Bray *et al. J Cheminform* (2020) 12:54
https://doi.org/10.1186/s13321-020-00451-6

Journal of Cheminformatics

EDUCATIONAL

Open Access

Intuitive, reproducible high-throughput molecular dynamics in Galaxy: a tutorial

Simon A. Bray[1], Tharindu Senapathi[2], Christopher B. Barnett[2*] and Björn A. Grüning[1*]

Abstract

This paper is a tutorial developed for the data analysis platform Galaxy. The purpose of Galaxy is to make high-throughput computational data analysis, such as molecular dynamics, a structured, reproducible and transparent process. In this tutorial we focus on 3 questions: How are protein-ligand systems parameterized for molecular dynamics simulation? What kind of analysis can be carried out on molecular trajectories? How can high-throughput MD be used to study multiple ligands? After finishing you will have learned about force-fields and MD parameterization, how to conduct MD simulation and analysis for a protein-ligand system, and understand how different molecular interactions contribute to the binding affinity of ligands to the Hsp90 protein.

Keywords: Galaxy, Molecular Dynamics, Reproducible

Introduction

Molecular dynamics (MD) is a commonly used method in computational chemistry and cheminformatics, in particular for studying the interactions between small molecules and large biological macromolecules such as proteins [1]. However, the barrier to entry for MD simulation is high; not only is the theory difficult to master, but commonly used MD software is technically demanding. Furthermore, generating reliable, reproducible simulation data requires the user to maintain detailed records of all parameters and files used, which again poses a challenge to newcomers to the field. One solution to the latter problem is usage of a workflow management system such as Galaxy [2], which provides a selection of tools for molecular dynamics simulation and analysis [3]. MD simulations are rarely performed singly; in recent years, the concept of high-throughput molecular dynamics (HTMD) has come to the fore [4, 5]. Galaxy lends itself well to this kind of study, as we will demonstrate in this paper, thanks to features allowing construction of complex workflows, which can then be executed on multiple inputs in parallel.

This tutorial provides a detailed workflow for high-throughput molecular dynamics with Galaxy, using the N-terminal domain (NTD) of Hsp90 (heat shock protein 90) as a case-study. Galaxy [2] is a data analysis platform that provides access to thousands of tools for scientific computation. It features a web-based user interface while automatically and transparently managing underlying computation details, enabling structured and reproducible high-throughput data analysis. This tutorial provides sample data, workflows, hands-on material and references for further reading. It presumes that the user has a basic understanding of the Galaxy platform. The aim is to guide the user through the various steps of a molecular dynamics study, from accessing publicly available crystal structures, to performing MD simulation (leveraging the popular GROMACS [6, 7] engine), to analysis of the results.

The entire analysis described in this article can be conducted efficiently on any Galaxy server which has the needed tools. In particular, we recommend using the Galaxy Europe server (https://cheminformatics.usega

*Correspondence: chris.barnett@uct.ac.za; gruening@informatik.uni-freiburg.de
[1] Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, Freiburg, Germany
[2] Department of Chemistry and Scientific Computing Research Unit, University of Cape Town, 7700 Cape Town, South Africa

laxy.eu) or the Galaxy South Africa server (https://galaxy-compchem.ilifu.ac.za). For users who wish to run their own Galaxy server locally, we provide a Docker container (https://quay.io/repository/galaxy/computational-chemistry-training) containing a full Galaxy installation, with all tools required for the tutorial preinstalled.

The tutorial presented in this article has been developed as part of the Galaxy Training Network [8] and its most up-to-date version is accessible online on the Galaxy Training Materials website [9], under the URL https://training.galaxyproject.org/training-material/topics/computational-chemistry/tutorials/htmd-analysis/tutorial.html.

What is high-throughput molecular dynamics?

Molecular dynamics (MD) is a method to simulate molecular motion by iterative application of Newton's laws of motion. It is often applied to large biomolecules such as proteins or nucleic acids. A common application is to assess the interaction between these macromolecules and a number of small molecules (e.g. potential drug candidates). This tutorial provides a guide to setting up and running a high-throughput workflow for screening multiple small molecules, using the open-source GROMACS tools provided through the Galaxy platform. Following simulation, the trajectory data is analyzed using a range of tools to investigate structural properties and correlations over time.

Why is Hsp90 interesting to study?

The 90 kDa heat shock protein (Hsp90) is a chaperone protein responsible for catalyzing the conversion of a wide variety of proteins to a functional form; examples of the Hsp90 clientele, which totals several hundred proteins, include nuclear steroid hormone receptors and protein kinases [10]. The mechanism by which Hsp90 acts varies between clients, as does the client binding site; the process is dependent on post-translational modifications of Hsp90 and the identity of co-chaperones which bind and regulate the conformational cycle [11].

Due to its vital biochemical role as a chaperone protein involved in facilitating the folding of many client proteins, Hsp90 is an attractive pharmaceutical target. In particular, as protein folding is a potential bottleneck to cellular reproduction and growth, blocking Hsp90 function using inhibitors which bind tightly to the ATP binding site of the NTD could assist in treating cancer; for example, the antibiotic geldanamycin and its analogs are under investigation as possible anti-tumor agents [12, 13].

In the structure which will be examined during this tutorial, the ligand of concern is a resorcinol, a common class of compounds with affinity for the Hsp90 N-terminal domain. It is registered in the PubChem database under the compound ID 135508238 [14]. As can be seen by viewing the PDB structure, the resorcinol part of the structure is embedded in the binding site, bound by a hydrogen bond to residue aspartate-93. The ligand structure also contains a triazole and a fluorophenyl ring, which lie nearer to the surface of the protein.

Methods: simulation

This section guides the reader through the step-by-step process required to prepare, run and analyze Hsp90. A brief explanation of the theory and purpose of each step is provided. Refer to the hands-on sections as these describe the task with tools and parameters to be carried out using Galaxy.

Get data

Create a new Galaxy history and then download a crystal structure for the Hsp90 protein from the Protein Data Bank (PDB). The structure is provided under accession code `6HHR` [16] and shows Hsp90 in complex with the resorcinol ligand (Fig. 1).

Hands-on 1: Data upload

1. Create a new history for this tutorial
2. Search Galaxy for the **Get PDB** tool. Request the accession code `6HHR`.
3. Rename the dataset to 'Hsp90 structure'

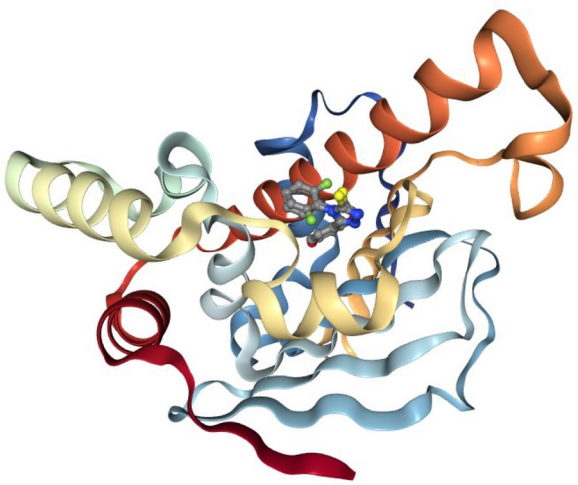

Fig. 1 Hsp90 cartoon view. Hsp90 cartoon with ligands in active site, rendered using the Galaxy NGL plugin [15]

Topology generation

The PDB structure is prepared for MD simulation in a process referred to as parameterization or topology generation.

GROMACS distinguishes between constant and dynamic attributes of the atoms in the system. The constant attributes (e.g. atom charges, bonds connecting atoms) are listed in the topology (TOP file), while dynamic attributes (attributes that can change during a simulation, e.g. atom position, velocities and forces) are stored in structure (PDB or GRO) and trajectory (XTC and TRR) files.

The PDB file includes neither parameters for simulations, nor the positions of hydrogen atoms. Therefore, before beginning simulation, this information must be calculated.

Extract protein and ligand coordinates

Parameterization is performed separately for the ligand and protein. The PDB file is separated into two sets of coordinates—one for the ligand and one for the protein—using the simple text manipulation tools integrated into Galaxy.

Hands-on 2: Separate protein and ligand coordinates

1. **Search in textfiles** with the following parameters:
   - *"Select lines from"*: 'Hsp90 structure'
   - *"that"*: `Don't Match`
   - *"Regular Expression"*: `HETATM`
2. Rename output to 'Protein (PDB)'
3. **Search in textfiles** with the following parameters:
   - *"Select lines from"*: 'Hsp90 structure'
   - *"that"*: `Match`
   - *"Regular Expression"*: `AG5E`
4. Rename output to 'Ligand (PDB)'

The PDB file is filtered twice: once for lines which do not match `HETATM`, which returns a PDB file containing only protein, not ligand and solvent; and once for lines which match the ligand's identity code `AG5E`, which returns a PDB file containing only the ligand.

Set up protein topology

The topology for the protein file is calculated with the **GROMACS initial setup** tool.

Hands-on 3: Generate protein topology

GROMACS initial setup with the following parameters:
- *"PDB input file"*: 'Protein (PDB)' file
- *"Force field"*: `AMBER99SB`
- *"Water model"*: `TIP3P`
- *"Generate detailed log"*: `Yes`

A force field is essentially a function to calculate the potential energy of a system, based on various empirical parameters (for the atoms, bonds, charges, dihedral angles and so on). There are a number of families of force fields; some of the most commonly used include CHARMM [17], AMBER [18], GROMOS [19] and OpenFF [20] (for a recent, accessible overview see [21]). One of the main AMBER force fields for protein modeling, `ff99SB`, was selected.

Apart from the force field, a water model was selected to model the solvent; a wide range of models exist for this purpose. The common TIP3P model is selected, which is an example of a 'three-site model'—so-called because the molecule is modeled using three points, corresponding to the three atoms of water (four- and five-site models include additional 'dummy atoms' representing the negative charges of the lone pairs of the oxygen atom) [22].

The tool produces four outputs: a GRO file (containing the coordinates of the protein), a TOP file (containing other information, including charges, masses, bonds and angles), an ITP file (which will be used to restrain the protein position in the equilibration steps later on), and a log file.

Please note all the GROMACS tools provided in Galaxy output a log file. These files provide useful information for debugging purposes.

Generate a topology for the ligand

The acpype [23] tool is used to generate a topology for the ligand. This provides a convenient interface to the AmberTools suite and creates the ligand topology in the format required by GROMACS.

Inspecting the contents of the ligand PDB file shows that it contains no hydrogen atoms. Hydrogens were added to the topology using the 'Compound conversion' tool (based on OpenBabel [24]).

Hands-on 4: Generate ligand topology

1. **Compound conversion**:
   - *"Molecular input file"*: 'Ligand (PDB)'
   - *"Output format"*: `Protein Data Bank format (pdb)`
   - *"Add hydrogens appropriate for pH"*: `7.0`
2. Rename the output file to 'Hydrated ligand (PDB)'
3. **Generate MD topologies for small molecules**:
   - *"Input file"*: 'Ligand (PDB)'
   - *"Charge of the molecule"*: `0`
   - *"Multiplicity"*: `1`
   - *"Force field to use for parameterization"*: `gaff`
   - *"Save GRO file?"*: `Yes`

The GAFF (general AMBER force field) is selected, which is a generalized AMBER force field [25] which can be applied to almost any small organic molecule.

Appropriate charge and multiplicity parameters are entered. The ligand studied is a simple organic molecule, with no charge; therefore, the charge is set to 0 and the multiplicity to 1. Alternative values for multiplicity need only be considered for more exotic species such as metal complexes or carbenes.

Next, the topologies are combined and the simulation box is defined.

Combine topology and GRO files

The separate topology and structure files for both protein and ligand are combined into a single set of files to continue with the simulation setup. A dedicated Galaxy tool is provided for this, using the Python library ParmEd [26].

Hands-on 5: Combine GRO and topology files

Merge GROMACS topologies with the following parameters:
- *"Protein topology (TOP) file"*: `TOP` file created by the **GROMACS initial setup tool**
- *"Ligand topology (TOP or ITP) file"*: `Topology` file created by the **acpype tool**
- *"Protein structure (GRO) file"*: `GRO` file created by the **GROMACS initial setup tool**
- *"Ligand structure (GRO) file"*: `Structure file (GRO format)` file created by the **acpype tool**

Note that, apart from this tool, the Galaxy platform also provides an integrated text editor for making more advanced changes to GROMACS topologies or configuration files.

Create the simulation box

The next step, once combined coordinate (GRO) and topology (TOP) files have been created, is to create a simulation box in which the system is situated.

Hands-on 6: Create simulation box

GROMACS structure configuration with the following parameters:
- *"Input structure"*: `System GRO file` (Input dataset)
- *"Configure box?"*: `Yes`
  - *"Box dimensions in nanometers"*: `1.0`
  - *"Box type"*: `Triclinic`
- *"Generate detailed log"*: `Yes`

This tool returns a new GRO structure file, containing the same coordinates as before, but defining a simulation box such that every atom is a minimum of 1 nm from the box boundary. A variety of box shapes are available to choose from: triclinic is selected, as it provides efficient packing in space and thus fewer computational resources need to be devoted to simulation of solvent.

Solvation

The next step is solvation of the newly created simulation box. Water was chosen as a solvent to in order to simulate biological conditions. Note that the system is charged (depending on the pH) and it is neutralized by the addition of the sodium and chloride ions (replacing existing water molecules) using the solvation tool.

Hands-on 7: Solvation

GROMACS solvation and adding ions with the following parameters:
- *"GRO structure file"*: output of **GROMACS structure configuration**
- *"System topology"*: `output`
- *"Generate detailed log"*: `Yes`

Energy minimization

After the solvation step, parameterization of the system is complete and preparatory simulations can be performed. The first of these is energy minimization,

which can be carried out using the 'GROMACS energy minimization' tool. The purpose of energy minimization is to relax the structure, removing any steric clashes or unusual geometry which would artificially raise the energy of the system.

Hands-on 8: Energy minimization

GROMACS energy minimization with the following parameters:

- *"GRO structure file."*: GRO output of **GROMACS solvation and adding ions**
- *"Topology (TOP) file."*: TOP output of **GROMACS solvation and adding ions**
- *"Parameter input"*: `Use default (partially customisable) setting`
  - *"Number of steps for the MD simulation"*: `50000`
  - *"EM tolerance"*: `1000.0`
- *"Generate detailed log"*: `Yes`
- Rename GRO output to `Minimized GRO file`

The EM tolerance here refers to the maximum force which will be allowed in a minimized system. The simulation will be terminated when the maximum force is less than this value, or when 50,000 steps have elapsed. The 'Extract energy components' tool is used to plot the convergence of the potential energy during the minimization.

Hands-on 9: Checking EM convergence

1. **Extract energy components with GROMACS** with the following parameters:
   - *"EDR file"*: EDR output of **GROMACS energy minimization**
   - *"Terms to calculate"*: `Potential`
   - *"Output format"*: `Galaxy tabular`
2. On the output tabular file, click on the 'Visualize this data' icon. This provides a range of visualization options. Select 'Line chart (jqPlot)'.
3. In the visualization window which appears, click on 'Select data.' Enter the following parameters:
   - *"Provide a label"*: `Energy potential`
   - *"Values for x-axis"*: `Column: 1`
   - *"Values for y-axis"*: `Column: 2`

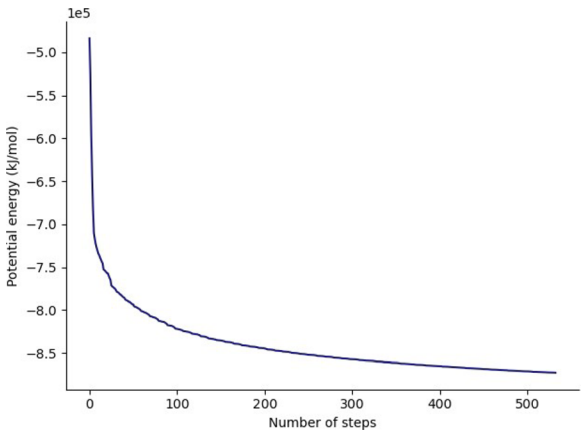

Fig. 2 Energy potential during the EM simulation

As seen in Fig. 2, the system first drops rapidly in energy, before slowly converging on the minimized state.

Equilibration

At this point equilibration of the solvent around the solute (i.e. the protein) is necessary. This is performed in two stages: equilibration under an NVT (or isothermal-isochoric) ensemble, followed by an NPT (or isothermal-isobaric) ensemble. Use of the NVT ensemble entails maintaining constant number of particles, volume and temperature, while the NPT ensemble maintains constant number of particles, pressure and temperature. Simulation under the NVT ensemble allows the solvent to be brought to the desired temperature, while simulation under the NPT ensemble brings the solvent to the correct pressure.

For equilibration, the protein is held in place while the solvent is allowed to move freely around it. This is achieved using the position restraint file (ITP) created during the system setup. This restraint does not prevent protein movement; rather movement is energetically penalized.

Hands-on 10: NVT equilibration

GROMACS simulation with the following parameters:

- *"GRO structure file"*: `Minimized GRO file` (from energy minimization step)
- *"Topology (TOP) file"*: TOP file produced by solvation step.
- In *"Inputs"*:
  - *"Position restraint (ITP) file"*: ITP file produced by initial setup step.
- In *"Outputs"*:
  - *"Trajectory output"*: `Return .xtc file (reduced precision)`
  - *"Structure output"*: `Return .gro file`
  - *"Produce a checkpoint (CPT) file"*: `Produce CPT output`
  - *"Produce an energy (EDR) file"*: `Produce EDR output`
- In *"Settings"*:
  - *"Parameter input"*: `Use default (partially customisable) setting`
    - *"Bond constraints (constraints)"*: `All bonds (all-bonds).`
    - *"Temperature /K"*: `300`
    - *"Step length in ps"*: `0.002`
    - *"Number of steps that elapse between saving data points (velocities, forces, energies)"*: `1000`
    - *"Number of steps for the simulation"*: `50000`
- *"Generate detailed log"*: `Yes`

Once the NVT equilibration is complete, it is worth using the 'Extract energy components' tool again to check whether the system temperature has converged on 300 K. This can be done as described for energy minimization, this time specifying `Temperature` under *Terms to calculate* rather than `Potential`. The plot should show the temperature reaching 300 K and remaining there, albeit with some fluctuation.

Having stabilized the temperature of the system with NVT equilibration, the pressure is stabilized by equilibrating using the NPT (constant number of particles, pressure, temperature) ensemble. The NPT simulation is continued from the NVT simulation by using the checkpoint (CPT) file saved at the end of the NVT simulation.

Hands-on 11: NPT equilibration

GROMACS simulation with the following parameters:

- *"GRO structure file"*: GRO output of **GROMACS simulation** (NVT equilibration)
- *"Topology (TOP) file"*: TOP file produced by solvation step.
- In *"Inputs"*:
  - *"Checkpoint (CPT) file"*: Output of **GROMACS simulation** (NVT equilibration))
  - *"Position restraint (ITP) file"*: ITP file produced by initial setup step.
- In *"Outputs"*:
  - *"Trajectory output"*: `Return .xtc file (reduced precision)`
  - *"Structure output"*: `Return .gro file`
  - *"Produce a checkpoint (CPT) file"*: `Produce CPT output`
  - *"Produce an energy (EDR) file"*: `Produce EDR output`
- In *"Settings"*:
  - *"Ensemble"*: `Isothermal-isobaric ensemble (NPT)`
  - *"Parameter input"*: `Use default (partially customisable) setting`
    - *"Bond constraints (constraints)"*: `All bonds (all-bonds).`
    - *"Temperature /K"*: `300`
    - *"Step length in ps"*: `0.002`
    - *"Number of steps that elapse between saving data points (velocities, forces, energies)"*: `1000`
    - *"Number of steps for the simulation"*: `50000`
- *"Generate detailed log"*: `Yes`

After the NPT equilibration is complete, 'Extract energy components' can be used again to view the pressure of the system. This is done as described for energy minimization, specifying `Pressure` under *Terms to calculate*. The plot should show convergence on 1 bar and remain there, although some fluctuation is expected.

Production simulation

The restraints are removed and a production simulation is carried out for 1 million steps. With a step size of 1 fs, this simulation represents a total time length of 1 ns. This is rather short compared to the state-of-the-art, but sufficient for the purposes of a tutorial.

Hands-on 12: Main simulation

GROMACS simulation with the following parameters:

- *"GRO structure file"*: Output of **GROMACS simulation** (NPT equilibration)
- *"Topology (TOP) file"*: Output of the solvation step
- In *"Inputs"*:
  - *"Checkpoint (CPT) file"*: Output of **GROMACS simulation** (NPT simulation))
- In *"Outputs"*:
  - *"Trajectory output"*: `Return .xtc file (reduced precision)`
  - *"Structure output"*: `Return .gro file`
  - *"Produce a checkpoint (CPT) file"*: `Produce CPT output`
- In *"Settings"*:
  - *"Ensemble"*: `Isothermal-isobaric ensemble (NPT)`
  - *"Parameter input"*: `Use default (partially customisable) setting`
    - *"Temperature /K"*: `300`
    - *"Step length in ps"*: `0.001`
    - *"Number of steps that elapse between saving data points (velocities, forces, energies)"*: `1000`
    - *"Number of steps for the simulation"*: `1000000`
- *"Generate detailed log"*: `Yes`

Methods: analysis

An analysis of the GROMACS simulation outputs (structure and trajectory file) will be carried out using Galaxy tools developed for computational chemistry [3] based on popular analysis software, such as MDAnalysis [27], MDTraj [28], and Bio3D [29]. These tools output both tabular files as well as a variety of attractive plots.

Convert coordinate and trajectory formats

Before beginning a detailed analysis, the structure and trajectory files generated previously need to be converted into different formats. The structural coordinates of the system in GRO format are converted into PDB format using the 'Convert coordinate and trajectory formats' tool (which is based on the 'editconf' GROMACS command). This PDB file will be used by most analysis tools as a starting structure. This tool can also be used to carry out initial setup (as discussed in the simulation methods section) for GROMACS simulations and to convert from PDB to GRO format. The trajectory file is converted from XTC to DCD format, as a number of tools (particularly those based on Bio3D) only accept trajectories in DCD format. This tool can also be used to interconvert between several other trajectory formats.

Hands-on 13: Convert coordinate and trajectory formats

1. **GROMACS structure configuration** with the following parameters:
   - *"Output format"*: `PDB file`
   - *"Configure box?"*: `No`
2. **MDTraj file converter** with the following parameters:
   - *"Output format"*: `DCD file`

RMSD analysis

The Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF) are calculated to check the stability and conformation of the protein and ligand through the course of the simulation. RMSD is a standard measure of structural distance between coordinate sets that measures the average distance between a group of atoms. The RMSD of the Cα atoms of the protein backbone is calculated here and is a measure of how much the protein conformation has changed between different time points in the trajectory. Note that for more complex systems, consider a more focused selection.

For the RMSD analysis of the ligand, the 'Select domains' parameter of the tool can for convenience be set to 'Ligand'; however, this automatic selection sometimes fails. Instead the 'Residue ID' is specified in the textbox provided. In this example the ligand's Residue ID is 'G5E'. The output is the requested RMSD data as a time series, the RMSD plotted as a time series and as a histogram (for example, see Fig. 3 in "Results and discussion" section).

Hands-on 14: RMSD Analysis: protein

RMSD Analysis with the following parameters:

- *"DCD trajectory input"*: output of **MDTraj file converter**
- *"PDB input"*: output of **GROMACS structure configuration**
- *"Select domains"*: `C-alpha`

Hands-on 15: RMSD Analysis: ligand using Residue ID

RMSD Analysis with the following parameters:

- *"DCD trajectory input"*: output of **MDTraj file converter**
- *"PDB input"*: output of **GROMACS structure configuration**
- *"Select domains"*: `Residue ID`
  - *"Residue ID"*: `G5E`

RMSF analysis

The Root Mean Square Fluctuation (RMSF) is valuable to consider, as it represents the deviation at a reference position over time. The fluctuation in space of particular amino acids in the protein are considered. The Cα of the protein, designated by `C-alpha`, is a good selection to understand the change in protein structure. Depending on the system these fluctuations can be correlated to experimental techniques including Nuclear Magnetic Resonance (NMR) and Mössbauer spectroscopy [30, 31]. The output from the tools is the requested RMSF data and the RMSF plotted as a time series (for example, see Fig. 5 in "Results and discussion" section).

Hands-on 16: RMSF Analysis

RMSF Analysis with the following parameters:

- *"DCD trajectory input"*: output of **MDTraj file converter**
- *"PDB input"*: output of **GROMACS structure configuration**
- *"Select domains"*: `C-alpha`

PCA

Principal component analysis (PCA) converts a set of correlated observations (movement of selected atoms in protein) to a set of principal components (PCs) which are linearly independent (or uncorrelated). Here several related tools are used. The PCA tool calculates the PCA in order to determine the relationship between statistically meaningful conformations (major global motions) sampled during the trajectory. The Cα carbons of the protein backbone are again a good selection for this purpose. Outputs include the PCA raw data and figures of the relevant principal components (PCs) as well as an eigenvalue rank plot (see Fig. 6) which is used to visualize the proportion of variance due to each principal component (remembering that the PCs are ranked eigenvectors based on the variance). Having discovered the principal components usually these are visualized. The PCA visualization tool creates trajectories of specific principal components which can be viewed in a molecular viewer such as VMD [32] or NGL viewer [15]. The PCA cosine content when close to 1 indicates that the simulation is not converged and a longer simulation is needed. For values below 0.7, no statement can be made about convergence or lack thereof.

Hands-on 17: PCA with BIO3D

PCA with the following parameters:

- *"DCD trajectory input"*: output of **MDTraj file converter**
- *"PDB input"*: output of **GROMACS structure configuration**
- *"Select domains"*: `C-alpha`

Hands-on 18: PCA visualization

PCA visualization with the following parameters:

- *"DCD trajectory input"*: output of **MDTraj file converter**
- *"PDB input"*: output of **GROMACS structure configuration**
- *"Select domains"*: `C-alpha`

Hands-on 19: Cosine content calculation

Cosine Content with the following parameters:

- *"DCD/XTC trajectory input"*: output of **MDTraj file converter**
- *"PDB/GRO input"*: output of **GROMACS structure configuration**

Hydrogen bond analysis

Hydrogen bonding interactions contribute to binding and are worth investigating, in particular persistent hydrogen bonds. All possible hydrogen bonding interactions between the two selected regions, here the protein and the ligand, are investigated over time using the VMD hydrogen bond analysis tool included in Galaxy. Hydrogen bonds are identified and in the output the total number of hydrogen bonds and occupancy over time is returned.

Hands-on 20: Hydrogen bond analyis

Hydrogen Bond Analysis using VMD with the following parameters:

- *"DCD/XTC trajectory input"*: output of **MDTraj file converter**
- *"PDB/GRO input"*: output of **GROMACS structure configuration**
- *"Selection 1"*: `protein`
- *"Selection 2"*: `resname G5E`

Results and discussion

After the completion of the simulation, the following questions arise: (1) is the simulation converged enough, and (2) what interesting molecular properties are observed. The timescale of motions of interest are in the picosecond to nanosecond range; these are motions such as domain vibration, hydrogen bond breaking, translation diffusion and side chain fluctuations. To observe meaningful conformational transitions of the protein μs sampling would be needed, but this is not the purpose here.

The PCA cosine content of the dominant motion related to PC1 is 0.93, indicating that the simulation is not fully converged. This is expected due to the short simulation length. For production level simulations, it is the norm to extend simulations to hundreds of nanoseconds in length, if not microseconds. A short simulation time of 1 ns was chosen as this tutorial is designed to be carried out on public webservers, which have finite computational resources to dedicate to training purposes.

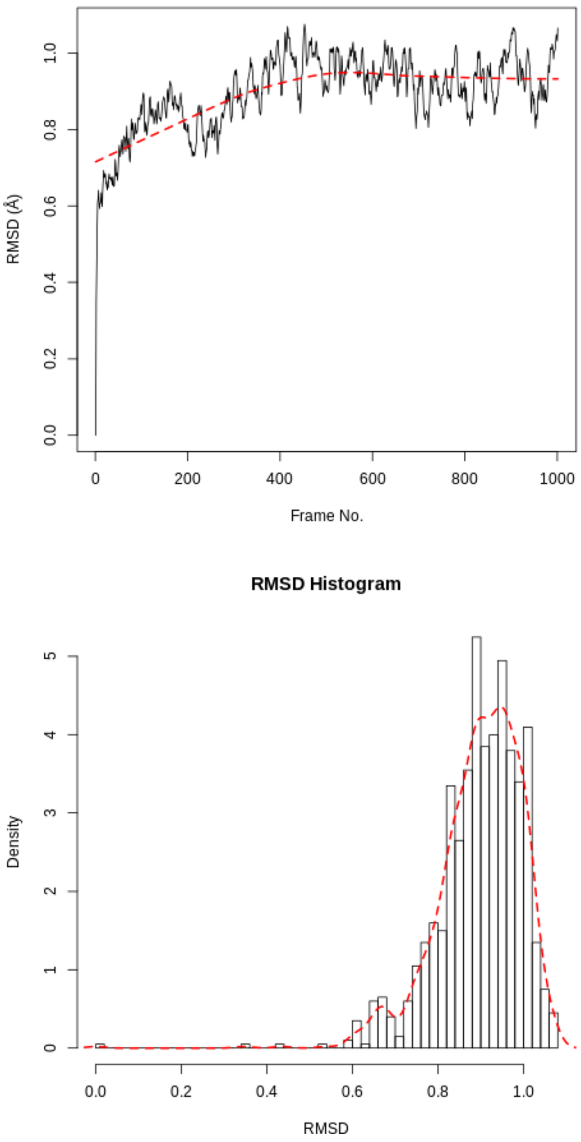

Fig. 3 RMSD for protein. RMSD time series and histogram for the protein

RMSD protein

The RMSD time series for the protein shows a thermally stable and equilibrated structure that plateaus at 1.0Å with an average RMSD between 0.8Å and 1.0Å. There are no large conformational changes during the simulation. The RMSD histogram confirms this, see Fig. 3. Note these graphs are automatically created by Galaxy as part of the tool's outputs.

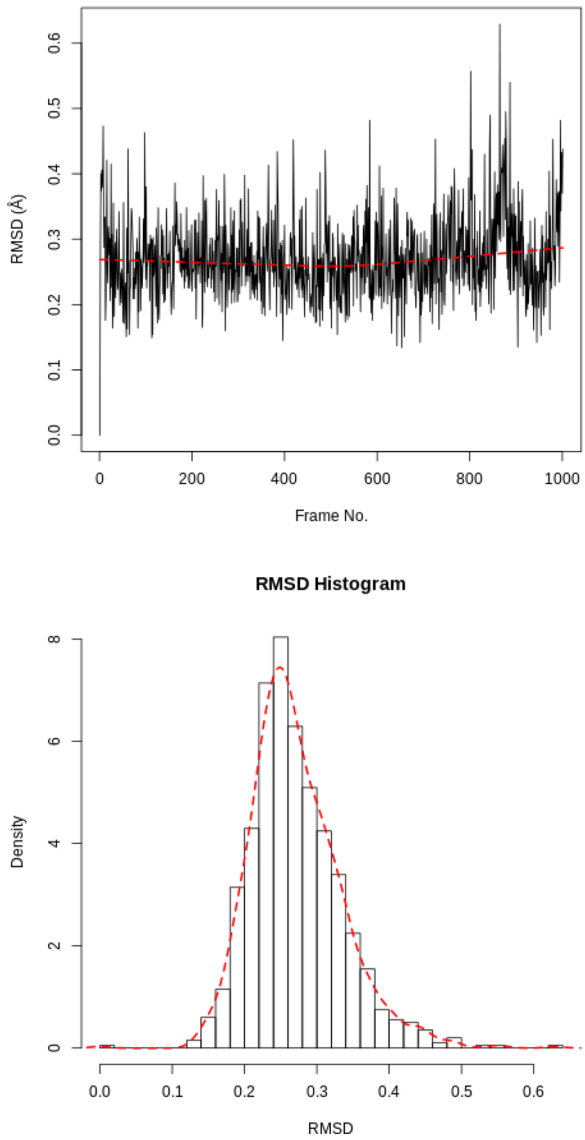

Fig. 4 RMSD for the ligand. RMSD time series and histogram for the ligand

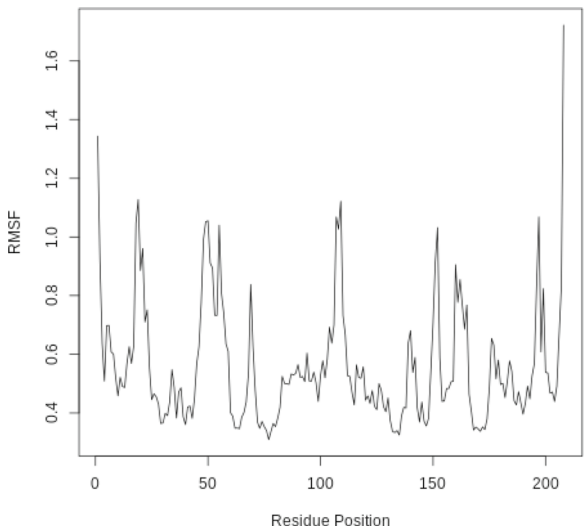

Fig. 5 RMSF for the protein. RMSF(Å) vs the residue position. Large fluctuations occur at various positions, which correspond to flexible loop regions on the surface of the protein

RMSD ligand

Calculating the RMSD of the ligand is necessary to check if it is stable in the active site and to identify possible binding modes. If the ligand is not stable, there will be large fluctuations in the RMSD.

In our case the ligand is stable with a single binding mode. The RMSD fluctuates around 0.3Å, with a slight fluctuation near the end of the simulation. This is more clearly seen in the histogram, see Figure 4. The conformation seen during simulation is very similar to that in the crystal structure and the ligand is stable in the active site.

RMSF

When considering the RMSF (Fig. 5), fluctuations greater than 1.0Å are of interest; for example see the fluctuations near residue positions 50, 110 and 160. Inspecting the structure with molecular visualization software such as VMD, these can be seen to correspond to flexible loop regions on the protein surface. In addition, very large fluctuations are seen for the C-terminus; this is common and no investigation is needed.

Note that the first few residues of this protein are missing in the PDB, and therefore residue position 0 in the RMSF corresponds to position 17 in the Hsp90 FASTA primary sequence. This is a fairly common problem that can occur with molecular modeling of proteins, where there may be missing residues at the beginning or within the sequence.

PCA

The first three principal components are responsible for 32.8% of the total variance, as seen in the eigenvalue rank plot (Fig. 6). The first principal component (PC1) accounts for 15.4% of the variance (see PC1 vs PC2 and eigenvalue rank plots in Fig. 6). Visualization of PC1 using VMD shows a rocking motion and wagging of the C-terminus.

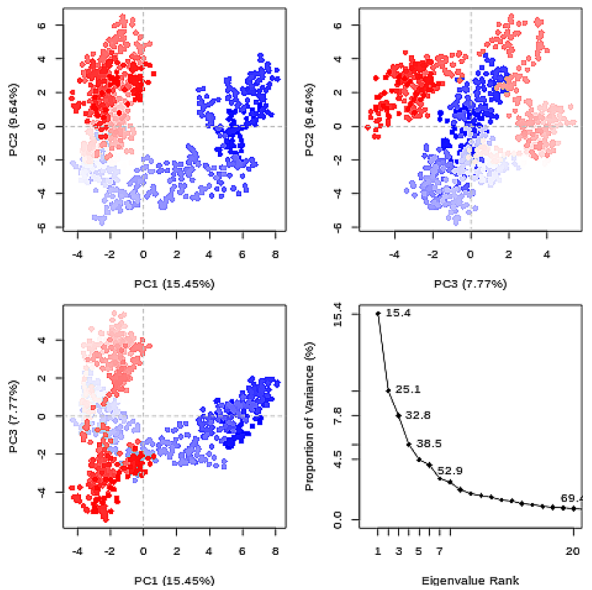

Fig. 6 Principal component analysis. PCA results which include graphs of PC2 vs PC1, PC2 vs PC3, PC3 vs PC1 colored from blue to red in order of time, and an eigenvalue rank plot (Scree plot). In the eigenvalue plot the cumulative variance is labeled for each data point

Hydrogen bonding

Multiple hydrogen bonds were identified between the active site of the protein and the ligand. The hydrogen bond between aspartate-93 and the ligand (as identified in the crystal structure) was found to be persistent, meeting the hydrogen bond criteria for 89.22% of the simulation. A hydrogen bond between the ligand and the carbonyl group of glycine-97 was found to have a 15.27% occupancy. Hydrogen bonding interactions with threonine-184, asparagine-51 and lysine-58 were also observed but these were not persistent and only present for a minority of the simulation. These values can be accessed from the 'Percentage occupancy of the H-bond' output of the hydrogen bond analysis tool.

High throughput workflows

Up until this step, Galaxy tools have been applied sequentially to datasets. This is useful to gain an understanding of the steps involved, but becomes tedious if the workflow needs to be run on multiple protein-ligand systems. Fortunately, Galaxy allows entire workflows to be executed with a single mouse-click, enabling straightforward high-throughput analyses.

The high-throughput capabilities of Galaxy are demonstrated by running the workflow detailed so far on a further three ligands [33–37].

Hands-on 21: High-throughput MD

1. Create a new history for running the high-throughput workflow and name it 'Hsp90 HTMD simulation'.
2. Upload the SD-file containing the new ligand structures from Zenodo [33] and rename it 'Ligands (SDF)'.
3. Import the simulation workflow from the European [34] or the South African Galaxy server [35].
4. Run the imported workflow with the following parameters:
   - *"SDF file with (docked) ligands"*: 'Ligands (SDF)' file.
5. Import the analysis workflow from the European [36] or the South African Galaxy server [37] (also available through Zenodo).
6. Run the imported workflow with the following parameters:
   - *"Send results to a new history"*: 'Yes'
   - *"History name"*: 'Hsp90 HTMD analysis'
   - *"GRO input"*: Collection of GRO files produced by simulation workflow
   - *"XTC input"*: Collection of XTC files produced by simulation workflow

This process runs the entire simulation and analysis procedure described so far on the new set of ligands. It uses Galaxy's collection [38] feature to organize the data; each item in the history is a collection (essentially a directory containing multiple individual datasets) containing one file corresponding to each of the input ligands.

Note that the SD-file needs to contain ligands with the correct 3D coordinates for MD simulation. The easiest way to obtain these is using a molecular docking tool such as Autodock Vina [39] or rDock [40]; tutorials and workflows are available for both of these from the Galaxy Training Network. As an example, the history in which the SD-file used in the HTMD workflow is generated (using AutoDock Vina) is provided [41].

Further information

Apart from manual setups or collections, there are several other alternatives which are helpful in scaling up workflows. Galaxy supports and provides training material for converting histories to workflows [42], using multiple histories [43], and the Galaxy Application Programming Interface (API) [44]. For beginners and users who prefer a visual interface, automation can be done

using multiple histories and collections with the standard Galaxy user interface.

If you are able to write small scripts, you can automate everything you have learned here with the Galaxy API. This approach allows interaction with the server to automate repetitive tasks and create more complex workflows (which may have repetition or branching). The simplest way to access the API is through the Python library BioBlend [45]. An example Python script, which uses BioBlend to run the GROMACS simulation workflow for each of a list of ligands, is given in the hands-on box below.

Hands-on 22: BioBlend script

```
from bioblend import galaxy

# Server and account details
API_KEY = 'YOUR USEGALAXY.EU API KEY'
gi = galaxy.GalaxyInstance(key=API_KEY,
    url='https://usegalaxy.eu/')

# ID for GROMACS workflow
workflow_id = 'adc6d049e9283789'

# Dataset IDs for ligands to dock
ligands = {
# ligand_name: dataset ID,
'lig1': '11ac94870d0bb33a79c5fa18b0fd3b4c',
# ...
}

# Loop over ligands, invoking workflow
for name, _id in ligands.items():
    inv = gi.workflows.invoke_workflow(
        workflow_id,
        inputs={
            '1': {'src': 'hda', 'id': _id}
        },
        history_name=f'HTMD run on {name}'
    )
```

Conclusion

This tutorial provides a guide on how to study protein-ligand interaction using molecular dynamics in Galaxy. Performing such analyses in Galaxy makes it straightforward to set up, schedule and run workflows, removing much of the difficulty from MD simulation. Thus, the technical barrier to performing high-throughput studies is greatly reduced. Results are structured in the form of Galaxy histories or collections, and include ready-plotted diagrams, which ensure data can be easily understood and reproduced if necessary. Apart from streamlining the process for existing MD users, this tutorial should also prove useful as a pedagogical guide for educating students or newcomers to the field.

After completing the tutorial, the user will be familiar at a basic level with a range of MD analysis techniques, and understand the steps required for a typical MD simulation. Thus, they will be equipped to apply these tools to their own problems.

Acknowledgements

We thank Bérénice Batut for helpful comments and discussions, Bert Droesbeke for assistance building the Docker container and Oleg Zharkov for assistance integrating the Galaxy text editor. In addition, we thank the entire Galaxy Training Network, as well as the European Galaxy and ilifu teams, for their support.

The European Galaxy server, which was used for the calculations described, is in part funded by Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and German Federal Ministry of Education and Research (BMBF grants 031 A538A/A538C RBC, 031L0101B/031L0101C de.NBI-epi (de.NBI)).

Authors' contributions

All authors contributed to writing Galaxy tools, creating workflows, writing training material, and writing the paper. All authors read and approved the final manuscript.

Funding

This work was supported by funding from the following organizations: S.A.B. was funded by the European Open Science Cloud (EOSC-Life) (Grant No. 824087); T.S. and C.B.B. were funded by the University of Cape Town's Research Committee (URC) and by the National Research Foundation of South Africa (Grant Numbers 115215 and 116362); and B.A.G was funded by the German Research Foundation for the Collaborative Research Center 992 Medical Epigenetics [SFB 992/1 2012 and SFB 992/2 2016].

Data and material availability

Data and materials are available on GitHub:
European Galaxy server (https://cheminformatics.usegalaxy.eu)
Galaxy Computational Chemistry South Africa server (https://galaxy-compchem.ilifu.ac.za)
Docker container providing a Galaxy installation with all required tools preinstalled. (https://quay.io/repository/galaxy/computational-chemistry-training)
Galaxy Training Network website (https://training.galaxyproject.org/topics/computational-chemistry/tutorials/htmd-analysis/tutorial.html)
Supplementary Material, including workflows and data used (https://github.com/galaxycomputationalchemistry/htmd-paper-sm)

Competing interests

The authors declare that they have no competing interests.

Received: 17 May 2020 Accepted: 27 July 2020
Published online: 10 September 2020

References

1. Berendsen HJC (2007) Simulating the physical world: hierarchical modeling from quantum mechanics to fluid dynamics. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511815348
2. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 46(Web Server issue):537–544. https://doi.org/10.1093/nar/gky379
3. Senapathi T, Bray S, Barnett CB, Grüning B, Naidoo KJ (2019) Biomolecular Reaction and Interaction Dynamics Global Environment (BRIDGE). Bioinformatics 35(18):3508–3509. https://doi.org/10.1093/bioinformatics/btz107
4. Harvey MJ, Fabritiis GD (2012) High-throughput molecular dynamics: the powerful new tool for drug discovery. Drug Discov Today 17(19):1059–1062. https://doi.org/10.1016/j.drudis.2012.03.017

5. Guterres H, Im W (2020) Improving protein-ligand docking results with high-throughput molecular dynamics simulations. J Chem Inf Model 60(4):2189–2198. https://doi.org/10.1021/acs.jcim.0c00057
6. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. https://doi.org/10.1016/j.softx.2015.06.001
7. Lemkul J (2019) From proteins to perturbed Hamiltonians: a suite of tutorials for the GROMACS-2018 molecular simulation package [article v1.0]. Living J Comput Mol Sci. https://doi.org/10.33011/livecoms.1.1.5068
8. Batut et al (2018) Community-driven data analysis training for biology. Cell Syst 6():752–7581. https://doi.org/10.1016/j.cels.2018.05.012
9. Galaxy Training: Computational chemistry. https://training.galaxyproject.org/training-material/topics/computational-chemistry/tutorials/htmd-analysis/tutorial.html.
10. Pearl LH, Prodromou C (2006) Structure and mechanism of the Hsp90 molecular chaperone machinery. Annu Rev Biochem 75(1):271–294. https://doi.org/10.1146/annurev.biochem.75.103004.142738
11. Schopf FH, Biebl MM, Buchner J (2017) The HSP90 chaperone machinery. Nat Rev Mol Cell Biol 18(6):345–360. https://doi.org/10.1038/nrm.2017.20
12. Stebbins CE, Russo AA, Schneider C, Rosen N, Hartl FU, Pavletich NP (1997) Crystal structure of an Hsp90–geldanamycin complex: targeting of a protein chaperone by an antitumor agent. Cell 89(2):239–250. https://doi.org/10.1016/s0092-8674(00)80203-2
13. Hermane J, Eichner S, Mancuso L, Schröder B, Sasse F, Zeilinger C, Kirschning A (2019) New geldanamycin derivatives with anti Hsp properties by mutasynthesis. Org Biomol Chem 17(21):5269–5278. https://doi.org/10.1039/c9ob00892f
14. PubChem: 3-(2,4-Dihydroxyphenyl)-4-(2-fluorophenyl)-1H-1,2,4-triazole-5-thione. Library Catalog: pubchem.ncbi.nlm.nih.gov. https://pubchem.ncbi.nlm.nih.gov/compound/135508238 Accessed 29 Apr 2020.
15. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW (2018) NGL viewer: web-based molecular graphics for large complexes. Bioinformatics 34(21):3755–3758. https://doi.org/10.1093/bioinformatics/bty419
16. Schuetz DA, Bernetti M, Bertazzo M, Musil D, Eggenweiler H-M, Recanatini M, Masetti M, Ecker GF, Cavalli A (2018) Predicting residence time and drug unbinding pathway through scaled molecular dynamics. J Chem Inf Model 59(1):535–549. https://doi.org/10.1021/acs.jcim.8b00614
17. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, Mackerell AD (2009) CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J Comput Chem. https://doi.org/10.1002/jcc.21367
18. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C (2015) ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. J Chem Theory Comput 11(8):3696–3713. https://doi.org/10.1021/acs.jctc.5b00255
19. Reif MM, Hünenberger PH, Oostenbrink C (2012) New interaction parameters for charged amino acid side chains in the GROMOS force field. J Chem Theory Comput 8(10):3705–3723. https://doi.org/10.1021/ct300156h
20. Mobley DL, Bannan CC, Rizzi A, Bayly CI, Chodera JD, Lim VT, Lim NM, Beauchamp KA, Slochower DR, Shirts MR, Gilson MK, Eastman PK (2018) Escaping atom types in force fields using direct chemical perception. J Chem Theory Comput 14(11):6076–6092. https://doi.org/10.1021/acs.jctc.8b00640
21. Lemkul JA (2020) Pairwise-additive and polarizable atomistic force fields for molecular dynamics simulations of proteins. In: Computational approaches for understanding dynamical systems: protein folding and assembly. p. 1–71. New York: Elsevier. https://doi.org/10.1016/bs.pmbts.2019.12.009.
22. Onufriev AV, Izadi S (2017) Water models for biomolecular simulations. Wiley Interdiscip Rev Comput Mol Sci 8(2):1347. https://doi.org/10.1002/wcms.1347
23. da Silva AWS, Vranken WF (2012) ACPYPE—AnteChamber PYthon parser interfacE. BMC Res Notes 5(1):367. https://doi.org/10.1186/1756-0500-5-367
24. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) OpenBabel: an open chemical toolbox. J Cheminf 3(1).
25. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general AMBER force field. J Comput Chem 25(9):1157–1174. https://doi.org/10.1002/jcc.20035
26. Swails J, Hernandez C, Mobley D, Nguyen H, Wang L, Janowski P (2016) ParmEd: Cross-program parameter and topology file editor and molecular mechanical simulator engine. Accessed 23 Jan 2020. https://parmed.github.io/ParmEd/html/index.html.
27. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O (2011) MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. J Comput Chem 32(10):2319–2327. https://doi.org/10.1002/jcc.21787
28. McGibbon R, Beauchamp K, Harrigan M, Klein C, Swails J, Hernández C, Schwantes C, Wang L-P, Lane T, Pande V (2015) MDTraj: a modern open library for the analysis of molecular dynamics trajectories. Biophys J 109(8):1528–1532. https://doi.org/10.1016/j.bpj.2015.08.015
29. Skjærven L, Yao X-Q, Scarabelli G, Grant BJ (2014) Integrating protein structural dynamics and evolutionary analysis with Bio3D. BMC Bioinf 15(1):399. https://doi.org/10.1186/s12859-014-0399-6
30. Berjanskii M, Wishart DS (2006) NMR: prediction of protein flexibility. Nat Protoc 1(2):683–688. https://doi.org/10.1038/nprot.2006.108
31. Kuzmanic A, Zagrovic B (2010) Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. Biophys J 98(5):861–871. https://doi.org/10.1016/j.bpj.2009.11.011
32. Humphrey W, Dalke A, Schulten K (1996) VMD—visual molecular dynamics. J Mol Graph 14:33–38
33. galaxycomputationalchemistry/htmd-paper-sm: Data and workflows—intuitive, reproducible high-throughput molecular dynamics in Galaxy: a tutorial. Zenodo. 2020. https://doi.org/10.5281/zenodo.3813283
34. Galaxy | Europe | Accessible history | Protein-ligand HTMD simulation. https://cheminformatics.usegalaxy.eu/u/sbray/w/protein-ligand-htmd-sim. Accessed 29 Apr 2020.
35. Galaxy | South Africa | Accessible History | Protein-ligand HTMD analysis. https://galaxy-compchem.ilifu.ac.za/u/sbray/w/protein-ligand-htmd-sim. Accessed 29 Apr 2020.
36. Galaxy | Europe | Accessible History | Protein-ligand HTMD analysis. https://cheminformatics.usegalaxy.eu/u/sbray/w/protein-ligand-htmd-analysis. Accessed 29 Apr 2020.
37. Galaxy | South Africa | Accessible History | Protein-ligand HTMD analysis. https://galaxy-compchem.ilifu.ac.za/u/sbray/w/protein-ligand-htmd-analysis. Accessed 29 Apr 2020.
38. Galaxy Training: Collections: Using dataset collection. https://galaxyproject.github.io/training-material/topics/galaxy-data-manipulation/tutorials/collections/tutorial.html. Accessed 29 Apr 2020.
39. Trott O, Olson AJ (2009) AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. https://doi.org/10.1002/jcc.21334
40. Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, Garmendia-Doval AB, Juhos S, Schmidtke P, Barril X, Hubbard RE, Morley SD (2014) rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. PLoS Comput Biol 10(4):1003571. https://doi.org/10.1371/journal.pcbi.1003571
41. Galaxy | Europe | Accessible History | Protein-ligand docking (6hhr). https://cheminformatics.usegalaxy.eu/u/sbray/h/protein-ligand-docking-6hhr. Accessed 29 Apr 2020.
42. Workflows: Extracting Workflows from Histories. https://galaxyproject.github.io/training-material/topics/galaxy-ui/tutorials/history-to-workflow/tutorial.html. Accessed 29 Apr 2020.
43. Galaxy Training: Histories: Understanding Galaxy history system. https://galaxyproject.github.io/training-material/topics/galaxy-ui/tutorials/history/tutorial.html. Accessed 29 Apr 2020.
44. Galaxy Training: Scripting Galaxy using the API and BioBlend. https://training.galaxyproject.org/training-material/topics/dev/tutorials/bioblend-api/slides.html. Accessed 29 Apr 2020.
45. Sloggett C, Goonasekera N, Afgan E (2013) BioBlend: automating pipeline analyses within Galaxy and CloudMan. Bioinformatics 29(13):1685–1686. https://doi.org/10.1093/bioinformatics/btt199

Publisher's Note