

使用 MolAICal 进行药物的 QSAR 计算

作者：Qifeng Bai (update 2021-10-16)

更多教程（含英文教程）请见如下：

MolAICal 官方主页：<https://molaical.github.io>

MolAICal 官方主页中国镜像：<https://molaical.gitee.io>

MolAICal 中文博客：<https://molaical.gitee.io/cntutorial.html>

1. 简介

药物的定量构象关系(QSAR)包含线性回归和分类，在本教程中选用 STAT3 蛋白靶点的药物分子作为研究对象；STAT3 是治疗癌症的一个重要蛋白靶点，研究 STAT3 药物的属性，有助于设计合理的抗癌药物。

2. 工具

2.1. 所需软件

1) MolAICal: <https://molaical.github.io>

国内镜像 MolAICal: <https://molaical.gitee.io>

2) Notepad++: <https://notepad-plus-plus.org>

说明：假如登陆不了 Notepad++ 的官方网址，可以使用百度直接搜索下载，Notepad++ 是一款免费的工具。

2.2. 操作所需的示例文件

1) 本教程所需的教程文件可以从以下网址下载：

<https://gitee.com/molaical/tutorials/tree/master/006-QSAR>

3. 步骤

MolAICal 提供了 2 个免费的分子描述符计算模块：PaDEL-Descriptor [1] 和 Mordred [2]。PaDEL-Descriptor 的许可是自由免费的；而 Mordred (Copyright (c) 2015-2017, Hiroto Moriawaki) 使用 BSD 3-Clause "New" 或 "Revised" 许可 (见: <https://github.com/mordred-descriptor/mordred/blob/develop/LICENSE>)。

3.1. 计算分子描述符

[切换到 006-QSAR/mordred](#)

选择 1: 使用 MolAICal 调用 Mordred 模块计算分子描述符，计算命令如下：

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

说明: “example.smi” 是包含分子 SMILES 字符串的文件。

运行命令之后，会生成两个文件，分别是“with3D-descriptors.csv”和“without3D-descriptors.csv”。其中“with3D-descriptors.csv”包含 2D 和 3D 的分子描述符，而“without3D-descriptors.csv”包含 2D 分子描述符但不包括 3D 分子描述符。

[切换到 006-QSAR/PaDEL](#)

选项 2: 使用 MolAICal 调用 PaDEL 模块计算分子描述符，计算命令如下：

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

这个命令将生成 2 个文件，分别是“2DDescriptor_md1.csv”和“3DDescriptor_md1.csv”。其中“2DDescriptor_md1.csv”包含 2D 分子描述符，而“3DDescriptor_md1.csv”包含 3D 分子描述符。

警告: “sdf” 是一个文件夹，里面放着 SDF 格式的文件。对于 **PaDEL** 分子描述的计算，必须在本地计算机上进行计算，目前远程机器调用不了 X11 window server，使用远程机器算 **PaDEL** 分子描述会报错。除此之外，**PaDEL** 计算分子描述符的时候，特别耗费内存，用户可以选取少量的分子（如 50 个分子）进行描述符的计算，然后在合并结果。更多详细命令的解释，请参考 MolAICal 的手册。

选项 3: 如果想尝试有限数量的描述符（例如，不超过 200 个描述符）或尝试 RDKit 中的描述符：

[切换到 006-QSAR/rdkit](#)

1) 输入 SMILES 文件并输出不含 3D 描述符的 RDKit 描述符 CSV 文件

```
#> molaical.exe -tool rdkit -i molecules.smi -o descriptors_no3D.csv
```

2) 输入 SMILES 文件并输出包含 3D 描述符的 RDKit 描述符 CSV 文件

```
#> molaical.exe -tool rdkit -i molecules.smi -o descriptors_3D.csv -n true
```

注: MolAICal 可从单个 SMILES 字符串输出描述符。更详细的信息请参考 MolAICal 手册。对于 RDKit 描述符的详细解释，请查阅谷歌搜索、RDKit 官方文档及源代码。例如，用户可在谷歌搜索“VSA_EState”描述符，或在 Linux 控制台中查阅 RDKit 的开源代码（如“rdkit-master”）：

```
#> grep -rn "VSA_EState"
```

随后可定位描述符“VSA_EState”对应的代码和文档。

3.2. 准备 QSAR 计算的文件

在本教程, 使用“3DDescriptor_md1.csv”文件进行计算。

1) 使用 Excel 打开“3DDescriptor_md1.csv”, 并且像图 1 一样设置参数:

	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										

图 1. 设置 QSAR 的参数。在本次教程中“title”和“number of molecular descriptor”分别是“PaDel data”和 431。图 1 是故意设置让用户知道这一块需要修改。

你必须在“3DDescriptor_md1.csv”中严格按照格式设置参数。第一行可以使用默认标题或者也可以使用你设置的任意标题。在第二行的第一个数字是用于 QSAR 计算的配体分子数, 第二行的第三个数字代表分子描述符的数量。第二行的其余数字可以使用默认数字或者其他任意数字, 这对 QSAR 的计算没有影响。在第三行上的字符“on”代表指定了训练集和验证集, 第四行是训练集的序号, 第五行是验证集的序号, 此序号对应文件“QSARMolDes.txt”底下配体的序号 (如图 1 所示)。如果第三行是“off”, 则使用留一验证法 (LOO) 进行 QSAR 的计算, 在这种情况下, 第四、五行的数字可以省略, MolAICal 自动使用留一法指定训练集与验证集进行运算 (请参考示例文件: “QSARMolDes_LOO.txt”)。除此之外, 序号“No.”要加到第一列, 序号应该从 1 开始而不是 0; “MolID”部分是分子的名称, 用户可以根据具体情况更改分子的名称, 分子名称不能有空格; 实验值如 pKd 等应该加到第三列中 (如图 1 所示)。

警告: PaDEL-Descriptor 和 Mordred 可能会在分子描述计算过程中生成字符而不是数字, 在这种情况下, 需要删掉这些包括字符的分子描述符, 不然会报无法识别的错误。

2) 通过 Excel 将“3DDescriptor_md1.csv”保存成“3DDescriptor_md1.txt”。但是这个文件的格式不是 UTF-8 的格式。因此, 需要将“3DDescriptor_md1.txt”转化成 UTF-8 的格式, Notepad++ 可以进行格式的转化。通过 Notepad++ 打开“3DDescriptor_md1.txt”。选择工具栏中的 Encoding→UTF-8, 最后保存成“QSARMolDes.txt” (见图 2)。

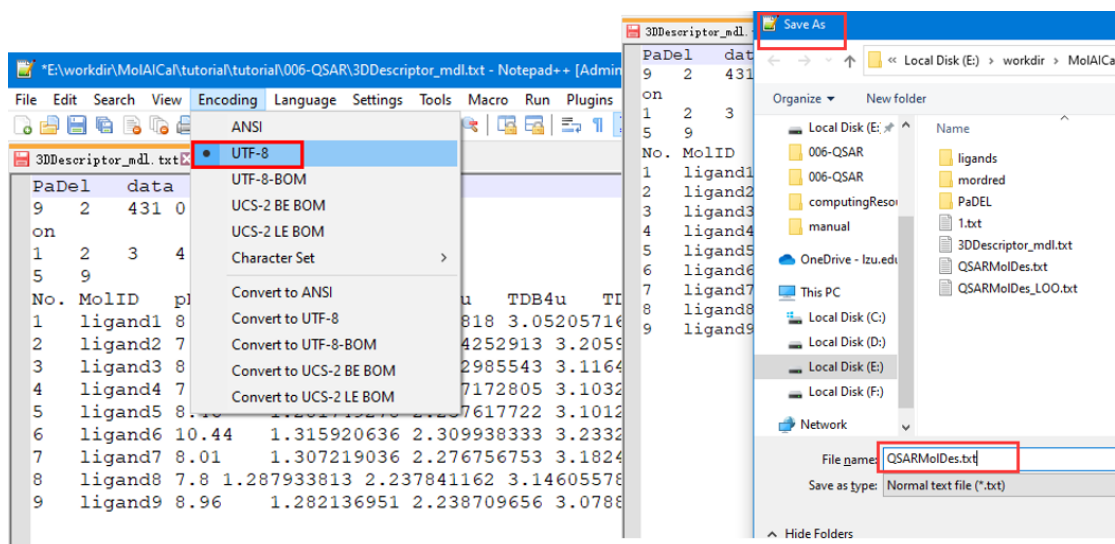


图 2. 将文件保存成 UTF-8 的格式

注意: 有时候, Excel 并不能把文件保存成 UTF-8 的格式。所以, 使用 Notepad++ 进行 UTF-8 的格式转化。假如用户的 Excel 可以将文件转化成 UTF-8 的格式, 则可以不使用 Notepad++。

3.3. QSAR 计算

运行如下命令:

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

或

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

假如你了解更多的 QSAR 参数, 请参考 MolAICal 的说明书。本教程仅仅包括 9 个配体。当 Q² 的运算值已经满足你的研究目的, 你可以通过“Ctrl + C”快捷键终止 MolAICal 的运行。最后的结果保存在“QSAROutFile.dat”文件中, 打开“QSAROutFile.dat”, 其具体运算结果的信息如下:

```
***** 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
```

注意：假如用户想解释分子描述符的物理化学意义等，可以访问以下链接，参考相关文档：
<https://gitee.com/molaical/documents/tree/master/manual/descriptors-instructions>

思考：若 Y 值对应于二分类或多标签分类的输出值，那么通过适当的数据处理，本教程可应用于二分类或多标签分类的训练任务。例如，pKd 值可替换为 0 和 1。训练完成之后，预测值大于 0.5 的被指定为标签 1，而预测值小于或等于 0.5 的被指定为标签 0。

参考文献：

1. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466-74.
2. Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. J Cheminform. 2018;10(1):4.