

Results of Molecular Docking

Huang Lichuang in Wie-Biotech

Contents

1 设计流程	1
1.1 已有数据	1
1.2 分子对接	2
2 分析和结果	2
2.1 中药和成分和靶点数据的预处理	2
2.1.1 整理来自 HERB 网站的数据	2
2.1.2 根据 BindingDB 数据库筛选药物和靶点的结合可能	2
2.2 获取药物分子结构数据和预处理	2
2.2.1 根据 BindingDB 提供的结合可能筛选药物分子	2
2.2.2 获取化合物分子的 SDF 数据	4
2.2.3 预处理化合物的结构数据	4
2.3 获取靶点蛋白数据	5
2.3.1 获取 MCC top 10 蛋白的 PDB ID	5
2.3.2 根据 BindingDB 数据提供的结合可能筛选靶点蛋白结构	6
2.3.3 获取靶点蛋白的 PDB 文件	6
2.3.4 根据种族过滤靶点蛋白	7
2.3.5 预处理靶点蛋白 PDB 文件	7
2.4 使用 AutoDock Vina 分子对接	7
2.4.1 所有分子和靶点蛋白结合的可能性	7
2.4.2 使用 AutoDock Vina 分子对接	7
Reference	9

1 设计流程

1.1 已有数据

已筛选的 Hubgenes; 药方的中药和相应成分, 以及对应靶点。

1.2 分子对接

步骤:

- 从已有数据得到成分和靶点数据, 根据筛选的 Hubgenes 预处理, 得到可能的药物分子和靶点的结合。
- 从 Uniprot 网站 (<https://www.uniprot.org/>) 或 RCSB PDB (<https://www.rcsb.org/>) 获取靶点蛋白的结构。
- 从 PubChem <https://pubchem.ncbi.nlm.nih.gov/> 获取药物分子结构。
- 以 AutoDock Vina 1.2.0¹ (<http://vina.scripps.edu>), 在 Python 下使用, 批量对接多个分子和靶点。
- 将结果可视化或输出成表格。

2 分析和结果

以下分析和上述设计思路可能有所不同。

除了 AutoDock Vina (<http://vina.scripps.edu>), 还会用到 ADFR 工具组 (<https://ccsb.scripps.edu/adfr/>), Python 的 Meeko, 或其它工具。

流程请参考文献² 或者 https://autodock-vina.readthedocs.io/en/latest/docking_basic.html。

2.1 中药和成分和靶点数据的预处理

2.1.1 整理来自 HERB 网站的数据

从网站 <http://herb.ac.cn/Download/> 获取成分信息。

药方的化合物成分信息概览, 共有 1097 个唯一成分: (对应文件为 `data.cpd`s)

2.1.2 根据 BindingDB 数据库筛选药物和靶点的结合可能

BindingDB 数据库记录了化合物和成分的亲和信息: <https://www.bindingdb.org/rwd/bind/chemsearch/marvin/Download.jsp>。

下载相关数据, 以供后续筛选化合物和靶点。

2.2 获取药物分子结构数据和预处理

2.2.1 根据 BindingDB 提供的结合可能筛选药物分子

PubChem CID 连接到 PubChem 数据库³。

根据整理的药方成分的 PubChem CID 信息和 BindingDB 数据的 PubChem CID 信息过滤化合物。

注意: 药方统计有 1097 个唯一成分, 而包含 PubChem ID 的化合物共有 7 个。

以下为包含 PubChem ID 的化合物的概览:

```
## # A tibble: 796 x 3
##   Ingredient.id Ingredient.name PubChem_id
##   <chr>         <chr>         <int>
## 1 HBIN000177    10-o-acetylgeniposide    6324916
```

```
## 2 HBIN000280 11,13-Eicosadienoic acid, methyl ester 5365674
## 3 HBIN000328 1,1,6-trimethyl-2H-naphthalene 121677
## 4 HBIN000391 11-deoxyglycyrrhetic acid 12305517
## 5 HBIN000392 11-deoxyglycyrrhetinic acid 12305517
## 6 HBIN000573 1,2,3,4,6-pentagalloylglucose 65238
## 7 HBIN000643 1,2,4-benzenetriol 10787
## 8 HBIN001070 1,3,5-trihydroxyxanthone 5281663
## 9 HBIN001080 13657-68-6 14106072
## 10 HBIN001307 13-Tetradecenyl acetate 521718
## # i 786 more rows
```

这些数据结合 BindingDB 数据进一步过滤。

以下为 BindingDB 中记录有上述化合物的条目,包含了靶点蛋白的信息:(对应文件为 BindingDB_data_filter_by_herb_com

```
## # A tibble: 3,909 x 2
##   PubChem_id pdb_id
##   <int> <chr>
## 1 689043 "1IVO,1M14,1M17,1MOX,1XKK,1Z9I,2GS2,2GS6,2ITW,2ITX,2ITY,2J5E,2J5F,2J6M,2N5S,2RF9,2RGP,
## 2 5280443 "1G3N,1JOW,1XO2,2EUF,2F2C,3NUP,3NUX,4AUA,4EZ5,4TTH,5L2I,5L2S,5L2T,6OQL,6OQ0"
## 3 5280343 "1G3N,1JOW,1XO2,2EUF,2F2C,3NUP,3NUX,4AUA,4EZ5,4TTH,5L2I,5L2S,5L2T,6OQL,6OQ0"
## 4 5281607 "1G3N,1JOW,1XO2,2EUF,2F2C,3NUP,3NUX,4AUA,4EZ5,4TTH,5L2I,5L2S,5L2T,6OQL,6OQ0"
## 5 5280863 "1G3N,1JOW,1XO2,2EUF,2F2C,3NUP,3NUX,4AUA,4EZ5,4TTH,5L2I,5L2S,5L2T,6OQL,6OQ0"
## 6 5280443 "1UNG,1UNH,1UNL,3OOG,7VDP,7VDQ,7VDR,7VDS"
## 7 5280343 "1UNG,1UNH,1UNL,3OOG,7VDP,7VDQ,7VDR,7VDS"
## 8 5281607 "1UNG,1UNH,1UNL,3OOG,7VDP,7VDQ,7VDR,7VDS"
## 9 5280863 "1UNG,1UNH,1UNL,3OOG,7VDP,7VDQ,7VDR,7VDS"
## 10 5280443 ""
## # i 3,899 more rows
```

将上述数据结合 MCC top 10 靶点蛋白的注释数据(包含 PDB ID)进一步过滤。

(靶点蛋白的 PDB ID 的获取,参考 2.3.1)

现在,已有化合物(PubChem CID)和对应的靶点蛋白(PDB ID)的信息:

```
## $`370`
## [1] "1CXW" "1GEN" "1RTG"
##
## $`370`
## [1] "1C5G" "1DVN" "1LJ5" "4AQH"
##
## $`73111`
## [1] "1C5G" "1DVN" "1LJ5" "4AQH"
##
## $`5280704`
```

```
## [1] "2AZ5" "3ALQ" "3IT8" "3L9J" "4G3Y" "4TWT" "5M2I" "5M2J" "5M2M" "5MU8" "5WUX" "5YOY" "600Y" "600Z"
## [19] "6X86" "7JRA"
##
## $`66065`
## [1] "1IL6" "1P9M" "2IL6" "4CNI" "4J4L" "4NI7" "4NI9" "4O9H" "4ZS7"
##
## $`66065`
## [1] "2AZ5" "3ALQ" "3IT8" "3L9J" "4G3Y" "4TWT" "5M2I" "5M2J" "5M2M" "5MU8" "5WUX" "5YOY" "600Y" "600Z"
## [19] "6X86" "7JRA"
##
## $`6476139`
## [1] "1CXW" "1GEN" "1RTG"
##
## $`5280343`
## [1] "1CXW" "1GEN" "1RTG"
##
## $`5280343`
## [1] "1L6J" "20VX" "20VZ" "20W0" "20W1" "20W2" "4H1Q" "4H82" "4HMA" "4WZV" "5TH6" "5TH9" "6ESM"
##
## $`689043`
## [1] "1L6J" "20VX" "20VZ" "20W0" "20W1" "20W2" "4H1Q" "4H82" "4HMA" "4WZV" "5TH6" "5TH9" "6ESM"
##
## $`689043`
## [1] "1CXW" "1GEN" "1RTG"
##
## $`689043`
## [1] "1L6J" "20VX" "20VZ" "20W0" "20W1" "20W2" "4H1Q" "4H82" "4HMA" "4WZV" "5TH6" "5TH9" "6ESM"
##
## $`5280343`
## [1] "1L6J" "20VX" "20VZ" "20W0" "20W1" "20W2" "4H1Q" "4H82" "4HMA" "4WZV" "5TH6" "5TH9" "6ESM"
```

2.2.2 获取化合物分子的 SDF 数据

通过 PubChem ID 使用 PubChem API 获取官方.sdf 文件 (<https://pubchem.ncbi.nlm.nih.gov/docs/pug-rest/>)⁴。

所有.sdf 文件被整合。(对应文件为 `./SDFs/all_compounds.sdf`)

2.2.3 预处理化合物的结构数据

使用 Python 的 Meeko 转化.sdf 数据为.pdbqt。

共有 7 个化合物被成功转化。(对应文件为 `./pdbqt`)

2.3 获取靶点蛋白数据

2.3.1 获取 MCC top 10 蛋白的 PDB ID

使用 R BiomaRt 获取 MCC 筛选的 Top 10 蛋白的 PDB ID。

结果如下：(对应文件为 MCC_tops_PDB_ID.csv)

```
## # A tibble: 215 x 2
##   hgnc_symbol pdb
##   <chr>      <chr>
## 1 TNF        4y6o
## 2 TNF        3l9j
## 3 TNF        1a8m
## 4 TNF        1tnf
## 5 TNF        2az5
## 6 TNF        2e7a
## 7 TNF        2tun
## 8 TNF        2zjc
## 9 TNF        2zpx
## 10 TNF       3alq
## # i 205 more rows
```

注意，一个蛋白对应多种结构，对应有多个 PDB ID：

```
## Data of 10
##   +++ Protein  1 +++
##   IL6
##   Sum: 12
##   pdb: 1p9m, 5fuc, 1alu, 1il6, ...
##
##   +++ Protein  2 +++
##   IL1B
##   Sum: 59
##   pdb: 3o4o, 4dep, 1hib, 1ilb, ...
##
##   +++ Protein  3 +++
##   TNF
##   Sum: 39
##   pdb: 4y6o, 3l9j, 1a8m, 1tnf, ...
##
##   +++ Protein  4 +++
##   MMP9
##   Sum: 25
##   pdb: 1gkc, 1gkd, 1itv, 1l6j, ...
```

```

##
##   +++ Protein    5 +++
##   CXCL8
##       Sum: 18
##       pdb: 1ilp, 1ilq, 6xmn, 6lfm, ...
##
##   +++ Protein    6 +++
##   TGFB1
##       Sum:
##       No data.
##   +++ Protein    7 +++
##   MMP2
##       Sum: 11
##       pdb: 3ayu, 1ck7, 1cxw, 1eak, ...
##
##   +++ Protein    8 +++
##   IL10
##       Sum: 8
##       pdb: 1j7v, 1y6k, 6x93, 1ilk, ...
##
##   +++ Protein    9 +++
##   ICAM1
##       Sum: 14
##       pdb: 1mq8, 3tcx, 1d3e, 1d3i, ...
##
##   +++ Protein   10 +++
##   SERPINE1
##       Sum: 29
##       pdb: 5brr, 5z1z, 3pb1, 1oc0, ...

```

2.3.2 根据 BindingDB 数据提供的结合可能筛选靶点蛋白结构

由于同一个蛋白对应多个名称，根据 BindingDB 提供的结合可能筛选，减少计算量。

此步骤已经在 @ref{par} 中同步实现。

2.3.3 获取靶点蛋白的 PDB 文件

使用 RCSB PDB 提供的 API 获取.pdb 文件。共有 49 个。<https://www.rcsb.org/docs/programmatic-access/batch-downloads-with-shell-script>

(对应文件为 ./protein_pdb)

2.3.4 根据种族过滤靶点蛋白

PDB 文件中记录有种族信息，根据种族（人种）过滤靶点蛋白（Regex match: “ORGANISM_SCIENTIFIC: HOMO SAPIENS;”）。

过滤前有 215 个文件，过滤后有 49 个文件。

2.3.5 预处理靶点蛋白 PDB 文件

使用 ADFR 工具给受体蛋白加氢并转化为.pdbqt 文件。成功获取 48 个文件。

(对应文件为 ./protein_pdbqt)

2.4 使用 AutoDock Vina 分子对接

2.4.1 所有分子和靶点蛋白结合的可能性

结合 2.2.1 得到的对应关系以及最终获得的化合物和靶点蛋白的.pdbqt 文件，共有以下 121 结合可能：

```
## # A tibble: 121 x 2
##   Ligand Receptor
##   <chr> <chr>
## 1 370 1cxw
## 2 370 1gen
## 3 370 1rtg
## 4 370 1c5g
## 5 370 1dvn
## 6 370 1lj5
## 7 370 4aqh
## 8 73111 1c5g
## 9 73111 1dvn
## 10 73111 1lj5
## # i 111 more rows
```

2.4.2 使用 AutoDock Vina 分子对接

该步骤包括使用 ADFR 工具计算 affinity maps <https://ccsb.scripps.edu/adfr/>。

尽管已经通过多种方式筛选了化合物和蛋白的结合，依然有 121 种可能性。

vina 的一次计算时间约 0.5 分钟到 60 分钟不等；由于无法在短时间内全部计算完，此处设定了计算时间限制（120 秒），超出时间限制将被强制取消。

所有可能都被计算，中途可能强制取消。结果文件和计算需要的分子或蛋白信息都被存储。(对应文件为 ./vina_space) 子目录的命名规则为：“PubChem ID” + “_into_” + “PDB ID”。子目录下的更多文件的解释请参考：https://autodock-vina.readthedocs.io/en/latest/docking_basic.html

在 121 次计算中：

- 成功计算（53 次）

- 或时间限制或软件原因失败 (68 次)。

对接的结果概览 (Affinity 单位为 kcal/mol, 值越低, 结合程度越好): (对应文件为 ./results_of_batch_docking.csv)

```
## # A tibble: 53 x 6
##   PDB_ID PubChem_id Affinity Ingredient.id Ingredient.name hgnc_symbol
##   <chr>   <chr>         <dbl> <chr>         <chr>         <chr>
## 1 6x81    66065         -6.13 HBIN017919 bergenin      TNF
## 2 6ooz    66065         -6.11 HBIN017919 bergenin      TNF
## 3 6x82    66065         -5.62 HBIN017919 bergenin      TNF
## 4 6ooy    66065         -5.53 HBIN017919 bergenin      TNF
## 5 1cxw    6476139        -5.38 HBIN020365 Chlorogenic acid methyl ester MMP2
## 6 4ni7    66065         -5.23 HBIN017919 bergenin      IL6
## 7 1rtg    6476139        -5.09 HBIN020365 Chlorogenic acid methyl ester MMP2
## 8 6x86    66065         -4.83 HBIN017919 bergenin      TNF
## 9 6x83    66065         -4.70 HBIN017919 bergenin      TNF
## 10 4zs7   66065         -4.50 HBIN017919 bergenin      IL6
## # i 43 more rows
```

可视化如下 (根据靶点蛋白去重复): (对应文件为 ./figs/Docking_Affinity.pdf)

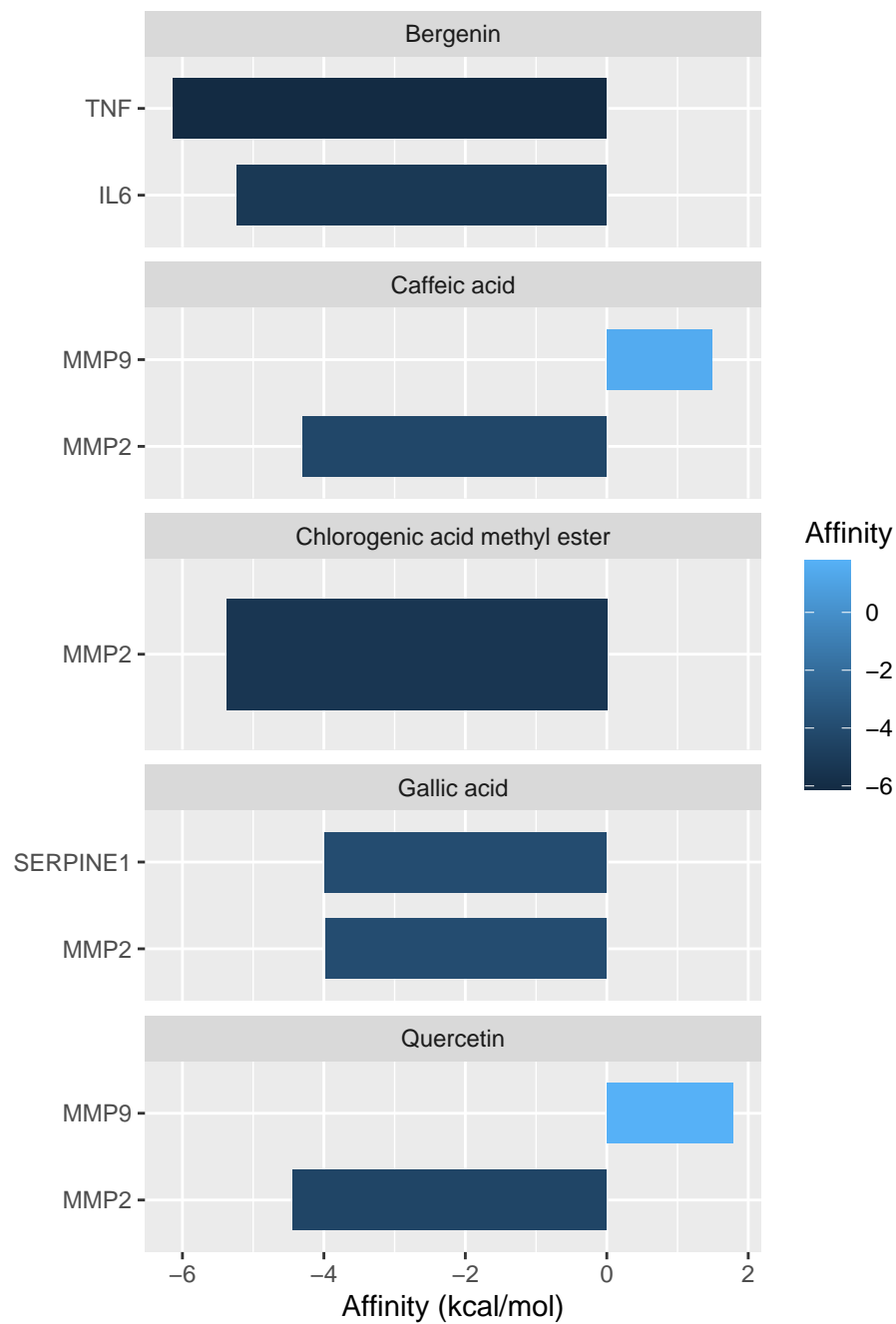


Figure 1: Molecular Docking Affinity

Reference

1. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling* **61**, 3891–3898 (2021).
2. Forli, S. *et al.* Computational proteinligand docking and virtual drug screening with the autodock suite.

Nature Protocols **11**, (2016).

3. Kim, S. *et al.* PubChem substance and compound databases. *Nucleic Acids Research* (2015).

4. Kim, S., Thiessen, P. A., Cheng, T., Yu, B. & Bolton, E. E. An update on pug-rest: RESTful interface for programmatic access to pubchem. *Nucleic Acids Research* **46**, (2018).