NMRlipids databank

Anne Kiirikki[1,*] **and O. H. Samuli Ollila**[1,*]

[1]University of Helsinki, Institute of Biotechonology, Helsinki, Finland
[*]samuli.ollila@helsinki.fi

ABSTRACT

We present a databank of lipid bilayer simulations from the NMRlipids open collaboration project.

Introduction

The demand for sharing and reusing of MD simulation data is increasing, but practical solution remains unclear due to unsolved issues in data storage and indexing. Here we present a solution for lipid bilayers based on overlay databank structure.

Results

Quality evaluation of force fields

The quality of each simulation is evaluated against NMR order parameters. Indexing of experimental data to automatize evaluation is currently in progress by Anne Kiirikki.

Water permeability across membranes

Averaging water density profiles over all simulations in the databank enables us to calculate the barrier for the water penetration through lipid bilayers.

Spin relaxation rates of confined water close to membranes

Usefulness of the databank beyond MD simulation experts is demonstrated by analysing water spin relaxation times close to bilayers which are used in MRI imaging.

Discussion

Methods

Structure of the databank

NMRlipids databank is a overlay databank composed of index files containing information on the location of original data and all essential information on simulations needed in further use and automatic analysis. Each simulation system is identified using the hash of original trajectory and topology file. The index files are stored in folder structure based on the identity codes of the simulations.

Analysis of the data in databank

Index files in the datanbank contain all the essential information to perform analyses from the simulations. Systems under interest can be selected automatically browsing the index files and filtering the desired properties.

Analysis results can be most conveniently stored to a new datanbank with identical indexing as the original databank. The results can be then easily shared and browsed indentically to the original databank.

Indexing the simulation data

AddData.py is a script that builds a database that contains a dictionary file and analysis data of each simulation. The dictionary file contains information about the simulation. The script also calculates order parameters of all CH bonds of the lipids in the simulation. To add a simulation it must be first uploaded to Zenodo (www.zenodo.org). The trajectory and topology files of the simulation are downloaded to the working directory from Zenodo but these are not saved into the database. To add a simulation to the database the user has to give some essential information about the simulation. This is done by writing a info file (*INFO.py) which is passed to AddData.py.

AddData.py requires GROMACS and MDAnalysis library to be installed.

Run AddData.py as follows:

python3 AddData.py exampleINFO.py

There is also a script runAddData.sh that can be used to loop over several info files to add many simulations at one go.

A simulation dictionary contains information about the simulation. Some of the information is provided by the user. The numbers of lipid molecules, solvent and ions are automatically read from the files and so are the simulation temperature and trajectory length. All this information is saved to a file named README.yaml.

Compulsory user input

The following parameters are compulsory and have a strict format.

DOI

DOI is the DOI access number of the simulation. Use the version DOI on Zenodo.

MAPPING

MAPPING provides the name of the mapping file for a lipid. It must be given as a string of alternating lipid name and the corresponding mapping file name separated by a comma. The existing mapping files are in the directory named "mapping_files". If a mapping file of a certain lipid does not exist the user must construct it.

The purpose of a mapping file is to circumvent the problem caused by different atom naming conventions used by different force fields. The first column of a mapping file contains general atom names. The second column contains the name of the atom as it is in the force field. If the lipid consists of several residues which is the case with some AMBER force fields, then a third column is needed which contains the name of the residue to which each atom belongs to.

SOFTWARE

SOFTWARE stands for the name of the software used for running the simulation. The options are GROMACS, AMBER, NAMD, CHARMM and OPENMM. So far, only simulations run with GROMACS are accepted by the script.

TRJ

TRJ stands for the name of the trajectory file.

TPR

TPR stands for the name of the topology file.

PREEQTIME

PREEQTIME means the time used for pre-equilibrating the system. This should be in nano seconds.

TIMELEFTOUT

TIMELEFTOUT stands for the length of the simulation in nano seconds to be left out at the beginning of the simulation. This value is used in the script to leave out the given length from the analysis!

UNITEDATOM

The handling of united atom simulations is enabled by a separate script called buildH_calcOP.py. In case of a united atom simulation, the user has to give the names of the lipids and the corresponding names of those lipids as they are in the dic_lipids.py dictionary used by buildH_calcOP.py script. In case of an all atom simulation this can be given as an empty string.

Molecule names

The databank has default names for lipids, ions and solvent. The user must also provide the names of the molecules, ions and solvent that are used in the simulation to match the names used by the databank. The names provided by the user must be the same as in the tpr file. If a lipid consists of more than one residue the name of the head group residue is given. If the name of the molecule is not in the databank it needs to be added. The names of the molecules are listed in dictionaries called lipids_dict and molecules_dict which are in the script. If a new molecule is added it needs to be added to molecule_numbers_dict and molecule_ff_dict too.

dir_wrk

The user has to give the path of the working directory.

Free form user input
The form of how the values of these parameters are written is not essential for the script to work properly.

SYSTEM
SYSTEM is name of the system.

FF
FF is the name of the forcefield used in the simulation.

FF_SOURCE
FF_SOURCE tells where the forcefield parameters are taken from.

FF_DATE
FF_DATE is the date when the forcefield parameters were created. The format is day/month/year.

Individual force field names for molecules
The user can also give forcefield names to different molecules individually. These parameters are named as FFPOPC, FFPOT, FFSOL etc.

Automatically analyzed parameters
The following parameters are read automatically from the trajectory and topology files.

Molecule numbers
Numbers of lipid molecules (NPOPC, NPOPG, etc.) per membrane leaflet are calculated by determining on which side of the center of mass of the membrane the center of mass of the head group of each lipid molecule is located.

Numbers of other molecules such as solvent and ions (NSOL, NPOT, NSOD, etc.) are read from the topology file.

Temperature
Temperature of the simulation is read from the topology file.

Trajectory length
The length of a trajectory is read from the trajectory file.

Acknowledgements

Author contributions statement

Must include all authors, identified by initials, for example: A.A. conceived the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the manuscript.

Additional information