Structure Search

S. Kim, J. Cuadros

November 15th, 2019

Objectives

- Learn various types of structure searches including identity search, similarity search, substructure and super structure searches.
- Learn the optional parameters available for each search type.

Using PUG-REST, one can perform various types of structure searches (https://bit.ly/2lPznCo (https://bit.ly/2lPznCo)), including: - identity search - similarity search - super/substructure search - molecular formula search

As explained in a PubChem paper (https://bit.ly/2kirxky (https://bit.ly/2kirxky)), whereas structure search can be performed in either an 'asynchronous' or 'synchronous' way, it is highly recommended to use the synchronous approach.
The synchronous searches are invoked by using the keywords prefixed with 'fast', such as **fastidenity**, **fastsimilarity_2d**, **fastsimilarity_3d**, **fastsubstructure**, **fastsuperstructure**, and **fastformula**.

In this task, we will use some cheminformatics packages to ease some processes. In R, some options are `rcdk`, `ChemmineR` and `ChemmineOB`. In Python, a useful package is `RDKit`; in R, we'll make use of it online version, the Beaker API of ChEMBL (https://chembl.gitbook.io/chembl-interface-documentation/web-services (https://chembl.gitbook.io/chembl-interface-documentation/web-services)).

1. Identity Search

PUG-REST allows you to search the PubChem Compound database for molecules identical to the query molecule. PubChem's identity search supports different contexts of chemical identity, which the user can specify using the optional parameter, "identity_type". Here are some commonly-used chemical identity contexts. - **same_connectivity**: returns compounds with the same atom connectivity as the query molecule, ignoring stereochemistry and isotope information. - **same_isotope**: returns compounds with the same isotopes (as well as the same atom connectivity) as the query molecule. Stereochemistry will be ignored. - **same_stereo**: returns compounds with the same stereochemistry (as well as the same atom connectivity) as the query molecule. Isotope information will be ignored. - **same_stereo_isotope**: returns compounds with the same stereochemistry AND isotope information (as well as the same atom connectivity). This is the default.

The following code cell demostrates how these different contexts of chemical sameness affects identity search in PubChem.

```
if(!require("httr")) {
  install.packages(("httr"), repos="https://cloud.r-project.org/",
         quiet=TRUE, type="binary")
  library("httr")
}
```

```
## Loading required package: httr
```

```
if(!require("jsonlite")) {
  install.packages(("jsonlite"), repos="https://cloud.r-project.org/",
         quiet=TRUE, type="binary")
  library("jsonlite")
}
```

```
## Loading required package: jsonlite
```

```
if(!require("png")) {
  install.packages(("png"), repos="https://cloud.r-project.org/",
         quiet=TRUE, type="binary")
  library("png")
}
```

```
## Loading required package: png
```

```
if(!require("grid")) {
  install.packages(("grid"), repos="https://cloud.r-project.org/",
       quiet=TRUE, type="binary")
  library("grid")
}
```

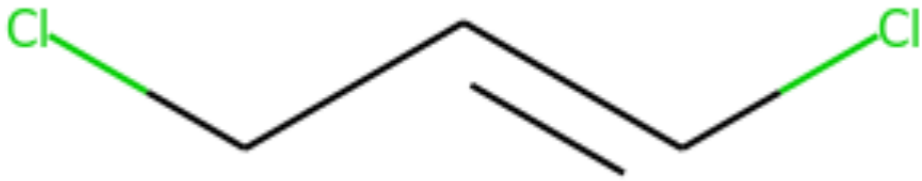
```
## Loading required package: grid
```

```
if(!require("gridExtra")) {
  install.packages(("gridExtra"), repos="https://cloud.r-project.org/",
       quiet=TRUE, type="binary")
  library("gridExtra")
}
```

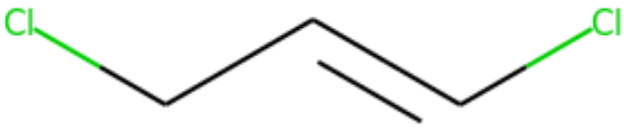
```
## Loading required package: gridExtra
```

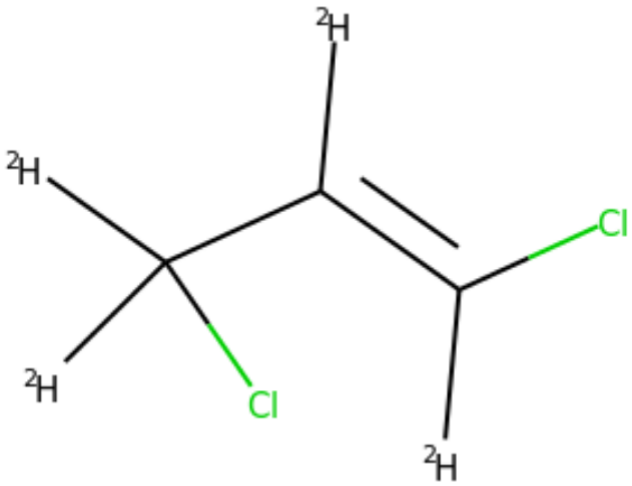
```
prolog <- "https://pubchem.ncbi.nlm.nih.gov/rest/pug"
smiles <- "C(/C=C/Cl)Cl"
options <- c('same_stereo_isotope',
             'same_stereo',
             'same_isotope',
             'same_connectivity') # same_stereo_isotope is the default


for(opt in options) {
    print(paste ("#### Identity_type:", opt))
    url <- paste(prolog,
                 "/compound/fastidentity/smiles/",
                 "property/isomericsmiles/csv?smiles=",
                 URLencode(smiles,reserved = T),
                 "&identity_type=",
                 opt, sep="")
    dfChem <- read.table(url, sep=",", header=T)
    print(dfChem)


    url_img <- paste("https://www.ebi.ac.uk/chembl/api/utils/smiles2image",
        "?size=300&engine=rdkit",sep="")
    res  <- POST(url_img,
                body=list(smiles=paste(dfChem[,2],collapse="\n")))
    img <- readPNG(res$content, native=TRUE)
    grid.arrange(rasterGrob(img))

    Sys.sleep(.5)
}
```

```
## [1] "#### Identity_type: same_stereo_isotope"
##     CID IsomericSMILES
## 1 24726   C(/C=C/Cl)Cl
```

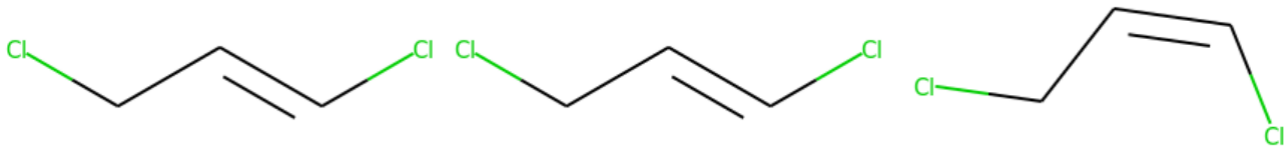



```
## [1] "#### Identity_type: same_stereo"
##         CID                         IsomericSMILES
## 1     24726                           C(/C=C/Cl)Cl
## 2 102602172 [2H]/C(=C(/[2H])\\Cl)/C([2H])([2H])Cl
```

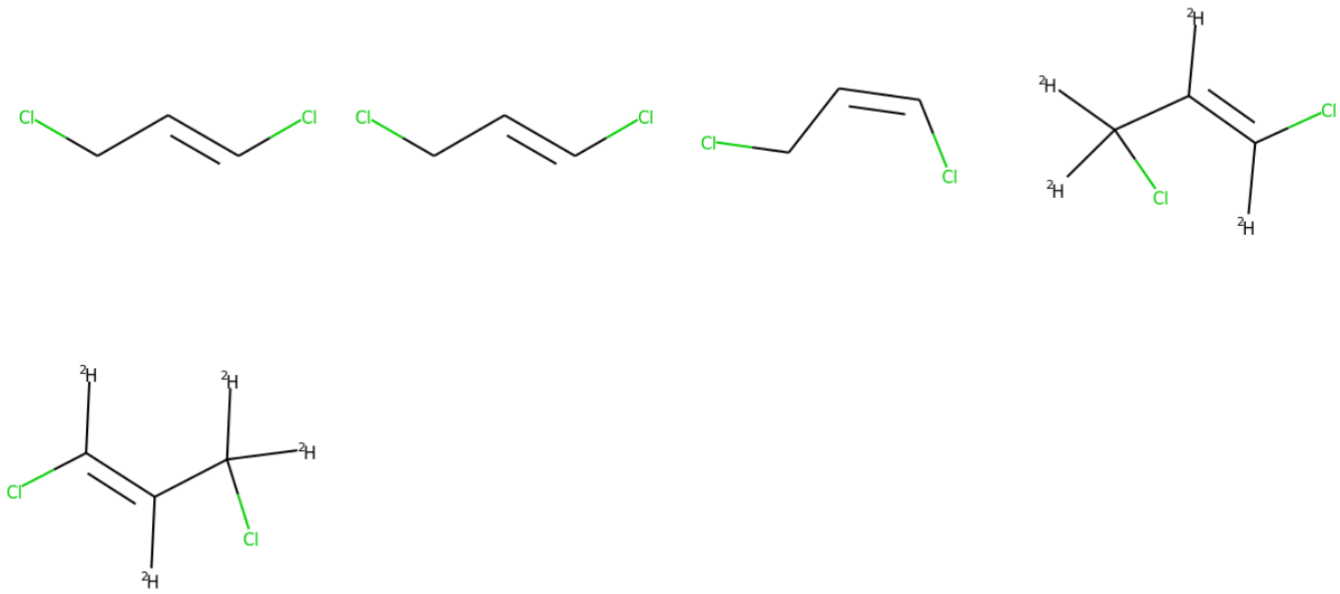





```
## [1] "#### Identity_type: same_isotope"
##       CID IsomericSMILES
## 1   24726   C(/C=C/Cl)Cl
## 2   24883     C(C=CCl)Cl
## 3 5280970  C(/C=C\\Cl)Cl
```




```
## [1] "#### Identity_type: same_connectivity"
##         CID                         IsomericSMILES
## 1     24726                           C(/C=C/Cl)Cl
## 2     24883                             C(C=CCl)Cl
## 3   5280970                           C(/C=C\\Cl)Cl
## 4 102602172 [2H]/C(=C(/[2H])\\Cl)/C([2H])([2H])Cl
## 5 131875718      [2H]C(=C([2H])Cl)C([2H])([2H])Cl
```



Exercise 1a: Find compounds that has the same atom connectivity and isotope information as the query molecule.

```
query <- "CC1=CN=C(C(=C1OC)C)C[S@](=O)C2=NC3=C(N2)C=C(C=C3)OC"
```

For each compound returned from the search, retrieve the following information. - CID - Isomeric SMILES string - chemical synonyms (for simplicity, print only the five synonyms that first occur in the name list retrieved for each compound) - Structure image

```
# Write your code here
```

2. Similarity search

PubChem supports 2-dimensional (2-D) and 3-dimensional (3-D) similarity searches. Because molecular similarity is not a measurable physical observable but a subjective concept, many approaches have been developed to evaluate it. Detailed discussion on how PubChem quantifyies molecular similarity, read the following LibreTexts page:

Searching PubChem Using a Non-Textual Query (https://bit.ly/2lPznCo (https://bit.ly/2lPznCo))

The code cell below demonstrates how to perform 2-D and 3-D similarity searches.

```
mydata <- list(smiles="C1COCC(=O)N1C2=CC=C(C=C2)N3C[C@@H](OC3=O)CNC(=O)C4=CC=C(S4)Cl")
url <- paste(prolog,
             "/compound/fastsimilarity_2d/smiles/cids/txt?Threshold=99",
             sep="")
res <- POST(url,body=mydata)
cids <- unlist(strsplit(rawToChar(res$content),"\n",fixed=T))

print(paste("# Number of CIDs:", length(cids)))
```

```
## [1] "# Number of CIDs: 29"
```

```
print(cids)
```

```
##  [1] "9875401"   "6433119"   "11524901"  "68152323"  "25190310"
##  [6] "25164166"  "123868009" "56598114"  "25255944"  "11994745"
## [11] "25190129"  "25190130"  "25190186"  "25190187"  "25190188"
## [16] "25190189"  "25190190"  "25190248"  "25190249"  "25190250"
## [21] "25190251"  "25190252"  "25190311"  "25255845"  "25255945"
## [26] "25255946"  "49849874"  "56589668"  "133687098"
```

It is worth mentioning that the parameter name “Threshold” is **case-sensitive**. If “threshold” is used (rather than “Threshold”), it will be ignored and the default value (0.90) will be used for the parameter. [As a matter of fact, all optional parameter names in PUG-REST are case-sensitive.]

```
url1 <- paste(prolog, "/compound/fastsimilarity_2d/smiles/cids/txt?Threshold=95", sep="")
url2 <- paste(prolog, "/compound/fastsimilarity_2d/smiles/cids/txt?threshold=95", sep="")
# "threshold=95" is ignored.

res1 <- POST(url1,body=mydata)
res2 <- POST(url2,body=mydata)
cids1 <- unlist(strsplit(rawToChar(res1$content),"\n",fixed=T))
cids2 <- unlist(strsplit(rawToChar(res2$content),"\n",fixed=T))

print(paste("# Number of CIDs:", length(cids1), "vs.", length(cids2)))
```

```
## [1] "# Number of CIDs: 166 vs. 766"
```

It is possible to run 3-D similarity search using PUG-REST. However, because 3-D similarity search takes much longer than 2-D similarity search, it often exceeds the 30-second time limit and returns a time-out error, especially when the query molecule is big.

In addition, for 3-D similarity search, it is **not** possible to adjust the similarity threshold (that is, the optional “Threshold” parameter does not work). 3-D similarity search uses a shape-Tanimoto (ST) of >=0.80 and a color-Tanimoto (CT) of >=0.50 as a similarity threshold. Read the libreTexts page for more details (https://bit.ly/2lPznCo (https://bit.ly/2lPznCo)).

```
mydata <- list(smiles="CC(=O)OC1=CC=CC=C1C(=O)O")
url <- paste(prolog,
             "/compound/fastsimilarity_3d/smiles/cids/txt",
             sep="")
res <- POST(url,body=mydata)
cids <- unlist(strsplit(rawToChar(res$content),"\n",fixed=T))

print(paste("# Number of CIDs:", length(cids)))
```

```
## [1] "# Number of CIDs: 21424"
```

Exercise 2a: Perform 2-D similarity search with the following query, using a threshold of 0.80 and find the macromolecule targets of the assays in which the returned compounds were tested. You will need to take these steps.

- Run 2-D similarity search using the SMILES string as a query (with Threshold=80).
- Retrieve the AIDs in which any of the returned CIDs was tested “active”.
- Retrieve the gene symbols of the targets for the returned AIDs.

```
query <- "[C@@H]23C(=O)[C@H](N)C(C)[C@H](CCC1=COC=C1)[C@@]2(C)CCCC3(C)C"
```

```
# Write your code here
```

3. Substructure/Superstructure search

When a chemical structure occurs as a part of a bigger chemical structure, the former is called a substructure and the latter is referred to as a superstructure (https://bit.ly/2lPznCo (https://bit.ly/2lPznCo)). PUG-REST supports both substructure and superstructure searches. For example, below is an example for substructure search using the core structure of antibiotic drugs called cephalosporins as a query (https://en.wikipedia.org/wiki/Cephalosporin (https://en.wikipedia.org/wiki/Cephalosporin)).

```
mydata <- list(smiles="C12(SCC(=C(N1C([C@H]2NC(=O)[*])=O)C(=O)O[H])[*])[H]")
url <- paste(prolog,
             "/compound/fastsubstructure/smiles/cids/txt?Stereo=exact",
             sep="")
res <- POST(url,body=mydata)
cids <- unlist(strsplit(rawToChar(res$content),"\n",fixed=T))

print(paste("# Number of CIDs:", length(cids)))
```

```
## [1] "# Number of CIDs: 21824"
```

An important thing to remember about substructure search is that, if the query structure is not specific enough (that is, not big enough), it will return too many hits for the PubChem server can handle. For example, if you perform substructure search using the “C-C” as a query, it will give you an error, because PubChem has ~96 million (organic) compounds with more than two carbon atoms and most of them will have the “C-C” unit. Therefore, if you get an “time-out” error while doing substructure search, consider providing more specific structure as an input query.

Exercise 3a: Below is the SMILES string for a HCV (Hepatitis C Virus) drug (Sovaldi). Perform substructure search using this SMILES string as a query, identify compounds that are mentioned in patent documents, and create a list of the patent documents that mentioning them.

- Use the default options for substructure search.
- Use the “XRefs” operation to retrieve Patent IDs associated with the returned compounds.
- For simplicity, ignore the CID-Patent ID mapping. (That is, no need to track which CID is associated with which patent document.)

```
query <- "C[C@@H](C(=O)OC(C)C)N[P@](=O)(OC[C@@H]1[C@H]([C@@]([C@@H](O1)N2C=CC(=O)NC2=O)(C)F)
O)OC3=CC=CC=C3"
```

```
# Write your code here
```

4. Molecular formula search

Strictly speaking, molecular formula search is not structure search, but its PUG-REST request URL is constructed in a similar way to structure searches like identity, similarity, and substructure/superstructure searches.

```
query <- "C22H28FN3O6S"  # Molecular formula for Crestor (Rosuvastatin: CID 446157)
url <- paste(prolog, "/compound/fastformula/",
             query, "/cids/txt", sep="")
cids <- readLines(url)

print(paste("# Number of CIDs:", length(cids)))
```

```
## [1] "# Number of CIDs: 179"
```

It is possible to allow other elements to be present in addition to those specified by the query formula, as shown in the following example.

```
query <- "C22H28FN3O6S"  # Molecular formula for Crestor (Rosuvastatin: CID 446157)
url <- paste(prolog, "/compound/fastformula/",
             query, "/cids/txt?AllowOtherElements=true", sep="")
cids <- readLines(url)

print(paste("# Number of CIDs:", length(cids)))
```

```
## [1] "# Number of CIDs: 200"
```

Exercise 4a: The general molecular formula for alcohols is $C_nH_{(2n+2)}O$ [for example, CH4O (methanol), C2H6O (ethanol), C3H8O (propanol), etc]. Run molecular formula search using this general formula for n=1 through 20 and retrieve the XLogP values of the returned compounds for each value of n. Print the minimum and maximum XLogP values for each n value.

```
# Write your code here
```