Working with PhiTools

The PhiTools approach

Manipulating collections of compounds with multiple and complex annotations, obtained from heterogeneous source can be cumbersome.

Here we proposed a simple approach based on simple idea: *"convert SDFiles in tables, manipulate as much as possible with a standard spreadsheet, then add back the structures"*

The collection of programs included in PhiTools aim to provide simple and effective tools for implementing this approach, linking back and forth tables and chemical structures, as well as providing auxiliary tools for common tasks.

A typical workflow is as follows:

[names/SMILES]->**getSDF**->[SDFile] -> **extractData** ->[Table]->spreadsheet->[Table']->**addSDFtoData**->[SDFile']

In addition, PhiTools contains tools for joining data tables (**join**) and visualizing SDFiles in a web browser (**viewSDF**)

In this document we will illustrate some PhiTools functionalities using a simple example.

We will start with two tab-separated CSV containing different data for two collections of compounds. We want to obtain a unique SDF file containing all the common compounds, combining the information from these CSV files and incororating 3D structures

Joining tables

We match elements from both tables with the same values in the field 'database_substance_id':

```
join -a fileA.csv -b fileB.csv --id 'database_substance_id' -o fileA+B.csv
```

INPUT:

fileA.csv (3 rows)

database_substance_id	compound_id	db_description	pharmacological_action
AZ_GGA_204328833	4797	eTox Lhasa 2015.1	Class III antiarrhythmic agent
AZ_GGA_203287776	4801	eTox Lhasa 2015.1	Antibiotic - Inhibitor of the beta-subunit of DNA gyrase (GyrB)
GGA_PFI_0343701	4803	eTox Lhasa 2015.1	null

fileB.csv (5 rows)

smiles	database_substance_id	molecular_weight
CCC[S@](=O)CCCN(CC)C[C@H](O)COc1ccc(cc1)C#N	AZ_GGA_204328833	352.492
CS(=O)(=O)c1ccc(cc1Cl)[C@@H](CC1CCCC1)C(=O)Nc1cnccn1	Roche_PC_RO0505082	407.9
Cc1[nH]c(C(=O)N[C@@H]2CCN(C[C@@H]2F)c2ncc(s2)C(O)=O)c(Cl)c1Cl	AZ_GGA_203287776	421.274
CN(C)C1CCN(CC1)C(=O)c1cc(Cc2n[nH]c(=O)c3CCCCc23)ccc1F	AZ_GGA_204990230	412.5
Cc1cc(CCN2CCN(CC2)c2nsc3ccccc23)cc2c1NC(=O)CC2(C)C	GGA_PFI_0343701	434.597

OUTPUT:

fileA+B.csv

smiles	database_substance_id	molecular_weight	pharmacological_action	compound_id	db_description	pharmacological_action
CCC[S@](=O)CCCN(CC)C[C@	AZ_GGA_204328833	352.492	Class III antiarrhythmic ag	4797	eTox Lhasa 2015.1	Class III antiarrhythmic agen
Cc1[nH]c(C(=O)N[C@@H]2C	AZ_GGA_203287776	421.274	Antibiotic - Inhibitor of the	4801	eTox Lhasa 2015.1	Antibiotic - Inhibitor of the be
Cc1cc(CCN2CCN(CC2)c2nsc	GGA_PFI_0343701	434.597	null	4803	eTox Lhasa 2015.1	null

Now, a new file named **fileA+B.csv** is present in the current folder.

Obtaining a SDFile

We can generate an SDFile from the SMILES included in the table. If the CSV has a header line you must include '--header'. The field with the SMILES must be present in the first column. If this is not the case, you can reorder the columns easily using a spreadsheet program or shell commands (cut, paste, awk, etc...):

```
getSDF -s fileA+B.csv --id=database_substance_id -o output.sdf --header
```

The file **output.sdf** contains only the structures of the compounds, but not the rest of the annotations. These can be easily added using

```
addDataToSDF -f output.sdf -d fileA+B.csv --id=database_substance_id -o final.sdf
```

InChI and InChiKey, computed from the included 2D structure can be also added. This only requires to type

```
addInchi -f final.sdf  -o final_Inchi.sdf
```

Extract information

Once the annotated SDFile is ready to use we can do different extraction of the data. This can be done using the extractData command, selecting the field to show:

```
extractData -f final_Inchi.sdf  --field=inchi
```

OUTPUT:

```
InChI=1S/C18H28N2O3S/c1-3-11-24(22)12-5-10-20(4-2)14-17(21)15-23-18-8-6-16(13-19)7-9-18/h6-9,17,21H,3-5,10-12,14-15H2,1-2H3/t17-,24-/m0/s1
InChI=1S/C15H15Cl2FN4O3S/c1-6-10(16)11(17)12(20-6)13(23)21-8-2-3-22(5-7(8)18)15-19-4-9(26-15)14(24)25/h4,7-8,20H,2-3,5H2,1H3,(H,21,23)(H,24,25)/t7-,8+/m0/s1
InChI=1S/C25H30N4OS/c1-17-14-18(15-20-23(17)26-22(30)16-25(20,2)3)8-9-28-10-12-29(13-11-28)24-19-6-4-5-7-21(19)31-27-24/h4-7,14-15H,8-13,16H2,1-3H3,(H,26,30)
```

To print all the information in a tabular (tab separated) format, we just type

```
extractData -f final_Inchi.sdf --table
```

OUTPUT:

```
cas_number     cdk_title      common_name    compound_id    database_substance_id db_description db_name
        db_version     inchi   m       molecular_formula     molecular_weight      pharmacological_action
        query_text     smiles  software_version      source  subst_id       substance_status
        vitic_legacy_recno
null            null    4797    AZ_GGA_204328833      eTox Lhasa 2015.1     ETOX_LHASA     2015.1.0
        InChI=1S/C18H28N2O3S/c1-3-11-24(22)12-5-10-20(4-2)14-17(21)15-23-18-8-6-16(13-19)7-9-18/h6-9,17,21H,3-5,10-12,14-15H2,1-2H3/t17-,24-/m0/s1     CCC[S@](=O)CCCN(CC)C[C@H](O)COc1ccc(C#N)cc1  null
        352.492 Class III antiarrhythmic agent       null    CCC[S@](=O)CCCN(CC)C[C@H](O)COc1ccc(cc1)C#N
        2.5.0   null    AZ_GGA_204328833      eTox non-confidential null
null            null    4801    AZ_GGA_203287776      eTox Lhasa 2015.1     ETOX_LHASA     2015.1.0
        InChI=1S/C15H15Cl2FN4O3S/c1-6-10(16)11(17)12(20-6)13(23)21-8-2-3-22(5-7(8)18)15-19-4-9(26-15)14(24)25/h4,7-8,20H,2-3,5H2,1H3,(H,21,23)(H,24,25)/t7-,8+/m0/s1
```

…

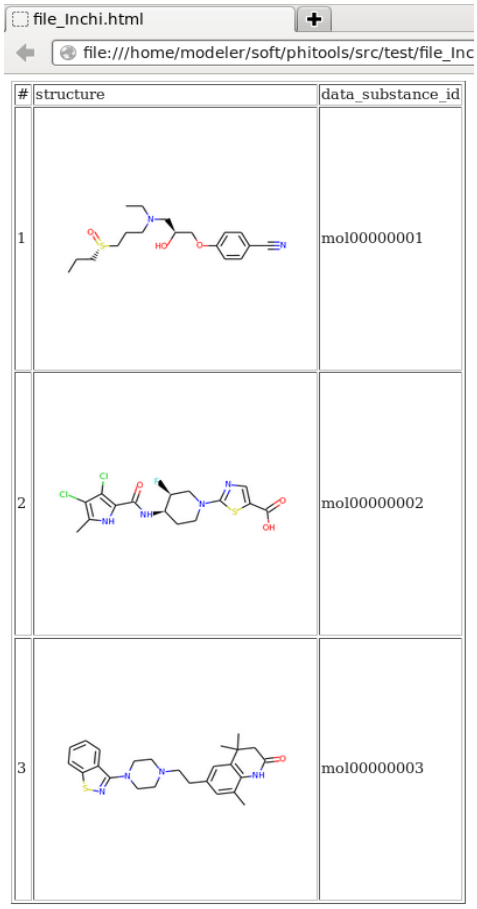
This output can be redirected to a new file using the ">" operator

```
extractData -f final_Inchi.sdf --table > extraction.csv
```

Visualizing SDFiles

The molecular structures inside a SDFile can be visualized as a table using a common browser. This only requires to type

```
viewSDF -f final_Inchi.sdf -o output_Inchi.html
```



To open the output file output_Inchi.html on a web browser, you can click on the generated file or type

```
firefox final_Inchi.html
```