Tutorial of Simple QSAR calculation by MolAICal

Qifeng Bai
Email: molaical@yeah.net
Homepage: https://molaical.github.io
Lanzhou University
Lanzhou, Gansu 730000, P. R. China

1. Introduction

The quantitative structure-activity relationship (QSAR) models are regression or classification models used in drug design. In this tutorial, the simple regression model of QSAR is introduced based on the ligands of signal transducer and activator of transcription 3 (STAT3) protein which is considered as a potential drug target of cancer. Here, the MolAICal soft package (https://molaical.github.io) is employed for this tutorial.

2. Materials

2.1. Software requirement

1) MolAICal: https://molaical.github.io
2) Notepad++: https://notepad-plus-plus.org

Note: Chinese users can download Notepad++ from searching Baidu if they cannot access the official site of Notepad++

2.2. Example files

1) All the necessary tutorial files are downloaded from:
https://github.com/MolAICal/tutorials/tree/master/006-QSAR

3. Procedure

MolAICal supplies two modules for molecular descriptor calculation based on PaDEL-Descriptor [1] and Mordred [2]. PaDEL-Descriptor is free for all (e.g. personal, academic, non-profit, non-commercial, government, commercial, etc) to use. Mordred (Copyright (c) 2015-2017, Hirotomo Moriwaki) uses the BSD 3-Clause "New" or "Revised" License (see: https://github.com/mordred-descriptor/mordred/blob/develop/LICENSE).

3.1. Calculate molecular descriptor

Go to 006-QSAR/mordred

Option 1: calculating molecular descriptor in Mordred by MolAICal. The command is as below:

```
#> molaical.exe -tool mordred -i example.smi
```

Note: “example.smi” is a file that contains molecular SMILES strings.

It will generate two files named “with3D-descriptors.csv” and “without3D-descriptors.csv”. “with3D-descriptors.csv” contains 2D and 3D molecular descriptor, while “without3D-descriptors.csv” contains 2D molecular descriptor without 3D molecular descriptor.

Go to 006-QSAR/PaDEL

Option 2: calculate molecular descriptor in PaDEL by MolAICal

```
#> molaical.exe -tool padel -f sdf -i sdf
```

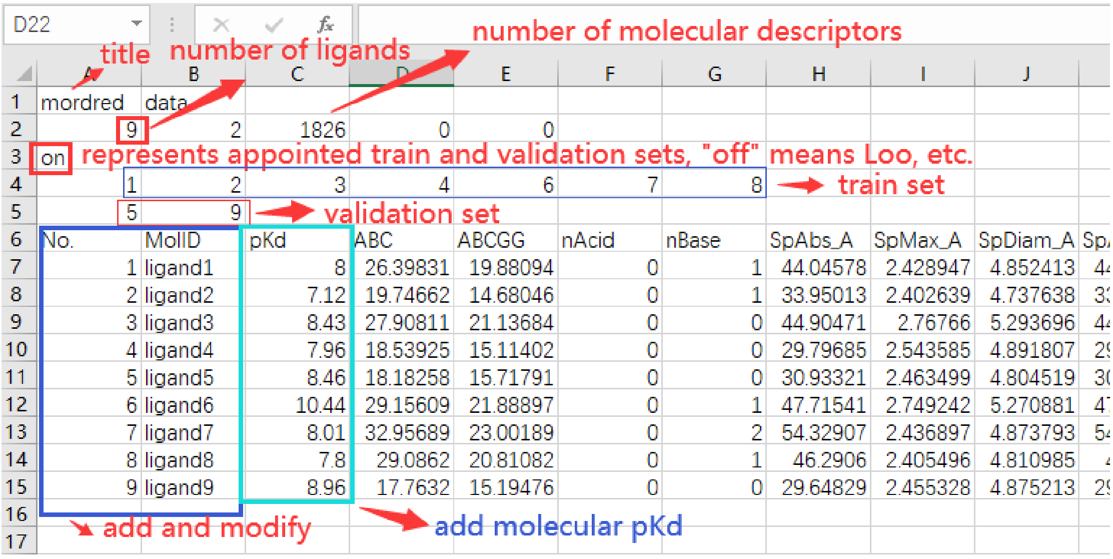
It will produce two files named “2DDescriptor_mdl.csv” and “3DDescriptor_mdl.csv” which contain 2D and 3D molecular descriptors, respectively.

Warning: “sdf” is a folder that contains SDF format molecular files. For **PaDEL** descriptor calculation, it should **run on the local machine,** if the job is run on the remote machine, it may be no results because it needs to start X11 window server that cannot be started by default in the remote machine. Besides, PaDEL occupies much memory when calculating descriptors. So users can calculate the suitable number of molecules at one time, and then, merge all the results. For example, 50 molecules are calculated at one time. For more commands about PaDEL-Descriptor, please see MolAICal manual.

3.2. Prepare files for QSAR

Here, “3DDescriptor_mdl.csv” which is produced by PaDEL-Descriptor is employed for this tutorial.

1) Open “3DDescriptor_mdl.csv” with Excel and set parameters as shown in Figure 1.


D22
title
number of ligands
number of molecular descriptors
represents appointed train and validation sets, "off" means Loo, etc.
train set
validation set
add and modify
add molecular pKd

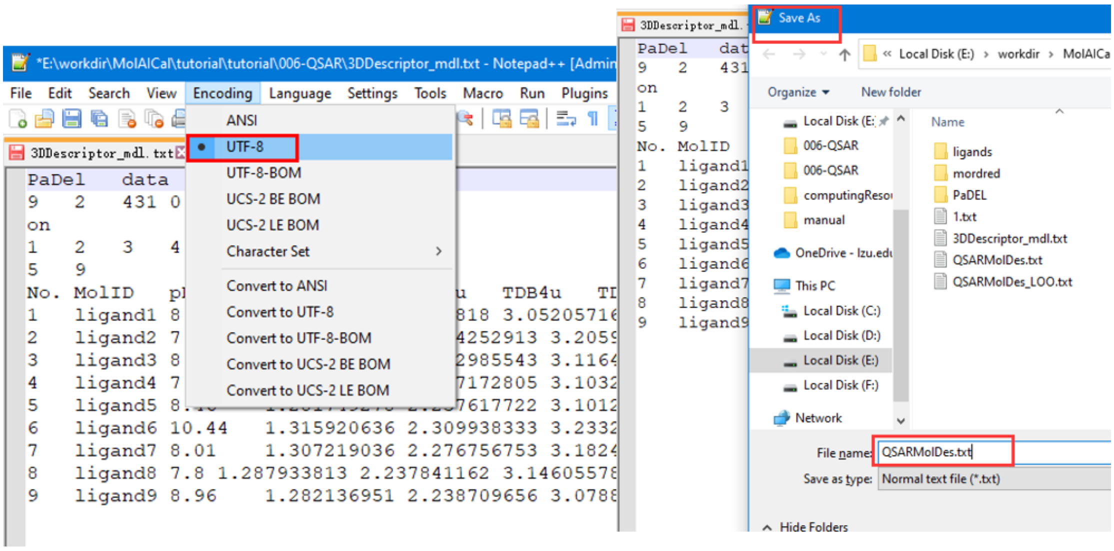
	A	B	C	D	E	F	G	H	I	J
1	mordred	data								
2	9	2	1826	0	0					
3	on									
4	1	2	3	4	6	7	8			
5	5	9								
6	No.	MolID	pKd	ABC	ABCGG	nAcid	nBase	SpAbs_A	SpMax_A	SpDiam_A
7	1	ligand1	8	26.39831	19.88094	0	1	44.04578	2.428947	4.852413
8	2	ligand2	7.12	19.74662	14.68046	0	1	33.95013	2.402639	4.737638
9	3	ligand3	8.43	27.90811	21.13684	0	0	44.90471	2.76766	5.293696
10	4	ligand4	7.96	18.53925	15.11402	0	0	29.79685	2.543585	4.891807
11	5	ligand5	8.46	18.18258	15.71791	0	0	30.93321	2.463499	4.804519
12	6	ligand6	10.44	29.15609	21.88897	0	1	47.71541	2.749242	5.270881
13	7	ligand7	8.01	32.95689	23.00189	0	2	54.32907	2.436897	4.873793
14	8	ligand8	7.8	29.0862	20.81082	0	1	46.2906	2.405496	4.810985
15	9	ligand9	8.96	17.7632	15.19476	0	0	29.64829	2.455328	4.875213
16										
17										

Figure 1. Set parameter for QSAR. Here, “title” and “number of molecular descriptor” in tutorial material are “PaDel data” and 431. This is deliberately to tell the users that these values can be changed.

Users must set the parameters in “3DDescriptor_mdl.csv” strictly. Users can use any title or default title in the first line. The first number in the second line must be the number of ligands for QSAR. The third number in the second line must be the number of molecular descriptors. The other numbers in the second line can be arbitrary numbers. The character “on” in the third line means train and validation sets are appointed. The numbers in the fourth line are the sequence numbers of train set from the below No. of ligands. The numbers in the fifth line are the sequence numbers of validation set from the below No. of ligands (see Figure 1). If “off” is chosen, it means leave-one-out (LOO) cross-validation is used for QSAR calculation, in this case, the fourth and fifth lines for train and validation sets can be omitted (See file “QSARMolDes_LOO.txt”). In addition, “No.” should be added in the first column, and the first number should be increased from 1 rather than 0. “MolID” is the ligand names which should not contain any space. Users can modify ligands names in the light of user requirements. Here, this tutorial uses modified ligand names. The experimental values such as pKd should be added in the following column (see Figure 1).

Warning: PaDEL-Descriptor and Mordred may generate characters rather than numbers in some items of molecular descriptor. Please delete the molecular descriptor values that contain characters.

2) Save “3DDescriptor_mdl.csv” to a file named “3DDescriptor_mdl.txt” by Excel. But this file is not in UTF-8 format. To convert “3DDescriptor_mdl.txt” to UTF-8 format for studying, Notepad++ is used for format conversion. Open “3DDescriptor_mdl.txt” by Notepad++. Select Encoding→UTF-8, and save as this file named “QSARMolDes.txt” (see Figure 2).



```
PaDel   data
9   2   431 0
on
1   2   3   4
5   9
No. MolID   pI                              u   TDB4u   TI
1   ligand1 8                               818 3.05205716
2   ligand2 7                               4252913 3.2059
3   ligand3 8                               2985543 3.1164
4   ligand4 7                               7172805 3.1032
5   ligand5 8.                              7617722 3.1012
6   ligand6 10.44   1.315920636 2.309938333 3.2332
7   ligand7 8.01    1.307219036 2.276756753 3.1824
8   ligand8 7.8 1.287933813 2.237841162 3.14605578
9   ligand9 8.96    1.282136951 2.238709656 3.0788
```

Figure 2. Saving file in UTF-8 format

Notice: Sometimes, Excel cannot save a file in UTF-8 format. So Notepad++ is used for UTF-8 format conversion. If users’ Excel can change the file in UTF-8 format, Notepad++ can be omitted.

3.3. QSAR calculations

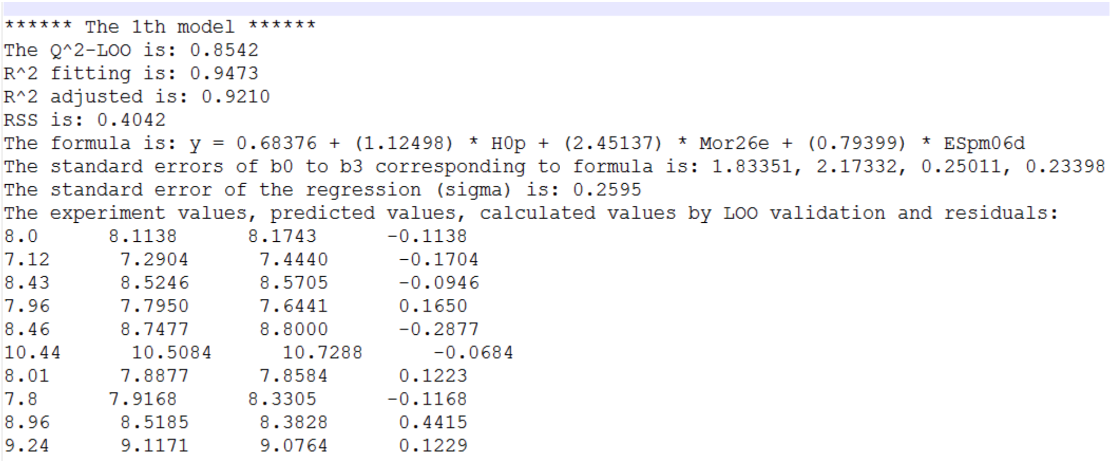
Running command as below:

```
#> molaical.exe -qsar GA -i QSARMolDes.txt
```

Or

```
#> molaical.exe -qsar GA -i QSARMolDes_LOO.txt
```

If you want to know more parameters for QSAR, please check the manual of MolAICal. This tutorial just contains 9 ligands. When Q2 is enough for your research, you can stop the QSAR running by keyboard shortcut “Ctrl + C”. The results are stored in the file “QSAROutFile.dat”. Open “QSAROutFile.dat” and the information is as below:



```
****** The 1th model ******
The Q^2-LOO is: 0.8542
R^2 fitting is: 0.9473
R^2 adjusted is: 0.9210
RSS is: 0.4042
The formula is: y = 0.68376 + (1.12498) * H0p + (2.45137) * Mor26e + (0.79399) * ESpm06d
The standard errors of b0 to b3 corresponding to formula is: 1.83351, 2.17332, 0.25011, 0.23398
The standard error of the regression (sigma) is: 0.2595
The experiment values, predicted values, calculated values by LOO validation and residuals:
8.0      8.1138      8.1743      -0.1138
7.12      7.2904      7.4440      -0.1704
8.43      8.5246      8.5705      -0.0946
7.96      7.7950      7.6441      0.1650
8.46      8.7477      8.8000      -0.2877
10.44      10.5084      10.7288      -0.0684
8.01      7.8877      7.8584      0.1223
7.8      7.9168      8.3305      -0.1168
8.96      8.5185      8.3828      0.4415
9.24      9.1171      9.0764      0.1229
```

Notice: if users want to explain the molecular descriptors, they can refer to the document in https://github.com/MolAICal/documents/tree/master/manual/descriptors-instructions

References

1. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466-74. Epub 2011/03/23. doi: 10.1002/jcc.21707. PubMed PMID: 21425294.
2. Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. J Cheminform. 2018;10(1):4. Epub 2018/02/08. doi: 10.1186/s13321-018-0258-y. PubMed PMID: 29411163; PubMed Central PMCID: PMCPMC5801138.