

COVID-19 主蛋白酶同源建模

生信 2001 张子栋 2020317210101

Table of Contents

COVID-19 主蛋白酶同源建模

背景

实验步骤

获取目标蛋白质序列和参考 PDB 文件

在 MOE 中进行序列对比和构建模型

使用 AMBER 进行动力学模拟

准备输入文件

`min_wat.in` 溶剂最小化

`md_water.in` 溶剂 MD - 模拟退火

`min_all.in` 体系最小化

`md_all.in` 体系 MD - MD 运行

`md_back.in` 侧链/溶剂MD - MD运行

准备提交可执行文件

提交并运行动力学

轨迹分析

去除水

RMSD

温度

总能量

模型优化

计算平均结构

挑选代表结构

导出代表结构

评估模型的合理性

总结

背景

当没有实验手段测定的结构存在时，同源建模时用于近似蛋白质 3D 结构的一种技术手段。该技术基于的原理是蛋白质的三级结构比氨基酸序列更保守，因此具有高序列同源性的蛋白质也应该具有高度的结构同源性。被建模的蛋白质称为目标蛋白，用于模拟目标的具有实验确定结构的蛋白质称为模版蛋白。模版蛋白和目标蛋白必须具有高度序列同源性，据此可以推测他们的 3D 结构非常相似。目标蛋白的氨基酸序列映射到模版蛋白的 3D 结构以产生近似模型，这就是同源模型。

实验步骤

获取目标蛋白质序列和参考 PDB 文件

1. 访问 [RCSB PDB: Homepage](#) 获取目标蛋白(#6LU7) FASTA 序列
2. 访问 [Protein BLAST](#)
 - 粘贴 FASTA 序列到 BLAST 搜索框
 - 选择 PDB 作为搜索数据库，选择 blastp 作为搜索算法
 - 点击底部 BLAST 开始搜索
3. 选择模板

- 这里选择符合度低的 7VTC_A
4. 下载 FASTA 序列和 PDB 文件
 5. 将 6UL7 和 7VTC_A 的 FASTA, PDB 文件放入 `homology_modeling` 目录

在 MOE 中进行序列对比和构建模型

- 删除除目标序列外的其他杂质
- 删除水分子
- 存为新的文件

使用 AMBER 进行动力学模拟

```
1 # 运行 tleap 程序
2 tleap
3
4 # 导入合适力场文件 这里使用 AMBER99 力场
5 source oldff/leaprc.ff99SB
6
7 # 导入有机力场
8 source leaprc.gaff
9
10 # 导入水的力场文件 因为会使用水对蛋白进行溶剂化处理
11 source leaprc.water.tip3p
12
13 # 导入模型的 PDB 文件
14 mol = loadpdb final.pdb
15
16 # 添加 8.0 埃水盒子
17 solvatebox mol TIP3PBOX 8
18
19 # 保存 *.crd 和 *.top 文件
20 saveamberparm mol final_HOH.prmtop final_HOH.inpcrd
21
22 # 使用 ambpdb 查看 PDB 文件
23 ambpdb -p final_HOH.prmtop -c final_HOH.inpcrd > final_HOH.pdb
```

如果残基数超过 10000 会重新从 0 开始计数
protein: 1-306 water: 307-12039

准备输入文件

以下操作使用到了 Vim, 输入时要先按 `i` 进入插入模式, 否则会出现输入内容的丢失

`min_wat.in` 溶剂最小化

`bellymask` 参数填入水分子范围

```

1 Constant Volume Minimization
2 # Control section
3 &cntrl
4 ntwe = 500, ntwx = 500, ntp = 500,
5 nsnb = 25, dielc = 1, cut = 8.0,
6 ntb = 1,
7 maxcyc = 10000, ntmin = 1, ncyc = 5000, dx0 = 0.01, drms = 0.0001,
8 ntp = 0,
9 ibelly = 1, ntr = 0,
10 imin = 1,
11 bellymask = ':307-12039', iwrap = 1,
12 /
13 END

```

md_water.in 溶剂 MD - 模拟退火

bellymask 参数填入水分子范围

```

1 Dynamic Simulation with Constant Pressure
2 # Control section
3 &cntrl
4 ntwx = 500, ntp = 500,
5 ntt = 1, temp0 = 300.0, tempi = 5.0, tautp = 1.0,
6 dielc = 1, cut = 8.0,
7 ntb = 2, ntc = 2, ntf = 2,
8 nstlim = 250000, dt = 0.0020,
9 ntp = 1, taup = 0.1, comp = 44.6, pres0 = 1.0,
10 ibelly = 1, ntr = 0,
11 imin = 0, irest = 0, ntx = 1, nmropt = 1,
12 bellymask = ':307-12039', iwrap = 1,
13 /
14 &ewald
15 skinnb = 2.0,
16 nbflag = 1,
17 /
18 &wt
19 type = 'TEMP0', istep1 = 1, istep2 = 12500, value1 = 5.0, value2 = 300.0,
20 /
21 &wt
22 type = 'TEMP0', istep1 = 12501, istep2 = 237500, value1 = 300.0, value2 =
23 300.0,
24 /
25 &wt
26 type = 'TEMP0', istep1 = 237501, istep2 = 250000, value1 = 300.0, value2 =
27 5.0,
28 /
29 &wt
30 type = 'END'
31 /
32 END

```

min_all.in 体系最小化

```
1 Constant Volume Minimization
2 # Control section
3 &cntrl
4 ntwe = 500, ntwx = 500, ntp = 500,
5 nsnb = 25, dielc = 1, cut = 8.0,
6 ntb = 1,
7 maxcyc = 10000, ntmin = 1, ncyc = 5000, dx0 = 0.01, drms = 0.0001,
8 ntp = 0,
9 ibelly = 0, ntr = 0,
10 imin = 1,
11 iwrap = 1,
12 /
13 END
```

md_all.in 体系 MD - MD 运行

```
1 Dynamic Simulation with Constant Pressure
2 # Control section
3 &cntrl
4 ntwx = 500, ntp = 500,
5 ntt = 1, temp0 = 300.0, tempi = 5.0, tautp = 1.0,
6 dielc = 1, cut = 8.0,
7 ntb = 2, ntc = 2, ntf = 2,
8 nstlim = 762500, dt = 0.0020,
9 ntp = 1, taup = 0.1, comp = 44.6, pres0 = 1.0,
10 ibelly = 0, ntr = 0,
11 imin = 0, irest = 0, ntx = 1, nmropt = 1,
12 /
13 &ewald
14 skinnb = 2.0,
15 nbflag = 1,
16 /
17 &wt
18 type = 'TEMP0', istep1 = 1, istep2 = 25000, value1 = 5.0, value2 = 300.0,
19 /
20 &wt
21 type = 'TEMP0', istep1 = 25001, istep2 = 762500, value1 = 300.0, value2 =
22 300.0,
23 /
24 &wt
25 type='END'
26 /
27 END
```

md_back.in 侧链/溶剂MD - MD运行

```
1 Dynamic Simulation with Constant Pressure
2 # Control section
3 &cntrl
4 ntwx = 500, ntp = 500,
5 ntt = 1, temp0 = 300.0, tempi = 5.0, tautp = 1.0,
6 dielc = 1, cut = 8.0,
7 ntb = 2, ntc = 2, ntf = 2,
```

```

8  nstlim = 400000, dt = 0.0020,
9  ntp = 1, taup = 0.1, comp = 44.6, pres0 = 1.0,
10 ibelly = 0, ntr = 1,
11 imin = 0, irest = 0, ntx = 1, nmropt = 1,
12 restraintmask = '@CA,C,O,N', restraint_wt=100,
13 iwrap = 1,
14 /
15 &ewald
16 skinnb = 2.0,
17 nbflag = 1,
18 /
19 &wt
20 type = 'TEMP0', istep1 = 1, istep2 = 25000, value1 = 5.0, value2 = 300.0,
21 /
22 &wt
23 type = 'TEMP0', istep1 = 25001, istep2 = 400000, value1 = 300.0, value2 =
24 300.0,
25 /
26 &wt
27 type = 'REST', istep1 = 1, istep2 = 40000, value1 = 0.1, value2 = 1.0,
28 /
29 &wt
30 type = 'REST', istep1 = 40001, istep2 = 400000, value1 = 1.0, value2 = 1.0,
31 /
32 &wt
33 type='END'
34 /
35 END

```

准备提交可执行文件

run.com

```

1  # 1. 溶剂能量最小化
2  sander -O -i min_wat.in -o min_wat.out -p nameSol.prmtop -c nameSol.inpcrd -
3  r min_wat.rst
4
5  # 2. 模拟退火
6  sander -O -i md_wat.in -o md_wat.out -p nameSol.prmtop -c min_wat.rst -r
7  md_wat.rst -x md_wat.crd
8
9  # 3. 整个体系能量最小化
10 sander -O -i min_all.in -o min_all.out -p nameSol.prmtop -c md_wat.rst -r
11 min_all.rst
12
13 # 4. 整个体系升温并进行1500 ps动力学
14 sander -O -i md_all.in -o md_all.out -p nameSol.prmtop -c min_all.rst -r
15 all_prod.rst -x all_prod.crd
16
17 # 5. 固定蛋白骨架后进行50ps升温然后进行750ps动力学
18 sander -O -i md_back.in -o md_back.out -p nameSol.prmtop -c min_all.rst -r
19 back_prod.rst -ref min_all.rst -x back_prod.crd &

```

赋予文件执行权限 `chmod +x run.com`

提交并运行动力学

忽略输出运行文件

```
1 | nohup ./run.com &
```

使用 `top` 查看进程是否在运行

在实验过程中, 这里出现了一点错误, 一个文件中的水分子的范围 错误地写成了 `307-12093`, 应为 `307-12039`

如果进程较多, 可以使用 `pa -aux | grep <userName>` 直接查看

轨迹分析

可能是因为在创建文件时出现了错误, 这里并没有跑出结果.

使用了助教的运行结果

去除水

- 在 `md_simulation` 目录下, 创建 `traj_strip_wat.in` 文件:

```
1 | # Strip water of trajectory files
2 | trajin all_prod.crd
3 | strip :WAT
4 | autoimage
5 | trajout all_prod_nowat.crd
6 | run
7 | exit
```

- 运行:

```
1 | cpptraj -p final_HOH.prmtop -i traj_strip_wat.in > traj_strip_wat.out
```

- 创建 `top_strip_wat.in` 文件:

```
1 | # Strip water of top files
2 | parmstrip :WAT
3 | parmwrite out final_nowat.prmtop
4 | run
5 | exit
```

- 运行:

```
1 | cpptraj -p final_HOH.prmtop -i top_strip_wat.in > top_strip_wat.out
```

RMSD

在 `hm_analysis` 目录下: `mkdir md_analysis`

- 创建 `rmsd.in`

```

1 # Caculate RMSD of protein
2 trajin ../../md_simulations/all_prod_nowat.crd
3 rms first mass @CA,C,N
4 rmsd protein-rmsd :1-303&!(@H=) out rmsd.dat
5 run
6 exit

```

- 保存 data

```

1 cpptraj -p ../../md_simulations/final_nowat.prmtop -i rmsd.in > rmsd.out

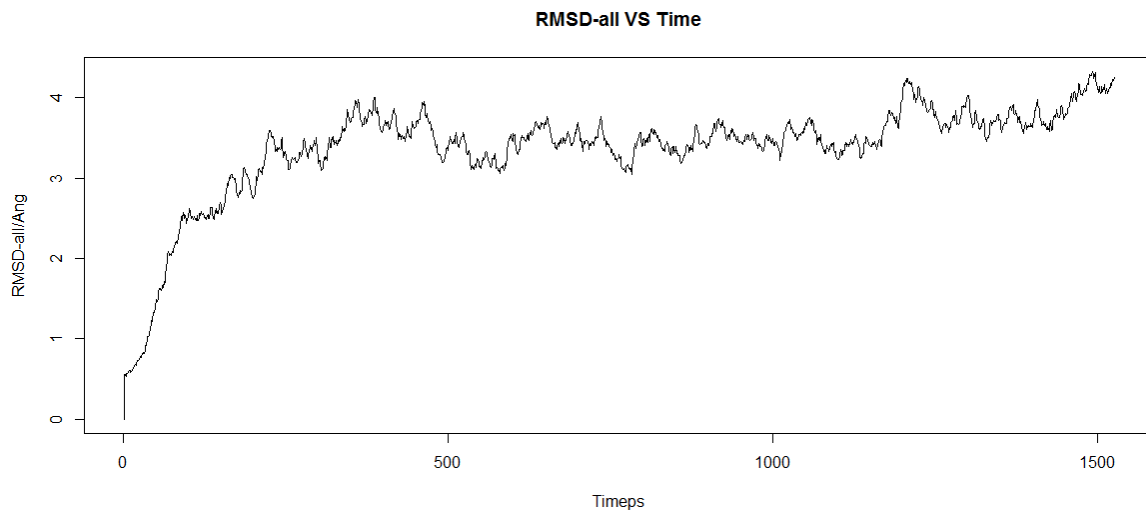
```

- 使用 R 作图

```

1 data <- read.table(file.choose())
2 plot(data, type = "l", xlab="Time/ps", ylab = "RMSD-all/Ang", main = "RMSD-all
VS Time")

```



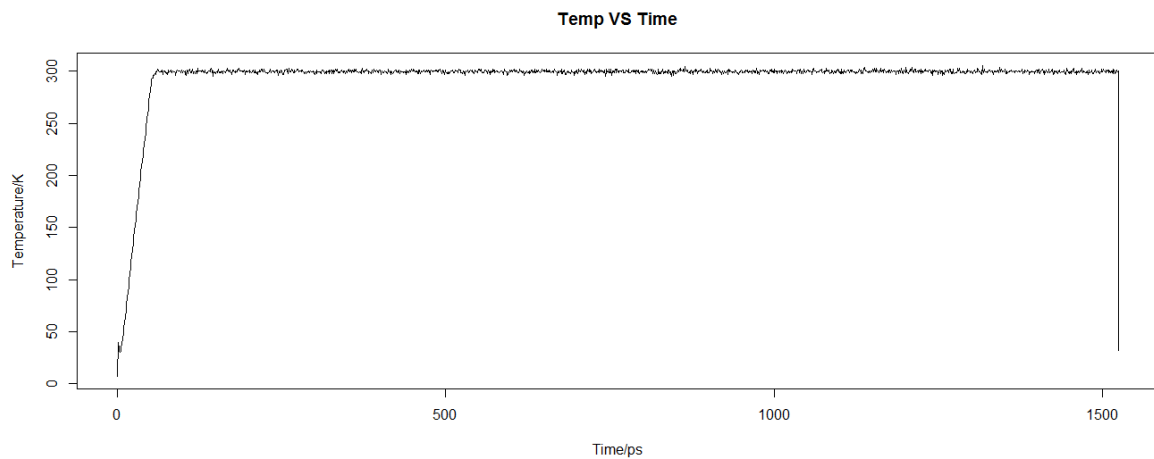
温度

运行 `grep TIME ../../md_simulations/md_all.out > temp_time.txt` 并作图

```

1 data <- read.table(file.choose())
2 time <- data$V6
3 temperature <- data$V9
4 plot(time, temperature, type = "l", main = "Temp VS Time", xlab = "Time/ps",
ylab = "Temperature/K")

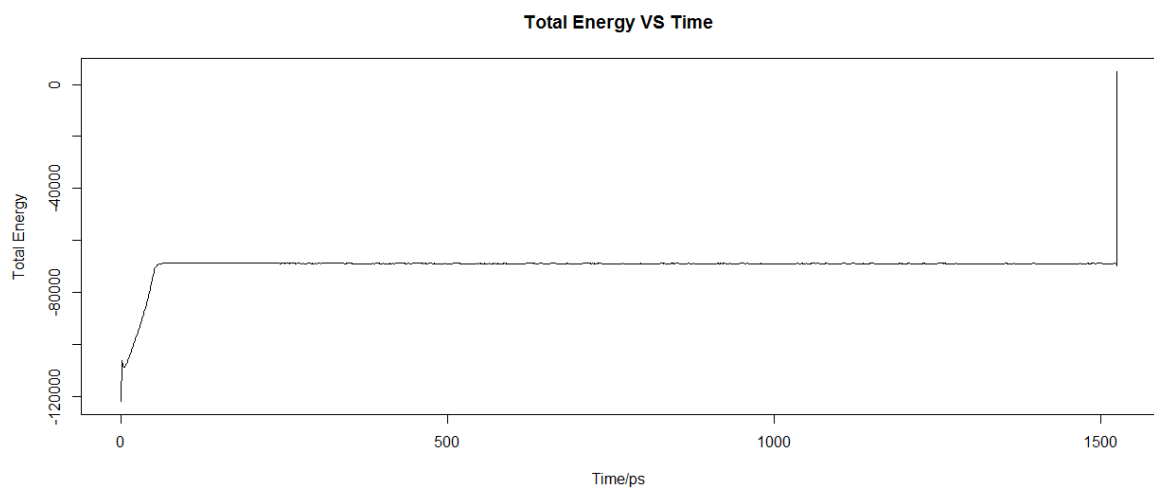
```



总能量

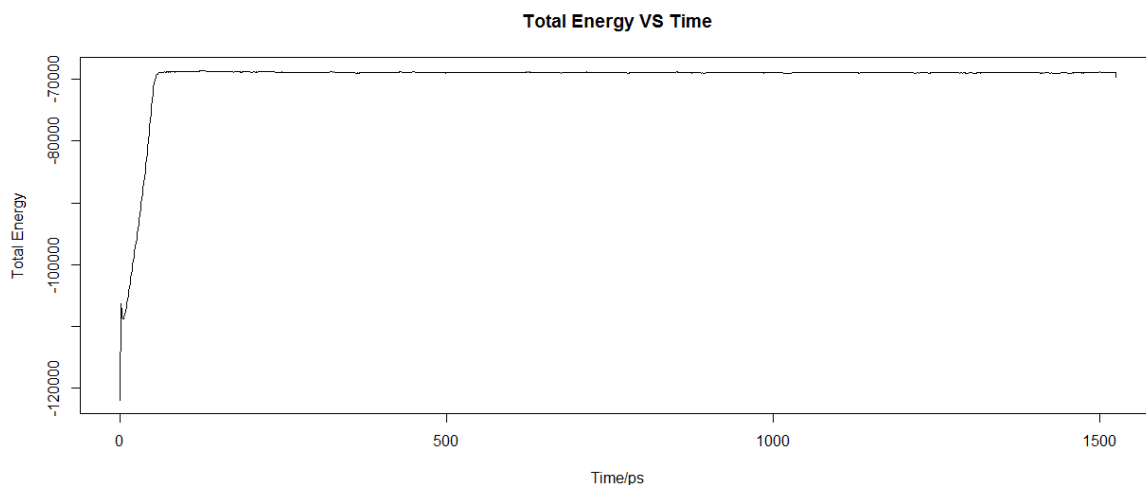
运行 `grep Etot ../../md_simulations/md_all.out > Etot_time.txt` 并作图

```
1 etot_time <- read.table(file.choose())
2 temp_time <- read.table(file.choose())
3 time <- temp_time$V6
4 temperature <- temp_time$V9
5 etot <- etot_time$V3
6 plot(time, etot, type = "l", main = "Total Energy VS Time", xlab = "Time/ps",
  ylab = "Total Energy")
```



- 有一点明显错误, 去除后重新作图

```
1 time <- time[-1528]
2 etot <- etot[-1528]
3 plot(time, etot, type = "l", main = "Total Energy VS Time", xlab = "Time/ps",
  ylab = "Total Energy")
```

模型优化

与仅有一个结构表示的最初的同源模型结构不同，优化模型结构是由动力学轨迹(一系列结构)来进行描述的，为了在轨迹中获取单个代表性的结构数据我们需要计算优化模型的平均结构。

计算平均结构

`md_analysis` 目录下, 编辑 `get_avg_struct.in`:

```
1 # Get average structure
2 trajin ../../md_simulations/all_prod_nowat.crd
3 average traj_avg_struct.pdb :1-303 start 50 stop 3050 offset 1 pdb
4 run
5 exit
```

- 运行

```
1 cpptraj -p ../../md_simulations/final_nowat.prmtop -i get_avg_struct.in >
  get_avg_struct.out
```

挑选代表结构

编辑 `rmsd_from_avg.in`:

```
1 # Calculate RMSD from average structure
2 trajin ../../md_simulations/all_prod_nowat.crd
3 reference ./traj_avg_struct.pdb
4 rms reference out rmsd_from_avg.txt :1-303
5 run
6 exit
```

- 运行

```
1 cpptraj -p ../../md_simulations/final_nowat.prmtop -i rmsd_from_avg.in >
  rmsd_from_avg.out
```

导出代表结构

编辑 `get_refine_struct.in`:

```
1  #Get snapshot most like the average structure
2
3  trajin ../../md_simulations/all_prod_nowat.crd 1383 1383 1
4  trajout struct_like_avg.pdb pdb
5  run
6  exit
```

- 运行

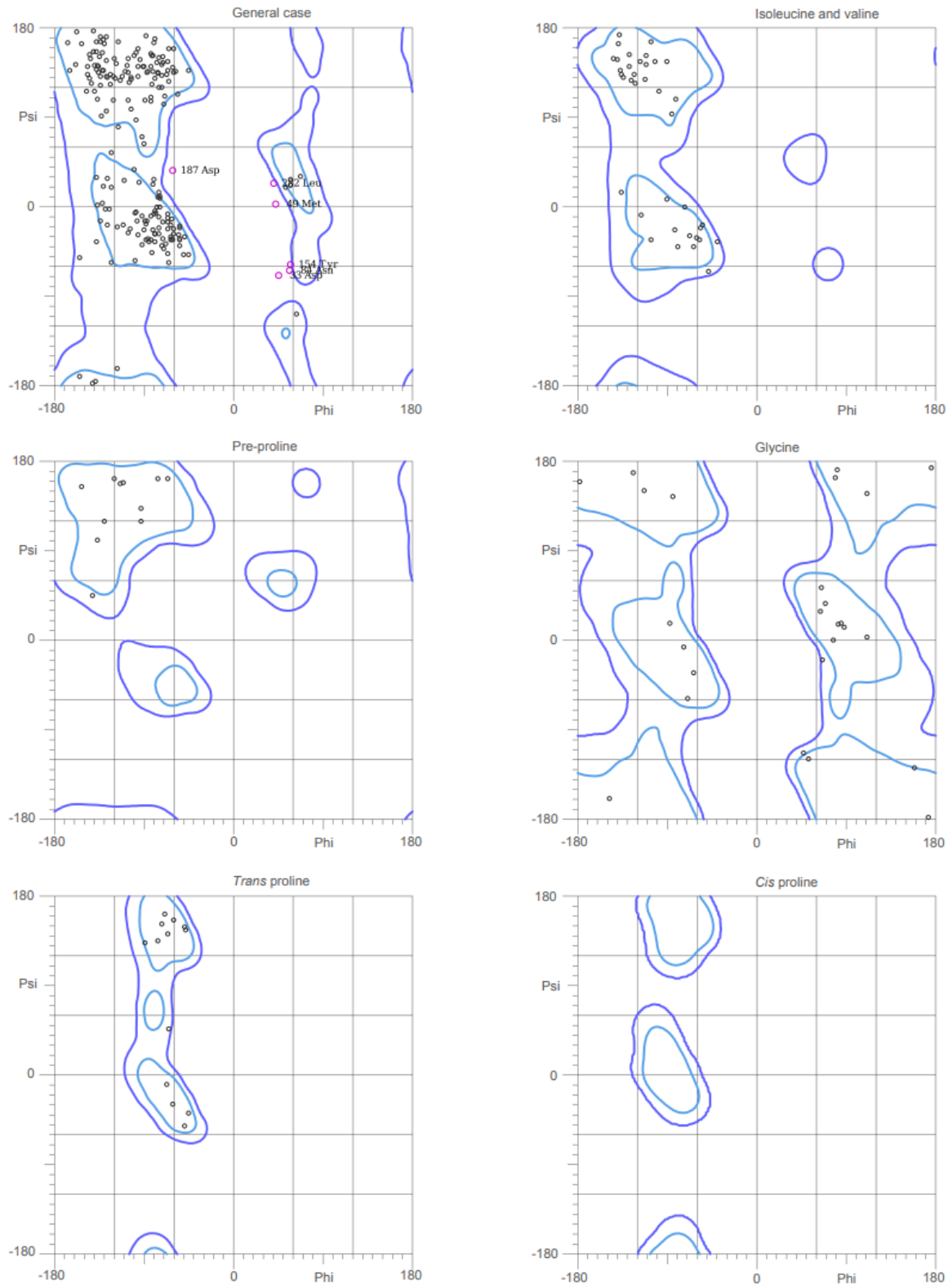
```
1  cpptraj -p ../../md_simulations/final_nowat.prmtop -i get_refine_struct.in >
   get_refine_struct.out
```

评估模型的合理性

在页面 [Main page - MolProbity \(duke.edu\)](#) 上获取拉氏构象图

MolProbity Ramachandran analysis

struct_like_avg.pdb, model 1



- 从拉氏构象图可以看到, 有部分氨基酸的二面角不在经验支持的范围内

总结

本次课程中, 通过同源建模和分子动力学模拟这两种方法, 我们完成了对 6LU7 蛋白的空间结构模拟和分析. 对这类实验的过程有了初步的了解. 因为背景知识的局限, 本次分析对蛋白结构的具体差异挖掘比较浅显, 后续还可以对导致空间结构差异的残基进行更加深入的分析.

