UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS

MÁSTER EN BIOLOGÍA COMPUTACIONAL

DEPARTAMENTO DE BIOTECNOLOGÍA-BIOLOGÍA VEGETAL

Implementation of Molecular Dynamics workflow in Scipion and its validation.

MASTER THESIS

Author: Pedro Febrer Martínez

Tutors: Nuria Campillo Martin
Carlos Óscar Sorzano

September 2021

Acknowledgements

First of all, I would like to thank the two people who have given me the opportunity to carry out this work with them and who have helped me at all times when I needed it, Nuria Campillo Martin and Carlos Óscar Sorzano.

Next, I would like to thank the colleagues I have had during this work at the CNB. To Carlos Roca for his invaluable help with everything related to Molecular Dynamics and to Alberto Parra for his selfless help at all times with Scipion. Thanks also to the CNB staff, for always being available to help.

Thanks to my parents, my brother and my grandmother, my friends and Pilar for their unconditional support and the facilities they have given me during all these months, even in the hardest moments.

Abbreviations

1H8F	→ Glycogen Synthase Kinase 3β identifier (Protein)
6TCU	→ Glycogen Synthase Kinase 3β identifier (Protein)
AMBER	→ Assisted Model Building with Energy Refinement (Software)
Arg	→ Arginine (Amino acid)
ASP	→ Aspartic Acid (Amino acid)
CG	→ Conjugate Gradient (Algorithm)
CHARMM	→ Chemistry at HARvard Macromolecular Mechanics (Software)
CNB	→ Spanish National Center for Biotechnology
Cryo-Em	→ Cryogenic Electron Microscopy
CSIC	→ Spanish National Research Council
Cys	→ Cysteine (Amino acid)
GLU	→ Glutamic Acid (Amino acid)
GRACE	→ GRaphing, Advanced Computation and Exploration of data (Software)
GROMACS	→ GROningen MAchine for Chemical Simulations (Software)
GROMOS	→ GROningen MOlecular Simulation (Force field)
GSK-3	→ Glycogen Synthase Kinase 3 (Protein)
GUI	→ Graphical User Interface
HB2	→ Hydrogen from Protein Data Bank file (Atom)
Leu	→ Leucine (Amino acid)
MD	→ Molecular Dynamics
NAMD	→ Nanoscale Molecular Dynamics (Software)
NMR	→ Nuclear Magnetic Resonance
NPT	→ Constant Mol, Pressure and Temperature (Isothermal-Isobaric ensemble)
NVT	→ Constant Mol, Volume and Temperature (Canonical ensemble)
O	→ Oxygen (Atom)
OPLS-aa	→ Optimized Potentials for Liquid Simulations – all atoms (Force field)
PBC	→ Periodic Boundary Conditions
PDB	→ Protein Data Bank
RMSD	→ Root-mean-squared deviation
RMSF	→ Root-mean-squared fluctuation
TIP3P	→ Transferable Intermolecular Potential with 3 Points (Water force field)
TIP4P	→ Transferable Intermolecular Potential with 4 Points (Water force field)
TIP5P	→ Transferable Intermolecular Potential with 5 Points (Water force field)
TRAPP	→ TRAnsient Pockets in Proteins (Software)
Tyr	→ Tyrosine (Amino acid)
URL	→ Uniform Resource Location
Val	→ Valine (Amino acid)

Abstract

Molecular dynamics (MD) is a widely used computational technique to simulate the movement of atoms in a system of molecules and macromolecules. This technique allows the study of biological compounds without having to perform laboratory experiments, enabling the discovery of new properties based on the dynamics of the protein structure. Since proteins are mobile entities and can present different conformations over time, it is important to have tools for the observation of this movement and the analysis of their properties depending on it. GROMACS is a free software, designed to produce MD simulations and their analysis, but it requires a great amount of specific prior knowledge for its correct use. This makes MD a difficult technique to access for many scientists. For this reason, in this work, a plug-in of the GROMACS software has been designed for a workflow management platform, Scipion, in order to carry out simple and effective MD simulations. Scipion, in addition to facilitating the use of the technique, allows its integration with other structural biology software, giving the possibility of creating complete workflows. It has been designed so that its results and inputs are interoperable with other software. The plug-in has been validated with a use case: the comparative study of the glycogen synthase kinase 3 (GSK-3) paralogs. The two paralogs, GSK-3α and GSK-3β differ only in one amino acid in their catalytic pocket, making it difficult to discover selective drugs needed for the treatment of diseases such as acute myeloid leukemia, Alzheimer's disease or diabetes mellitus.

Keywords: Molecular Dynamics, GROMACS, Scipion, Plug-in, GSK-3, structural biology

1. Introduction

The field of structural biology has made great advances in recent decades, and this progress has led to the emergence of numerous tools for the analysis of biological molecules and systems. Often, there is no time to learn all the techniques, as they require very specific knowledge and time. One of the solutions to this problem is the implementation of these programs in software management platforms that facilitate their use and integration with other software.

One of the most challenging techniques to use is molecular dynamics (MD). This computational technique allows the motion simulation of the atoms of a molecule or a set of molecules over time. To do so, the interactions between each of the atoms of the system are simulated by solving Newton's equations of motion for each type of atom and interaction[1].

1.1. Molecular dynamics

A few years ago, it was thought that proteins were rigid entities with restricted conformational changes. Nowadays, it is known that proteins are highly dynamical, and their conformations affect their biological functions. Before MD, the only way to know the dynamics of a protein was with complex and costly lab experiments. MD allows the study of protein dynamics without major expenses and much faster[2].

Simulations carried out with MD are used in a wide range of research fields and their applications are numerous as it is used in materials engineering, physicochemical studies, drug design and docking[3], peptide structure prediction, *ab initio* protein folding, protein structure refinement, etc[4].

To see the importance of this computational technique today, it can be looked the number of publications per year containing the term "Molecular dynamics" (Figure 1). It can be seen that the number of publications increases over the years and even with an exponential trend, with more than 50,000 publications during the year 2020.

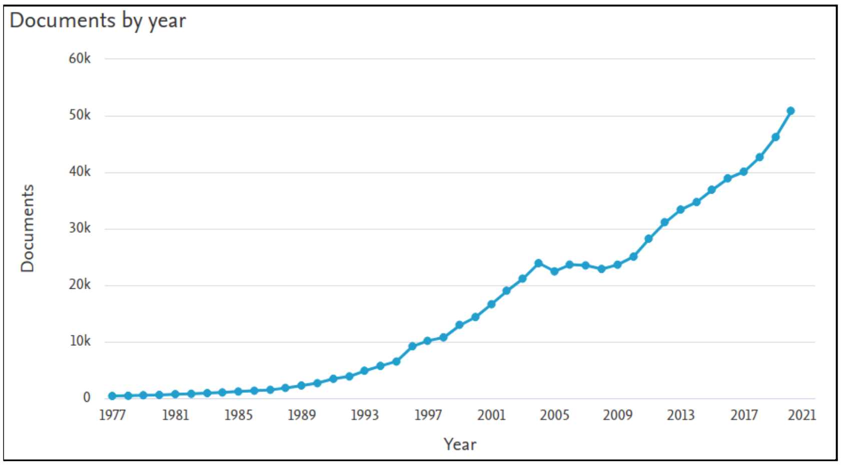


Figure 1. Number of publications by year with the term "Molecular dynamics" from 1977 to 2020. Graph generated by Scopus (Retrieved on 05/09/2021).

By areas, it can be seen that MD is a multidisciplinary tool, as it is present in scientific papers on chemistry, physics and astronomy, materials science, biochemistry, engineering, etc. (Figure 2).

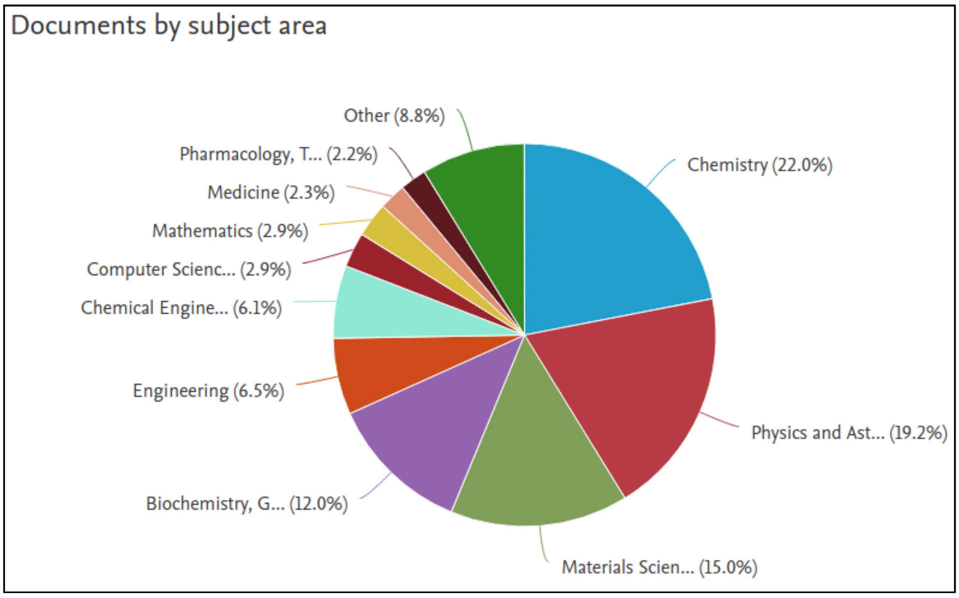


Figure 2. Percentage of documents by subject area containing the term “Molecular dynamics” from 1977 to 2020. Graph generated by Scopus (Retrieved on 05/09/2021).

MD, by combining the position of the atoms obtained from a structure file (PDB) with a specific set of mathematical equations called “force field”, is able to obtain the potential energy and force of the atoms and therefore their velocities and trajectories.
The set of equations that make up the force field has been obtained by parametrising semi-empirical quantum mechanical calculations or by fitting experimental data obtained from various techniques such as electron and X-ray diffraction, NMR, or neutron spectroscopy, among others[5].

The equations must be sufficiently precise for each type of atom and interaction, and simple enough to avoid large computational costs for the simulation. For proteins, certain force fields are used, such as OPLS-aa, CHARMM27, GROMOS43A1, AMBER96, etc., and each of them has different equations and constants specific to different types of proteins and interactions. For water molecules, which are very abundant in protein simulations, they have their own force field, which can be less complex, such as TIP3P, or more complex, such as TIP4P and TIP5P. The more complex the force field, the more reliable the results, but also the higher the computational cost.
To simulate the dynamics, periodic boundary conditions (PBC) are normally used to obtain a system without physical limits in which the protein can be without interaction with other proteins and at the same time is not a finite system.

The resulting product of a MD simulation is a trajectory file and a structure file that can be visualised and analysed to determine specific properties that occur during the simulation.

1.2. Applications of Molecular Dynamics

MD studies have evolved to the point of becoming a versatile technique capable of simulating various processes occurring at the macromolecular level. In structural biology, the field that concerns this work, the applications of MD can be divided between allosteric regulation of proteins, docking and drug design, and structure refinement[6].

In docking and drug design, MD is applied to understand how small molecules or drugs bind to proteins, since proteins aren't static entities, and by analysing protein-ligand binding only with a PDB structure, the elasticity of the molecules is lost and therefore possible interactions[6].

For protein structure refinement, MD is used to eliminate possible unnatural constrictions caused by the crystallization method used, or to simulate the protein in different conditions from those of crystallization. In MD, the protein can be subjected to changes in temperature, pressure, solvent, pH, etc. to obtain protein structures in their most natural state.

In addition, it is possible to obtain structures by sequence homology models that haven't been crystallized and optimize the structure with MD. An example of this is GSK-3α structure, obtained by sequence homology with GSK-3β, since it hasn't been possible to obtain its structure by crystallization.

MD has also proved to be effective for the study of allosteric regulation of proteins, as it allows to visualise the conformational changes that allosteric proteins undergo when binding a ligand, which can range from the movement of a few atoms to radical changes in the quaternary structure[7].

1.3. Molecular Dynamics workflow.

The workflow of MD to search the stability of a protein along the time carry out several steps (Figure 3). All the steps prior to the MD itself are to prepare the system so that it is a natural state and no abrupt changes occur at the start of the simulation. The system always begins with a Protein Data Bank (PDB)[8] file with the position of the atoms of the protein, obtained through different techniques like cryogenic electron microscopy (cryo-EM), homology modelling, NMR, X-Ray Crystallography, machine learning[9] etc. This file should contain only the atoms of the protein, being previously removed ligands, water molecules and hydrogens of the protein.

The first step of MD is the generation of the system. In this step, all the files needed to start the simulation are created, the force fields (set of equations with specific constants) to be used for the protein and the solvent are chosen, the system is solvated inside a box, and the ions are added. Most of the time, periodic boundary conditions for each box are chosen.
The second step is to perform an energy minimization process. This step is performed to find an energetic minimum of the protein to relax it in case it is in an unnatural conformation due to crystallization conditions.

The third step is to perform equilibration at constant volume, temperature and number of atoms (NVT), also called isothermal-isochoric or canonical equilibration. It consists of allowing the solvent to move but applying energetic restrictions to the heavy atoms of the protein so that they cannot move, but the solvent can adapt to it. This step is necessary because if it is not carried out, when the dynamics start, this adaption of the solvent to the protein will occur and is not a real situation and should therefore be avoided.

The fourth step consists of equilibration at constant pressure, temperature and number of atoms (NPT). It serves the same purpose as the previous equilibration and is also necessary.

The fifth step is the MD itself, in which the system no longer has artifacts, artificial conformation, or unwanted solvent-protein interactions. In this case, the result is a simulation of the trajectory of all protein atoms over time that can be observed with a visualizer and analysed.

The sixth step is the analysis of the simulation. In this final step, the trajectories of the simulation are extracted, modified in case of errors and certain values are extracted such as root-mean-square deviation (RMSD), root-mean-square fluctuation (RMSF), the accessible surface of the volume, the formation of pockets, etc.

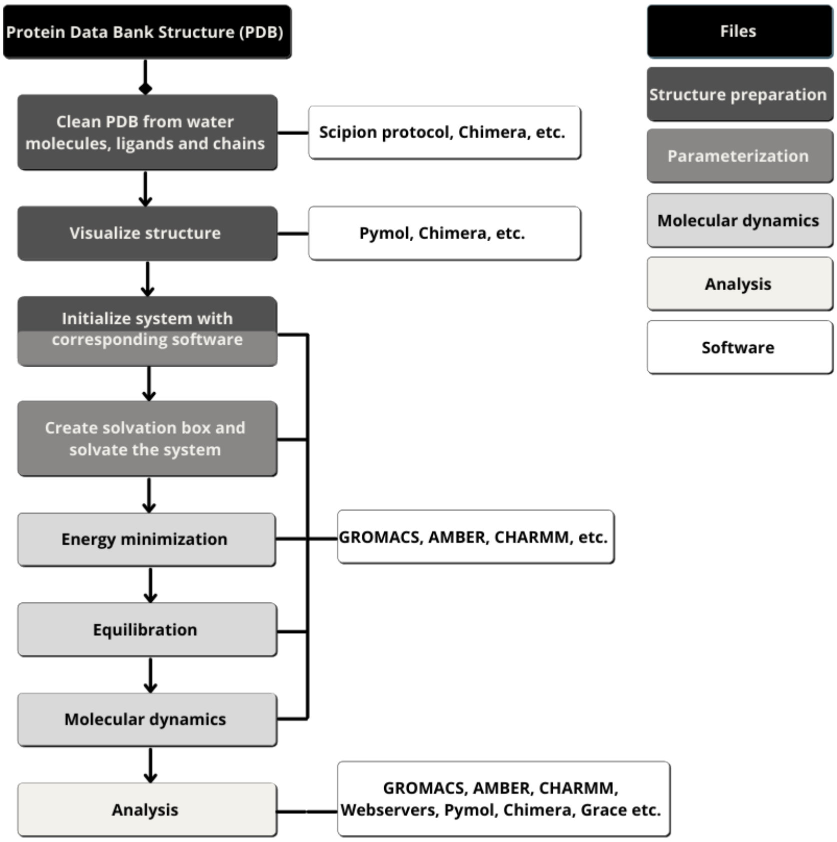

Protein Data Bank Structure (PDB)
Clean PDB from water molecules, ligands and chains
Scipion protocol, Chimera, etc.
Visualize structure
Pymol, Chimera, etc.
Initialize system with corresponding software
Create solvation box and solvate the system
Energy minimization
GROMACS, AMBER, CHARMM, etc.
Equilibration
Molecular dynamics
Analysis
GROMACS, AMBER, CHARMM, Webservers, Pymol, Chimera, Grace etc.
Files
Structure preparation
Parameterization
Molecular dynamics
Analysis
Software

Figure 3. MD workflow.

1.4. Scipion as an integrator platform

To perform convention MD, hundreds of parameters and dozens of files are used. Keeping track of all this is important, as each parameter can influence the result, and there is no point in performing MD without a clear purpose. Each purpose will determine the use of certain parameters, input files and analysis. The Scipion platform, an open source workflow management software with a plug-in structure, is capable of integrating diverse software, such as GROMACS in this case, to perform MD and keep a record of each step, as well as offering small guides during the process. Scipion emerged in 2016 as a tool for the integration of software dedicated to cryo-EM. Over the last few years, new plug-ins have been developed for new structural biology programs.

The aims of Scipion are varied. Firstly, it aims to bring together several structural biology data processing programs in one site, not only for ease of use, but also to be able to compare results from different software and combine them with each other.

Secondly, Scipion produces a record of every parameter, file and command used, thus facilitating further analysis of the results, while allowing the experiment to be reproduced in its entirety[10].

2. Objectives

Into this context the objectives of this work are:
Firstly, the implementation of the GROMACS software for MD on the Scipion workflow management platform. The purposes of this implementation are:

- To facilitate access to this computational tool for researchers who do not have the necessary computer skills to carry out the simulations.
- To be able to access the results and procedures used and obtain a record of all of them in order to be able to replicate, modify or publish them.
- Integrate another software into Scipion to be able to create more complex workflows, combining the new plug-in with previous plug-ins already implemented.

Secondly, the validation of the plug-in implemented by carrying out MD for the comparative study of GSK-3α, and GSK-3β kinases, exploiting an Asp-Glu switch, for the subsequent design of selective drugs for each of them[11].

3. Plug-in development

In this section will be discussed the technical aspects relative to the plug-in and its integration, in addition to the technicalities about Scipion.

All the plug-in will be developed in Python and Shell Scripting, and it will be available to consult in Github at the URL “https://github.com/Pefema/scipion-em-gromacs-TFM”

3.1 Scipion main aspects

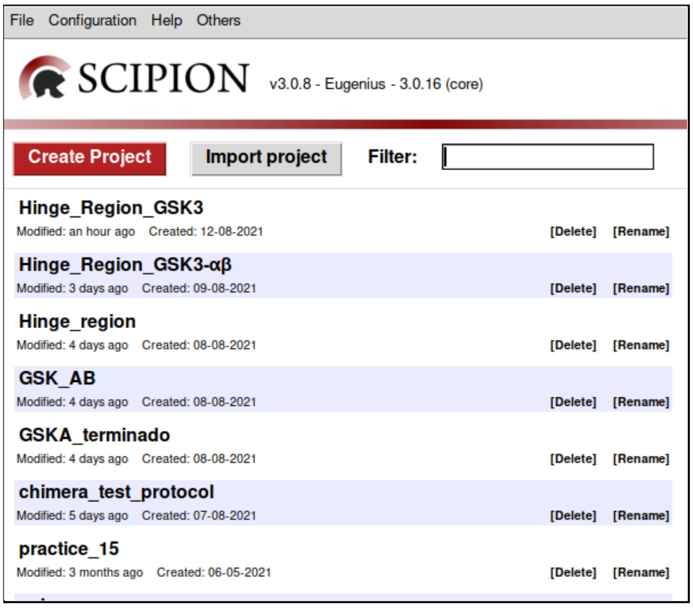
Scipion is a software management platform, developed at the National Center of Biotechnology (CNB, CSIC). This platform allows the execution of reusable, standardized, reproducible and traceable structural biology protocols. The procedures are implemented in plug-in format and are interoperable.

Scipion works on the basis of protocols, which are like boxes that perform a relatively high-level function. Each protocol requires an input and produces an output, which can be used by another protocol or downloaded directly from the platform. Protocols are organized by process of software, and the output of each protocol is produced in the most standard way possible so that it can be used by as much software as possible to promote interoperability.

Scipion has integrated many open source software and licensed software packages (there are over 70 plug-ins[12]). Free software can be automatically installed and used by any user, and licensed software requires the user to have the license on their computer in order to use it. It is important to note that Scipion has many programs to do the same work because it also has protocols to choose the best results from different software or even combine them.

3.2. Scipion's GUI

The first Scipion GUI consists of a project window where the user can choose, import or create a project for a specific work. (Figure 4)



Hinge_Region_GSK3
Modified: an hour ago Created: 12-08-2021

Hinge_Region_GSK3-αβ
Modified: 3 days ago Created: 09-08-2021

Hinge_region
Modified: 4 days ago Created: 08-08-2021

GSK_AB
Modified: 4 days ago Created: 08-08-2021

GSKA_terminado
Modified: 4 days ago Created: 08-08-2021

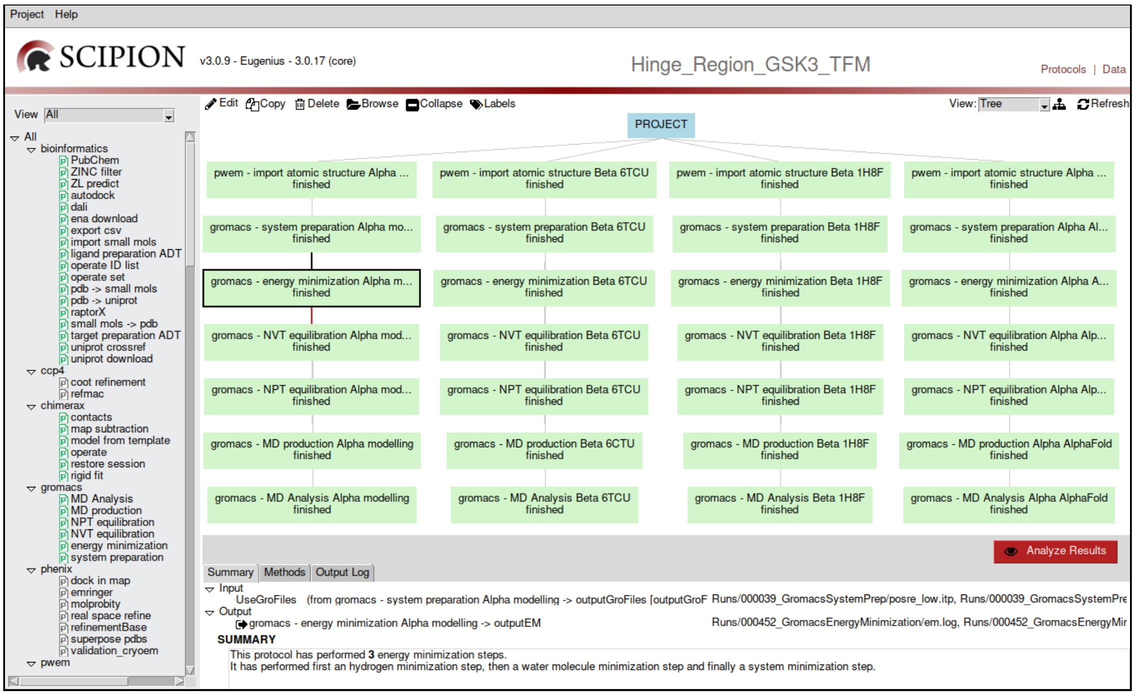
chimera_test_protocol
Modified: 5 days ago Created: 07-08-2021

practice_15
Modified: 3 months ago Created: 06-05-2021

Figure 4. Scipion's project window.

Scipion's main GUI is divided in 3 sections (Figure 5). In the left section, the user will choose which protocol he/she wants to use. The protocols are divided by plug-in, functions, etc. The right upper section consists of the project itself, and is the place where all the protocols will be displayed. It follows a workflow distribution, with all related protocol connected by lines.

All files generated or used by a protocol are stored in the button “Browse” and can be displayed.

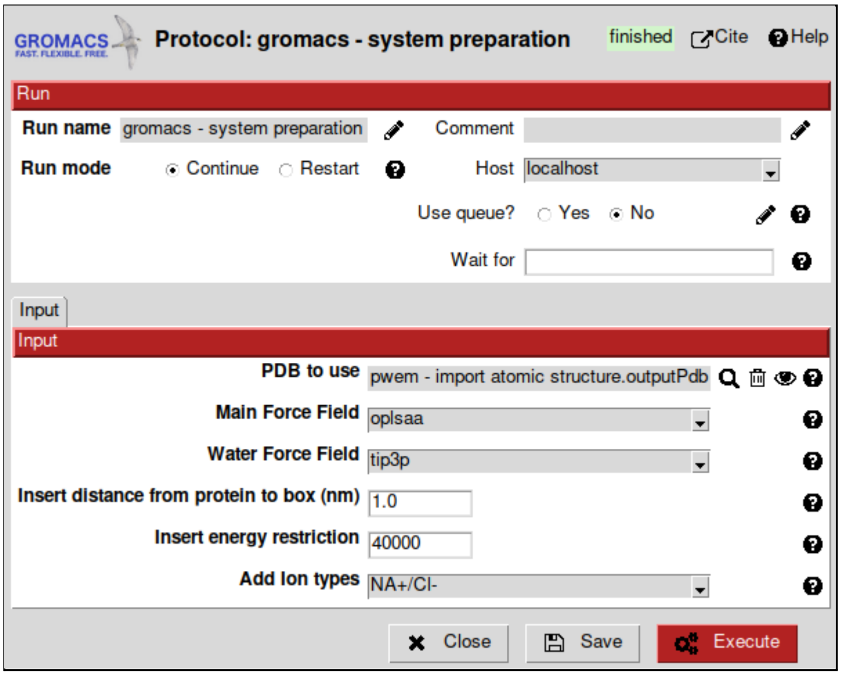

Project Help
SCIPION v3.0.9 - Eugenius - 3.0.17 (core)
Hinge_Region_GSK3_TFM
Protocols | Data
View All
Edit Copy Delete Browse Collapse Labels
View: Tree
Refresh
PROJECT
pwem - import atomic structure Alpha ... finished
pwem - import atomic structure Beta 6TCU finished
pwem - import atomic structure Beta 1H8F finished
pwem - import atomic structure Alpha ... finished
gromacs - system preparation Alpha mo... finished
gromacs - system preparation Beta 6TCU finished
gromacs - system preparation Beta 1H8F finished
gromacs - system preparation Alpha Al... finished
gromacs - energy minimization Alpha m... finished
gromacs - energy minimization Beta 6TCU finished
gromacs - energy minimization Beta 1H8F finished
gromacs - energy minimization Alpha A... finished
gromacs - NVT equilibration Alpha mod... finished
gromacs - NVT equilibration Beta 6TCU finished
gromacs - NVT equilibration Beta 1H8F finished
gromacs - NVT equilibration Alpha Alp... finished
gromacs - NPT equilibration Alpha mod... finished
gromacs - NPT equilibration Beta 6TCU finished
gromacs - NPT equilibration Beta 1H8F finished
gromacs - NPT equilibration Alpha Alp... finished
gromacs - MD production Alpha modelling finished
gromacs - MD production Beta 6CTU finished
gromacs - MD production Beta 1H8F finished
gromacs - MD production Alpha AlphaFold finished
gromacs - MD Analysis Alpha modelling finished
gromacs - MD Analysis Beta 6TCU finished
gromacs - MD Analysis Beta 1H8F finished
gromacs - MD Analysis Alpha AlphaFold finished
All
bioinformatics
PubChem
ZINC filter
ZL predict
autodock
dali
ena download
export csv
import small mols
ligand preparation ADT
operate ID list
operate set
pdb -> small mols
pdb -> uniprot
raptorX
small mols -> pdb
target preparation ADT
uniprot crossref
uniprot download
ccp4
coot refinement
refmac
chimerax
contacts
map subtraction
model from template
operate
restore session
rigid fit
gromacs
MD Analysis
MD production
NPT equilibration
NVT equilibration
energy minimization
system preparation
phenix
dock in map
emringer
molprobity
real space refine
refinementBase
superpose pdbs
validation_cryoem
pwem
Analyze Results
Summary Methods Output Log

```
▽ Input
    UseGroFiles   (from gromacs - system preparation Alpha modelling -> outputGroFiles [outputGroF  Runs/000039_GromacsSystemPrep/posre_low.itp, Runs/000039_GromacsSystemPre
▽ Output
    gromacs - energy minimization Alpha modelling -> outputEM                                    Runs/000452_GromacsEnergyMinimization/em.log, Runs/000452_GromacsEnergyMir
  SUMMARY
    This protocol has performed 3 energy minimization steps.
    It has performed first an hydrogen minimization step, then a water molecule minimization step and finally a system minimization step.
```

Figure 5. Scipion’s project page with GROMACS workflow.

Finally, the left lower section represents the information section, and there it is displayed a summary of the protocol, the methods it has used and de output log with all the steps the protocol has made, and the output generated by the internal software.

Each protocol has its own GUI asking the user the necessary parameters to run it (Figure 6). It also has a common section (run) with some information regarding the name, comments, and other Scipion parameters.



Protocol: gromacs - system preparation

finished

Run

Run name: gromacs - system preparation

Comment:

Run mode: ● Continue ○ Restart

Host: localhost

Use queue? ○ Yes ● No

Wait for:

Input

PDB to use: pwem - import atomic structure.outputPdb

Main Force Field: oplsaa

Water Force Field: tip3p

Insert distance from protein to box (nm): 1.0

Insert energy restriction: 40000

Add Ion types: NA+/Cl-

Figure 6. Example of Scipion's protocol (GROMACS). System preparation protocol.

3.3. Software to wrap.

The software that is going to be wrapped is GROMACS[13], in order to do the process of MD, and GRACE[14], in order to extract visual information from the analysis protocol.

GROMACS (GROningen MAchine for Chemical Simulations) is a software to perform MD, mainly developed for protein, lipid and nucleic acid dynamics simulations. It is free and open-source, and it was originally developed in the University of Groningen. Although being open-source, it is one of the fastest and most popular MD software available. It doesn't have a GUI, so it can only be called from the command line. The version used for this work is GROMACS 2020.4.

Some GROMACS commands produces .xvg files, which are intended to be processed by GRACE software, as it is written in the GROMACS manual. GRACE (GRaphing, Advanced Computation and Exploration of data) is a free 2D graph plotting tool. It has a GUI and also can be called from the command line. It is commonly used by other MD software like, NAMD[15] or Visual Molecular Dynamics.

3.4. Protocol basics.

Scipion's protocols consist of a window in which all the parameters and input files necessary to launch the program specified in that box are added. Each protocol has some common parameters (Run) and other specific parameters for each protocol (Input) as well as a description of its functions (help) and its bibliographic reference (cite).

In the Run section, there are several fixed parameters:

- **Run name**: Name of the protocol. It gives by default a predefined name that can be changed.
- **Comment**: Parameter to add a comment if desired.
- **Run mode**: Parameter to choose whether the protocol should continue from the last executed step or from the beginning.
- **Wait for**: Parameter to determine the start of the protocol after the end of a previous one.
- **GPU IDs (optional)**: If present in the protocol and the GPU usage and identifier is specified, it will be used to run the protocol program.
- **Expert level (optional)**: If present, it allows switching between Normal and Advanced parameters in the Run section.

The parameters in the input field will be the ones specifically used by the protocol to perform the task and call the execution program.
Each parameter has a small guide explaining its characteristics and what it will be used for. In addition, in certain cases, the files introduced as parameters have the option of being displayed (eye).

Different types of parameters can be requested in a protocol: integers (Integer), booleans (Boolean), floating numbers (Float), paths to files (Path), lists of options (Enum) and Scipion objects (Pointer), among others.

Each file can be stored as a Scipion object. In order to do so, such an object must be created, which will have certain characteristics and will depend on its nature. These objects are the ones that can be called with the Pointer parameter in other steps or protocols. They can be grouped together to form sets of objects with the same characteristics or that will be used by the same protocol later on. The objects stored in a set can be retrieved, calling them with a function.

3.5. Workflow to wrap

For the development of the GROMACS plug-in in Scipion, 6 protocols will be produced (Figure 7).

1. The first protocol will be used for the generation of the system. In this protocol, several files will be produced. The first file will be a PDB analogue in GROMACS format (gro), with all atoms defined with the chosen force field. In this file the hydrogens will have already been added, in case they were previously absent. A topology file (topol.top) will also be produced, which will be used on multiple occasions and will be explained in detail later. Finally, a file is produced for the positional constraints of the heavy atoms, to be used in the energy minimization and equilibration protocols (posre.itp). These three files are the first generated files that will be used in the following protocols. For the first protocol, it has been decided to link all the commands that prepare the system: Generation of the basic files for the use of GROMACS (pdb2gmx), definition of the box for the subsequent solvation (editconf), solvation of the system (solvate), generation of the tpr format file (grompp) and the application of the tpr file for the addition of ions to the system (genion).

2. The second protocol consists of an energetic minimization of the system to relax the system, in case it has constraints that modify its state. For this protocol, the commands are used to generate the tpr file (grompp) and to produce the energy minimization (genion). There will be from 1 to 3 energy minimization steps to be carried out, depending on the user decision.

3. The third protocol consists of an equilibration of the system at constant volume, temperature and number of atoms (NVT). It is also called isothermal-isochoric or canonical equilibration. The aim of this equilibration is to stabilize the temperature of the system without causing changes in the protein, making the solvent added in the first step with the "solvate" command adapt to the protein without causing unwanted changes. This is done by using the energy constraint file generated with the "pdb2gmx" command to prevent the movement of heavy atoms (all atoms other than hydrogen). The movement of these atoms is allowed, but only after overcoming a significant energy penalty. It is a necessary step, because the added solvent might be optimized within itself, but not with the protein.

4. The fourth protocol consists of another equilibration of the system, but this time at constant pressure, temperature and number of atoms (NPT). It is also called isothermal-isobaric ensemble. The aim of this equilibration is to stabilize the pressure, and thus, the density of the system.

5. The fifth protocol consists of the production of MD. The previous steps have made it possible to obtain a system with the chosen solvent, ions and force field, energetically minimized, and therefore, relaxed of stresses produced by crystallization and equilibrated at the desired temperature and density. In this case, the protocol will simulate the movement of the protein under stable conditions, which will allow the natural movement of the protein to be observed during the desired simulation time, which may vary depending on the objective of the simulation.

6. The sixth and last protocol consists of the analysis and post-processing of the MD. For post-processing, a command has been used to centre the protein in the box, in case it has moved out of the box during the simulation, and to fix the protein backbone to avoid twists in the visualization (trjconv). Several commands have also been introduced to perform various analyses such as the calculation of the RMSD (rms), the RMSF (rmsf), the surface area accessible to the solvent throughout the simulation (sasa), the number of hydrogen bonds of a specific residue formed and destroyed during the simulation (hbond), the analysis of the main cluster of frames produced (cluster), or tools to obtain the reference structure of the first frame of the dynamics (trjconv), or the reduction of frames for better handling of the dynamics (trjconv).

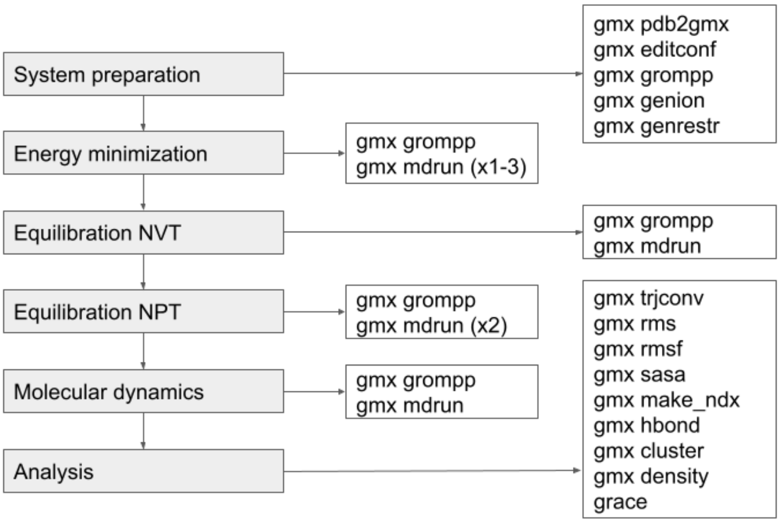


Figure 7. MD workflow implemented in Scipion. The grey boxes represent the protocols, and the white boxes the commands used in each protocol.

3.6. Parameters and files

The parameters chosen to show to the user are a few compared to all the parameters GROMACS have. This is due to simplification as well as assuring that the protocol will work and will perform a correct MD study.

If the protocol needs to be upgraded and show more parameters it can be easily done.
The chosen parameters and options are the ones that make a significant change in the MD of the protein.

The files produced by the GROMACS Scipion protocol are stored as Scipion objects and then retrieved when needed. The files are named with an identifier (first PDB filename) followed by the specific file descriptive name and file type. There have been used or produced 12 file types by GROMACS (Table 1).

Table 1. File types used or produced by GROMACS and description (GROMACS manual).

File type	**Description**
gro	Contains a molecular structure in Gromos87 format (GROMACS file).
top	Stands for topology. It contains information relative to the molecule within a simulation.
itp	File extension included in topology files. It usually contains information for position restrains.
tpr	Portable binary run input file. It contains the starting structure of the simulation, the molecular topology and all the simulation parameters.
pdb	File extension for Protein Data Bank files containing the description of atom's positions in a molecular structure.
trr	Contains the trajectory of a simulation (Coordinates, velocities, forces and energies).
log	Logfile generated by GROMACS, usually in human-readable format.
edr	Portable energy file. It contains all energy terms collected during energy minimization.
xtc	Portable format for trajectories. The trajectories are stored using a reduced precision algorithm which makes several data transformations for space optimization.
xvg	GRACE files. GRACE software uses these files to plot information and results.
cpt	Portable checkpoint file, with complete state of the simulation.
ndx	Index file, which contains some user definable sets of atoms.

4. Validation

In order to validate the GROMACS MD plug-in introduced in Scipion, it has been used in a case study. A comparative analysis of GSK-3α and GSK-3β. (Glycogen Synthase kinase 3)

4.1. Introduction

GSK-3 is a very interesting therapeutic target for its implication in a variety of diseases such as non-insulin-dependent diabetes mellitus[16], bipolar disorder[17], Alzheimer's disease[17], and cancer[18–20].

Some drugs have been found to be effective to inhibit the action of both GSK-3α and GSK-3β, but the inhibition of both proteins produces the stabilization and nuclear translocation of β-catenin, a GSK-3 substrate, leading to drug toxicities.

Doble *et al*[21]*.* discovered that the knock-out of only one of the GSK-3α paralog does not increase β-catenin and Guezguez *et al*[22]*.* determined that this knock-out inhibits the formation of acute myeloid leukemia in mice.

GSK-3 paralogs have a high homology in their binding site and therefore is it difficult to discover a paralog-selective inhibitor drug. The only difference present in the paralogs is an Asp-Glu (Asp133 (β) → Glu99 (α)) amino acid switch in the binding site, which should be exploited in order to design selective inhibitors[11].

It is therefore considered a good case for the comparative MD analysis of these two paralogs with the implemented plug-in in Scipion. Performing MD could show some differences which may not have been seen in other structural biology analysis.

4.2. Materials and methods.

In this section it will be exposed the materials and methods used to perform the MD with the Scipion plug-in.

4.2.1. Protein structures

The protein structures used for the analysis have been extracted from the Protein Data Bank (PDB) and Modelarchive[23].

In the case of GSK-3α, the protein could not be obtained from the PDB as it hasn't been crystallized. The protein structure was obtained from Modelarchive, and it was produced in 2006 by homology modelling of its paralog GSK-3β structure, using MODELLER 7V7[24].

The structure from which GSK-3α was produced is the protein with identifier 1H8F[25] from the Protein Data Bank, and it has been used for the simulation. It was obtained in 2001 by X-ray diffraction with a resolution of 2.80 Angstroms.

In order to obtain some more results from the analysis, it has been also obtained the GSK-3α from the recently released AlphaFold Protein Structure Database[9]. The structures from AlphaFold DB have been obtained from the prediction of the 3D protein's structure from its amino acid sequence, using artificial intelligence. Its accuracy is compared with experimentally obtained protein structures. It is an AI system developed by DeepMind.

It has also been added, to perform MD, another GSK-3β protein with better resolution than 1H8F. It is the structure with identifier 6TCU[26] from the PDB, with a resolution of 2.14 Angstroms, obtained with X-ray diffraction on 2019.

4.2.2. Structure preparation

The structures have been cleaned of duplicated chains, solvent molecules and heteroatoms in all cases. All molecules are complete and do not need to be modelled with additional software.

The AlphaFold protein have been cleaned of non-confident amino acid sequences, which were the C-terminus and N-terminus of the protein.

4.2.3. System preparation for GROMACS

To start the simulation, the file was introduced in Scipion with the *Import atomic structure* protocol. First, the MD was performed for GSK-3β and then for GSK-3α with the same parameters.

With the *System preparation* protocol of GROMACS, the PDB file of the previous protocol was inserted and a series of parameters were selected:

- **PDB to use:** The clean PDB file has been inserted.
- **Main Force Field:** The OPLS-aa force field was chosen, as it is a fairly generic force field and is highly optimized for proteins without ligands.

- **Water Force Field:** TIP3P has been chosen, as it is a very optimized force field for water molecules, with much lower computational costs than TIP4P and TIP5P and for the required simulation it is not necessary to use a more complex water force field.
- **Insert distance from protein to box (nm):** 1 nm has been chosen for this parameter, as the box has a margin of 1 nm, the protein will always be 2 nm away from its own image and will not produce undesired interactions. Increasing the margin also increases the number of solvent molecules needed to fill the box, and therefore the computational time.
- **Insert energy restriction:** For the energy restriction applied to the heavy atoms of the protein in later protocols, 40000 KJ/mol/nm² has been chosen.
- **Add Ion types:** In this case K+/Cl- has been chosen, although in this case the ion pair chosen is not very relevant.

From the protocol, the necessary files have been generated so that GROMACS can start the simulation in the subsequent protocols.

4.2.4. Energy minimization

The *Energy minimization* protocol has been used to carry out the energy minimization. This step is important so that the protein structure does not have highly energetic conformations at the beginning of the simulation that could disturb the protein state. For this protocol, the parameters previously predefined by Scipion have been chosen.

- **Use default parameters:** Yes, as no input file will be provided.
- **Energy minimization steps:** Hydrogen+Water+System. In this case, 3 energy minimization steps have been carried out. The first one on the hydrogens of the system, the second one on the water molecules and the third one on the heavy protein atoms.
- **Algorithm for hydrogen energy minimization:** Steepest descent. It has been chosen for the first one because the minimization of hydrogens is not a step that needs to be very precise, and the steepest descent algorithm is less time-consuming than conjugate gradient.
- **Algorithm for waters energy minimization:** Steepest descent. It has been chosen for the same reason as the previous parameter.
- **Algorithm for system energy minimization:** Conjugate gradient. It has been chosen because as it is the minimization of the whole system, and it is convenient to find the minimum energy in a more precise way.
- **Nstcsgsteep for CG algorithm in System minimization:** 1000. A steepest descent step is added every 1000 steps so that the local minimum is reached faster.
- **Number of steps (nsteps):** 50000 steps, which represents the maximum number of steps to find the energy minimum and is sufficient for the minimization given the small size of the protein.
- **Input Gro File:** The object created in the previous protocol is inserted, representing the set of all generated files.

Once energy minimization is complete, NVT equilibration can be performed without risk of the protein being in high-energy positions produced by the crystallization process or by homology modelling (GSK-3α).

4.2.5. NVT equilibration.

Equilibration at constant volume is carried out with the *NVT equilibration* protocol.
The parameters entered are as follows:

- **Input GROMACS file set:** The files generated by the first protocol are entered.
- **Input em_gro File:** The set of files generated in the *Energy minimization* protocol is entered.
- **Coulomb distance cut-off (rcoulomb):** 1.0 Angstroms. This value should be as low as possible and reasonable for the simulated system. It is decisive for the MD result. The lowest recommended value by GROMACS has been chosen.
- **Number of steps (nsteps):** 150000 steps.
- **Time step for integration (dt):** 0.002, which multiplied by 150000 is 300 ps.
- **Number of Lincs iterations (lincs_iter):** 1. This represents the number of iterations to correct for rotational elongation in LINCS. For normal runs, a single step is sufficient. 2 would be more accurate but computationally expensive.
- **Interpolation order for PME (pme_order):** 4. Equivalent to cubic interpolation. For larger systems, it could be 6, 8 or 10.
- **Temperature coupling (tcoupl):** V-scale is used to use velocity rescaling with a stochastic term. It is similar to Berendsen, but the stochastic term is used to generate a suitable canonical set.
- **Reference temperature in Kelvin (ref_t):** 300 Kelvin as it is almost a physiological temperature.
- **Generation seed (gen_seed):** -1 in order to be selected randomly.

After the NVT equilibration, the files will be saved in an object for the next protocol.

4.2.6. NPT Equilibration.

NPT equilibration takes place with the *NPT equilibration* protocol. The parameters chosen are the same as in the previous protocol except for the following:

- **Number of steps (nsteps):** 300000 steps
- **Time step for integration (dt):** 0.001 dt, which will produce a 300 ps equilibration.

The time step for integration has been reduced as GROMACS has failed for higher time steps. By halving the time step for integration, the number of steps has been doubled so that the equilibration time is the same in NVT and NPT and sufficient in both cases. Once 1 NPT equilibration with energy constraints and 1 equilibration with fewer constraints have been produced, the protein is ready for MD without pressure and volume effects that could cause artefacts during the first picoseconds of the simulation.

4.2.7. Molecular Dynamics.

The *MD production* protocol is used to produce the MD. The parameters chosen for the simulation are the same as above, except for the following:

- **Number of steps (nsteps):** 5000000 steps.
- **Time step for integration (dt):** 0.002.

With the two previous parameters, a MD simulation of 100 nanoseconds is obtained.

Once the MD have been produced, the files can be analysed with the *MD analysis* protocol.

4.2.8. Results analysis.

The results of the dynamics will first be analysed with the methods offered by GROMACS.

- **Choose files source:** MD, to choose files from the previous protocol instead of from local files.
- **Perform analysis on NDX selection:** No, because the analysis will be performed on the whole protein except for the hbond.
- **trjconv analysis:** Yes, to fix the protein trajectory in case it has been left between boxes, as well as to fix the backbone in the centre of the protein.
- **rms analysis:** Yes, analysis of the RMSD of the protein.
- **rmsf analysis:** Yes, analysis of the RMSF of the protein.
- **sasa analysis:** Yes, analysis of the solvent-accessible surface of the protein.
- **hbond analysis:** Yes, analysis of the hydrogen bonds that are created and destroyed on a specific amino acid during the simulation.
- **Input residue to perform Hbond analysis:** 99 for GSK-3α protein and 133 for GSK-3β protein, corresponding to the amino acids that differentiate the two proteins at the ATP binding site and represent glutamine and aspartic acid, respectively.
- **Cluster analysis:** Yes, to obtain the most representative clusters of the MD, from which its reference structure can be obtained and to determine different conformations of the protein throughout the MD.
- **Enter cutoff for cluster analysis:** 0.1, is the cutoff chosen to perform the cluster analysis.
- **Density analysis:** Yes, to obtain the protein density along the MD.
- **Get Reference Structure from the first MD frame in PDB format:** Yes, the reference structure is used for further analysis such as analysing the pockets that are created and destroyed throughout the dynamics.
- **Reduce the number of frames:** Yes, to reduce the number of frames in the dynamics to reduce the final file size.
- **Extract frames every dt =:** 100, to extract 1 frame out of 100 from the MD.

The *MD analysis* protocol will produce some PNG images with the chosen analysis. Furthermore, the user will have all the files generated in case it is wanted some extra analysis or visual curation of the graphs.

4.2.9. TRAPP

Subsequently, once the reference structure and the reduced trajectory have been obtained, the TRAPP Webserver[27] (TRAnsient Pockets in Proteins) can be used to analyse the pockets that appear and disappear during the simulation. This is important because the hypothesis of the comparative study seeks to find differences in the binding pocket in order to subsequently be able to carry out a selective docking study. The reduction in the number of frames has been obtained because TRAPP has a file limit size of 500 MB.

TRAPP also requires the coordinates of the catalytic pocket to produce the simulation, and it has been chosen the spatial point between the HB2 atom of Leu98 and the O atom of Glu99

(Target) for the α-paralog and the HB2 atom of Leu132 and the O atom of Asp133 (Target) for the β-paralog.

The results of the TRAPP analysis produce a PyMOL[28] session file to visualize the pockets that are formed and destroyed during the MD 25%, 50% and 95% of the simulation.

4.2.10. PyMOL analysis

After performing a clustering analysis on the trajectories and extracting the most representative frame of the most representative cluster of each structure, it has been compared in PyMOL, and it has been calculated the RSMD of a selection of residues close to the target residue Asp133 for GSK-3β and Glu99 or Glu196 for GSK-3α. The RMSD will give a measure of difference between the distribution of the closest residues to the target one.

4.3. Hypothesis

In the binding site of both proteins, there is just one amino acid which changes in respect to the other paralog. In the case of GSK-3α, it has a Glutamic Acid while GSK-3β has an Aspartic Acid. The Glutamic Acid has one carbon more than the Aspartic Acid, which makes it longer. This extra carbon could allow the Glutamic Acid to make an interaction with a Lysine behind the pocket which could lead to a slight modification of the pocket and therefore, some selective drugs could be explored taking advantage of this difference. There also could be other interactions formed between the Glutamic Acid or the Aspartic Acid and other residues.

4.4. Results and discussion

Before analysing the simulation results, it should be noted that the product of 2 MD of 100 nanoseconds can produce non-significant results. It would have been desirable to run several simulations of each protein to obtain more conclusive results. It was not possible to run several simulations due to the duration of each simulation, which varied between 5 and 6 days. Once this observation has been made, we proceed to the interpretation of the results.

Several results have been extracted from the analyses carried out with GROMACS through Scipion and from other analyses carried out with files produced by the plug-in and introduced in other software (PyMOL, GROMACS, TRAPP).

First, control charts have been obtained to check that the energy minimization, NVT equilibration and NPT equilibration have been carried out correctly (Figure 8). These graphs have been generated automatically by the Scipion protocols. As can be seen from them, the system values remain stable. In the case of the pressure, the value fluctuates between -400 and 350 bar, but these values are to be expected for a non-gaseous system.

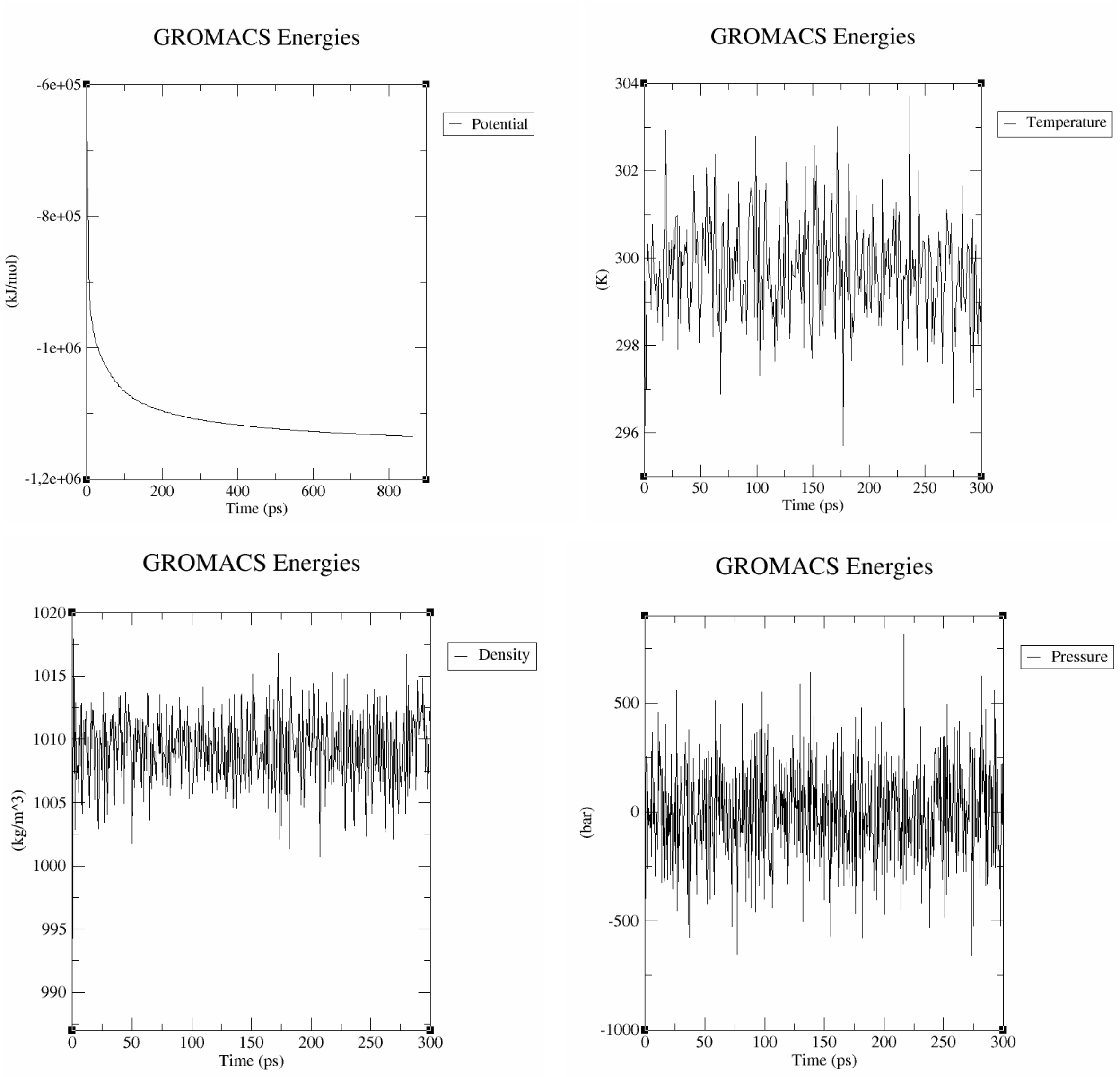


Figure 8. Four graphs representing the values of potential, temperature, density and pressure after energy minimization, NVT equilibration and NPT equilibration respectively from GSK-3α.

Once the dynamics had been completed, the *MD analysis* protocol processed the MD files and several other files were obtained.

First, a correction of the dynamics was carried out so that, in the event that the protein had partially left the box, it would be centred. Subsequently, the trajectory is partially fixed to the protein backbone so that the protein does not show movements that make it difficult to visualize, although the intrinsic movement of the protein produced during the dynamics is never modified.

An analysis of the files generated by the *MD analysis* protocol will now be carried out:

First, the RMSD is analysed along the dynamics (Figure 9). The root-mean-square-deviation of the protein atoms can be seen to remain relatively stable during the simulation.

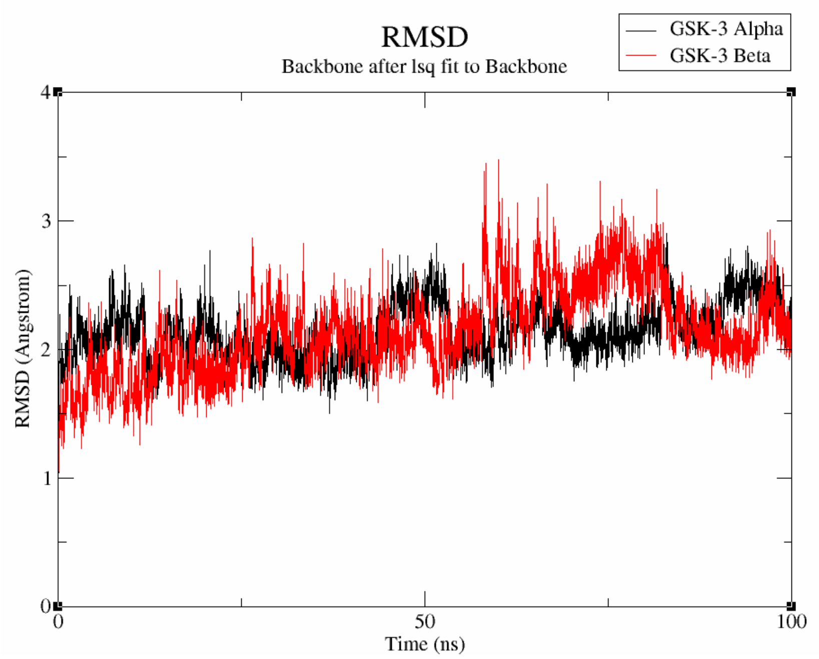


Figure 9. RMSD de GSK-3α y GSK-3β (1H8F).

The RMSF has also been produced (Figure 10). In order to compare both proteins, the X axis of them have been scaled. The results show how the RMSF of both is similar but with some regions that differ. The regions that differ are around amino acids 99, 108, 120, 175, 202, 212 and from 310 to 342. The closest regions that differ around the target amino acid are the target amino acid itself and Arg207 and Val108 amino acids. All the other regions are far from the target region, and they are mainly loops which seems to have a lot of mobility. Arg207 and Val108 could be further studied. They are 10 Angstroms away from Glu99, and they are part of an α-helix

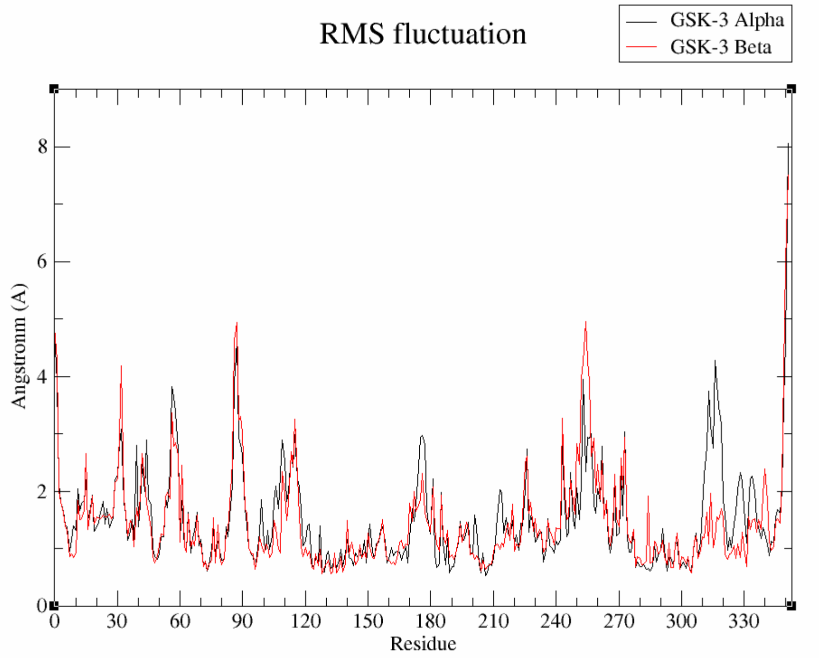


Figure 10. RMSF of GSK-3α and GSK-3β (PDB code 1H8F). Target amino acid is number 99 (Asp-Glu).

The analysis of hydrogen bonds taking place during the MD simulation between the target amino acids and the rest of the protein show how the number of hydrogen bonds between Glu99 with the rest of the residues remain almost stable in the case of GSK-3α but more unstable in the case of GSK-3β, showing ups and downs throughout the simulation (Figure 11).

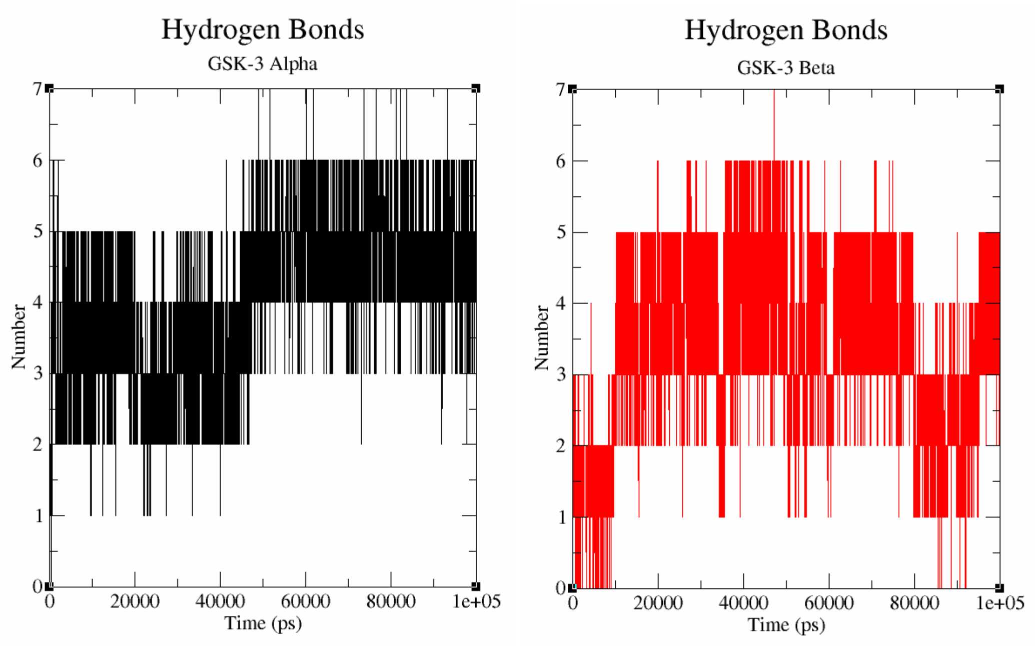


Figure 11. Number of Hydrogen Bonds of target amino acid (Glu99-Asp133) during the MD simulation of GSK-3α and GSK-3β (1H8F) respectively.

The TRAPP results have been performed successfully, and they show a difference between GSK-3α and GSK-3β. As can be seen in the figure 12, the TRAPP results show that in GSK-3α disappears a pocket and another pocket appear in the region formed between the amino acids Leu154, Cys165 and Val36, whereas in the case of GSK-3β two pockets are formed. One of them is next to the Tyr134 and the other one is next to the Cys199 like in GSK-3α.

Taking into account the results obtained, it could be said that it would be an interesting approach to identify new drugs taking into account the closeness of the formed pocket in TRAPP to the Tyr134, which seems to be a selective aspect of the paralog.

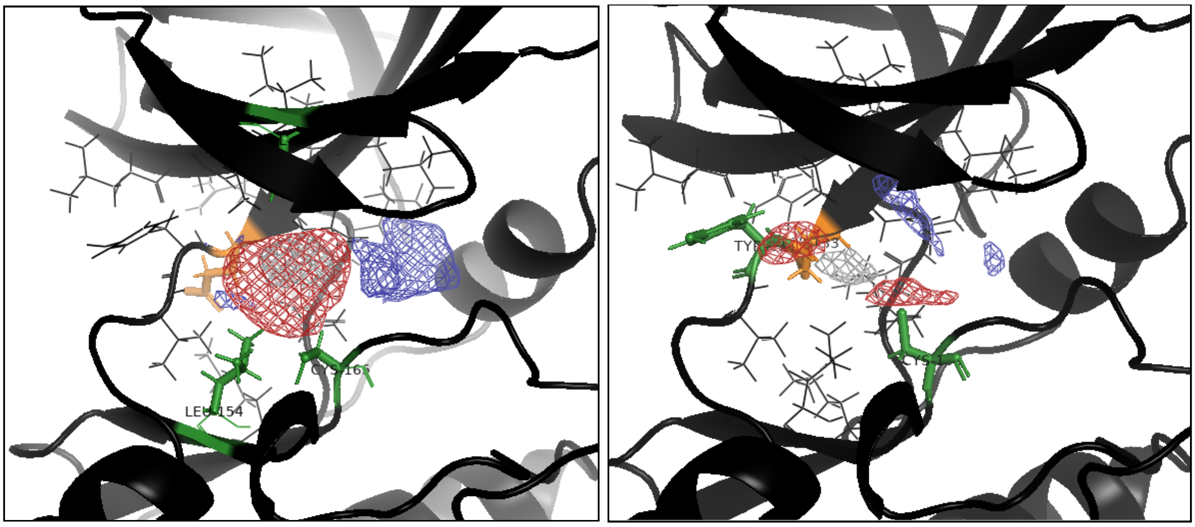


Figure 12. TRAPP analysis of GSK-3α and GSK-3β respectively. Red area indicates the formation of a pocket during the dynamics at least 50% of the time. Blue area indicates the destruction of the pocket during the dynamics at least 50% of the time. Green amino acids indicate closeness to the formed pocket. The target amino acid is represented in orange.

The analysis with PyMOL has been performed superposing GSK-3α and GSK-3β structures (Figure 13), selecting both target residues and all residues around them in a 4 Angstrom distance. Then, under that selection, it has been calculated the RMSD, and it has resulted in 0.792 Angstroms for 11 residues.

As it can be seen in figure 14, there are observable differences between the residues selected in TRAPP analysis for further study (Tyr100, Leu154 and Cys165). Tyr100 of GSK-3α and Tyr134 of GSK-3β are differently oriented, and they are the most different selected residue within structures.

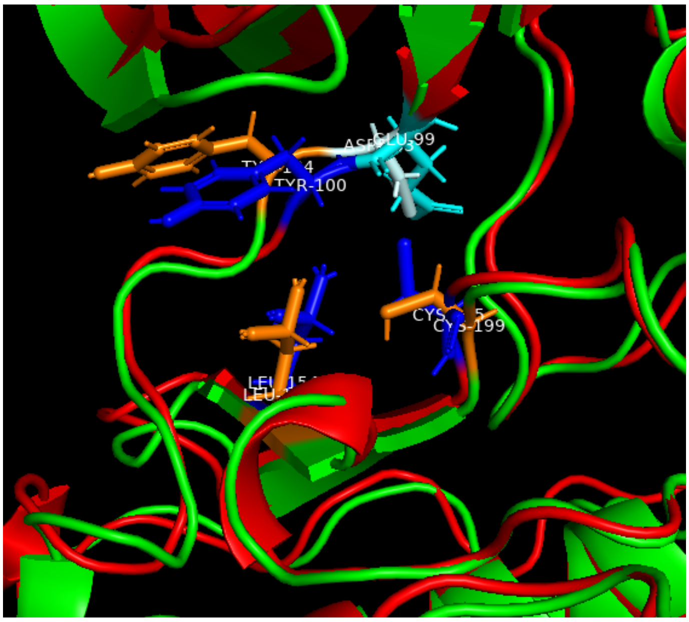


Figure 13. Superposed GSK-3α (red) and GSK-3β (green). The marked residues are Tyr100, Leu154 and Cys165 (dark blue, GSK-3α), Tyr134, Leu188 and Cys199 (orange, GSK-3β), Glu99 (cyan, GSK-3α) and Asp133 (gray, GSK-3β).

The results of the MD performed by the GROMACS plug-in in Scipion show that the two paralogs have differences, as it has been determined by the different analysis, enough to try to obtain new selective drugs for one of the paralogs.

The targets for the development of a new drug could be the Tyr134 for GSK-3β as it shows a formed pocket near to it in TRAPP analysis, and it is also oriented differently within the paralogs. It could also be studied a new drug targeting some amino acids like Leu154 or Val36 in GSK-3α.

From the results produced by the Scipion plug-in alone, several differences can already be observed, such as the number and stability of hydrogen bonds of the target amino acids or differences in the RMSF. Moreover, thanks to the plug-in, 4 MD could be prepared quickly and easily, and a record of each step and parameter used could be obtained.

4.5. GSK-3α from AlphaFold and GSK-3β 6TCU results

In order to have another point of view of the paralogs, two more structures have been used to perform some analysis. The structures that have been used are another GSK-3β with PDB Code 6TCU with better resolution, and another GSK-3α structure obtained with a machine learning approach, from the AlphaFold database.

In the hydrogen bond analysis (Figure 14) it can be observed that the hydrogen bonds formed within Glu99 and the rest of the protein are more stable in GSK-3α from AlphaFold in respect to the one from modelling. In the case of the 3D structure of 6TCU, the result is also unstable compared to GSK-3α but more stable than 1H8F.

In the figure it can be easily observed that GSK-3α has 6 hydrogen bonds almost all the time while GSK-3β has only 5 and with more variations.

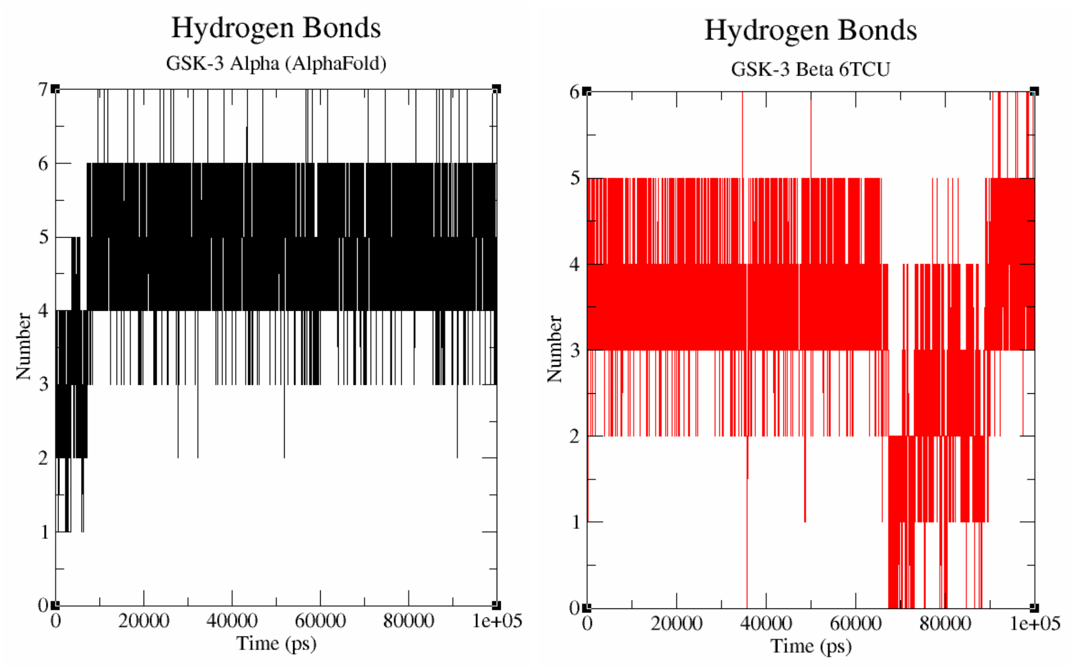


Figure 14. Number of Hydrogen Bonds with target amino acid (Glu-Asp) during the MD simulation of GSK-3α from AlphaFold and GSK-3β (6TCU)

5. Conclusion

The conclusions that can be drawn from the integration of the GROMACS plug-in in Scipion for MD are diverse:

Firstly, we have obtained a tool that allows, with basic knowledge of structural biology, to perform MD, a complex technique that would otherwise require very specific knowledge to be carried out. This technique could be used to complement structural biology studies and by integrating the GROMACS software, with a graphical user interface and guides at each step, its use is facilitated for non-computational scientists.

Secondly, to perform full MD, hundreds of parameters, protocols, input and output file types, etc. can be chosen. In this case, only the necessary parameters and steps have been introduced so that a simple, but complete and correct MD can be performed, minimizing the number of errors that may occur along the way. We have added the parameters that we believed to be relevant for the user, without introducing an excess of information that could overwhelm him/her.

Thirdly, by introducing the plug-in into Scipion, the results can be integrated with other software for greater use and versatility. Now it is possible to connect the processing of In addition, once other MD plug-ins using different software, such as AMBER, CHARMM or NAMD, are introduced, they can be integrated with each other and even compared for more accurate results.

Fourthly, the integration of the plug-in into Scipion has provided a system to easily manage all the MD results and obtain a record of each of the parameters used at any given time.

Fifthly, it is possible, through the different Scipion options, to automate the processes for carrying out numerous MD, managing them from a single place.

Regarding the conclusions of the validation, it can be concluded that the MD have been carried out correctly and the protocol has produced enough data to analyse the results. The results of the dynamics simulations allow us to identify a different network of hydrogen bonds in both GSK structures highlighting two possible different binding sites.

6. Future Work

Currently, the GROMACS plug-in allows the performance of MD for proteins without ligands. The ultimate goal of the GROMACS implementation is its complete use for the realization of MD of all kinds: Proteins alone, proteins with ligand, protein interaction, lipid bilayers, protein complexes, etc. The advantage of Scipion is that it allows protocols to be added and therefore functions can be added when required.

A future goal of the work is the implementation of other MD software in order to compare results between them and even make them interoperable with each other, to take advantage of each other's strengths. In addition, it is also necessary to add more protocols from the same plug-in in order to be able to perform more complete analyses. An example of a protocol to implement is that of performing protein-ligand MD to determine their affinity. The more protocols and types of analysis are added, the more number of parameters will have to be added.

In addition to software, Scipion allows the integration of other functionalities, such as TRAPP. Being it Web Servers, Scipion can be programmed to interact with it to launch the files and request results.

7. Bibliography

1. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol. 2002 99* **9**, 646–652 (2002).

2. Vlachakis, D., Bencurova, E., Papangelopoulos, N. & Kossida, S. Current State-of-the-Art Molecular Dynamics Methods and Applications. *Adv. Protein Chem. Struct. Biol.* **94**, 269–313 (2014).

3. Hansson, T., Oostenbrink, C. & Van Gunsteren, W. F. Molecular dynamics simulations. *Current Opinion in Structural Biology* vol. 12 190–196 (2002).

4. Geng, H., Chen, F., Ye, J. & Jiang, F. Applications of Molecular Dynamics Simulation in Structure Prediction of Peptides and Proteins. *Computational and Structural Biotechnology Journal* vol. 17 1162–1170 (2019).

5. González, M. A. Force fields and molecular dynamics simulations. *Collect. SFN* **12**, 169–200 (2011).

6. Hospital, A., Goñi, J. R., Orozco, M. & Gelpí, J. L. Molecular dynamics simulations: Advances and applications. *Advances and Applications in Bioinformatics and Chemistry* vol. 8 37–47 (2015).

7. Henzler-Wildman, K. A. *et al.* Intrinsic motions along an enzymatic reaction trajectory. *Nat. 2007 4507171* **450**, 838–844 (2007).

8. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

9. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

10. de la Rosa-Trevín, J. M. *et al.* Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. *J. Struct. Biol.* **195**, 93–99 (2016).

11. Wagner, F. F. *et al.* Exploiting an Asp-Glu 'switch' in glycogen synthase kinase 3 to design paralog-selective inhibitors for use in acute myeloid leukemia. *Sci. Transl. Med* **10**, 8460 (2018).

12. scipion-em. https://github.com/scipion-em.

13. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

14. Grace Home. https://plasma-gate.weizmann.ac.il/Grace/.

15. Phillips, J. C. *et al.* Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **153**, 044130 (2020).

16. K, M. *et al.* Glycogen synthase kinase 3alpha-specific regulation of murine hepatic glycogen metabolism. *Cell Metab.* **6**, 329–337 (2007).

17. T, K., B, S. & F, L. M. Small-Molecule Inhibitors of GSK-3: Structural Insights and Their Application to Alzheimer's Disease Models. *Int. J. Alzheimers. Dis.* **2012**, (2012).

18. PATEL, S. & WOODGETT, J. Glycogen Synthase Kinase-3 and Cancer: Good cop, bad cop? *Cancer Cell* **14**, 351 (2008).

19. AV, O. & DD, B. Targeting GSK-3: a promising approach for cancer therapy? *Future Oncol.* **2**, 91–100 (2006).

20. Tejeda-Muñoz, N. & Robles-Flores, M. Glycogen synthase kinase 3 in Wnt signaling pathway and cancer. *IUBMB Life* **67**, 914–922 (2015).

21. Doble, B. W., Patel, S., Wood, G. A., Kockeritz, L. K. & Woodgett, J. R. Functional Redundancy of GSK-3α and GSK-3β in Wnt/β-Catenin Signaling Shown by Using an Allelic Series of Embryonic Stem Cell Lines. *Dev. Cell* **12**, 957–971 (2007).

22. B, G. *et al.* GSK3 Deficiencies in Hematopoietic Stem Cells Initiate Pre-neoplastic State that Is Predictive of Clinical Outcomes of Human Acute Leukemia. *Cancer Cell* **29**, 61–74 (2016).

23. ModelArchive. https://www.modelarchive.org/.

24. A, S. & TL, B. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).

25. Dajani, R. *et al.* Crystal structure of glycogen synthase kinase 3β: Structural basis for phosphate-primed substrate specificity and autoinhibition. *Cell* **105**, 721–732 (2001).

26. RCSB PDB - 6TCU: Glycogen synthase kinase-3 beta (GSK3b) in complex with ligand 1. https://www.rcsb.org/structure/6TCU.

27. Stank, A. *et al.* TRAPP webserver: predicting protein binding site flexibility and detecting transient binding pockets. *Nucleic Acids Res.* **45**, W325–W330 (2017).

28. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.