Laboratory of Bioinformatics and Computational Intelligence

labic.utfpr.edu.br/

PathMolD-AB(Neighbourhood List Version): Generation of Spatiotemporal Pathways of Protein Folding Using Molecular Dynamics with a Coarse-grained Model and a Neighbourhood List method

Step-by-step: Generate your first protein folding pathway and analyze it

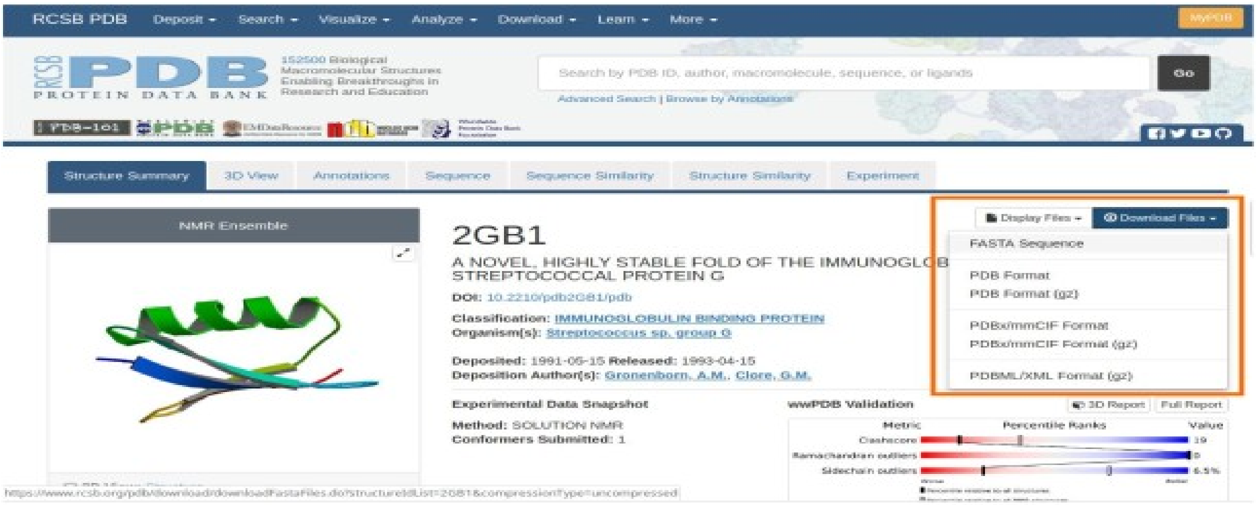
The purpose of this tutorial is to describe in a simplified way how to generate your first folding trajectory data and visualize the generated folding data for your analysis with PathMolD-AB(NL Version).
The steps required to accomplish this goal are presented below.

step1) Generate your protein AB sequence based on a FASTA file
step2) Set the protein information and the simulation parameters
step3) Compile and execute the protein folding simulation
step4) Realize the protein folding analysis

obs: words highlighted with yellow represent files and red represent line commands.

##
step1) generate your protein AB sequence based on a FASTA file
##

1.1)Download the FASTA file in the PDB database, the square in the figure below highlights the download location.



2GB1

A NOVEL, HIGHLY STABLE FOLD OF THE IMMUNOGLOB STREPTOCOCCAL PROTEIN G

DOI: 10.2210/pdb2GB1/pdb

Classification: IMMUNOGLOBULIN BINDING PROTEIN
Organism(s): Streptococcus sp. group G

Deposited: 1991-05-15 **Released:** 1993-04-15
Deposition Author(s): Gronenborn, A.M., Clore, G.M.

Experimental Data Snapshot

Method: SOLUTION NMR
Conformers Submitted: 1

wwPDB Validation

https://www.rcsb.org/pdb/download/downloadFastaFiles.do?structureIdList=2GB1&compressionType=uncompressed

In this sample, we download the 2GB1 fasta file from https://www.rcsb.org/structure/2GB1 as presented in the below Figure

1.2) Access the directory SRC_GPU_NL, then, execute the command below.
format

```
$ python AB_sequence.py<pdb id of the fasta file>
```

example

```
$ python AB_sequence.py2GB1
```

*Make sure that the file fasta is in the INPUT directory.
*The AB_sequence.py was developed in Python 2.7 language.
*Here, we use the Alberts AB classification.

The program will generate the AB sequence file of the 2GB1 protein in the INPUT directory.
example: 2GB1.txt

The content of the 2GB1.txt file is show below.

```
ABBBAAABABBABABBBBBAABAABABBAABBBABBBAABABABBBBABBBABABB
```

,where A and B represents the Hydrophobic and Polar amino acids,respectively.

###
step2) set the protein information and the simulation parameters
###

Insert the information about the protein and the simulation parameters into the input file. The file 2GB1_56.in(in the INPUT directory) is a sample of this input file for the protein 2GB1. The descriptions of each information are present below.

sequence: AB sequence of the protein (use step1 to get this information)
ProtLen: protein size, number of amino acids
LV: length of the 2D or 3D box
stepLimit: max number of steps
temperature: temperature
savepathways: y enable the function that saves amino acids coordinates in text files. This files can be used to show an example of a pathway
pathwaysstep: steps between generations of text coordinate files.

The content of the 2GB1_56.in file is show below.

```
Sequence = ABBBAAABABBABABBBBBAABAABABBAABBBABBBAABABABBBBABBBABABB
ProtLen = 56
LV = 112
stepLimit = 3000000
temperature = 0.1
savepathways = y
pathwaysstep = 3000
```

##
step 3) compile and execute the protein folding simulation
##

3.1) ONLY FOR CPU - Access the SRC_CPU directory and execute the command below to compile the MD program.

example

```
$ gcc -o a.out main_CPU.c func_MD_CPU.c -lm
```

3.2) Execute the MD program with the executable file (a.out) with the input file (2GB1.in) and a seed number, to generate a random initial structure of the simulation of the protein folding.

format

```
$ ./executable <input file> <seed>
```

example

```
$ ./a.out ../INPUT/2GB1_56.in 0 && mv pathways56_0.txt ../OUTPUT/pathways56_0.txt
```

after the execution of MD simulation, it will be generated a output file (pathway data) with the information about the protein structure, free energy and radius of gyration along the iterations (in our sample, the output file generated was pathways56_0.txt and we also move this file to the OUTPUT directory).

3.3) ONLY FOR GPU - Access the SRC_GPU directory and compile the parallel PathMolD-AB program

*The Parallel PathMolD-AB is written in C and CUDA.

example

```
$ nvcc --gpu-architecture = compute_61 --device -c main.c functions.cu
$ nvcc --gpu-architecture = compute_61 main.o functions.o
```

3.4) Execute the parallel PathMolD-AB program with the executable file (a.out) with the input file (2GB1.in), the output file(pathways56.txt), a seed number to generate a random initial structure of the simulation of the protein folding and the ID of the GPU that will run the simulation .

format

```
$ ./a.out <input file> <output file> <seed> > <GPU ID>
```

example

```
$ ./a.out ../INPUT/2GB1_56.in ../OUTPUT/pathways56 0 0
```

3.5) ONLY FOR GPU (METHOD WITH NEIGHBOURHOOD LIST) - Access the SRC_GPU_NL directory and compile the parallel PathMolD-AB(NL Version) program

*The Parallel PathMolD-AB(NL Version) is written in C and CUDA.

example

```
$ nvcc --gpu-architecture = compute_61 --device -c main.c functions.cu
$ nvcc --gpu-architecture = compute_61 main.o functions.o
```

3.6) Execute the parallel PathMolD-AB(NL Version) program with the executable file (a.out) with the input file (2GB1.in), the output file(pathways56_0.txt), a seed number to generate a random initial structure of the simulation of the protein folding and the ID of the GPU that will run the simulation

format

```
$ ./a.out <input file> <output file> <seed> > <GPU ID>
```

example

```
$ ./a.out ../INPUT/2GB1_56.in ../OUTPUT/pathways56 0 0
```

<input file>	: parameters of the simulation (generate by the step 2)
<output file>	: name of the output file/ pathway data
<seed>		: seed used to generate the initial structure of the protein
<GPU ID>	: ID of the GPU that will run the simulation. For computers with only one GPU, use the ID as zero.

*the root directory contains a makefile to run the programs.

```
$ make ab               # convert the AB sequence from the amino acid sequence
$ make cpu              # compile the sequential molecular dynamic version
$ make gpu              # compile the parallel molecular dynamic version
$ make gpu_nl           # compile the parallel molecular dynamic with neighbourhood list version
$ make run_cpu          # execute the sequential molecular dynamic version
$ make run_gpu          # execute the parallel molecular dynamic version
$ make run_gpu_nl       # execute the parallel molecular dynamic with neighbourhood list version
$ make vizualize        # generating the folding simulation video
```

Laboratory of Bioinformatics and Computational Intelligence

labic.utfpr.edu.br/

###
step 4) realize the protein folding analysis
###
*The pathway_print_multi-subplot.py script was developed in Python 2.7/3.6 language.

With the output file generated by the PathMolD-AB(NL Version) program (step 3), you can generate images of the protein folding trajectory with the program pathway_print_multi-subplot.py.

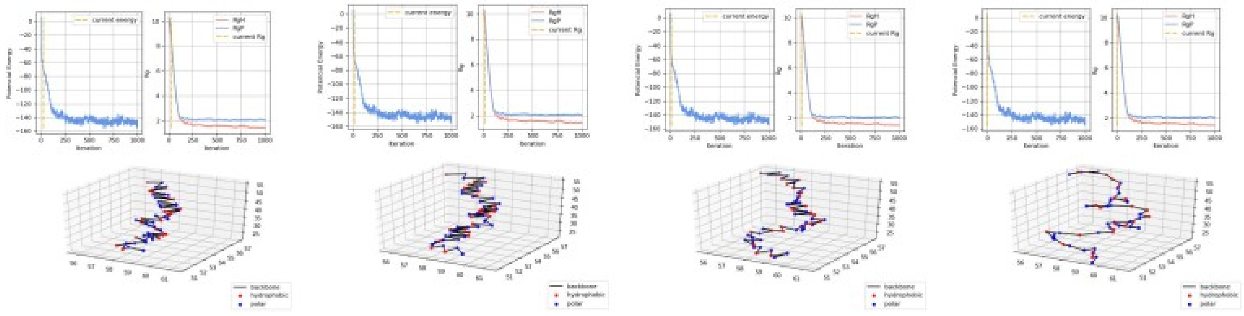
4.1) Before executing pathway_print_multi-subplot.py, you need to access the SRC_GPU_NL directory and set the parameters to read and save the files. Then, open pathway_print_multi-subplot.pyfile and change the variablespath_pathways , filename , filesequence and path_save .

path_pathways :folder of the protein pathway data
filename : name of the pathway data (generated by step 3)
filesequence : file containing the AB sequence of the protein (generated by step1)
path_save :folder where the images will be saved

4.2) Then in the same directory execute the command

```
$ python pathway_print_multi-subplot.py
```

Then, the program will produces images of the folding information at each iteration. The samples of the frame 0, 9, 18, and 27 are described below.



The image files of the folding process will be saved in the 'img' folder in this tutorial, and a video of protein protein folding will be generated called folding.mp4.

Thanks for using PathMolD-AB(NL Version)!!
----
for any doubt send a email to leandrotakeshihattori@gmail.com
Link for the Dataset of Spatiotemporal Pathways of Protein Folding:
https://mega.nz/#F!C5QkHQ6A!Ng2xowc2hVPoHHiSB7ww-w