

Molecular Dynamics simulations of polypeptides in water solution under physiological conditions. Application to fragments of SARS-CoV 3CL protease.

Adolfo Bastida

Universidad de Murcia (Spain)



Máster Universitario en Bioinformática

Modelado Molecular

Index

1 System

2 Build

- PyMOL

3 Build

- PyMOL
- PDB format

4 GROMACS

- MD package
- Force fields in Gromacs
- Gromacs modules
- Topology
- Periodic boundary conditions

5 Equilibration

- Input files
- Neutralization
- Equilibration

6 Production run

- Run
- Cartoons

7 Analysis

- Geometry
- Temperature and velocities
- Energy

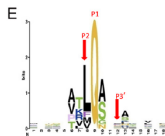


System

- Fragment of the sequence that are cleaved by SARS-CoV 3CL^{pro} (3C-like protease) in the SARS-CoV polyproteins

D

P2	P1	P3'	
RQCSGVTF	GKFKKIVKG		(nsp5/nsp6, C-terminus of 3CL ^{pro})
TSITSAYLV	SGFRKMAFP		(nsp4/nsp5, N-terminus of 3CL ^{pro})
PCIKVAVV	SMMSDVCT		(nsp4/nsp7)
MLDNRAVL	ATASEPSSL		(nsp7/nsp8)
RANSAVKL	NNSLSPVAL		(nsp8/nsp9)
SLAATVRL	AGNATSEVPA		(nsp9/nsp10)
DQLREPLM	SADASTFLN		(nsp10/nsp11, nsp10/nsp12)
MYTFHTVL	AVGACVLCN		(nsp12/nsp13)
PRHNVAEL	AEHVTGLPK		(nsp13/nsp14)
LSWTFTEL	SEENVAIVV		(nsp14/nsp15)
VETFFPKL	ASQANQGVV		(nsp15/nsp16)



model with charge). The unbiased Fo-Fc difference electron density map for omitted residues 303-310 (P4-P5) is contoured at 3.0 σ . The map was developed using PHENIX (14). The replaced active site residue (Ala145) is also indicated. **(D) The 11 sequences that are cleaved by SARS-CoV 3CL^{pro} in the SARS-CoV polyproteins. The first and second sequences are the C- and N-terminal cleavage sequences, respectively, of 3CL^{pro} itself (1).** (E) The consensus sequence for cleavage by 3CL^{pro}, analyzed by the sequence logo program (15), using the sequences shown in D and WebLogo version 2.8.2 (weblogo.berkeley.edu) (16).

T. Muramatsu et al, PNAS 113, 12997 (2016)

- Fragment KIVKG
 - K $\Rightarrow$ Lysine (lys)
 - I $\Rightarrow$ Isoleucine (ile)
 - V $\Rightarrow$ Valine (val)
 - G $\Rightarrow$ Glycine (gly)
- Convention: write the sequence from N- (left) to C-terminus (right).
- Terminal groups: ACE (CH₃CO-) and NME (-NHCH₃)

Working directory

```
> mkdir pymol  
> cd pymol  
> cp ../for-students/pymol/6lyr.pdb .
```

Pymol

```
> pymol 6lyr.pdb &
```

Modify the style

```
PyMOL > as lines  
PyMOL > as spheres  
PyMOL > as ribbon  
PyMOL > as cartoon  
PyMOL > as sticks  
PyMOL > as surface
```

PyMol default

File, Edit pymolrc

Filename of new pymolrc: /home/adolfo/.pymolrc Ok

Add: set auto_show_classified, 0 File, Save

PyMol Builder

- File, Reinitialize
- Builder, Protein, Secondary Structure: alpha helix, Ala, Create as new object, Ala (click 10 o more times) Done
- Press central button to move/Press right button to zoom in/out ...
- File, Reinitialize Everything
- Builder, Protein, Secondary Structure: parallel beta sheet, Ala, Create as new object, Ala (click 10 o more times) Done

Working directory

```
> mkdir 1-build  
> cd 1-build  
> cp ../for-students/pdb-file-charmm27/3clpro.pdb .
```

Pymol

```
> pymol 3clpro.pdb &
```

Review the sequence: , or , .

Protein Data Bank (PDB)

Through an internet information portal and downloadable data archive, the PDB provides access to 3D structure data for large biological molecules (proteins, DNA, and RNA). <https://www.rcsb.org>

PDB file

```
> less 3clpro.pdb
```

- ATOM
- Atom number
- Atom class (C is carbon atom in carbonyl groups in the backbone chain, CA is an alpha carbon atom,...)
- Residue name (three letter code)
- Residue number (from N terminal into C terminal)
- (x,y,z) coordinates in Amstrongs
- Occupancy. (X-ray)
- Temperature factor. (X-ray)
- Element name (4 characters): 1-2 Chemical Symbol



MD package

Paths

Add the following line to your `.bashrc` file

```
source /usr/local/gromacs/bin/GMXRC.bash
```

GROningen MAchine for Chemical Simulations

```
> gmx -version
```

Copyright (c) 1991-2000, University of Groningen, The Netherlands.

Copyright (c) 2001-2017, The GROMACS development team at:

Uppsala University, Stockholm University and the Royal Institute of Technology, Sweden.

check out <http://www.gromacs.org> for more information.



Force fields in Gromacs

Force fields (ff)

```
> ls -lita $GROMACS_DIR/share/gromacs/top |grep ff
```

AMBER (94, 99, 99sb, 99sb-ildn, 99, 03, GS)

CHARMM27

GROMOS(96) (43a1/2/3,53a5,53a6), Modif GROMOS87

OPLS-AA

Contributions:

http://www.gromacs.org/Downloads/User_contributions/Force_fields



Force fields in Gromacs

ff organization

```
> ls $GROMACS_DIR/share/gromacs/top/charmm27.ff
```

forcefield.doc	FF documentation
	Force Field parameters
forcefield.itp	Includes the whole FF topology (less and see)
ffnonbonded.itp	Non bonded terms (VdW)
ffnanonbonded.itp	
ffbonded.itp	Bonded terms (bonds, angles, dihedrals)
ffnabonded.itp	
gb.itp	Implicit solvation parameters
cmap.itp	CHARMM27 specific parameters
aminoacids.rtp	Residue Topology Database
	also for dna and rna (depends on the force field)



Gromacs modules

General syntax

```
> gmx program_name -flags (single precision)
> gmx_d program_name -flags (doble precision)

> gmx pdb2gmx -h
```

Options to specify input files:

-f [<.gro/.g96/...>] (eiwit.pdb) Structure file: gro g96 pdb brk ent esp tpr

Options to specify output files:

-o [<.gro/.g96/...>] (conf.gro) Structure file: gro g96 pdb brk ent esp

-p [<.top>] (topol.top) Topology file

Other options:

-[no]ter (no) Interactive termini selection, instead of charged (default)



Topology

Create topology

```
> gmx pdb2gmx -f 3clpro.pdb -o 3clpro.gro -p 3clpro.top -ter
```

Select the Force Field:

8: CHARMM27 all-atom force field (CHARM22 plus CMAP for proteins)

Select the Water Model:

1: TIP3P TIP 3-point, recommended

Select start terminus type for ACE-1:

2: None

Select end terminus type for CT3-7

4: None

You have successfully generated a topology from: 3clpro.pdb.
The Charmm27 force field and the tip3p water model are used.



Topology

Files

```
> less 3clpro.gro
```

Free text

Number of atoms

Residue Number, Residue name, atom class, atom number, (x,y,z)
coordinates in nm

```
> less 3clpro.top
```

Data required to evaluate the potential energy function: charges, bonded
atoms, angles, dihedrals,...



Periodic boundary conditions

Unit cell

```
> gmx editconf -f 3clpro.gro -o 3clpro-box.gro -bt cubic  
-box 3.0 3.0 3.0
```

-bt <enum> (triclinic) Box type for -box and -d: triclinic, cubic,...
-box <vector> (0 0 0) Box vector lengths (a,b,c)

Solvent molecules

```
> gmx solvate -cp 3clpro-box.gro -cs -o 3clpro-box-solv.gro  
-p 3clpro.top
```

Output configuration contains 2648 atoms in 857 residues

Volume : 27 (nm³)

Density : 981.765 (g/l)

Number of SOL molecules: 850

At 25°C water density is 0.997 kg/l



Periodic boundary conditions

Final result

```
> less 3clpro-box-solv.gro
```

contains now the water solvent molecules

```
> gmx editconf -f 3clpro-box-solv.gro -o 3clpro-box-solv.pdb
```

```
> pymol 3clpro-box-solv.pdb
```



Input files

Working directory

```
> cd ..  
> mkdir 2-equilibration  
> cd 2-equilibration  
> cp ../1-build/3clpro-box-solv.gro .  
> cp ../1-build/3clpro.top .
```




Input files

.mdp file

```
> cp ../for-students/equilibration-mdp/equiNVT.mdp .  
> less equiNVT.mdp  
integrator = md ; leap-frog integrator  
dt = 0.0005 ; 0.5 fs  
nsteps = 8000 ; 4 ps  
nstxout = 200 ; Save coordinates every 0.1 ps  
Tcoupl = v-rescale  
gen_vel = yes ; Should be the first equilibration  
gen_temp = 298.0 ; Temperature to generate corresponding  
Maxwell distribution  
gen_seed = 12345678 ; Random seed
```



Neutralization

.tpr file

```
gmx grompp -f equiNVT.mdp -c 3clpro-box-solv.gro -p  
3clpro.top -o 3clpro-a.tpr
```

```
NOTE 1 [file 3clpro.top, line 975]:  
System has non-zero total charge: 2.000000
```



Neutralization

Neutralization

```
> gmx genion -s 3clpro-a.tpr -p 3clpro.top -o 3clpro.gro -nn 2
```

-nn <int> (0) Number of negative ions

Select group: SOL

Replacing 2 solute molecules in topology file (3clpro.top) by 0
NA and 2 CL ions.

3clpro.gro now includes 2 Cl⁻ atoms

```
> less 3clpro.gro
```

Now the system is neutral

```
> gmx grompp -f equiNVT.mdp -c 3clpro.gro -p 3clpro.top -o  
3clpro.tpr
```

Check

```
> gmx editconf -f 3clpro.gro -o 3clpro.pdb
```

```
> pymol 3clpro.pdb
```



Equilibration

Equilibration run

```
> cp ../for-students/equilibration-mdp/submit_eck.sh .
> cat submit_eck.sh
#!/bin/bash
#
#SBATCH -p eck-q
#SBATCH -chdir=$HOME/pruebas/master/ejercicio/2-equilibration
#SBATCH -J equilibrado
#SBATCH -cpus-per-task=1
date
gmx mdrun -deffnm 3clpro -c 3clpro.g96 -nt 1
date
-deffnm <string> Set the default filename for all file options
-nt <int> (0) Total number of threads to start (0 is guess)
> sbatch submit_eck.sh
> queue
```



Equilibration

Equilibration run

Output:

- > less slurm-70.out (queue output)
- > less 3clpro.g96 (final positions and velocities)
- > less 3clpro.log (information)



Production run

Working directory

```
> cd ..  
> mkdir 3-run  
> cd 3-run  
> cp ../2-equilibration/3clpro.top .  
> cp ../for-students/g96-equilibrated/3clpro.g96 . (400 ps)  
> cp ../for-students/run-mdp/runNVT.mdp .
```

runNVT.mdp

```
> less runNVT.mdp  
  
nsteps = 4000 ; 2 ps  
nstxout = 2 ; Save coordinates every 1 fs  
gen_vel = no ; Should be the first equilibration
```



Production run

Run

```
> gmx grompp -f runNVT.mdp -c 3clpro.g96 -p 3clpro.top -o  
3clpro.tpr
```

```
> cp ../for-students/run-mdp/submit_eck.sh .
```

```
> cat submit_eck.sh
```

```
#!/bin/bash
```

```
#
```

```
#SBATCH -p eck-q
```

```
#SBATCH -chdir=$HOME/pruebas/master/ejercicio/3-run
```

```
#SBATCH -J equilibrado
```

```
#SBATCH -cpus-per-task=1
```

```
date
```

```
gmx mdrun -deffnm 3clpro -c 3clpro.g96 -nt 1
```

```
date
```

```
> sbatch submit_eck.sh
```

```
> squeue
```

Output files:

3clpro.trr (coordinates, velocities,)

3clpro.tpr (Structure+mass)



Production run

Cartoons

```
> gmx traj -f 3clpro.trr -s 3clpro.tpr -oxt cartoon.pdb
```

Select: 0 (System)

Output file: cartoon.pdb

```
> pymol cartoon.pdb
```

(lower right corner)

Roving_detail shows details of the environment centered on the screen

Hydrogen Bonds

```
> gmx traj -f 3clpro.trr -s 3clpro.tpr -oxt cartoon.pdb
```

Select: 1 (Protein)

Output file: cartoon.pdb

```
> pymol cartoon.pdb
```

(lower right corner)



Analysis

Working directory

```
> cd ..  
> mkdir 4-analysis  
> cd 4-analysis  
> cp ../3-run/3clpro.tpr .  
> cp ../3-run/3clpro.g96 .  
> cp ../3-run/3clpro.trr . (coordinates)  
> cp ../3-run/3clpro.edr . (energies)
```

GROMACS commands

<https://manual.gromacs.org/documentation/2019/user-guide/cmdline.html>

Commands by name

Commands by topic



Analysis

Radius of gyration

Definition:

$$R_{\text{gyr}}^2 = \frac{1}{M} \sum_i^N m_i (\mathbf{r}_i - \mathbf{R})^2 \quad \left\{ \begin{array}{l} M \text{ total mass} \\ N \text{ number of atoms} \\ m_i \text{ mass of } i\text{th atom} \\ \mathbf{r}_i \text{ position vector of } i\text{th atom} \\ \mathbf{R} \text{ center of mass vector} \end{array} \right.$$

```
gmx gyrate -f 3clpro.trr -s 3clpro.tpr -xvg none
```

```
Select 1 (Protein)
```

Plot:

```
> gnuplot
```

```
gnuplot > plot 'gyrate.xvg'
```



Analysis

Bond distances (I)

Create distances.ndx file

```
> cat distances.ndx
```

```
[CO-lys-2]
```

```
27 28
```

```
gmx distance -f 3clpro.trr -s 3clpro.tpr -n distances.ndx  
-oall -xvg none  
-xvg <enum> (xmgrace) Plot formatting: none, xmgrace, xmgr  
-oall [<.xvg>] (dist.xvg) (Optional) All distances as function of time
```

Select group: 0

Ctrl+d



Analysis

Bond distances (II)

Output:

CO-lys-2:

Number of samples: 2001

Average distance: 0.12314 nm

Standard deviation: 0.00191 nm

```
> less dist.xvg
```

Plot: > gnuplot

```
gnuplot> plot 'dist.xvg' with linespoints
```



Analysis

Bond distances (III)

```
> cat distances.ndx
```

```
[CO-lys-2]
```

```
27 28
```

```
[CACB-ILE-3]
```

```
31 33
```

```
gmx distance -f 3clpro.trr -s 3clpro.tpr -n distances.ndx
```

```
-oall -xvg none
```

```
Select group: 1
```

```
Ctrl+d
```



Analysis

Bond distances (IV)

Output:

CACB-ILE-3:

Number of samples: 2001

Average distance: 0.15340 nm

Standard deviation: 0.00342 nm

Plot: > gnuplot

```
gnuplot> plot 'dist.xvg' with linespoints
```



Analysis

Bond angles

```
> cat distances.ndx
```

```
[CA-C-O-lys-2]
```

```
9 27 28
```

```
gmx angle -f 3clpro.trr -n distances.ndx -ov -xvg none
```

```
Select group: 2 (Ctrl+d)
```

Output:

```
< angle > = 122.146
```

```
< angle^2 > = 14932.4
```

```
Std. Dev. = 3.57043
```

Plot:

```
> gnuplot
```

```
gnuplot> plot 'angaver.xvg' with linespoints
```



Analysis

Dihedral angles (I)

```
> gmx rama -f 3clpro.trr -s 3clpro.tpr -xvg none
> less rama.xvg
> grep LYS-2 rama.xvg | awk '{print $1}' | cat -n >
phi-lys-2.dat
> grep LYS-2 rama.xvg | awk '{print $2}' | cat -n >
psi-lys-2.dat
> wc -l phi-lys-2.dat
2001 phi-lys-2.dat
```

Plot:

gnuplot

```
gnuplot> plot 'phi-lys-2.dat' with linespoints
```

```
gnuplot> plot 'psi-lys-2.dat' with linespoints
```




Analysis

Dihedral angles (II)

```
> join phi-lys-2.dat psi-lys-2.dat > lys-2.dat  
> less lys-2.dat
```

Plot:

gnuplot

```
gnuplot> plot 'lys-2.dat' using 2:3 with linespoints
```



Analysis

Hydrogen bonds

```
> gmx hbond -f 3clpro.trr -s 3clpro.tpr -a 120 -r 0.18 -noda
-num num.svg -xvg none -nonitacc -hx hx.svg
-a Cutoff angle (degrees, Hydrogen - Donor - Acceptor)
-r Cutoff radius (nm, X - Acceptor, see next option)
-[no]da (yes) Use distance Donor-Acceptor (if True) or Hydrogen-Acceptor (False)
-num number of hydrogen bonds as a function of time.
-nonitacc OH and NH groups are regarded as donors, O is an acceptor always, N
is an acceptor by default, but this can be switched using -nitacc
-hx: the number of n-n+i hydrogen bonds as a function of time where n and n+i
stand for residue numbers and i ranges from 0 to 6.
Select 1 and 1 (Protein) for intramolecular
Select 1 and 13 (SOL) for intermolecular
> less num.svg
> less hx.svg
```



Analysis

Temperature

```
> gmx traj -f 3clpro.trr -s 3clpro.tpr -xvg none -ot
```

```
Select a group: 0
```

Plot:

gnuplot

```
gnuplot> plot 'temp.xvg' with linespoints
```

Making a histogram:

```
gnuplot> width = 2
```

```
gnuplot> hist(x,width)=width*floor(x/width)+width/2.0
```

```
gnuplot> set boxwidth width*0.9
```

```
gnuplot> set style fill solid 0.5
```

```
gnuplot> plot 'temp.xvg' u (hist($2,width)):(1.0) smooth  
freq w boxes lc rgb'green' notitle
```



Analysis

Velocities (I)

```
> gmx traj -f 3clpro.trr -s 3clpro.tpr -xvg none -ov -len  
-ov velocities veloc.xvg  
-ox coordinates coord.xvg  
-of force force.xvg  
-len Plot vector length  
Select a group: 1 (Protein)  
> less veloc.xvg  
x, y, and z velocities components and length for every atom in nm/ps  
Plot:  
gnuplot  
gnuplot> plot 'veloc.xvg' using 1:5
```



Analysis

Velocities (II)

Making a histogram:

```
gnuplot> n=100 #number of intervals
gnuplot> max=3. #max value
gnuplot> min=-3. #min value
gnuplot> width=(max-min)/n #interval width
gnuplot> hist(x,width)=width*floor(x/width)+width/2.0
gnuplot> set boxwidth width*0.9
gnuplot> set style fill solid 0.5 # fill style
gnuplot> plot 'veloc.xvg' u (hist($2,width)):(1.0) smooth
freq w boxes lc rgb'green' notitle
gnuplot> plot 'veloc.xvg' u (hist($3,width)):(1.0) smooth
freq w boxes lc rgb'green' notitle
```



Analysis

Velocities (III)

Maxwell-Boltzmann distribution:

$$f(v_x) = \left(\frac{m}{2\pi k_B T} \right) e^{-\frac{mv_x^2}{2k_B T}} \rightarrow \text{gaussian distribution}$$



Analysis

Total Energy

```
> gmx energy -f 3clpro.edr -s 3clpro.tpr -xvg none
```

```
Select a group: 15 (Total energy) (double enter)
```

```
Energy          Average    Err.Est.    RMSD Tot-Drift
```

```
-----
```

```
Total Energy -25421.9    44        150.922 110.342 (kJ/mol)
```

An error estimate of the average is given based on a block averages over 5 blocks using the full-precision averages.

Drift is calculated by performing a least-squares fit of the data to a straight line. The reported total drift is the difference of the fit at the first and last point.

Plot:

```
gnuplot
```

```
gnuplot> plot 'energy.xvg' with linespoints
```



Analysis

Kinetic Energy

```
> gmx energy -f 3clpro.edr -s 3clpro.tpr -xvg none
```

```
Select a group: 14 (Kinetic energy) (double enter)
```

```
Energy          Average    Err.Est.    RMSD Tot-Drift
```

```
-----  
Kinetic En. 9758.48      39      157.374 35.7389 (kJ/mol)
```

Test:

$$2644 \text{ atoms} \times \frac{3}{2} \underbrace{k_{BT}}_{2.478 \text{ kJ/mol}} = 9827 \text{ kJ/mol}$$