ST-CpHMD: Stochastic Titration Constant-pH Molecular Dynamics (version 4.1_GMX4.07)

www.itqb.unl.pt/simulation

September 14, 2020

Contents

1 License

2 Authors and citation

António M. Baptista Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Av. da República, 2780-157 Oeiras, Portugal.
baptista@itqb.unl.pt, www.itqb.unl.pt/~baptista

Miguel Machuqueiro Centro de Química e Bioquímica, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal.
machuque@ciencias.ulisboa.pt, mms.rd.ciencias.ulisboa.pt

Sara R. R. Campos Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Av. da República, 2780-157 Oeiras, Portugal.
scampos@itqb.unl.pt, www.itqb.unl.pt/simulation/members

Others (including testers) Catarina A. Carvalheda, Luís C. S. Filipe, Pedro R. Magalhães, Lucie Rocha, Vitor H. Teixeira, Diogo Vila-Viçosa

If you use the ST-CpHMD package in your work, please cite [2, 7, 9].

For help or bug reporting, contact us at cphmd@itqb.unl.pt.

3 Overview

This package, ST-CpHMD, implements the stochastic titration method developed by Baptista and co-workers to perform constant-pH molecular dynamics simulations. For details, see references [1–7].

Briefly, the method performs a molecular mechanics/molecular dynamics (MM/MD) simulation during which the protonation states of a titrable solute are regularly updated with new states provided by a combination of Poisson–Boltzmann (PB) free energy calculations

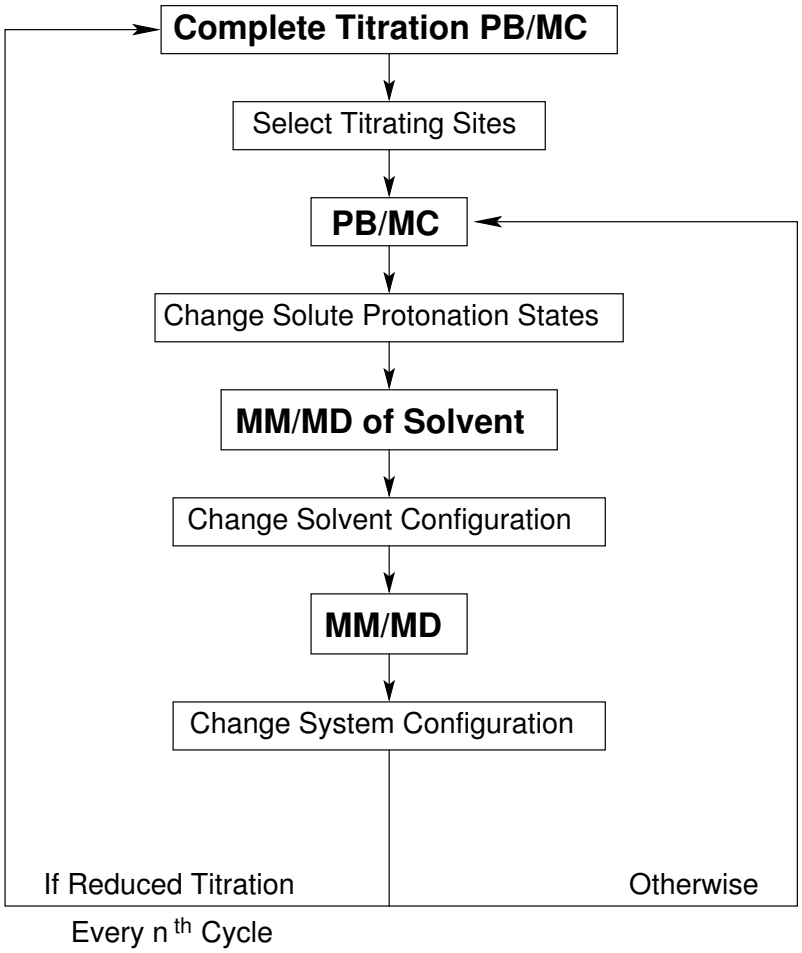

Figure 1: Scheme of the constant-pH MD algorithm.

and Monte Carlo (MC) sampling of protonation states. This alternation of MM/MD and PB/MC results in a three-step cycle that is repeated during the simulation (Figure 1):

1. PB/MC calculation to assign new protonation states to the solute titrable sites.

2. Short (e.g., 0.2 ps) MM/MD relaxation of the solvent (with "frozen" solute), which allows the solvent to adapt to the new protonation states.

3. Longer (e.g., 2 ps) MM/MD of the whole system, that generates a new sequence of structures with the assigned protonation states. Then return to step 1.

As discussed in ref. [1], this procedure generates a joint sample of structures and protonation states that is representative of a thermodynamic ensemble at a given pH value (a semi-grand ensemble).

To speed up the process, it is possible to use the *reduced titration* approach, also shown in Figure 1: every n^{th} cycle, a fixed protonation state is assigned to all the titrable sites whose

mean occupancies fall outside a predefined threshold. This creates an exclusion list with the sites that are titrating too far away from the pH of interest. See ref. [2] for details.

The current implementation uses proton tautomerism in the PB/MC calculations, where *tautomer* refers to the protonation state and geometry of a site. See references [4, 7, 8] for details.

This method does not require an explicit exchange of protons between the solute and the solvent. Instead, each titrable site has a permanent set of "dummy" hydrogen atoms required to produce all of its possible protonation states (including tautomeric forms), each of which is effectively realized by a specific combination of atomic charges and geometric parameters [1, 2]. This implies the existence of information and rules to produce the correct combinations, of two distinct kinds: (a) a modified MM force field with the protonation dependencies of the bonded and non-bonded parameters; (b) a set of PB-specific definitions of the titrable sites, with the protonation dependencies of the atomic charges and model pK_a values. See the description of directories `top` and `St-G54a7` in section 5. This distribution supports the titration of the side chains of Asp, Glu, Cys, Tyr, His, and Lys, besides the C- and N-terminal sites.

Even when a constant-pressure ensemble is selected, ST-CpHMD always runs the short MM/MD segment used for solvent relaxation (step 2 above) at constant volume. This avoids potential problems due to large "periodic solutes" that may loose contiguity upon changing box dimensions (e.g., a "frozen" lipid bilayer may get a gap near a box wall due to box expansion during solvent relaxation). Since any constant-volume ensemble distribution is a conditional case of the corresponding constant-pressure distribution (even at constant pH), any procedure that preserves the former distribution (such as the constant-volume solvent relaxation) also preserves the latter. Thus the overall sampling remains unaffected.

This version of ST-CpHMD can simulate proteins (or peptides, etc) in a membrane environment in contact with solvent, but it does *not* allow titrating lipids. For simplicity, the whole non-solvent part of the system is here designated as "solute". This version can also be used with solvent mixtures, but in that case make sure you know the dielectric constant of the mixture (to be given in the `pHmdp` file) and the pK_a of the model compounds in the mixture (you should provide your own `st` files). The files provided in the directory `St-G54a7` should be used only when the solvent is water.

The ST-CpHMD package uses parallelization when running the constant-pH MD simulations. This is done both at the CPU level (for MM/MD and PB/MC) and at the GPU level (for MM/MD only).

Required know-how The ST-CpHMD package is intended to be used by advanced users familiar with MM/MD and PB/MC methods, relying on external programs that implement those methods (see section 4). So, its use assumes these conditions:

- You should be well acquainted with performing standard MM/MD simulations (at fixed protonation). In particular, you should know how to use the GROMACS software package and have a good understanding of its molecular topologies.
- You should be well acquainted with performing standard PB/MC calculations (at

fixed structure). In particular, you should know how to use the meadTools and PETIT software packages to perform calculations with tautomerism.

If you do *not* fulfill these conditions, you will likely have difficulty using the ST-CpHMD package.

4 Installation and dependencies

To install the ST-CpHMD package, download its `tar` file from `www.itqb.unl.pt/simulation` and unpack it to a suitable path in your system.

In addition, you need also to have the following programs installed:

GROMACS 4.0.7 The software used for MM/MD. Actually, ST-CpHMD was only tested with a modified version of GROMACS 4.0.7 where ionic strength is implemented as an external parameter for the generalized reaction field (GRF) method; this version is available at our website: `www.itqb.unl.pt/simulation`. You can try different GROMACS versions (`www.gromacs.org`) at your own risk.

MEAD The software used for PB calculations. It used to be available at Don Bashford's webpage at `www.stjuderesearch.org/site/lab/bashford`, but that URL no longer exists. A copy of the original GPL-licensed distribution can be found at `www.itqb.unl.pt/simulation`.

meadTools 2.1 A set of tools to use MEAD, supporting tautomeric calculations. Available at `www.itqb.unl.pt/simulation`.

PETIT The program used to run Monte Carlo simulations of protonation states for a single-structure molecule. Available at `www.itqb.unl.pt/simulation`.

fixBox 1.2 A C program to properly assemble/center molecular systems inside a simulation box in an automated way, used here before running the PB calculations. It will be later made available as an independent program, but for now it is provided as part of the ST-CpHMD package, as an additional directory (see below).

The ST-CpHMD package uses scripts written in Bash and Gawk. So, make sure these are available in your system and are reachable from the assumed paths (see the "shebang" lines in those scripts). The package was tested on Linux systems only (mostly Ubuntu).

The Gnuplot package (`http://gnuplot.sourceforge.net`) is used in the tutorial to produce some plots, but you can replace it by other plotting software.

5 Package contents

`fixbox-1.2/` : Directory containing the fixBox program (see above). To install it, just compile the provided C source file, following the instructions in the program header. If you prefer, you can move the directory somewhere else, as long as you assign the correct path to the `$fixboxDIR` parameter in the `pHmdp` input file.

`LICENSE` : Text file with license.

`manual.pdf` : Package manual in PDF format (this file).

`scripts/` : Directory containing the package main scripts: `CpHMD.sh`, a Bash script with the main code; `update-top`, a Gawk script to update topologies.

`St-G54a7/` : Directory containing the `st` files for the GROMOS 54A7 force field, with model pK_a values in water calculated as explained elsewhere [7, 9]. These files are used by `meadTools` to perform the tautomeric PB calculations. If you want to use a solvent mixture, as in ref. [9], you must provide your own `st` files.

`templates/` : Directory containing examples of the input files that should be present at the location where the `CpHMD.sh` script will be run.

`tools/` : Directory containing tools useful for the setup of the system and the analysis of the simulations.

`top/` : Directory containing the force field files with the modifications required to perform constant-pH MD with tautomers. The currently supported force field is GROMOS 54A7 (using the file basename `G54a7pHt`). In addition to the usual GROMACS force field files, there is a dictionary (`dic`) file defining the protonation-dependent atomic charges, atom types and bonded parameters required by each site type; this `dic` file is used by `update-top` to update the topology file.

`tutorial/` : Directory with a tutorial that provides a full example of a constant-pH MD simulation, including file preparation, initialization, production and analysis. The model system is hen egg white lysozyme (as obtained from the PDB entry 4LZT), which is simulated in water at pH 7.0.

6 Input/output

6.1 File naming

Since this implementation consists of a cyclic series of stop-and-go runs of different programs (GROMACS, MEAD and PETIT), GROMACS checkpoint (`cpt`) files become largely useless. Instead, in order to avoid loosing data and to keep things manageable in case of crashed runs, it is convenient to run constant-pH MD simulations as a sequence of short blocks, as was typically done before checkpoint files were introduced in GROMACS. Thus, the adopted rationale is to define a run name string `<runname>` and a block index string `<index>` that are used to name both general files (usually `<runname>.<extension>`) and block-specific ones (usually `<runname>_<index>.<extension>`).

Both `<runname>` and `<index>` are inferred from the name of the block-specific `pHmdp` file (see below), which is assumed to be named as `<runname>_<index>.pHmdp`. So, if the file `lyso_001.pHmdp` is provided as input, the program will infer that `<runname>` is equal to `lyso` and that `<index>` is equal to `001`, deriving from that the names of other files. Thus,

it will read information from input files `lyso.mdp`, `lyso.mgm`, etc, which are common to the different simulation blocks, and will generate block-specific files `lyso_001.edr`, `lyso_001.gro`, etc. All these different files are described below in detail and listed in Table 1 for easy reference.

If your simulation is fast enough to be run in a single block (i.e., if you can easily afford to fully repeat it in case of a crash), you may discard the block index, which is then ignored. In that case, all `<runname>_<index>` occurences given here should be replaced with `<runname>`.

6.2 Input files

The input files required by the script `CpHMD.sh` are essentially of three types: those required for MM/MD (by GROMACS), those required for PB/MC (by meadTools/MEAD and PETIT), and a main parameter file where many other settings are done. The following input files are required to be explicitly present and to strictly adhere to the naming convention explained in section 6.1:

`<runname>_<index>.pHmdp` : This file is the single argument of the `CpHMD.sh` executable and contains all the parameter information needed to run the program. It must be edited according to the user's needs; e.g., take a copy of the example given in the `templates/` directory, rename it and change the parameters while keeping the syntax. The file uses Bash syntax, meaning that you can use all the features of Bash, such as parameter expansion, command expansion, etc. As explained in section 6.1, the name of this file is used to derive the general `<runname>` and the block-specific `<index>`.

The parameters in this file include physical parameters (e.g., `pH=7.0`), paths of required programs (e.g., `PetitDIR=/programs/petit-1.6`), input file locations (e.g., `GROin=MyGROFile.gro`), etc. A complete list and description of all parameters is given in section 7.1. When referred to in this manual, these parameters will be prefixed with the character $ (e.g., `$pH`, `$PetitDIR`, `$GROin`).

`<runname>.{ogm,mgm}` : These files[1] contain the grid information for MEAD calculations of the solute (`ogm`) and of the model compound (`mgm`). They have to be present in the running directory. They must be provided by the user if the parameter `$GridSize` is set to zero. Otherwise, they will be automatically generated (see section 7.2). The format of these files is the one used by the MEAD program (see MEAD's `README` file for more information).

`<runname>.sites` : This file is the `sites` file to be used by meadTools/MEAD. It has to be present in the running directory. You may use the program `makesites` in the `meadTools` package to produce this file. Keep in mind that when running the PB/MC calculations with tautomers, the `sites` file must include all alternative tautomers, differing from the traditional format (see MEAD's and meadTools's `README` files for more information).

[1] When convenient, multiple file names are here designated using the notation of Bash brace expansion. In particular, `<runname>.{ogm,mgm}` stands for the two files `<runname>.ogm` and `<runname>.mgm`.

<runname>.mdp : This file is the dynamics mdp file as used by GROMACS. It has to be present in the running directory. Some of the parameters defined in the pHmdp file will override parameters in this mdp file (e.g., see the mdp file given in directory templates/).

You may include ionic strength as an external parameter if you use our modified GROMACS 4.0.7 version (see section 4). In that case, this mdp file should have the following lines:

```
coulombtype = Generalized-Reaction-Field
ionicstrength = 0.09033 ; Overridden with the value from .pHmdp
```

<runname>.mdf : When $use_fixbox is set to y, this file must be present in the running directory. It is an input file for the fixBox program with all the instructions for assembling/centering the system and correcting PBCs before PB calculations. The user is responsible for providing these sets of instructions (which are case-dependent; see fixBox documentation for details). Do *not* include the solvent part in <runname>.mdf.

It should be noted that, in addition to these, other input files are implicitly given through the parameters in the pHmdp file. In particular, input files in GROMACS-formats gro, top and (optionally) ndx are specified through, respectively, the parameters $GROin, $TOPin and $NDXin, while the MM force field directory is given through $ffDIR. In addition, $StDIR indicates the directory containing the st files for the PB/MC calculations. See section 7.1 for details on these parameters.

6.3 Main output files

The main output files created by CpHMD.sh are essentially of three types: those referring to MM/MD (in GROMACS formats, produced either by grompp or mdrun), those referring to PB/MC, and an info file. Following the naming convention explained in section 6.1, the program produces these files:

<runname>_<index>.gro : A GROMACS-format gro file with the last structure frame generated by this simulation block.

<runname>_<index>.xtc : A GROMACS-format xtc file with the trajectory generated by this simulation block.

<runname>_<index>.edr : A GROMACS-format edr file with "energetic" info (energy, temperature, pressure, etc) about this simulation block.

<runname>_<index>.tpr : A GROMACS-format tpr file combining the system topology with the parameters, coordinates and velocities used to start the *last cycle* of this simulation block (see Figure 1). So, it does *not* correspond to the initial frame of the block trajectory, as would typically be the case in a standard GROMACS run.

<runname>_<index>.occ : File containing the protonation states (including tautomeric form) of all titrable sites, as generated at the end of the MC simulation (done with PETIT), which are then used by MM/MD during each simulation cycle (see Figure 1). It is an ASCII file with one line per cycle, and each line has one column per site, with the sites ordered as in the `<runname>.sites` input file. The protonation state of each site is represented by an integer number following the correspondence indicated in Table 2, and the key "empty" or "occupied" hydrogens for each tautomer are given in Table 3. So, each line corresponds to a full specification of the *global* protonation state used in that cycle.

<runname>_<index>.mocc : File containing the mean proton occupancies of all titrable sites, as obtained from the MC simulation (done with PETIT) during each simulation cycle (see Figure 1). It is an ASCII file with one line per cycle, and each line has one column per site, with the sites ordered as in the `<runname>.sites` input file. The mean proton occupancy of each site is a real number between 0 (fully deprotonated) and 1 (fully protonated). Note that there should good statistical agreement between these mean proton occupancies and the actual protonation states used during MM/MD, i.e., those in the `occ` file (e.g., see the plots produced in the `analysis/occ/` directory of the tutorial).

<runname>_<index>.info : File containing information about the user, host and start and end dates/times of this block, as well as a trace record of the start and end times/dates of the different steps of each cycle (see Figure 1).

It should be noted that the number of entries in the structural (e.g., `xtc`) output files is *not* necessarily the same as those in the protonation (e.g., `occ`) output files. For example, if you save frames twice every full (solute + solvent) MM/MD segment, which is itself performed once every cycle (see Figure 1), you will end up with two frames for each global protonation state. Understanding this relation is crucial to properly match structures with protonations during analysis (e.g., see the discussion of the use of the `statepdb` tool in the `analysis/occ/` directory of the tutorial).

6.4 Extra/debugging output files

Besides the main output files just described, `CpHMD.sh` produces some extra files intended mainly for debugging purposes or to trace in more detail the automated interface between MM/MD and PB/MC.

CpHMD_<runname>_<index>.mdp : File in GROMACS `mdp` format with the parameters used in the full system (solute + solvent) MM/MD step.

CpHMD_<runname>_<index>_SR.mdp : File in GROMACS `mdp` format with the parameters used in the solvent relaxation MM/MD step.

CpHMD_<runname>_<index>.ndx : File in GROMACS `ndx` (index) format containing the definitions of all relevant groups, including the group `PBMC` corresponding to the solute

part (protein, protein + membrane, etc) that is used in PB/MC calculations and is kept frozen during solvent relaxation. See also the discussion of the parameter $NDXin.

CpHMD_<runname>_<index>.log : File obtained by appending all log files produced by mdrun in the full system MM/MD steps.

CpHMD_<runname>_<index>.pqr : File obtained by appending all pqr files used in the PB/MC calculations (one per cycle). So, the number of frames in this pqr trajectory is equal to the number of lines in the occ and mocc files and, thus, potentially different from the number of frames in the xtc trajectory (see end of section 6.3).

CpHMD_<runname>_<index>.sites : This file is created when reduced titration is being used, containing the sites selected in each complete titration cycle and used in the subsequent cycles.

7 Parameters and settings

Most settings of CpHMD.sh are done through the parameters given in the pHmdp file. A few others are currently hard-wired in the code.

7.1 pHmdp parameters

The pHmdp file can contain the assignment of the following parameters (in alphabetic order):

bsize (default 10000) **:** Number of pseudo-site blocks to use in meadT (see meadTools documentation).

CpHDIR (mandatory) **:** Path of the ST-CpHMD package directory.

EffectiveSteps (mandatory) **:** Number of steps for the whole system dynamics. Corresponds to (and overrides) the parameter nsteps in the GROMACS mdp file.

EndCycle (mandatory) **:** Last cycle index. See example given for $InitCycle.

epsin (default 2) **:** Dielectric constant of molecular interior in PB calculations.

epssol (default 80) **:** Dielectric constant of the solvent in PB calculations.

ffDIR (default $CpHDIR/top) **:** Path to the directory with the force field files.

ffID (mandatory) **:** The force field ID string used in its files. In the current version it can take only the value G54a7pHt, corresponding to the provided force field (see section 5).

fixboxDIR (mandatory when $use_fixbox is y) **:** Path of the directory where the fixbox binary is located. See sections 4 and 5.

GridSize (default 0) **:** Parameter used to automatically generate the files <runname>.ogm and <runname>.mgm. When set to zero, those files are expected to be provided in the running directory. Otherwise, the files are created as described in section 7.2, using the (non-zero) value of GridSize as the dimension (number of nodes) of the solute grid in the first focusing level, assumed to be centered the solute geometric center and have a grid spacing of 1 Å; thus, GridSize should be large enough to make that first grid extend sufficiently beyond the solute. For more details, see MEAD's README file and section 7.2. **Warning:** Note that a non-zero value of GridSize will override any potentially existing <runname>.ogm and <runname>.mgm files.

GroDIR (mandatory) **:** Path of the GROMACS binaries. As mentioned in section 4, the current version of ST-CpHMD was tested only with a modified version of GROMACS 4.0.7 that includes the ionic strength in the molecular dynamics steps (with electrostatics being treated with a generalized reaction field), consistently with the continuum electrostatics part. This version is available at www.itqb.unl.pt/simulation. See also System preparation.

GROin (mandatory) **:** Initial gro file. The solvent molecules should appear in block at the end of the gro file, after the solute (e.g., protein), starting with the first residue whose name is equal to the parameter $SOL1st. The atoms of the solvent block thus identified are the ones subjected to MM/MD during the solvent relaxation stage (see Figure 1).

InitCycle (mandatory) **:** Index of the first cycle in this block. Together with $EndCycle, it allows keeping track of simulation time. For example, in a simulation with two blocks of 500 cycles, the first block has InitCycle=1 and EndCycle=500, and the second block has InitCycle=501 and EndCycle=1000.

ionicstr (mandatory) **:** Ionic strength in mol/L. This parameter is used in the PB calculations (MEAD) and in MM/MD when using GRF and our modified GROMACS 4.0.7 version where ionic strength is implemented as an external parameter (see section 4); it overrides the mdp parameter with the same name defined in this modified version.

mdrun (mandatory) **:** Command for running the GROMACS command mdrun. For instance, you may want to use mdrun="$GroDIR/mdrun" when using a single processor, and mdrun="mpirun.openmpi -np $nCPU_MD $GroDIR/mdrun_mpi" or mdrun="mpirun.lam -ssi rpi tcp C $GroDIR/mdrun_mpi" when using parallelization.

MeadDIR (mandatory) **:** Path of the MEAD binaries.

MToolsDIR (mandatory) **:** Path of the meadTools package directory. As indicated in section 4, the current version of ST-CpHMD works with meadTools 2.1.

nCPU (mandatory) **:** Number of CPUS that will run the PB/MC calculations.

nCPU_MD (default $nCPU) **:** Number of CPUS that will run the MD simulation.

NDXin (optional) **:** The index file of your system. When the NDXin variable is left empty or the file does not exist, an index file will be generated by make_ndx (GROMACS). Though this parameter is optional, bear in mind that the index file made by the CpHMD.sh program may not stand in all situations. This may cause the program to crash when running grompp (GROMACS). In that case, please provide your own index file; it can be useful when simulating solvent mixtures, for instance. The ndx file to be used by CpHMD.sh must include a group PBMC containing the set of solute atoms to be explicitly treated in the PB/MC calculations (protein, protein + membrane, etc); if an ndx file is provided and this group is not present, it is added assuming that it starts at the first atom and ends right before the first residue whose name is equal to the parameter $SOL1st. See also $GROin.

NRES (mandatory) **:** Number of the last residue in the merged protein. It is used to calculate the offset for stmodels (in meadTools) .

PetitDIR (mandatory) **:** Path of the PETIT package directory.

pH (mandatory) **:** pH of the solution.

PosRe (optional) **:** When using position restraints, it is convenient to remove them during the solvent relaxation step in which the protein and lipid are frozen. This parameter is a string signaling the lines to be absent from the mdp file during solvent relaxation, e.g. -DPOSRES.

RelaxSteps (mandatory) **:** Number of steps for the solvent relaxation dynamics. Corresponds to (and overrides) the parameter nsteps in the GROMACS mdp file.

RTFrequency (mandatory) **:** Reduced titration frequency, corresponding to number of cycles between two consecutive complete titrations. Only relevant if $RTThreshold is not 0.

RTThreshold (mandatory) **:** Reduced titration threshold. If 0, reduced titration will not be performed. Otherwise, if the fraction of the most populated protonation state for a certain site is bigger than $1 - $ $RTThreshold in the complete titration, this state will be attributed to this site, which will not be considered titratable until the next complete titration occurs.

seed (default 1234567) **:** Random seed used in program PETIT. It can be useful to generate different replicates.

SOL1st (default SOL) **:** The residue name of the first solvent molecule in the input gro file (given through $GROin). It starts the solvent block that remains implicit in the PB calculations and will be relaxed after the assignment of new protonation states to the solute (see also comments on $GROin).

StDIR (default $CpHDIR/St-G54a7) **:** Path of the st files directory.

temp (mandatory) **:** Temperature in Kelvin. It overrides the temperature (`ref_t`) in the `mdp` file.

TOPin (mandatory) **:** Topology of your system (`top` file). As in GROMACS, the included files must be present in the locations indicated in the `top` file. This topology must be created using the modified force-field residue names (see section System preparation).

use_fixbox (default `y`) **:** It can take the values `y` or `n` (yes/no). If `y`, the program `fixbox` is used to correctly center the system before the PB calculations and you need to provide the file `<runname>.mdf` with the necessary instructions for `fixbox`. If `n` is chosen, the tool `trjconv` with option `-pbc mol` is used to correct the PBC (this is adequate to treat only a single-chain solute).

7.2 Other settings

Some settings of the PB/MC calculations are currently pre-defined:

- When the `mgm` and `ogm` files are automatically generated (see `$GridSize`), two levels of focusing are used with spacings of 1 and 0.25 Å. This is the content of the `mgm` file:

  ```
  ON_GEOM_CENT 61 1.0
  ON_CENT_OF_INTR 65 0.25
  ```

  The `ogm` file will depend on the value of `$GridSize`:

  ```
  ON_GEOM_CENT $GridSize 1.0
  ON_CENT_OF_INTR 65 0.25
  ```

  For details on the format of these files, see MEAD's `README` file.

- The molecular surface is defined by a rolling probe of radius 1.4 Å and the Stern layer is 2 Å. For details, see MEAD's `README` file.

- Each MC run performs 10^5 MC steps, after 10^3 steps of equilibration. Each MC step consists of a cycle of random choices of state for all sites and pairs of sites with a coupling above 2 pK_a units. For details, see PETIT's `README` file.

8 Usage of ST-CpHMD

8.1 System preparation

Before starting ST-CpHMD production runs it is expected that the system (`gro` and `top` files) has been correctly initialized (please consult the GROMACS manual). It is necessary that the `gro` and `top` files account for the presence of all proton positions. This can be easily made by performing the GROMACS setup steps in the presence of the modified constant-pH MD force field, present in the `CpHMD/top` directory. The initial `pdb` file that is given to `pdb2gmx`

must contain the names of the titrable residues correctly altered (HIX, ASX, GLX, etc). You should use the tool update-top to edit the top file according to the desired protonation states.

In the current version, a correction is needed to have the C-terminus correctly built after pdb2gmx (if not capped). That correction can be done on the top file using the scripts fixCTER and fixCTER_angle in directory tools. You can also try to make this correction by simply editing the top file. You should have in the end: i) the second oxygen of the C-terminus carboxyl correctly named O2, and ii) the angle defined by C, CA, and O2 of the C-terminus corrected according to the assigned charge state (and the atom type of O2).

After this correction, the top file can be used as usual in the following minimization/equilibration steps and, later on, it can be given as the input file $TOPin in <runname>.pHmdp.

If PME is chosen over GRF/RF for long-range electrostatic treatment, the system should contain an adequate amount of counter-ions that brings the total charge closer to neutrality at a given pH. Obviously, one needs to run a short ST-CpHMD simulation in order to determine an average charge of the protein for each pH value. This means that exact electroneutrality would be ensured indirectly through PME's background charge.

8.2 Running constant-pH simulations

After initializing the system and preparing all the input files needed (see section 6.2), you are ready to run the constant-pH simulations. The program that runs the simulations is called CpHMD.sh and you can find it inside the directory scripts (see section 5). Simply run the following command in the presence of all necessary input files.

```
CpHMD.sh <runname>.pHmdp
```

8.3 Typical workflow

1. Run pdb2gmx (GROMACS) creating gro and top files accounting for all proton positions.

   a) Have the modified force field in your directory

   ```
   ln -s <CpHMD_PATH>/top/* .
   ```

   b) Change the name of the titrating residues in your pdb file to the name of ST-CpHMD building blocks. In this example, Asp and Glu will be titrating.

   ```
   sed 's/ASP/ASX/g;s/GLU/GLX/g' <your_protein>.pdb \
                                 > <your_protein_ready>.pdb
   ```

   You can also use the tool convert-pdb in directory tools together with the file convert.def.

c) Run pdb2gmx as usual, using the pdb file made in the previous step.

d) Correct the C-terminus in the topology as explained above (System preparation).

2. Perform a PB/MC calculation at the desired pH value to determine initial protonation states. You can use the tool update-top, in directory scripts, to update the topology file according to the assigned new charge states and the rules in the dictionary (dic) file (see section 5). update-top expects an input file with the assigned protonation states in the 1st column, residue number in the 2nd and residue name in the 3rd, e.g.

```
# Create a file with the final MC states outputed by petit
# (don't be scared of the following command lines, the result
# is shown next and you can make this as you please):

r2x="/NTPRO/{b};/NTGLY/{b};s/NT.*/NT/;s/CT.*/CT/"

gawk '/^f /{for(i=2;i<=NF;i++)print $i}' petit_out \
     | paste - <your_protein>.sites \
     | gawk '{print gensub(/([[:upper:]]+).*/, "\\1","g")}' \
     | sed "$r2x" \
     > <your_protein>.states

# Check <your_protein>.states.
# Residue name in agreement with force field file
# dictionary.dic.

head -n 5  <your_protein>.states
3          1 NT
3          1 LYS
4          7 GLU
3         13 LYS
1         15 HIS


#Update the topology:

update-top <your_protein>.states ffG54a7pHt.dic \
           <your_protein>.top > <your_protein_ready>.top
```

3. Run the minimization and equilibration steps as usual, in the presence of the modified force field.

4. Write the parameters file <runname>.pHmdp.

5. Run CpHMD.sh in the presence of <runname>.sites, <runname>.mdp, <runname>.mdf and, optionally, <runname>.{o,m}gm (see section 6).

```
CpHMD.sh <runname>.pHmdp
```

8.4 Tutorial

A full example of the use of ST-CpHMD, including some analyses, can be found in the directory `tutorial/`.

9 Additional information

Table 1: Input and output files of ST-CpHMD

File name	Description
Input files:	
`<runname>_<index>.pHmdp`	main parameter file
`<runname>.ogm`	grid file for solute
`<runname>.mgm`	grid file for model compounds
`<runname>.sites`	sites to be titrated
`<runname>.mdp`	MM/MD parameters
`<runname>.mdf`	"centering" definitions
Main output files:	
`<runname>_<index>.tpr`	`tpr` for last cycle in block
`<runname>_<index>.gro`	`gro` with block's last frame
`<runname>_<index>.edr`	`edr` info
`<runname>_<index>.xtc`	`xtc` trajectory
`<runname>_<index>.occ`	protonation states
`<runname>_<index>.mocc`	mean proton occupancies
`<runname>_<index>.info`	general info file
Extra/debugging output files:	
`CpHMD_<runname>_<index>.mdp`	`mdp` for full system MM/MD
`CpHMD_<runname>_<index>_SR.mdp`	`mdp` for solvent relaxation MM/MD
`CpHMD_<runname>_<index>.ndx`	`ndx` used in this block
`CpHMD_<runname>_<index>.log`	append of full-system `log` files
`CpHMD_<runname>_<index>.pqr`	`pqr` trajectory
`CpHMD_<runname>_<index>.sites`	sites selected from full PB/MC

Table 2: Residue and occ designations for tautomeric and non-tautomeric assignments.

Residue	Residue Name	occ designation						
		tautomeric					non-tautomeric[1]	
		tau1	tau2	tau3	tau4	charged	neutral	charged
His	HIX	0	1	–	–	2	0	1
Lys	LYX	0	1	2	–	3	0	1
Asp	ASX	0	1	2	3	4	0	1
Glu	GLX	0	1	2	3	4	0	1
Cys	CYX	0	1	2	–	3	0	1
Tyr	TYX	0	1	–	–	2	0	1
Arg	ARX[1]	0	1	2	3	4	0	1
Termini[2]								
Nter	–	0	1	2	–	3	0	1
NterP	–	0	1	–	–	2	0	1
Cter	–	0	1	2	3	4	0	1

[1] Not present in this version.

[2] Nter, NterP and Cter are, respectively, the N-terminus, the proline N-terminus, and the C-terminus.

Table 3: Key atom names for correct assignment.

Residue[2]	key atoms[1]			
	tau 1	tau 2	tau 3	tau 4
Nter	H1	H2	H3	–
NterP	H1	H2	–	–
His	HE2	HD1	–	–
Lys	HZ1	HZ2	HZ3	–
Cter	HO11	HO21	HO12	HO22
Asp	HD11	HD21	HD12	HD22
Glu	HE11	HE21	HE12	HE22
Cys	HG1	HG2	HG3	–
Tyr	HH1	HH2	–	–

[1] Key atoms are the positions where the proton is either introduced (for the anionic sites: Tyr, Cys, Glu, Asp and Cter) or removed (for the cationic sites: Nter, NterP, His and Lys).

[2] Nter, NterP and Cter are, respectively, the N-terminus, the proline N-terminus, and the C-terminus.

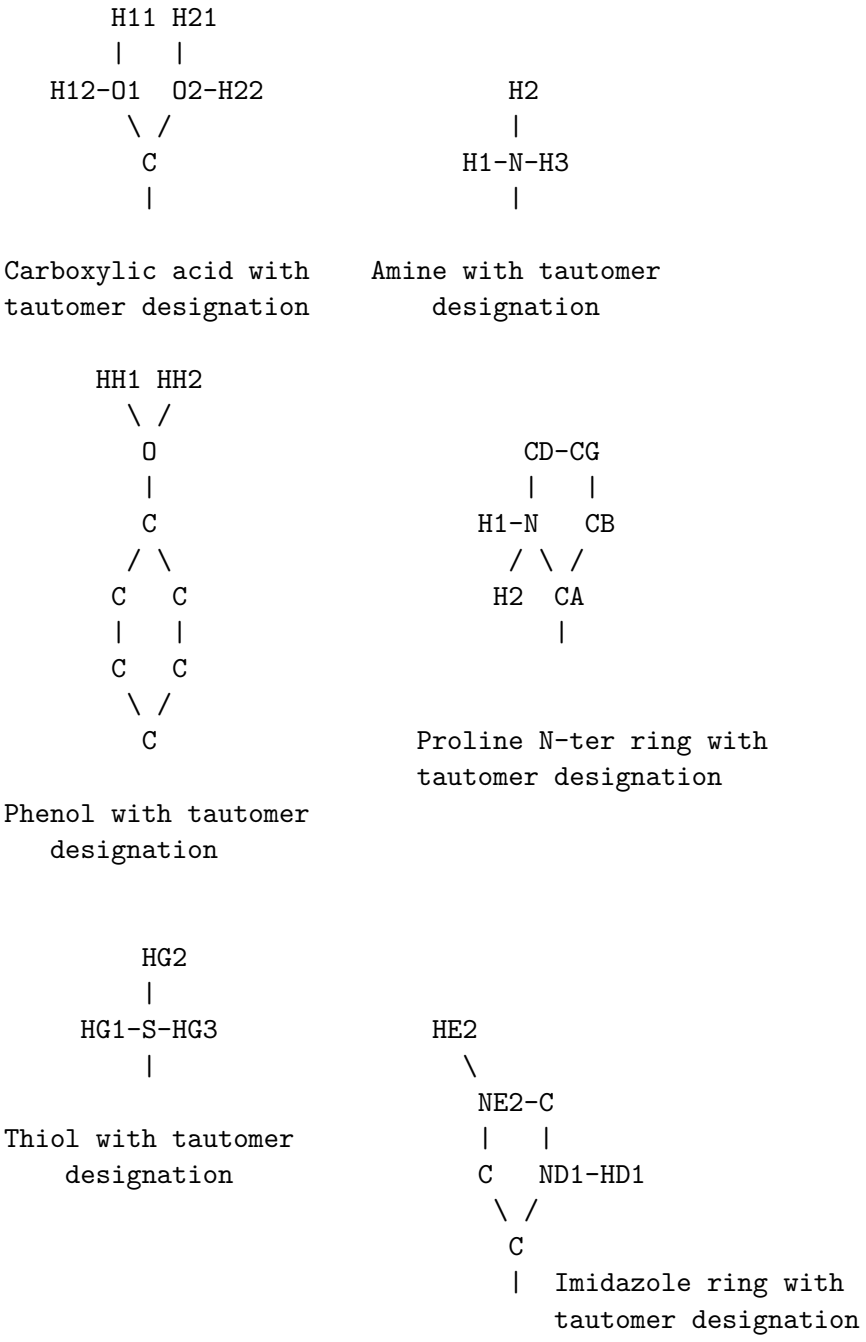

Figure 2: Graphic representation of generic tautomeric sites.

References

[1] Baptista, A.M., Teixeira, V.H., Soares, C.M. (2002) "Constant-pH molecular dynamics using stochastic titration", J. Chem. Phys. 117, 4184–4200.

[2] Machuqueiro, M. and Baptista, A.M. (2006) "Constant-pH molecular dynamics with ionic strength effects: protonation-conformation coupling in decalysine", J. Phys. Chem B 110, 2927–2933.

[3] Machuqueiro, M. and Baptista, A.M. (2007) "The pH-dependent conformational states of kyotorphin: a constant-pH molecular dynamics study", Biophys. J. 92, 1836–1845.

[4] Machuqueiro, M. and Baptista, A.M. (2008) "Acidic range titration of HEWL using a constant-pH molecular dynamics method", Proteins Struct. Funct. Bioinf. 72, 289–298.

[5] Machuqueiro, M. and Baptista, A.M. (2009) "Molecular dynamics at constant pH and reduction potential: application to cytochrome c_3", J. Am. Chem. Soc. 131, 12586–12594.

[6] Campos, S.R.R., Machuqueiro, M. and Baptista, A.M. (2010) "Constant-pH molecular dynamics simulations reveal a beta-rich form of the human prion protein", J. Phys. Chem. B 114, 12692–12700.

[7] Machuqueiro, M. and Baptista, A.M. (2011) "Is the prediction of pK_a values by constant-pH molecular dynamics being hindered by inherited problems?", Proteins Struct. Funct. Bioinf. 79, 3437–3447.

[8] Baptista, A.M., Soares, C.M. (2001) "Some theoretical and computational aspects of the inclusion of proton isomerism in the protonation equilibrium of proteins", J. Phys. Chem. B 105, 293–309.

[9] Carvalheda, C.A., Campos, S.R.R., Machuqueiro, M. and Baptista, A.M. (2013) "Structural Effects of pH and Deacylation on Surfactant Protein C in an Organic Solvent Mixture: A Constant-pH MD Study", J. Chem. Inf. Model. 53, 2979–2989; Correction in: J. Chem. Inf. Model. 2015, 55, 206.