2.8.1: R Assignment 2A

Exploring Chemical Identity

J. Cuadros

August 5th, 2019

Downloadable Files

L02_ChemicalIdentity.Rmd
RL02_ChemIdentity.pynb

- You can use the R-studio you created in section 1.4 or the Jupyter hub at LibreText: https://jupyter.libretexts.org (see your instructor if you do not have access to the hub).
- This page is an html version of the above R file.
  - If you have questions on this assignment you should use this web page and the hypothes.is annotation to post a question (or comment) to the 2019OLCCStu class group. If you are not on the discussion group you should contact your instructor for the link to join.
- .pynb is a Jupyter Notebook that opens with an R Kernal

Objectives

- Understand the problem of chemical identity.
- Explore some chemical substances identifiers.
- Understand the layered model of the InChI as a model to chemical identity.

1. The Problem of Chemical Identity

Since the XVIII century, when chemists started to understand chemical substances have a fixed composition, we have faced the need to identify and discriminate them. In the analogic word, we use names (both traditional and systematic) and formulas to fullfill this role. In the digital world, registry numbers and line notations are commonly used for substance identification.

Chemistry information and documentation recurrently face the problem of having to classify substance records (data pieces) in a systematic way. But as we will explore in this activity, this is trickier than it seems.

Let's start facing the problem. You are given 18 substance records (that's not much, PubChem holds more than 200 milions of these, https://pubchemdocs.ncbi.nlm.nih.gov/statistics) and you are asked to decide which correspond to the same substance (and why).

Here are the records:

- https://pubchem.ncbi.nlm.nih.gov/substance/227885365
- https://pubchem.ncbi.nlm.nih.gov/substance/242744695
- https://pubchem.ncbi.nlm.nih.gov/substance/329830556
- https://pubchem.ncbi.nlm.nih.gov/substance/341141642
- https://pubchem.ncbi.nlm.nih.gov/substance/341193751
- https://pubchem.ncbi.nlm.nih.gov/substance/342898240
- https://pubchem.ncbi.nlm.nih.gov/substance/355138175
- https://pubchem.ncbi.nlm.nih.gov/substance/355178551
- https://pubchem.ncbi.nlm.nih.gov/substance/369730804
- https://pubchem.ncbi.nlm.nih.gov/substance/376125581
- https://pubchem.ncbi.nlm.nih.gov/substance/376145687
- https://pubchem.ncbi.nlm.nih.gov/substance/376602811
- https://pubchem.ncbi.nlm.nih.gov/substance/383210891
- https://pubchem.ncbi.nlm.nih.gov/substance/383219135
- https://pubchem.ncbi.nlm.nih.gov/substance/383428756
- https://pubchem.ncbi.nlm.nih.gov/substance/384452886
- https://pubchem.ncbi.nlm.nih.gov/substance/384647147

- https://pubchem.ncbi.nlm.nih.gov/substance/385112115

Exercise 1a: Browse these records and make a table that includes, for each record, the given name, its molecular formula and its structural formula.

```
sids <- c(227885365, 242744695, 329830556, 341141642, 341193751, 342898240, 355138175
          376602811, 383210891, 383219135, 383428756, 384452886, 384647147, 385112115)

paste("# Number of SIDs:", length(sids) )

pugrest <- "https://pubchem.ncbi.nlm.nih.gov/rest/pug"
pugoper <- "cids"
pugout  <- "txt"

pugin <- paste("substance/sid/", paste(sids[1:length(sids)],collapse=","),sep="")
url <- paste(pugrest,pugin,pugoper,pugout,sep="/")
cids <- readLines(url)

cids <- (unique(cids))
paste("# Number of CIDs:", length(cids) )

pugin <- paste("compound/cid/", paste(cids[1:length(cids)],collapse=","),sep="")
pugoper <- "property/IUPACName,MolecularFormula,IsomericSMILES"
pugout  <- "csv"
url <- paste(pugrest,pugin,pugoper,pugout,sep="/")

df <- read.table(url,sep=",",header=TRUE)
print(df)
```

run | restart | restart & run all

Exercise 1b: How many different substances do you think there are in this set? How would you classify them?

2. Digital Identifiers

In the digital world, identity is usually associated with the use of an identifier; when two identifiers coincide when we can say that both data pieces belong to the same entiy.

For substance, and besides registry numbers, two common identifiers are SMILES and InChI.

Exercise 2a: For each record, check if the data providers included any SMILES or InChI information. Collect this information when available.

Exercise 2b: For each record, use a molecular drawing program to compute the SMILES, the standard InChI and the InChIKey. Make a table with them. If you don't have a molecular drawing program at hand, you may consider using MolView (http://molview.org/) or the drawing tool included in the Chemical Identifier Resolver (https://cactus.nci.nih.gov/chemical/structure).

Exercise 2c: Compare the provider identifiers with the computed ones. Do you see any differences?

Exercise 2d: Would you reconsider the classification you decided in exercise 1b?

3. The InChI Layered Notation and Identity Matching

If you look carefully at the InChI for the different records, you will notice that some of the identifiers are more similar then others. Some match completely, while some others may match only for some of the layers, especially for the main layer. Sometimes, we consider to be the same substance, any substance where the InChI main layer is coincident. For other applications, some other layers need to be taken into account; for instance, stereochemical information is critical in health-related uses.

Exercise 3a: Classify the records according to the main layer of the InChI.

Exercise 3b: Classify the records again according to the full InChI.

Substance records in PubChem are grouped into compound records. This information appears in each one of the elements of the set.

Exercise 3c: Compare the classification used in PubChem with the InChI-based classifications done in 3a and 3b.