organic chemistry built since 1970.

The full set of RInChI and RInChIKeys has been calculated over a weekend with only 239 reactions failing the conversions. After that the database was checked for duplicates using the Short-RInChIKey and the Web-RInChIKey

	Reactions	Unique reactions with multiple entries in SPRESI	Total unique reactions
SPRESI_RINCHI	4.564.718		
Short-RInChIKey: Total number of reactions with at least one duplicate	418.540	176.454	4.322.632
Web-RInChIKey: Total number of reactions with at least one duplicate	455.572	191.382	4.300.528
Delta	37.032	14.928	-22.104

The analysis of the duplicate results showed that most of the reactions with identical Web-RInChIKey and differing Short-RInChIKey differ in the direction parameter B(ackward) or F(orward).

Example: Web-RInChIKey=AAAOSLOCRWISFZQDE-MLNOZKJYGJAWFSA

Short-RInChIKey=SA-**B**UHFF-XJIWICTZTS-SFRVIVNSLX-HEMHJVSKTP-NXJJT-MFOWU-MUHFF-ZZZ	Short-RInChIKey=SA-**F**UHFF-SFRVIVNSLX-XJIWICTZTS-HEMHJVSKTP-MQXDD-NSIAJ-MUHFF-ZZZ
RXN Regno = 531159	RXN Regno = 531157
	

The standard reaction exchange formats only know forward reactions so that equilibrium reactions must be entered twice. While the RInChI is the only representation handling equilibrium reactions, duplicate checks based on Short-InChIKeys and Web-RInChIKeys provide a fast and simple way to detect equilibrium reactions in reaction databases.

Availability

- Download (under open-source agreement)
  - http://www.inchi-trust.org/downloads/
  - RInChI libraries for Linux and Windows
  - Application example
    - RInChI cmdline tool
    - Example applications to use molfiles with reaction roles.
- Test Web-site provided by University of Cambridge, UK including test database with one million reactions
  - http://www-rinchi.ch.cam.ac.uk/, Accessed May 2018
- Development Web-Site
  - http://proteax.dk/rinchi/rinchi.py/demo, Accessed May 2018
- Additional tools built for RInChI testing
  - Oracle cartridge for RInChIs based on Oracle 11
  - Pipeline Pilot components

Plans for upcoming releases

- Additional input formats: InChI (with component roles), reaction smiles, UDM (currently further developed by Pistoia Alliance)
- Extension of the direction layer to failed reactions.
- Enhanced representation of stereochemistry and tautomers to reflect reactions with changes in terms of stereochemistry and/or tautomerism.
- Integration of Mixture InChIs (MInChIs) handling mixtures and formulations
- Generic reactions depending on the further development of InChI
- Additional layer /ProcAuxInfo handling the conditions of reactions as proposed by the Lapkin group at the University of Cambridge (Jacob P-M, Lan Y, Goodman JM, Lapkin AA (2017) J Cheminform 9:23)
- Support for additional operating systems: MAC
- Adaption to further releases of InChI (currently InChI 1.04)

Publications

- International chemical identifier for reactions (RInChI)
  G. Grethe, G. Blanke, H. Kraut and J. M. Goodman *J. Cheminformatics* 2018, **10**, 22.
  DOI: 10.1186/s13321-018-0277-8
- International chemical identifier for reactions (RInChI)
  G. Grethe, J. M. Goodman and C. H. G. Allen *J. Cheminformatics* 2013, **5**, 45.
  DOI: 10.1186/1758-2946-5-45

Special acknowledgements are due to the RInChI Working Group for their contributions. We are grateful to Alan McNaught and Steve Heller from the InChI Trust for initializing and supporting the project. Financial support from the IUPAC Division VIII Subcommittee and the Royal Society of Chemistry is very much appreciated. Chad Allen, Matthew Morton, James Davies, Rudi Pisa, Ben Hammond, Nicholas Parker, James Apthorp, Bryn Reinstadler and Duncan Hampshire are thanked for their contributions to the program. We are also grateful to NextMove Software, FIZ Chemie (Wiley) Berlin, InfoChem and Elsevier for providing trial datasets to test the program.
Working group: Colin Batchelor (Royal Society of Chemistry), Gerd Blanke (StructurePendium), Jonathan Goodman (University of Cambridge), Guenter Grethe, Jan Holst Jensen (BioChemFusion), Hans Kraut (InfoChem), Alexander Lawson, Henry Matuszczyk (InfoChem), Martin Schmidt (FIZ Chemie), Keith Taylor (Lederac Consulting).
Authors: [1]) Dr. Guenter Grethe, Poway, CA 92064, US; [2]) StructurePendium Technologies GmbH, Essen, Germany; [3]) University of Cambridge, Department of Chemistry, Cambridge, UK; [4]) InfoChem Gesellschaft für chemische Information mbH, Munich, Germany
Funding: InChI Trust, IUPAC, Royal Society of Chemistry.
Downloads: The Reaction InChI (RInChI) is available on the InChI Trust website: http://www.inchitrust.org/ Accessed May 2018