UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN INFORMACIJSKE TEHNOLOGIJE

ZAKLJUČNA NALOGA
(FINAL PROJECT PAPER)

# HOMOLOGNO MODELIRANJE IN MOLEKULSKE SIMULACIJE KAMELJIH NANOTELES KOT POTENCIALNIH TERAPEVTSKIH ORODIJ PRI ZDRAVLJENJU SARKOMA
# (HOMOLOGY MODELING AND MOLECULAR SIMULATIONS OF THE CAMELID NANOBODIES AS THERAPEUTIC TOOLS IN THE TREATMENT OF SARCOMA)

KATJA PRAČEK

2018
ZAKLJUČNA NALOGA (FINAL PROJECT PAPER)
PRAČEK

UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga
(Final project paper)

# Homologno modeliranje in molekulske simulacije kameljih nanoteles kot potencialnih terapevtskih orodij pri zdravljenju sarkoma

(Homology modeling and molecular simulations of the camelid nanobodies as therapeutic tools in the treatment of sarcoma)

Ime in priimek: Katja Praček
Študijski program: Bioinformatika
Mentor: doc. dr. Andrej Perdih
Somentor: dr. Sara Fortuna

Koper, januar 2018

**Ključna dokumentacijska informacija**

Ime in PRIIMEK: Katja PRAČEK

Naslov zaključne naloge: Homologno modeliranje in molekulske simulacije kameljih nanoteles kot potencialnih terapevtskih orodij pri zdravljenju sarkoma

Kraj: Koper

Leto: 2018

Število listov: 29 Število slik: 25 Število tabel: 4

Število prilog: 8 Št. strani prilog: 7

Število referenc: 31

Mentor: doc. dr. Andrej Perdih

Somentor: dr. Sara Fortuna

Ključne besede: sarkom, nanotelo, VHH, homologno modeliranje, simulacija molekulske dinamike

Izvleček:

Sarkom je vrsta raka, ki nastane v vezivnih tkivih in njegovo napredovanje je odvisno od prisotnosti kompleksa med proteinoma Twist1 in p53. V predhodnih študijah so identificirali štiri nanotelesa, ki bi lahko ovirala to interakcijo. Le-ta so enodomenska protitelesa kameljega izvora s poznano aminokislinsko sekvenco, njihove 3D strukture pa še niso znane. V zaključni nalogi smo zgradili tridimenzionalne strukture identificiranih nanoteles, ki predstavljajo izhodišče za razvoj potencialnih protirakavih učinkovin, ki bi delovale kot inhibitorji protein-protein interakcije med Twist1 in p53. Pri delu smo uporabili različne *in silico* metode kot so homologno modeliranje, molekulska mehanika in simulacije molekulske dinamike. V PDB bazi podatkov smo najprej poiskali 3D strukture s podobnim aminokislinskim zaporedjem in izbrali najoptimalnejše za homologno modeliranje. Te smo poravnali z izbranimi strukturami, naredili ustrezne modifikacije ter zgradili začetne 3D modele nanoteles. 3D modele smo nato optimizirali s programskim paketom GROMACS. Za vsako strukturo smo izvedli energijsko minimizacijo ter simulacije molekulske dinamike. Analiza molekulske dinamike s pomočjo RMSD in RMSF parametrov je pokazala, da so vse strukture relativno stabilne ter identificirala dele, ki so izkazovali večjo fleksibilnost; to so bile bile predvsem nekatere zanke, območja z ostrimi zavoji in deli, ki so bili mutirani. Rezultati zaključne naloge bodo uporabni za nadaljnje študije vezave; kako ta nanotelesa interagirajo s proteinom p53.

**Key words documentation**

Name and SURNAME: Katja PRAČEK

Title of the final project paper: Homology modeling and molecular simulations of the camelid nanobodies as therapeutic tools in the treatment of sarcoma

Place: Koper

Year: 2018

Number of pages: 29 Number of figures: 25 Number of tables: 4

Number of appendix: 8 Number of appendix pages: 7

Number of references: 31

Mentor: Assist. Prof. Andrej Perdih, PhD

Co-Mentor: Sara Fortuna, PhD

Keywords: sarcoma, nanobody, VHH, homology modeling, molecular dynamics simulation

Abstract:

Sarcoma is a type of cancer found in connective tissue. Its progression has been shown to depend on the Twist1:p53 protein-protein interaction. Four antibodies protein were previously identified that could interfere with this interaction. These proteins were single domain antibodies (camelid origin) with an identifiable amino acid sequence. Their 3D structures have not been experimentally determined. The *in silico* methods (e.g. homology modeling, molecular mechanics, and molecular dynamics simulations) were used to construct and validate 3D structures of the identified nanobodies. Using the Protein Data Bank (PDB), we then selected and chose the nanobodies with similar amino acid sequence and 3D structures. We than aligned and fitted these sequences with selected structures and performed appropriate modifications to construct 3D models of these nobodies. The generated models were then optimised using GROMACS package. We first minimised and equilibrated each system and then ran a production molecular dynamics (MD) simulations to analyse the stability of the obtained homology models. RMSD and RMSF analytical parameters confirmed that the modeled structures were relatively stable. It also identified which parts of the structures were more flexible. As expected, the results included loops, areas with sharp curves, and parts that were mutated. The following results will be further useful for subsequent docking studies to determine the interaction between these nanobodies and p53 protein as a next step in Twist1:p53 protein protein inhibitor design.

# ACKNOWLEDGEMENT

I would like to thank my supervisor Assist. Prof. Andrej Perdih, PhD (National Institute of Chemistry, Ljubljana) for professional support, encouragement and for many useful advices.
I would like to express my sincere gratitude to my co-supervisor Sara Fortuna, PhD (University of Trieste and SISSA, Italy) for generous help and lectures connected with the project, her spiritual support, availability and quick response.
I am grateful to Assoc. Prof. Ario de Marco, PhD (University of Nova Gorica) for experimental results, organization of the project and to be there for any question about the biological part. I would also like to thank Elisa Mazzega, PhD (University of Nova Gorica) for providing experimental results.
Finally, I would like to thank my family and boyfriend Dejan for the support.

# CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF SUPPLEMENTARY FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| 3D | Three Dimensional |
| AMBER | Assisted Model Building with Energy Simulation |
| BMRB | Biological Magnetic Resonance Bank |
| CDR | Complementarity-Determining Region |
| IgG | Immunoglobulin G |
| MD | Molecular Dynamics |
| NMR | Nuclear Magnetic Resonance |
| NPT | Constant number of particles, pressure and temperature |
| NVT | Constant number of particles, volume and temperature |
| PDB ID | Protein Data Bank Identification code |
| PDB | Protein Data Bank |
| PPI | Protein-protein interaction |
| RCSB | Research Collaboratory for Structural Bioinformatics |
| RMSD | Root Mean Square Deviation |
| RMSF | Root Mean Square Fluctuation |
| VHH | Single domain antibody |
| wwPDB | Worldwide Protein Data Base |

# 1 INTRODUCTION

## 1.1 Biological background

Sarcomas a type of cancer with multiple targets in connective tissues such as nerves, lymph vessels, blood vessels, fat, muscles and other (Figure 1). The cancer is primarily detected when a bump or swelling occurs in the connective tissue. Once the cancer is large enough and presses on nerves throughout the body, then the symptoms are detected. The symptoms however do not always occur, as sarcoma can also be found in symptom-lacking patients. Sarcoma is considered a rare disease, yet their collective incidence is relevant especially among young patients. The common prognosis of these cancers encourages many scientists to study this aggressive disease. When treating the disease, the properties of each sarcoma sub-type must be understood and considered in order to have the beneficial results.

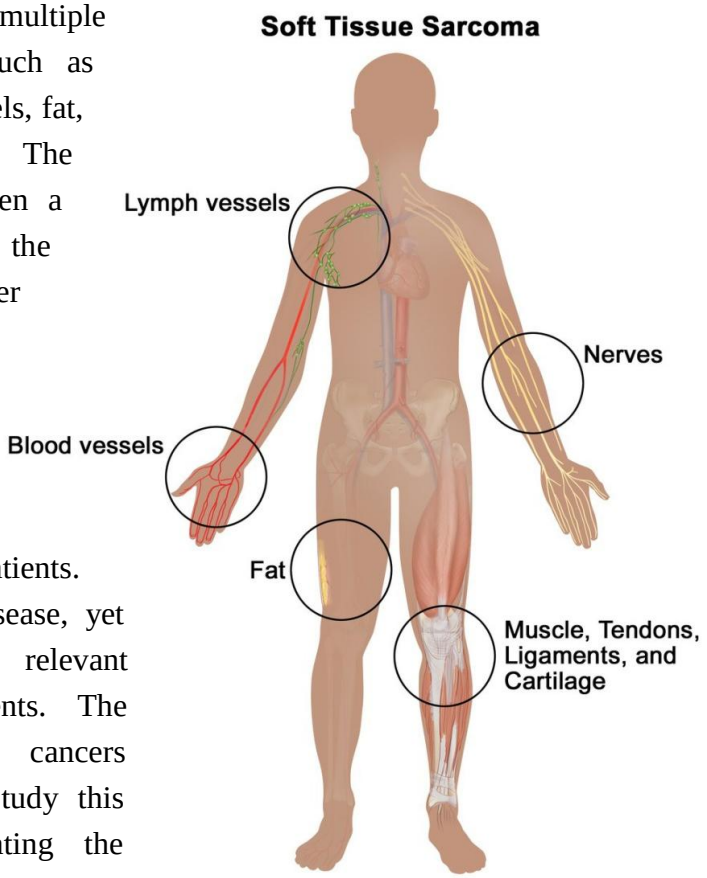



**Figure 1 Body parts where the sarcoma cancer can evolve.**[1]

Rhabdomyosarcoma is an example of a sub-type of sarcoma and is most frequently found in children. It progresses in skeletal muscles and is usually noticed when lumps located near the body surface appear. This enables its clinical identification [18, 21].

Another sub-type is a gastrointestinal stromal cancer. It is found in the stomach and is diagnosed mostly in adults. It can be benign or malignant and is usually detected in late stages as this cancer type lacks early symptoms.

Many studies show that p53 protein malfunction occurs at the developing stages of sarcoma cancer. This protein is an important transcription factor which helps to regulate the cell cycle arrest, apoptosis, senescence, fertility and metabolism. p53, as well as other tumour-suppressor proteins, can be prevented to do its job by gene mutations or mechanism of sequestration by direct interaction with other molecules [17].

[1] **Source: https://www.cancer.gov/types/soft-tissue-sarcoma**

When a protein binds to another protein partner we define this as a protein-protein interaction (PPI) [1]. These protein-protein interactions are extremely important as they regulate the biological processes (such as intercellular communication, metabolism, transport and programmed cell death, or apoptosis). The interface of a protein that participates in the binding includes residues that have specific affinity properties to bind the partner and these regions are referred in the literature as hot spots. In the protein-protein complex the hot spots of the involved proteins look highly complementary. Forces that are active between the complementary hot spots are predominantly non-covalent interactions such as hydrogen bonds, ionic interactions, Van der Waals interactions or hydrophobic interactions. If a PPI is involved in progression of an illness condition, it is often hypothesized that the illness could be treated by disrupting this particular interaction. One option to prevent the interaction is to find another molecule (which can also be a protein or a small molecule) which binds to one of the proteins hot spots preventing the interaction with the other protein. These compounds are called PPI inhibitors[2].
A Twist1 protein was found bound to a p53 protein [21], consequently disabling its function and preventing its posttranslational modification leading to a degradation of p53. Thus, it was proposed that disrupting the Twist1:p53 complex could be beneficial in the treatment of sarcoma cancer. The study [21] argued that the Twist1 protein binds to the p53 C-terminus via the Twist box domain of Twist1 (these are the last 23 residues of Twist1 protein [7]).

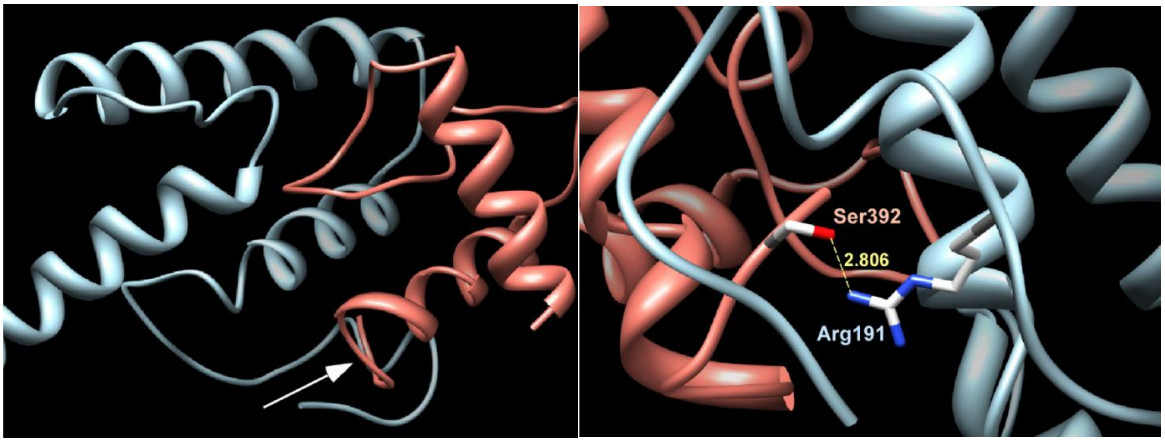



**Figure 2 Suggested model of PPI between p53 (red) and Twist1 (blue) proteins. Binding site is pointed with white arrow. Right: enlargement of this PPI with key residues name and distance between them in angstroms. Based on reference [21].**

*In silico* molecular docking further confirmed this hypothesis. Figure 2 represents the Twist1:p53 PPI. The figure on the left depicts the interaction while the side of interaction is shown with the arrow. The right figure shows the enlarged PPI hot spot with the residues that form hydrogen bonds between proteins (Serine392 of p53 protein and Arginine191 of Twist1 protein). The distance between these two residues is 2.806 Å [21].

A previous study conducted from our colleagues at the University of Nova Gorica attempted to inhibit this Twist1:p53 PPI by using nanobodies that bind to the p53 protein [16]. A nanobody (called also VHH) corresponds to the heavy-chain variable domain of the *Camelidae* Immunoglobulin G (IgG) the only main difference is that it lacks the light chains. Figure 3 shows a comparison between the human and camelid antibody. The heavy chain variable domain of the *Camelidae* IgG is the exclusive paratope that binds to the antigen. Consequently, VHH can be isolated and its binding ability to interact with various antigens remains the same as it is for the corresponding IgG [26]. VHH is more manageable for the *in silico* modeling and molecular dynamic simulation than a whole IgG [26]. It is also easier and inexpensive to produce in microbial cells.

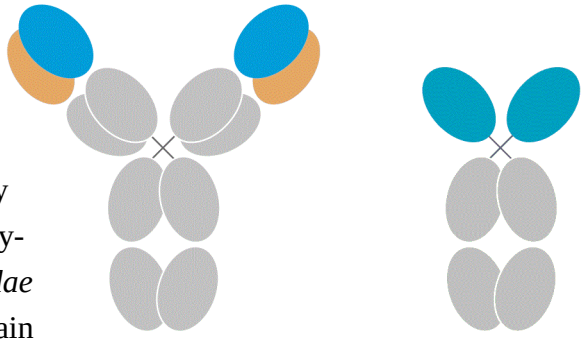


**Figure 3 Comparison between the human (left) and camelid antibodies (right). On the right side is represented the heavy chain antibody and on the left side a conventional IgG antibody. Heavy chain domains are coloured blue and light chain domains in orange.[2]**

The structure of a VHH has three recombinant loops (complementarity-determining regions (CDRs) and conserved regions called a framework. The CDR regions are the one that bind on antigen. Because there are many different antigens to bind, the CDR regions vary from antibody to antibody to adjust antigen's paratope. The third CDR loop on VHH is normally longer than the corresponding loop in conventional antibodies and compensates with its large surface for the three missing CDRs present on the IgG's light chain. Figure 4 shows an example of a VHH.

**Figure 4 Structure of nanobody of Lama Glama antibody. Framework blue and CDRs yellow. (PDB ID: 2X1O)**

To isolate the nanobodies which could bind to p53 protein, our collaborators at University of Nova Gorica have panned a pre-immune VHH phage display library [16]. The binders were subcloned to share the same framework (that has been proven being simpler) and partially humanized to make it less immunogenic [16].

In our case, the library used for panning has about $5\times10^7$ different clones and enabled to isolate VHHs that bind on p53 protein C-terminus or at least in its vicinity and prevent the interaction with Twist1. VHHs with four unique sequences were finally recovered and there was the need to determine their exact (3D structure) epitope on p53.

[2] Source: https://www.dovepress.com/cr_data/article_fulltext/s107000/107194/img/IJN-107194-F01.jpg

# 2 AIM

Sarcoma is a particular cancer which grows in the connective tissue and the protein-protein interaction between p53 protein and Twist1 protein plays a role in cancer progression. For the treatment of Sarcoma s molecule which disrupts the protein-protein interaction could be useful as a therapeutic tool in the treatment of sarcoma. Nanobodies (or VHHs) are very useful molecules which prevent protein-protein interactions. VHH are single-domain antibodies of Camelid origin that, like a whole antibody, bind selectively to a specific antigen. Our colleagues at the University of Nova Gorica have experimentally identified four primary sequences of possible protein binders to the same p53 C-terminus region where Twist1:p53 protein-protein interaction surface is located.
The aim of this work is to predict the three dimensional (3D) structures of the four identified sequences by homology modeling. We will also then study the stability of the obtained 3D models using molecular dynamics (MD) simulations.

The specific goals in more detail are as follows:

- to define the 3D structure of all sequences that are potential therapeutic candidates for sarcomas using homology modeling;
- to test the stability of the produced 3D models with MD simulations;
- to learn how to use different molecular modeling programs and corresponding computer software (e.g. Protein Data Bank, GROMACS, Linux terminal work, etc).

This approach will allow the first steps to a more detailed atomistic knowledge of the important PPI, while providing crucial information for the next development stages of efficient inhibition of Twist1:p53 PPI.

# 3 METHODS

## 3.1 Homology modeling

The nanobody structure prediction is the first step towards the determination of the putative binding surfaces on the p53 protein to enable the modeling of the inhibition mechanism by which nanobodies prevented Twist1:p53 interaction.

Understanding the 3D structure of a protein can provide information about its biological function [13] and enables further studies on the binding ability to a particular molecule. This can further be explored by molecular docking.

To determine the 3D structure of a "target" protein one can use experimental methods such as x-ray crystallography or nuclear magnetic resonance (NMR). But these two methods have some disadvantages, such as long time duration, high price and sometimes protein inability to crystallize for the former or protein being too big for the latter. Another way is to determine the structure *in silico* by homology modeling [8]. Homology modeling predicts the 3D structure of a protein by comparing its amino acid sequence with amino acid sequence of a protein whose structure is known: the template molecule [30]. Thankfully, many scientists who are researching protein structures have deposited their results in the Protein Data Bank (PDB) (represented in section 2.1). This structural information is required for the homology modeling [8].

The method is established on two premises. First is that an amino acid sequence determines the structure of a protein and in theory we could determine the structure just by knowing the sequence [6] and second is that through evolution the structure of proteins have a tendency to remain constant to maintain their function even when the sequence mutates. This implies that proteins can have similar structures despite the difference in their protein sequences [4], [24].

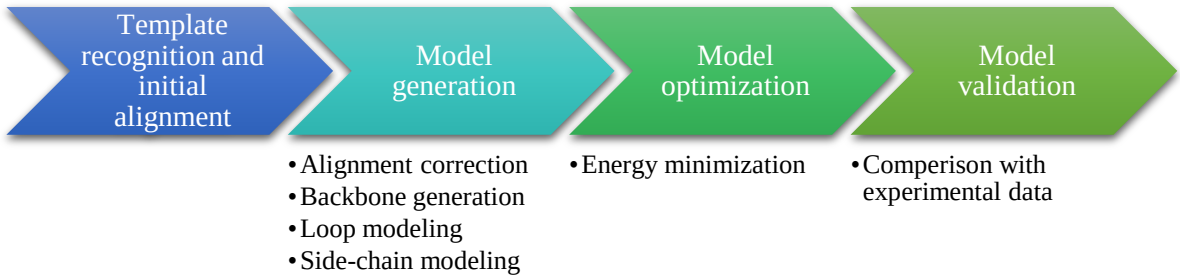



**Figure 5 Steps of homology modeling.**

To begin with homology modeling, the first step is to find a 3D template structure with a similar amino acid sequence. The PDB is an online database where the 3D structures of biomacromolecules are collected, thus it is the best resource to find a good template. Sequence of the chosen template structure is aligned with given sequence. The following step is model generation. This includes (i) alignment correction, where the alignment is

made also based on template structure or with the help of a third sequence, (ii) backbone generation, where backbones of different templates can be chosen, (iii) loop modeling where insertions or deletions that occur on the secondary structured parts are shifted to the disorganised regions as turns and loops, and (iv) side chain modeling where side chains of amino acid can be rotated if they clash with other residues or the backbone. By following this procedure an initial 3D structure is generated without considering the underlying laws of physics for this particular target. The conformation could have very large potential energy and could be very unstable. This can be corrected by minimizing the potential energy of the initial model. The minimization is done by iteratively adjusting the conformation and by recalculating the potential energy until its value converges. Through these steps of energy minimization a more realistic and stable conformation is constructed. The minimization process relies on a careful description of the intra/intermolecular interactions. These are defined by molecular mechanics (section 2.2).

### 3.1.1 The Protein Data Bank

The PDB is an on-line archive that collects 3D structure of biomacromolecules (such as proteins, DNA, RNA) on an atomic level determined by experimental methods [23]. The database is directed by Worldwide Protein Data Base (wwPDB[3]) that has three regional centres; in United States of America (Research Collaboratory for Structural Bioinformatics (RCSB)[4] PDB) [3], in Europe (PDB in Europe or PDBe[5]) [29] and in Japan (PDB Japan or PDBj[6]) [12]. Two additional centres that contain archived macromolecules determined with NMR are located in United States of America BMRB[7] (Biological Magnetic Resonance Bank) [27] and in Japan (PDBj-BMRB[8]). wwPDB collects, comments, validates, and standardizes data form researchers all over the world. Archived data are freely available to all users [23].

RCSB PDB website [23] allow its users to search, browse, and download protein structural information. One can search the database by PDB ID (Protein Data Bank Identification code), authors of the corresponding publication, macromolecule name, sequence, or ligands. Advanced searching is also possible as users can choose between many different query type (e.g. Sequence (BLAST/FASTA/PSI-BLAST)) and set search parameters. For every molecule the FASTA and pdb file can be downloaded. FASTA file, one can find an amino acid sequence of particular molecule but in the pdb file coordinates of single atom of molecule are also indicated. From its creation in 1971, the protein bank currently contains more than 130,000 experimental structures [22].

[3] http://wwpdb.org
[4] http://rcsb.org
[5] http://pdbe.org
[6] http://pdbj.org
[7] http://www.bmrb.wisc.edu
[8] http://bmrbdep.pdbj.org/

### 3.1.2 BLAST

BLAST [19] (Basic Local Alignment Search Tool) is an algorithm built to search similar sequences (amino or nucleotide acids) in a database. For example, searching sequences in PDB to find protein structure with similar sequences.

It can be also used for sequence alignment, which includes the smallest number of mutations which must be done to make the sequences the same (represented in Figure 6).

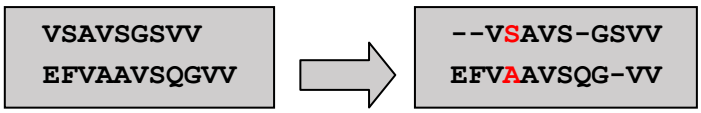



**Figure 6 Alignment of two amino acid sequences.**
**Gaps are indicated by "-" and substitutions are highlighted in red.**

### 3.1.3 Swiss PDB Viewer

Swiss PDB Viewer [9] is a visualization and modeling tool freely accessible on-line at http://www.expasy.org/spdbv/. The application allows the alignment and to superpose structures. It further allows protein structural modifications, such as mutations, insertions and deletions, amino acids displacements and side chain rotations. As inputs it accepts molecular coordinate files, such as a pdb files, as well as amino acid sequences in fasta format. Output of program can be also a pdb file.

## 3.2 Molecular mechanics

Molecular mechanics [14] relies on the classical description of atoms, molecules and their interactions. Atoms are treated as soft spheres connected with harmonic bonds and interact with intramolecular and intermolecular interactions. The total potential energy is calculated as a sum of intramolecular and intermolecular potential energy ($E_{tot} = E_{intra} + E_{inter}$). Intermolecular interactions depend on the individual distance between bonded atoms, the angle between three connected atoms and the (torsion) angle of plains of four consecutively connected atoms. The potential energies of these parameters can be calculated using spring equations. For instance, the bond energy depends on how much a bond is longer or shorter than its equilibrium length and on harmonic bond stiffness. The basic equation describing the bond energy is

$$E_{bond} = \sum_{bond} \frac{k}{2}(r - r_0)^2,$$

where $k$ is bond stiffness, $r$ is bond length and $r_0$ is the equilibrium bond length. This and other intermolecular interactions are illustrated in Figure 7, while the sum of their potential energies gives the energy of intramolecular interaction: $E_{intra} = E_{bond} + E_{angle} + E_{torsion}$.

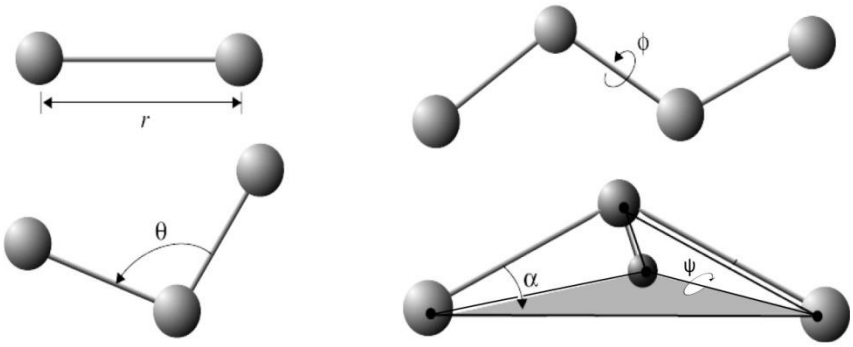



**Figure 7 Representation of the bonded interactions. Left up is distance between atoms, left down is angle between atoms and right picture represent torsion angle.[9]**

The total potential energy depends also on the intermolecular or non-bonded terms: Van der Waals and electrostatic potentials. The latter can be described by the Coulomb, dipole-dipole, or H-bond and the former with Lennard-Jones potential. Potential energy function for non-bonded interaction is written as: $E_{inter} = E_{VdW} + E_{electrostatic}$ .

Each term of the potential energy is calculated from the atomic positions. Furthermore, the equations require parameters (stiffness of springs, equilibrated bond lengths, angles...) that can be obtained either from experiments or from quantum mechanical calculations.

For different molecular systems different functions and parameters are used to obtain better predictions. Functional forms and parameter sets for a certain molecular system are collectively called a "force field". But because of transferability, similar molecules can be modeled using the same force field. For example force field AMBER (Assisted Model Building with Energy Simulation) is designed for processing proteins and nucleic acids. There are also force fields that are designed for calculating whichever molecule, however, they are not as reliable as if a specific force field is utilized. This is why more force fields are available. The principal force field groups are AMBER, CHARMM (Chemistry at HARvard Macromolecular Mechanics), GROMOS (GROningen Molecular Simulation package), OPLS (Optimized Potential for Liquid Simulations), and MMFF (Merck Molecular Force Field).

## 3.3 Molecular dynamics

Molecules in nature are not static entities but rather they are constantly vibrating, rotating and changing their conformation. In proteins this is especially true for their corresponding side chains. Using molecular dynamic methods (one of the principal methods of molecular simulations), we attempted to represent and describe the scale of these motions. Using Newton's second law ($F = ma$), the corresponding equations of motion are used to calculate velocity and acceleration (depends on the integration algorithm) of every atom to then determine their new positions. This process is then repeated numerous times. The size

[9] Source: http://www.ks.uiuc.edu/Training/TutorialsOverview/namd/namd-tutorial-unix-html/bondstretch.jpg

of the time step has to be smaller than even fastest molecular motion, namely the OH vibration in $H_2O$ molecule, and it is usually between 0.5 and 1 femtoseconds (1 fs = $10^{-15}$ s). The result of this method is a trajectory that contains structures taken at successive time intervals during the simulation.

### 3.3.1 Adding water and ions

Every molecule in nature is surrounded by an environment influencing its properties. Therefore, if one wants to simulate the behaviour of a molecule, a realistic molecular environment is needed. In general, biological molecules such as proteins are dissolved in water, and thus in every simulation many water molecules are added to the system. If a protein has an overall positive or negative charge instead of being neutral, ions are added to neutralise the total charge of the system. After adding water and ions, the system needs to be minimised.

### 3.3.2 Periodic boundary conditions

To simulate the environment along with a target molecule the system becomes very large (with large number of atoms) thus the computer must calculate enormous numbers of interactions. This slows down the process of simulation, and can even make it impossible to simulate. A way to simulate bulk properties and have a lesser number of particles is to employ periodic boundary conditions [14]. This means that target molecule and environmental molecules are put in a box that could be any shape. The simplest is cubic. By simulating only one box with periodic boundary conditions, it is like if the box was copied by all sides of original box (so 26 copies). Therefore, if a molecule goes outside the box it appears at the other side of the box. This is represented in Figure 8 but with the difference that in the figure periodic boundary condition for two dimensions is shown for simplicity and not for three as it is implemented in program for molecular simulation.

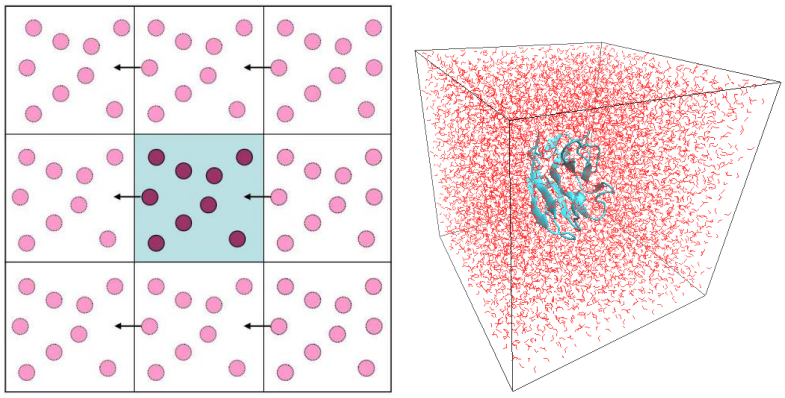


**Figure 8 Periodic boundary condition and box of solvent. Right picture represents how molecule traverse in periodic boundary conditions in two dimensions and left picture represents protein (blue) (target molecule) in box of solution in this case water (red).[10]**

[10] Source:
https://www.researchgate.net/profile/Paraskevi_Gkeka/publication/292609006/figure/fig6/AS:354670401867778@1461571421154/Figure-23-Periodic-boundary-conditions-A-two-dimensional-representation-of-the.jpg

For the simulated system it is recommended to construct the smallest cube as possible, yet big enough so the molecule cannot interact with itself. The accuracy is better if the cube is larger.

### 3.3.3 Equilibration stage of the MD simulation

Equilibration [14, 15] is a method to prepare molecular system for MD. It optimises solvent and ions with the solute, because it is possible that the solvent molecules are minimized just among them and not necessarily with the protein. Equilibration can be achieved by bringing the system to the chosen conditions for a certain amount of time. This is generally done with the MD simulation having the following conditions: constant number of particles, volume and temperature (also known as NVT ensemble). The actual temperature is not constant in the beginning as it has to increase from zero to the target temperature. After a certain amount of MD simulation time it should stabilize. The second equilibration can be conducted under the NPT ensemble (constant number of particles, pressure and temperature). Under these conditions pressure and density should stabilise, therefore the system is put in a more realistic environment and its behaviour along the MD will be more realistic.

### 3.3.4 Molecular simulations with GROMACS

GROMACS [28] is a very fast open source code for the molecular simulations which is used to prepare molecular system, run energy minimization, equilibration, MD simulation and do analysis of the results. The two quantitative parameters used to analyse the MD trajectory the Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF) will be described in section 2.3.6.

Software package GROMACS is focused to simulate biological (macro) molecules in their aqueous or membrane environment. The simulations can be run on a single processor or on parallel computer systems to enable simulations of a larger systems [28]. This molecular simulation toolkit was developed by the team of Prof. Herman Berendsen from the Department of Biophysical Chemistry at University of Groningen in the early 1990s. The abbreviation GROMACS means Groningen MAchine for Chemical Simulation.

### 3.3.5 Visualization with VMD

Visual molecular dynamics (VMD) [10] is a program that visualizes molecules whose atomic coordinates are written in a text file (such as pdb and gro files).

Molecules can be represented with different drawing methods. One of them is “NewCartoon” which shows

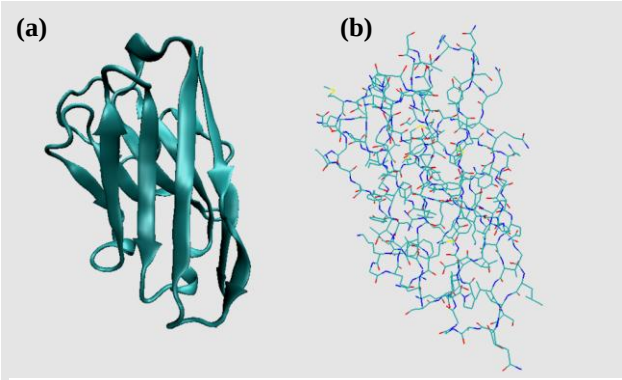



**Figure 9 VMD molecular visualization styles: (a) “NewCartoon” and (b) “Lines” drawing method.**

only the backbone and the organized secondary structure elements α–helix and β–sheets (Figure 9a). The latter are represented with a wider line so that one can easily visualize the structural conformation of the protein under investigation. To see all the atoms and their connectivity the "Lines" draving method can be used (Figure 9b).In VMD one can load also the GROMACS trr trajectory files that contain coordinates of different conformations of a molecule along time and visualise the animation of a MD trajectory.

### 3.3.6 Analysis of molecular simulation

Through simulation, all observed molecules were changing their positions. To find if the simulations have stable conformations we must first see if the structures have been changing over time. This can be done quantitatively by calculating the RMSD between initial structure conformation and conformations produced during the simulation.
At every time step, the RMSD is calculated as follows:

$$RMSD(t) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} {d_{i,t}}^2}$$

Where $n$ is total number of atoms, $d_i$ is distance between $i$ atom from reference and current structure (conformation) in the $t$-time step and the summation is taken over every atom.
The RMSD can be studied as a function of the simulation time to see whether the molecule is stable or not. Example of the graphs can be found in the Results section (Figure 20, Figure 21) together with comments about stability in section 3.3.
In addition, some parts of the protein structure are more flexible during molecular dynamic simulation than others. With RMSF parameter [5] one can see and quantity the movement of an individual atom in a given amount of MD simulation time. It is calculated so that the root mean square of distances between the atom in initial structure and same atom in every time step is calculated and this is calculated for every atom in the molecule:

$$RMSF(i) = \sqrt{\frac{1}{T}\sum_{t=0}^{T} d_{t,i}}$$

Where $i$ is atom number, $d_t$ is distance between the $i$ atoms form reference structure and structure at time $t$, $t$ is current time step, $T$ is number of total time steps and summation is taken over all time steps. Further examples and comments can be found in Result section (Figure 22-25 in section 3.3).
The obtained data was plotted with Gnuplot [11], a program that visualizes data and mathematical functions in many types of 2D and 3D plots. It has a command line interface and it is freely available. In this work we used Gnuplot version 5.0.1.

# 4 RESULTS

Our collaborators at the University of Nova Gorica, provided us with the four sequences of nanobodies that bind protein p53: namely A5, B3, C9 and D11 sequences (Figure 10).Sequence D11 is the longest as it contains 130 amino acids while the other three sequences A5, B3 and C9 have 121 amino acids.

The framework sequence is the same in all four chosen nanobodies (contain 96 amino acids). All sequences differ just in the CDRs regions and the length is different only for D11 third CDR region (shown in Figure 10). Our task was to determine the 3D structure of these four proteins.


**Figure 10 Alignment of the nanobodies sequences identified by our collaborators. CDRs are highlighted by red boxes, framework regions are not highlighted.**

## 4.1 Template recognition and initial alignment

Homology modeling was done to predict the 3D structure of the four proteins for which only the primary sequence is known was done by homology modeling. The first step was finding a good template for each protein.

In the Protein Data Bank we searched for proteins – nanobodies with the most similar sequences. Because the CDR regions are variable we looked in the PDB database for nanobodies with very similar sequence of framework. On the RCSB website www.rcsb.org/ we advanced searched, looking for the framework shared among all molecules. We searched using BLAST by setting E-Value Cutoff to 10.0 and Sequence Identity Cutoff to 80%. The results of this query were obtained in 45 sequences. Because framework sequence is the same in all four given sequences also the result is the same for all sequences. Their PDB ID and alignments with D11 are shown in Figure S1.

We aligned them in order to find a suitable template for the D11 sequence (in continuation also for other sequences). We first eliminated the sequences containing deletions bigger than one amino acid and insertions bigger than two amino acids. Figure S2 depicts the

sequences used to satisfy these requirements are provided. Second elimination was about substitutions in the non CDR regions. We aimed to keep the sequence with the smaller number of substitutions in the conserved regions of the framework. Table 1 shows the substitutions in the four framework regions and the sum of them. Sequences with the lowest number of substitutions are 4KRP, 4DKA, 4DK3 and 4DK6. Their structures available in the PDB were chosen to be used for the homology modeling.

| D11 | The number of substitutions | | | | SUM |
|---|---|---|---|---|---|
| | FR1 | FR2 | FR3 | FR4 | |
| 4KRP:B | 4 | 2 | 5 | 0 | **11** |
| 4DKA:A | 4 | 1 | 6 | 0 | **11** |
| 4DK3:A | 4 | 1 | 6 | 0 | **11** |
| 4DK6:A | 4 | 1 | 6 | 0 | **11** |
| 5HGG:S | 4 | 2 | 5 | 1 | 12 |
| 5HDO:A | 4 | 2 | 5 | 1 | 12 |
| 3RJQ:B | 4 | 1 | 8 | 0 | 13 |
| 3R0M:A | 4 | 1 | 8 | 0 | 13 |
| 4EIZ:C | 5 | 3 | 6 | 0 | 14 |
| 4EJ1:C | 5 | 3 | 6 | 0 | 14 |
| 1MVF:A | 5 | 2 | 7 | 1 | 15 |
| 3JBC:7 | 5 | 3 | 6 | 1 | 15 |

**Table 1 D11 alignment analysis. The columns show the number of substitutions in each of the four separate framework regions (FR1-4), sum of them, number of insertions and sum of substitutions and insertions for sequences of template candidates extracted from the PDB database and with compatible CDR length aligned to the experimental sequence D11. Chosen structures are marked with the red rectangle.**

When downloading the 4KRP [25] crystal structure, we discovered that the sequence of amino acid is not the same as in the fasta format. The header of pdb file revealed that the atomic position of the missing residues were not determined in the crystallographic structure. Figure S3 depicts the alignment of 4KRP sequence from FASTA file with sequence from crystallographic structure is provided. Missing residues at the end are not important because C- and N- terminus are generally free to move but the 8 missing amino acids between residue number 12 and 20 prevented us to use the 4KRP structure as a template for homology modeling.

The other three structures (4DKA, 4DK3 and 4DK6) have the same sequence (they were also published in the same paper [20]) except few additional amino acids in the end of 4DKA – which is not expected to influence the 3D structure.

We aligned the 45 sequences already selected from the PDB database with the A5 sequence (alignment provided in Figure S4). We then eliminated sequences with the deletions longer than one amino acid and insertions longer than four amino acids. As a result, 10 sequences remained (see Figure S5). Because A5, B3 and C9 have the same length and the same framework sequence they have also the same number of insertions or deletions if they are aligned to the same sequence. This is why we did not have to align all

45 sequences again with B5 and C9 sequences. We aligned B5 and C9 just with sequences that satisfied previous conditions (see Figure S6 and Figure S7).

| A5 | Number of substitutions | | | | SUM subst. | Number of insertions | SUM sub & ins |
|---|---|---|---|---|---|---|---|
| | FR1 | FR2 | FR3 | FR4 | | | |
| **4KSD:B** | 4 | 4 | 5 | 0 | 13 | 0 | **13** |
| **1OL0:A** | 4 | 3 | 3 | 1 | 11 | 4 | 15 |
| **2X1O:A** | 5 | 3 | 5 | 0 | 13 | 4 | 17 |
| **3TPK:A** | 4 | 4 | 6 | 0 | 14 | 4 | 18 |
| **4PPT:A** | 5 | 4 | 6 | 0 | 15 | 4 | 19 |
| **4POY:A** | 5 | 4 | 6 | 0 | 15 | 4 | 19 |
| **5DA4:A** | 5 | 3 | 6 | 0 | 14 | 5 | 19 |
| **5DA0:B** | 5 | 3 | 6 | 0 | 14 | 5 | 19 |
| **1BZQ:K** | 5 | 4 | 6 | 1 | 16 | 4 | 20 |
| **2P4A:B** | 5 | 4 | 6 | 1 | 16 | 4 | 20 |

**Table 2 A5 alignment analysis. The columns show the number of substitutions in each of the four framework regions (FR1-4), sum of them, number of insertions and sum of substitutions and insertions for sequences of template candidates extracted from the PDB database and with compatible CDR length aligned to the experimental sequence A5. Chosen structure is marked with the red rectangle.**

In Table 2, we compared 10 sequences from the PDB aligned with the A5 sequence. We also computed the number of substitutions for the separate parts of the framework, their sum, the number of insertions and the sum of substitutions and insertions. The last column shows, the nanobody sequence described by the 4KSD [31] PDB structure aligned with A5 sequence has the lowest number of substitutions and insertions and was therefore chosen to be most appropriate for the homology modeling. Because A5, B3 and C9 share the same framework and the same length, they are also sharing the number of substitutions in the frameworks and the number of insertions or deletions. This means that the most suitable structure for the homology modeling of B3 and C9 sequences is also 4KSD crystal structure.

Table 3 represents the percentages of identity given for the alignments of sequences from laboratory at University of Nova Gorica and sequences of the template. The percentages tell which sequences are similar enough to be used as templates for the homology modeling. In Figure 11 we outline which crystallographic structures which will be used as templates for each sequence when generating their homology models.

| Alignment | D11 x 4KDA | A5 x 4KSD | B3 x 4KSD | C9 x 4KSD |
|---|---|---|---|---|
| Identity | 67.2% | 71.4% | 71.4% | 71.4% |

**Table 3 Percentage of identity between the experimental sequences and their respective selected template.**


**Figure 11 Templates used for each experimental sequence.
In the first line are the labels of the sequences from the lab and below them
the PDB codes of the structures selected as templates for homology modeling.**

## 4.2 Homology model generation

Raw A5, B3 and C9 sequences were fitted with the 4KSD structure, while the raw sequence of D11 used 4DKA, 4DK3 and 4DK6 structures in program Swiss PDB Viewer 4.1.0. In Figure 12 is raw sequence (left) that will be fitted with selected template structure (right). Because in all obtained raw sequences one amino acid needed to be inserted, but the program was not able to do it automatically, it was inserted manually with the same program. Tyrosine was inserted as the 55$^{th}$ amino acid of A5, Alanine as the 61$^{st}$ amino acid of B3, Lysine as the 59$^{th}$ amino acid of C9 and Tyrosine as the 59$^{th}$ amino acid of D11. Because of clashes two Tryptophans in new generated structure of A5 sequence, were rotated with using same program. Figure 13 represents the 3D homology model of the A5 sequence made with the program Swiss PDB Viewer (shown with the highlighted insertion (red) and rotations (green and blue)). Figure 14 demonstrates the part of the structure where two Tryptophans were rotated. Left side of the Figure shows the structure before, while the right side is depicting the structure after rotation.

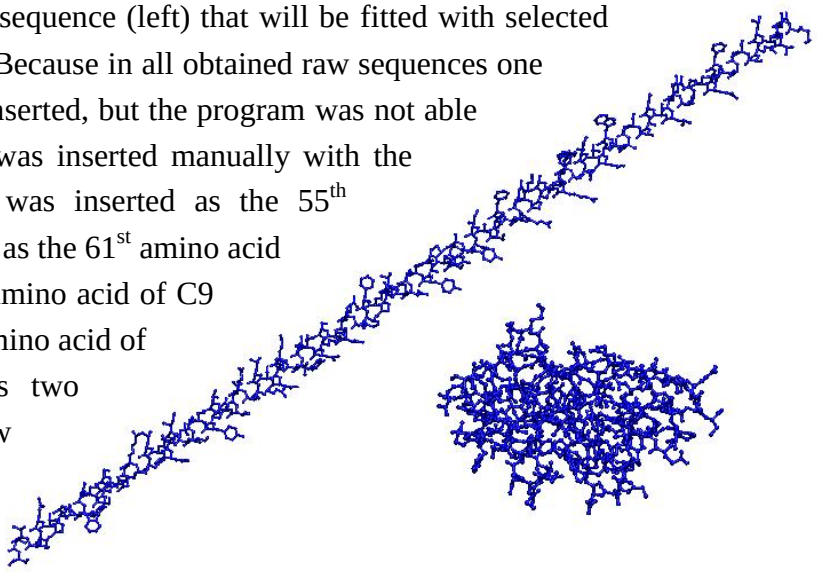


**Figure 12 Raw sequence from the lab D11 and crystal structure of 4DKA.**

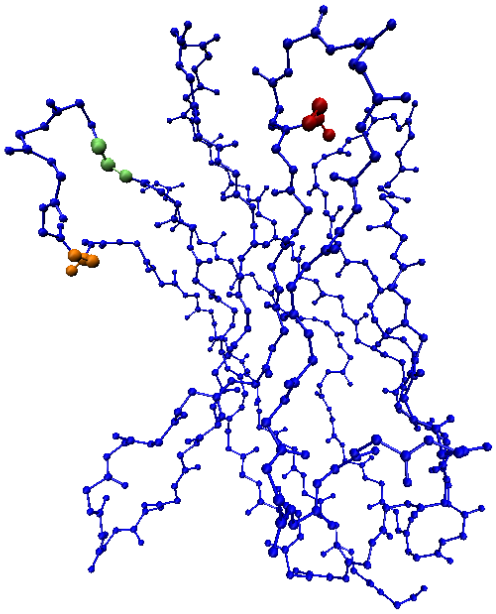


**Figure 13 Modification made with the Swiss PDB Viewer.
Backbone of the A5 sequence superposed to the 4KSD (blue) with highlighted insertion (red) and rotations (green and orange).**

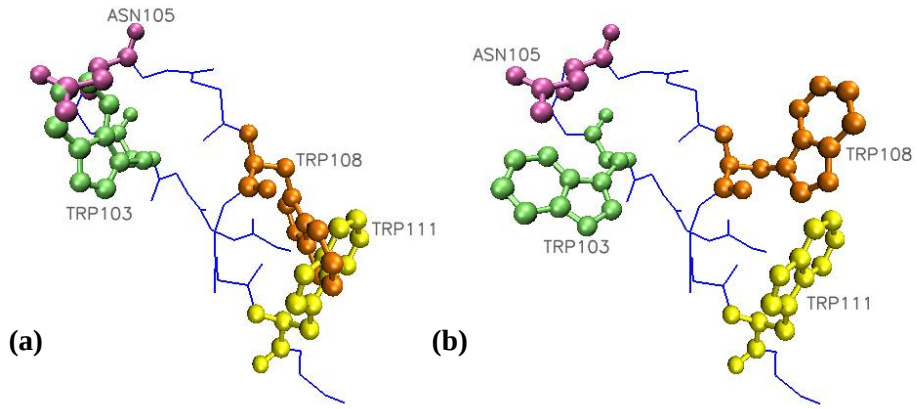



**Figure 14 Rotation of two Tryptophans.
In both pictures is the same part (resID 100-112) of A5 sequence superposed to 4KSD (a) before and right (b) after rotation of 103$^{rd}$ and 108$^{th}$ residue.**

## 4.3 Homology model optimization

Optimization occurs after the 3D models have been generated via homology modeling. This was done by energy minimization, MD simulations, and subsequent analysis of the newly generated structures. All the operations were executed with GROMACS 5.0.2. by following the "GROMACS Tutorial: Lysozyme in water" [15].

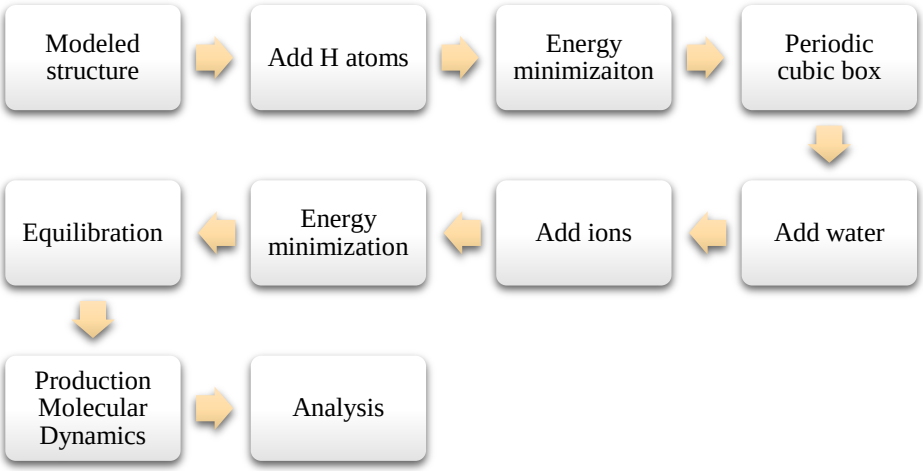



**Figure 15 Flowchart of steps performed in GROMACS.**

Following steps were performed for all models. The input files were first generated according to force field AMBER99SB-ILDN (2010) and hydrogen atoms were added. We then created the periodic cubic box and placed the molecule in the centre and at least 0.7 nm from the box edge. At this point *in vacuuo* energy minimization was done. This first minimization, without solvent, is not included in the original GROMACS tutorial because the tutorial starting structure is more realistic (it has been experimentally determined and downloaded from the PDB database) while our structures have been constructed by homology and need more initial structure preparation. Energy step size was set to 0.005 and maximum number of minimization steps to perform to 50,000. Other conditions for the energy minimization are provided in Figure S8. We then added water molecules with the help of a file spc216.gro which is automatically provided with GROMACS and includes the coordinates of a water cluster. Ions were added by using the ions.mdp file which was prepared in tutorial. We then chose to substitute some solvent molecules with ions and performed another minimization but that time with the .mdp file from the tutorial. Figure 16 represents the potential energy of D11 structure through steps of minimization; similar graphs were obtained for other structures under study (data not shown). One can see the energy decreasing as the minimization progress converges to its equilibrium value. The right side of Figure 16 shows overlapped structures are depicted; the blue before and the red after minimization. This provided first indications how much the structure has changed, in this step. We observed predominantly light displacements of the side chain atoms.

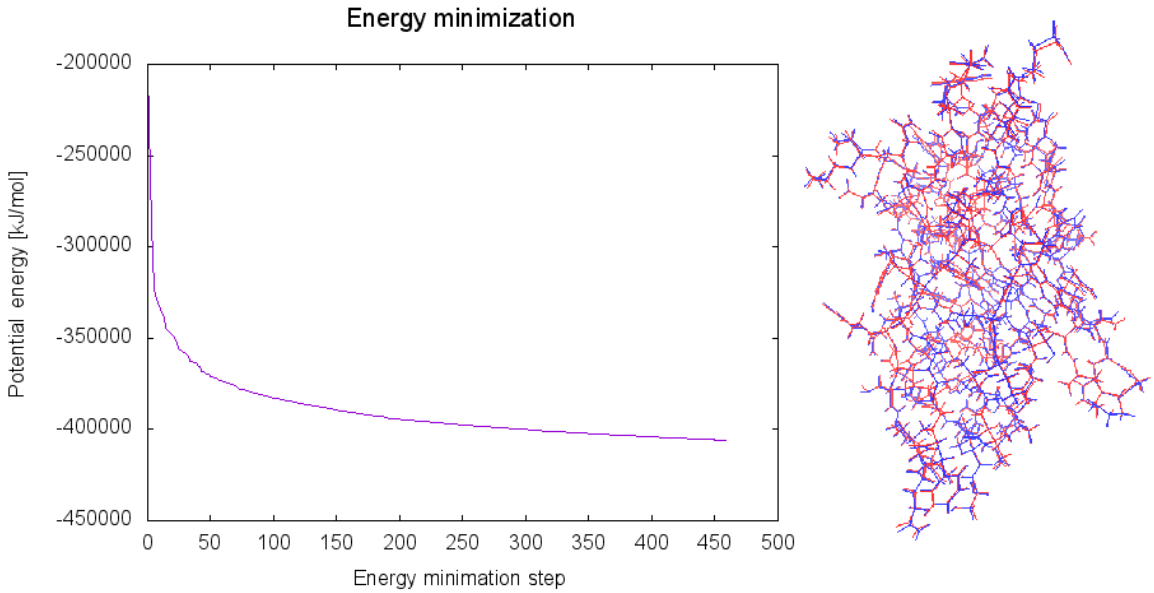



**Figure 16 Energy minimization of the structure D11.
Plot shows potential energy through steps of minimization (left)
and molecule D11 before (blue) and after (red) minimization (right).**

The next step was to equilibrate the MD simulation in which under NVT ensemble the temperature was risen to 300 K with 100 ps total simulation time. GROMACS collects data during the MD equilibration stage, allowing us to observe the temperature vs. time. Figure 17 represents the data for the structure D11. Here it can be observed that the temperature quickly reached the required target quantity and stabilised. Graphs for the remaining other structures under study showed similar behaviour and are not shown here. Also following the MD equilibration obtained graphs are presented just for structure D11 while for the other structures due to similar behaviour we decided not to explicitly include them.

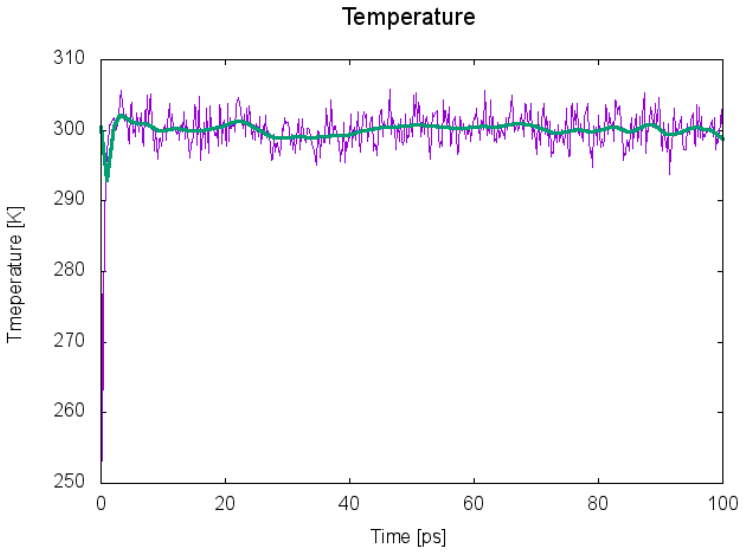



**Figure 17 Temperature along simulation time during the equilibration of D11 structure:
sampled points (violet) and Bezier curve (green).**

The second equilibration was done under the NPT ensemble. This time we were interested in pressure and density. In Figure 18 (where pressure is plotted along time), we can conclude the pressure has stabilised during the equilibration.

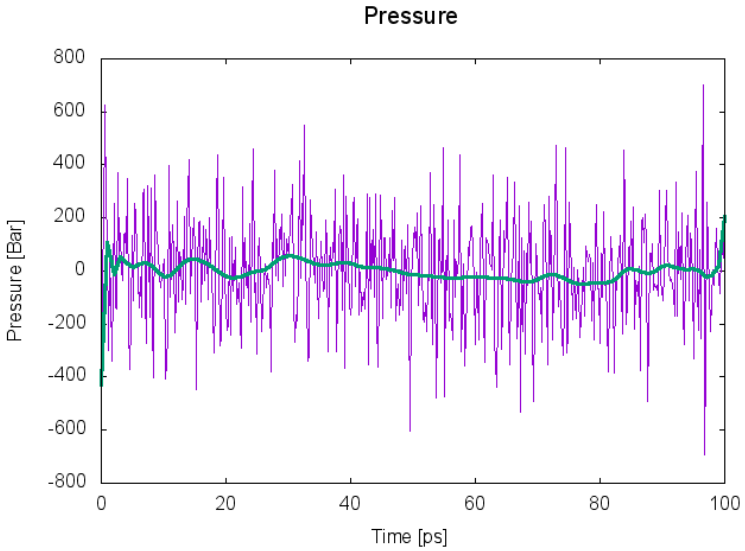



**Figure 18 Pressure along the simulation time during equilibration of D11 under NPT ensemble: sampled points (violet) and Bezier curve (green).**

In addition, data of density vs. time were collected during the equilibration for every system. For one of the D11 system the data are represented in Figure 19.

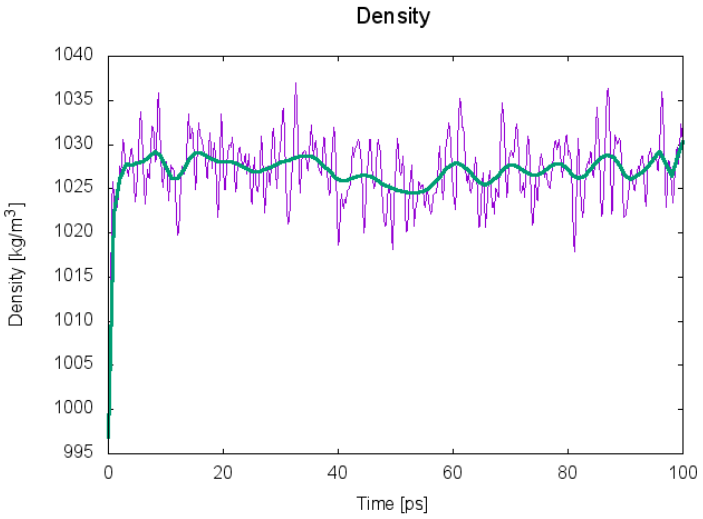



**Figure 19 Density along simulation time during equilibration of D11 under NPT ensemble: sampled points (violet) and Bezier curve (green).**

The next step was the production MD simulation where the simulation of protein was done for 250 ns with the 2 fs time step. MD calculations were carried out within the cluster at the National Institute of Chemistry in Ljubljana. The result was a trajectory, which we visualised and analysed.

## 4.4 Analysis of the production MD trajectories

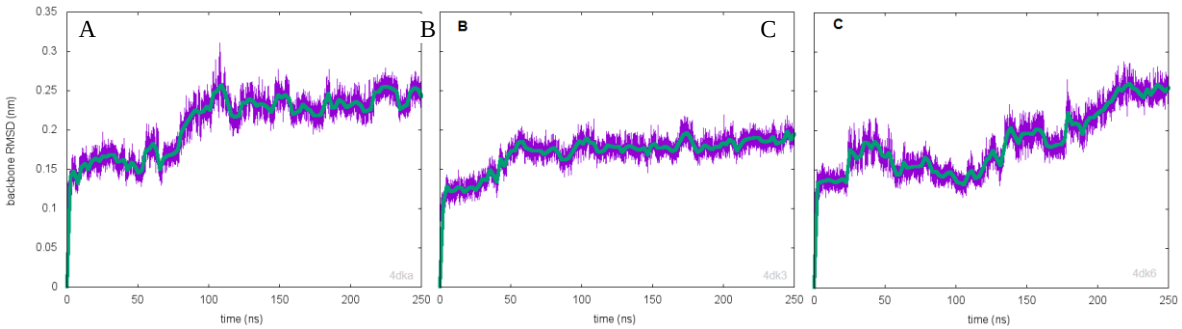



**Figure 20 Time-dependant RMSD Graphs. Backbone RMSD (nm) plotted vs. time (ns) of structures with sequence D11 modeled after target structure (A) 4DKA, (B) 4DK3 and (C) 4DK6. Sampled points (violet) and running average over 100 samples (green).**

Figure 20 represents three graphs each depicting the flexibility of each generated D11 model protein backbone during the MD simulation. The first (Figure 20A) and second (Figure 20B) graphs show the structures which reached equilibrium: first after 100 ns and second after 50 ns simulation time. The first has larger oscillations and higher equilibrated RMSD value (first 0.23 nm, second 0.18 nm) than the second one. The third RMSD graph (Figure 20C) has the tendency to diverge which, indicates instability. It might stay stable after 220 ns onwards but this information can only be confirmed with longer MD simulation. The graphs (Figure 20) show that the structures modeled by 4DKA and 4DK3 templates reached equilibrium. To choose the most suitable structure we looked at the quality of the template (Table 4). Because the crystal structure quality parameters indicate better quality for the 4DKA structure we will use this one for further analysis.

| Crystal structure | Resolution (Å) | Wilson B-factor |
|---|---|---|
| **4DKA** | 1.97 | 25.7 |
| **4DK6** | 2.65 | 33 |
| **4DK3** | 2.76 | 63 |

**Table 4 Quality control indicators of crystal structures.**

In Figure 21 we compared the RMSD graphs of all four modeled sequences: three modeled by the 4KSD and one (D11) by the 4DKA template structure. Structures of the A5, B3, C9 and D11 reach equilibrium after 50, 130, 20 and 100 ns respectively. The A5 RMSD is the most regular (Figure 21-A5). The B3 structure RMSD value (Figure 21-B3) tells us that the structure has changed its conformation at least two times: at about 130 ns and 220 ns of MD simulation time. The C9 structure also changed its conformation during the MD simulation (Figure 21-B3): first at about 120 ns and then again after further 100 ns. The last 10 ns cannot fully be speculated about reaching the equilibration as the graph is decreasing. We would also note that the D11 structure RMSD (Figure 21-D11) is the same as in Figure 20A. The graphs from Figure 21 further suggest that the modeled structures are good 3D representations for the experimentally identified VHH sequences.

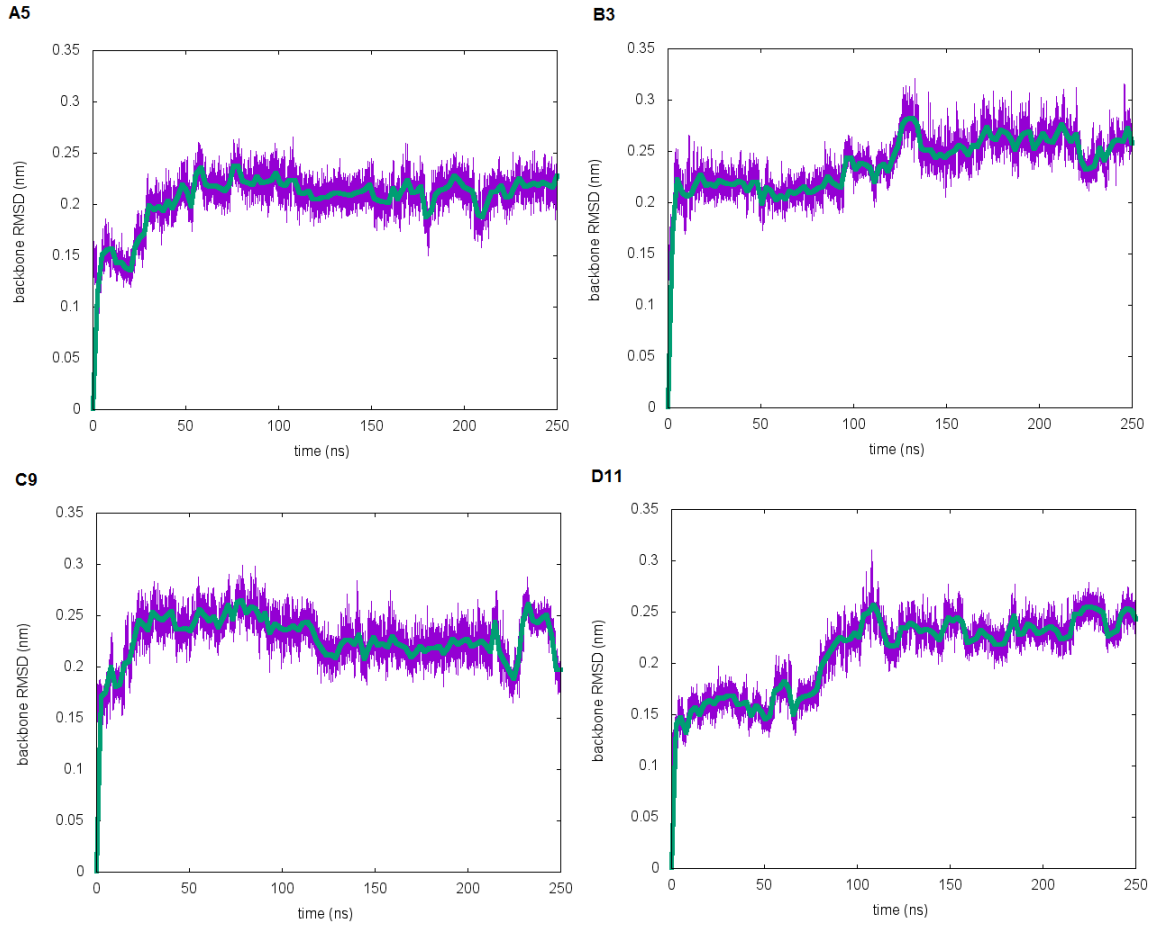



**Figure 21 Graphs of RMSD values obtained from the MD trajectories. Backbone RMSD is plotted for structures of sequences A5, B3, C9 and D11 during MD simulation. Sampled points (violet) and Bezier curve (green).**

Figures 22 - 25 represents the RMSF plots of four modeled structures during the same production stage of the MD simulations. Higher values of the RMSF correspond to more mobile atoms or regions. As we had initially expected the most flexible regions are those that are unstructured or mutated. Atoms that are part of the residues that were mutated are labelled in yellow and we can observe in these Figures that most peaks correspond to these mutations.

We further observed that if a mutation region is longer, RMSF peak is higher and those regions are fluctuating more throughout the MD simulation. This is because the structures from the PDB exist in nature, but by performing mutations (substitution, insertion or deletion of amino acid or rotation of residue) we deform the structures and possibly put the new atoms in unstable positions. The longest mutated regions in our structures are CDRs (length around 6, 9 and 17 amino acid substitutions). In addition, they are also unstructured which increases the fluctuation of these regions even more.

When we analysed all four RMSF graphs we could observe the five peaks are present in these plots. Here, we ignore the beginning and the end fluctuations of these structures because the residues are free to move and have naturally high RMSF values. Among the central peaks, three of them corresponded to CDRs, but two did not.

The question that followed was: why do we have peaks in the protein regions that are fully or slightly mutated? We visualised the structures and established that these regions form very sharp turns. In all the structures the peak located at atom numbers a little under 600 corresponded to a region that is not mutated and it can be seen in each structure at their bottom left corner. The other one (atom numbers about 1100) is higher than the previous peak sometimes even higher than the peak corresponding to CDRs. The region is unstructured and it forms a sharp turn as the previous one. But beside of its unstructured structure it was additionally mutated and this might explain why the peak in this graph is so high. (The turn is at the back of the structure figure added along graphs and it cannot be clearly seen.)

The A5 RMSF (Figure 22) unveils that the CDR regions are not fluctuating the most during simulation but the region with sharp turn and mutations is more mobile it reaches RMSF value 0.22 nm. The CDR peaks of the B3 RMSF (Figure 23) graph are the highest among peaks in the graph and among peaks from other graphs. Their RMSF values are 0.28 nm, 0.45 nm and 0.54 nm. In C9 RMSF (Figure 24) the CDR peaks exceed 0.2 nm, including a peak with the RMSF value 0.24 nm that is not corresponding to the CDRs. Finally, in the D11 structure RMSF value (Figure 25) we observe the peak of the third CDR is the highest (0.35 nm) compared to the other peaks of this graph. This is probably due to the fact that this region is the longest mutated region.

However, through energy minimization, equilibration and MD simulation the suitable position of atoms and overall stable structures were generated thus showing a high probability that the proteins we have modeled using homology modeling followed by MD simulations really have structures strongly similar to the structures in nature. We hope that in the future structural studies will provide a direct experimental validation of our models.

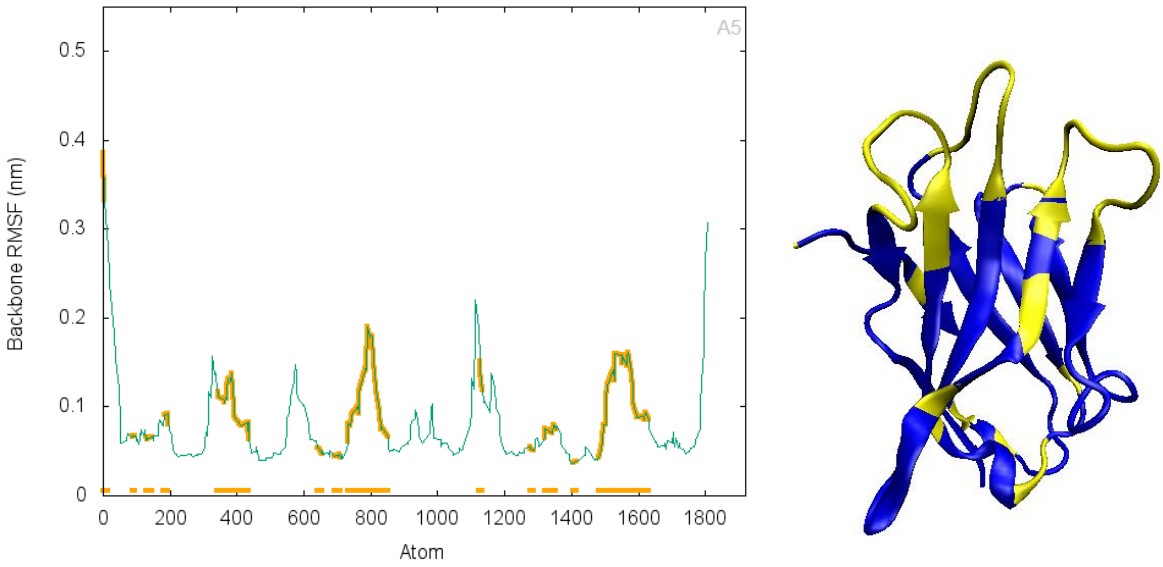



**Figure 22 RMSF plot vs. atom number of the A5 structure (left) and its visualization. On both figures mutated regions are labelled in yellow.**

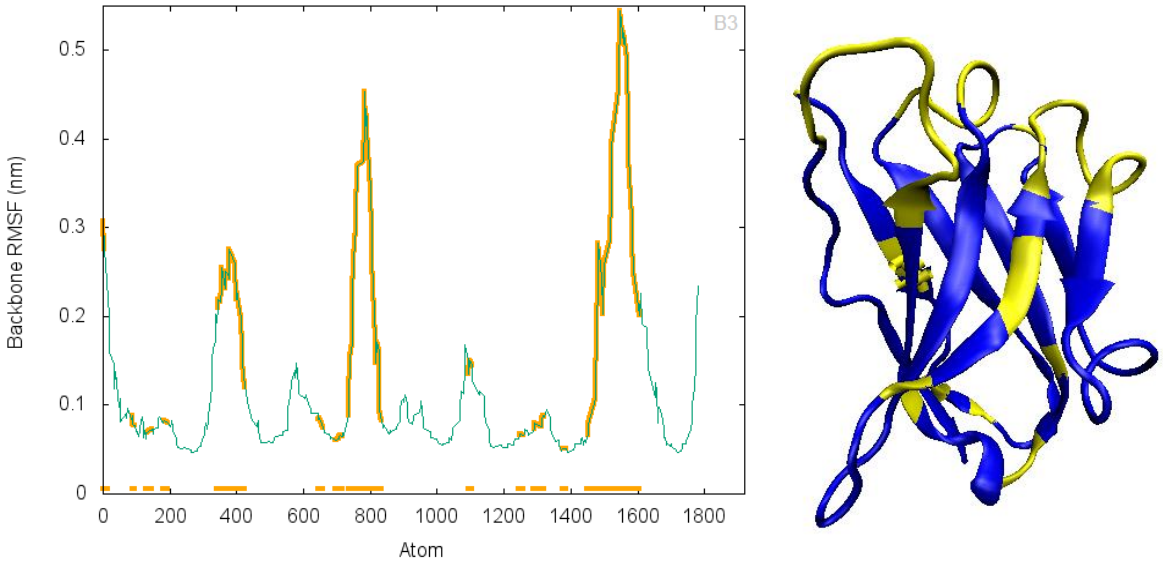



**Figure 23 RMSF plot vs. atom number of the B3 structure (left) and its visualization. On both figures mutated regions are labelled in yellow.**

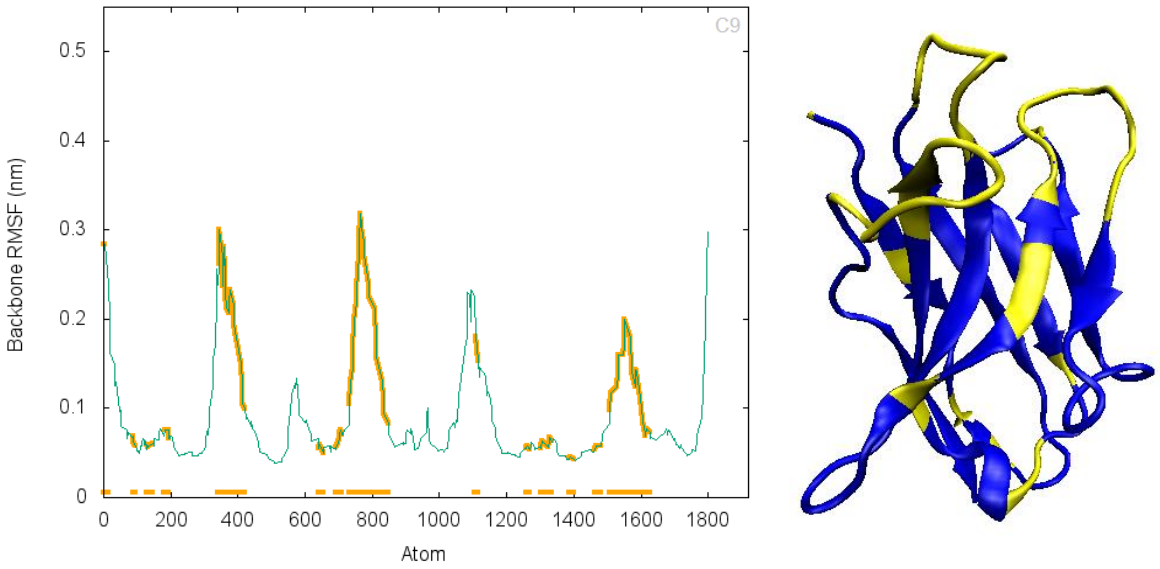



**Figure 24 RMSF plot vs. atom number of the C9 structure (left) and its visualization. On both figures mutated regions are labelled in yellow.**

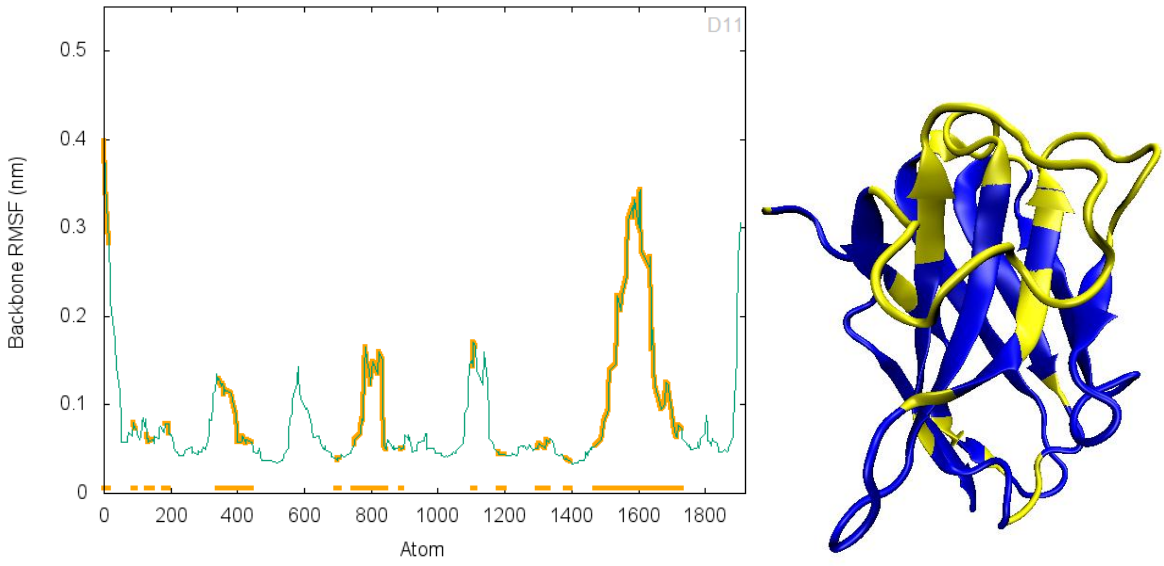



**Figure 25 Plot of RMSF according to atom number of D11 structure (left) and its visualization. On both figures are mutated regions labelled in yellow.**

# 5 CONCLUSION

Sarcoma is type of cancer with multiple targets in the bodies' connective tissues. Although considered a rare cancer its collective incidence is relevant especially among young patients and new approaches to therapy need to be discovered.

The PPI between the Twist1 and p53 proteins were shown to be one of the most important factors enabling aggressive cancer Sarcoma progression and a relevant target of inhibitors of this PPI interaction. Form previous research work we received four amino acid sequences of nanobodies (single domain antibodies) A5, B3, C9 and D11 that could interfere with this interaction, by binding to the p53 protein potentially paving the way for further inhibitors design.

The main goal of this project was to predict and analyse 3D structures of these identified four sequences. In order to achieve this goal we used homology modeling approach coupled with molecular simulations based on empirical force fields to generate and analyse the produced 3D structure. Thus we have also utilized a number of tools for structure prediction and data analysis such as PDB, BLAST, Swiss PDB Viewer, GROMACS, VMD and Gnuplot.

In the first step we identified appropriate predetermined structures from the PDB database, aligned them with our sequences, and selected the initial template structures that have the minimum number of mutations – especially we were avoiding potential deletions and insertions. Then the four sequences were fitted to their corresponding initial templates and some additional modifications were made. After that, we generated full 3D homology models by adding H-atoms, water molecules to solvate the system, ions to neutralize the charge, and set up the periodic boundary conditions to get better results during molecular simulations. We used energy minimization, equilibration and production MD simulation. The last stage of our work was to analyse the obtained trajectories using the RMSD and RMSF parameter.

Our results confirmed that all four produced 3D models possess acceptable conformations and are thus reasonable starting points for further research. These structures will be then used in molecular docking to identify the structure of each p53/VHH complex, for all four A5, B3, C9 and D11 structures. We believe the results obtained can be used as the first step to develop a new way of treating Sarcoma by inhibiting the Twist1 and p53 PPI.

# 6 POVZETEK NALOGE V SLOVENSKEM JEZIKU

Sarkom je vrsta raka, ki se razvije v vezivnem tkivu. Čeprav je relativno redko rakavo obolenje, predvsem njegova agresivnost in pogostejše pojavljanje pri mladih spodbuja raziskovalce k preučevanju in odkrivanju novih protirakavih učinkovin [18, 21]. V sarkomski celici, ima kompleks med proteinoma p53 in Twist1 dokazano vlogo pri razvoju raka [21], saj protein Twist1 preko proteinske interakcije inaktivira protein p53, kar je posledično direktno povezano z zaviranjem nastanka tumorja [17]. Molekulo, ki bi preprečila to protein-protein interakcijo med p53 in Twist1, bi tako lahko potencialno uporabili kot terapevtsko orodje pri zdravljenju sarkoma [21].

Kolegi iz Univerze v Novi Gorici so v predhodnih raziskavah in rešetanjih za identifikacijo inhibitorjev te Twist1:p53 protein-protein interakcije eksperimentalno identificirali štiri proteinske sekvence, za katere so pokazali, da se vežejo na isto mesto, na C-terminalni regiji proteina p53, kjer je prisotna protein-protein interakcija s Twist1 proteinom. Zato predvidevajo, da bodo ti proteini preprečili vzpostavitev neželenega protein-protein kompleksa. Identificirane sekvence so pripadale različnim nanotelesom; to so enodomenska protitelesa, ki jih najdemo npr. pri kamelah in lahko specifično interagirajo z izbranim antigenom. V primerjavi z običajnimi protitelsi, kamelja protitelesa nimajo lahkih variabilnih domen, ki bi pomagale pri interakciji z antigenom, ampak samo težka variabilna domena vpliva na afiniteto z antigenom. Zaradi tega izolirana domena ohrani afiniteto, ki jo ima celotno pripadajoče protitelo. Izolacija nanoteles je relativno lahko izvedljiva in njihova produkcija je enostavnejša in cenejša v primerjavi s klasičnimi protitelesi , zato je to rešetanje potekalo v sintetični knjižnici nanoteles [26].

V tej nalogi smo z namenom, da bi dobili boljši atomistični vpogled v predhodne eksperimentalne rezultate zgradili in ovrednotili tridimenzionalne (3D) strukture identificiranih nanoteles.

Najprej smo generirali tridimenzionalne homologne modele sekvenc nanoteles, ki so bili predhodno identificirani, da lahko ovirajo protein-protein interakcijo med Twist1:p53. Štiri sekvence, iz izvedenih eksperimentov so predstavljale našo začetno točko za računalniško podprto homologno modeliranje. Najprej smo v proteinski bazi struktur (Protein Data Bank - PDB) poiskali dostopne strukture bioloških molekul, ki kažejo najvišjo stopnjo podobnosti z našimi identificiranimi sekvencami. Nato smo jih poravnali in na podlagi te poravnave v več stopnjah redukcije izbrali najboljše predloge (ang. template) in s pomočjo njih, zgradili začetne 3D modele nanoteles z uporabo homolognega modeliranja. Zaradi insercij in delecij smo posledično nekatere aminokislinske ostanke dodatno vstavili oziroma izbrisali ter zasukali nekatere aminokislinske ostanke, kjer je prihajalo v začetnih modelih do steričnih trkov.

V nadaljevanju smo stabilnost in konformacijsko mobilnost izgrajenih homolognih modelov študirali z molekulskimi simulacijami in uporabo empiričnih polj sil (t.i.

molekulske mehanike). Molekulska mehanika opisuje atome kot krogle z definiranim radijem, ki jih povezujejo prožne vzmeti (to je približek za kovalentne vezi) in interagirajo med seboj preko intramolekularnih in intermolekularnih interakcij ki jih opišemo z elektrostatskim Coulombovim členom in nepolarnim van der Waalsovim členom. Bolj kot so atomi, vezi in torzijski kot v prisiljenem neravnotežnem položaju glede na eksperimentalno določene vrednosti za posamezne atomske tipe, večja je potencialna energija sistema [14].

Tako smo najprej izvedli energijsko minimizacijo za homologne modele nanotelesc. Minimizacijo smo ponovili, ko smo sistemu dodali še molekule vode in ione, da je bil celoten sistem elektronevtralen in pri računu upoštevali še periodične robne pogoje. Te smo uvedli v simulacije, da smo kar najbolj realno za izbrani sistem opisali naravno okolje. Molekule se v različnih okoljih (npr. vodni milje) namreč različno obnašajo, zato je pomembno, da pri modeliranju le-teh modeliramo tudi njihovo okolje [14]. Vendar bi modeliranje celotnega okolja vsebovalo preveliko število delcev in posledično, bi se potreben čas za izvedbo simulacij preveč povečal. Zato smo tarčne molekule dali v virtualno škatlo, napolnjeno z molekulami vode in upoštevali periodične robne pogoje. Te omogočajo, da se molekule, ki med simuliranjem gredo ven iz škatle, ponovno pojavijo na drugem koncu škatle. Tako smo ohranili relativno majhno, konstantno količino delcev, ki so bili računsko obvladljivi za energijsko minimizacijo in kasnejše simulacije [14].

Stabilnost vseh generiranih minimiziranih struktur smo nadalje preverili še z molekulsko dinamiko. Najprej smo izvedli ekvilibracijo, kjer smo opazovali, kako so se pod različnimi pogoji spreminjajo tlak, temperatura in gostota sistema. Za tako pripravljene molekulske sisteme smo izvedli produkcijsko molekulsko dinamiko v dolžini 250ns. Kot rezultat smo dobili trajektorije, ki predstavljajo časovno zaporedje konformacij proteinov generiranih s to metodo. Sledila je analiza dobljenih MD trajektorij, pri kateri smo opazovali in s RMSD in RMSF parametri tudi kvalitativno določili, koliko povprečno se posameznemu proteinu spreminja konformacija glede na čas simulacije in povprečno koliko se vsak atom proteinov premakne med simulacijo. Rezultati so potrdili stabilnost proteinov, in tudi pokazali kateri deli proteinov so med simulacijo bolj fleksibilni. Izkazalo se je, da so to predvsem deli, ki tvorijo zanke in deli, ki smo jih mutirali, kar je bilo tudi pričakovano.

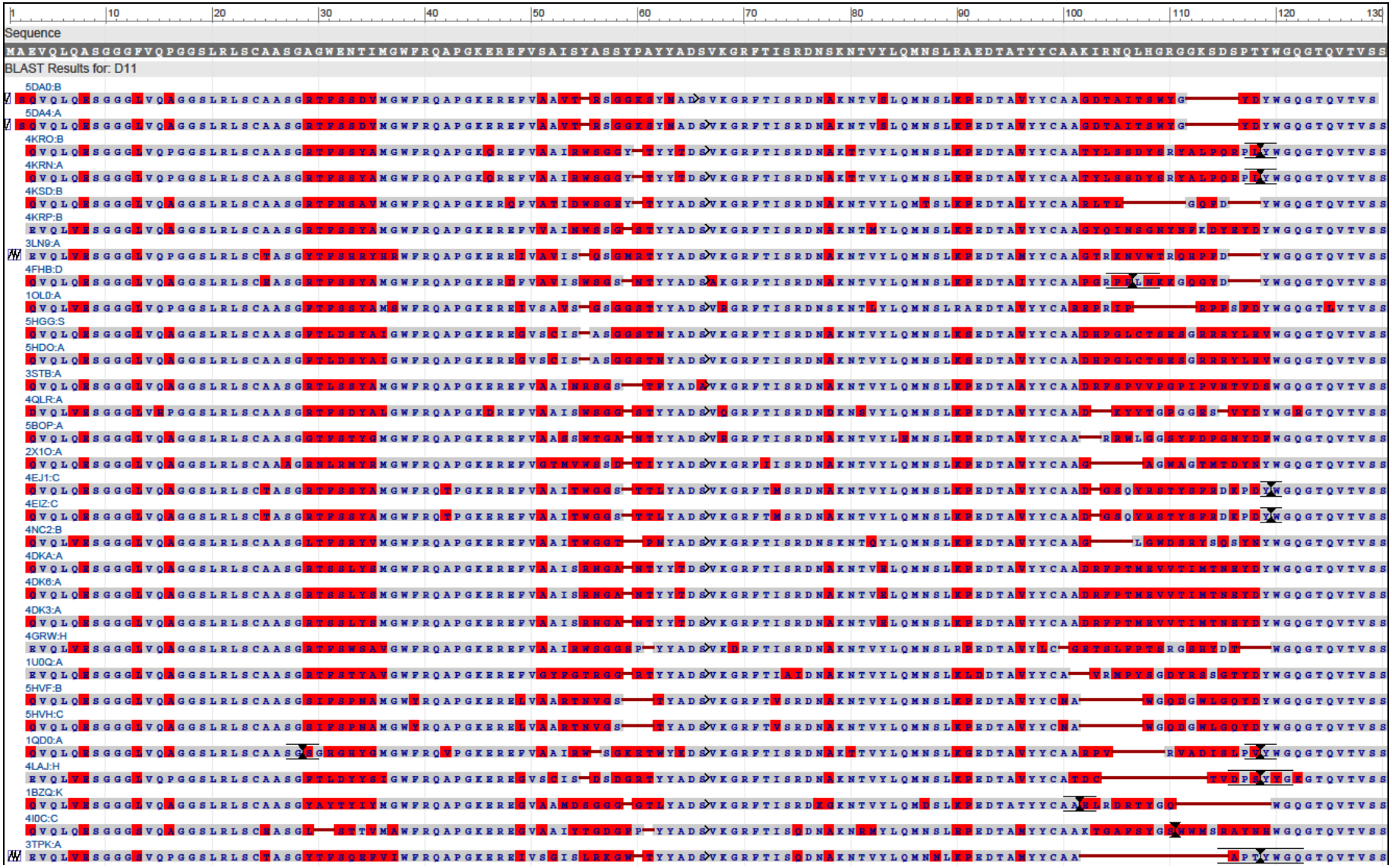
Rezultati te zaključne naloge bodo uporabni za nadaljnje študije, saj bo z metodo molekulskega sidranja natančnejše pokazano kje, s kakšno konformacijo in kakšno afiniteto identificirana nanotelesa interagirajo s proteinom p53. To pa bo podalo prve smernice, v kateri smeri naj gre razvoj inhibitorjev protein-protein interakcije med p53 in Twist1, ki bo morda dolgoročno pripeljal tudi do novih protirakavih zdravilnih učinkovin za zdravljenje sarkoma.

# 7 BIBLIOGRAPHY

1. M. R. Arkin and J. A. Wells, *Small-molecule inhibitors of protein-protein interactions: progressing towards the dream.* Nature reviews. Drug discovery, 2004. **3**(4): p. 301.
2. M. R. Arkin and J. A. Wells, *Small-molecule inhibitors of protein–protein interactions: progressing towards the dream.* Nature reviews Drug discovery, 2004. **3**(4): p. 301-317.
3. H. M. Berman, et al., *The protein data bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-242.
4. C. Chothia and A. M. Lesk, *The relation between the divergence of sequence and structure in proteins.* The EMBO journal, 1986. **5**(4): p. 823.
5. B. S. Daan Frenkel, *Understanding Molecular Simulation*. 2002: Mpg Books Ltd, Bodmin, Cornwall (Great Britain).
6. C. J. Epstein, R. F. Goldberger and C. B. Anfinsen. *The genetic control of tertiary protein structure: studies with model systems*. in *Cold Spring Harbor symposia on quantitative biology*. 1963: Cold Spring Harbor Laboratory Press.
7. R. P. Gajula, et al., *The twist box domain is required for Twist1-induced prostate cancer metastasis.* Molecular Cancer Research, 2013: p. molcanres. 0218.2013.
8. J. B. Gu, Philip E, *Structural bioinformatics*. Vol. 44. 2009: John Wiley & Sons (Hoboken, New Jersey).
9. N. Guex and M. C. Peitsch, *SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling.* Electrophoresis, 1997. **18**(15): p. 2714-2723.
10. W. Humphrey, A. Dalke and K. Schulten, *VMD: visual molecular dynamics.* Journal of molecular graphics, 1996. **14**(1): p. 33-38.
11. P. K. Janert, *Gnuplot in Action; Understanding Data with Graphs*. 2016: Manning Publications (Shelter Island, NY).
12. A. R. Kinjo, et al., *Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format.* Nucleic Acids Research, 2011: p. gkr811.
13. R. A. Laskowski, J. D. Watson and J. M. Thornton, *ProFunc: a server for predicting protein function from 3D structure.* Nucleic Acids Research, 2005. **33**(suppl_2): p. W89-W93.
14. A. R. Leach, *Molecular modelling: principles and applications*. 2001: Pearson education (United Kindom).
15. J. A. Lemkul. *GROMACS Tutorial: Lysozyme in Water*. 2008-2015 [cited 2017 7 Nov]; Available from: http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/lysozyme/index.html.
16. A. Monegal, et al., *Immunological applications of single-domain llama recombinant antibodies isolated from a naive library.* Protein Engineering, Design & Selection, 2009. **22**(4): p. 273-280.
17. C. Muñoz-Fontela, et al., *Emerging roles of p53 and other tumour-suppressor genes in immune regulation.* Nature Reviews Immunology, 2016. **16**(12): p. 741-750.

18. National Cancer Institute. *Soft Tissue Sarcoma—Patient Version*. n.d. [cited 2013 7 Nov]; Available from: https://www.cancer.gov/types/soft-tissue-sarcoma.
19. National Center for Biotechnology Information. *BLAST: Basic Local Alignment Search Tool*. 2017 [cited 2017 7 Nov]; Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi.
20. Y.-J. Park, et al., *The structure of the C-terminal domain of the largest editosome interaction protein and its role in promoting RNA binding by RNA-editing ligase L2*. Nucleic Acids Research, 2012. **40**(14): p. 6966-6977.
21. S. Piccinin, et al., *A "twist box" code of p53 inactivation: twist box: p53 interaction promotes p53 degradation.* Cancer cell, 2012. **22**(3): p. 404-415.
22. Protein Data Bank. *RCSB Protein Data Bank*. 2017 [cited 2017 7 Nov]; Available from: http://www.rcsb.org/pdb/home/home.do.
23. P. W. Rose, et al., *The RCSB protein data bank: integrative view of protein, gene and 3D structural information.* Nucleic Acids Research, 2017. **45**(D1): p. D271-D281.
24. C. Sander and R. Schneider, *Database of homology-derived protein structures and the structural meaning of sequence alignment.* Proteins: Structure, Function, and Bioinformatics, 1991. **9**(1): p. 56-68.
25. K. R. Schmitz, et al., *Structural evaluation of EGFR inhibition mechanisms for nanobodies/VHH domains.* Structure, 2013. **21**(7): p. 1214-1224.
26. M. A. Soler, A. De Marco and S. Fortuna, *Molecular dynamics simulations and docking enable to explore the biophysical factors controlling the yields of engineered nanobodies.* Scientific Reports, 2016. **6**.
27. E. L. Ulrich, et al., *BioMagResBank.* Nucleic Acids Research, 2008. **36**(suppl 1): p. D402-D408.
28. D. Van Der Spoel, et al., *GROMACS: fast, flexible, and free.* Journal of computational chemistry, 2005. **26**(16): p. 1701-1718.
29. S. Velankar, et al., *PDBe: improved accessibility of macromolecular structure data from PDB and EMDB.* Nucleic Acids Research, 2016. **44**(D1): p. D385-D395.
30. V. Vyas, et al., *Homology modeling a fast tool for drug discovery: current perspectives.* Indian journal of pharmaceutical sciences, 2012. **74**(1): p. 1.
31. A. B. Ward, et al., *Structures of P-glycoprotein reveal its conformational flexibility and an epitope on the nucleotide-binding domain.* Proceedings of the National Academy of Sciences, 2013. **110**(33): p. 13386-13391.

# SUPPLEMENTS


Sequence
BLAST Results for: D11
5DA0:B
5DA4:A
4KRO:B
4KRN:A
4KSD:B
4KRP:B
3LN9:A
4FHB:D
1OL0:A
5HGG:S
5HDO:A
3STB:A
4QLR:A
5BOP:A
2X1O:A
4EJ1:C
4EIZ:C
4NC2:B
4DKA:A
4DK8:A
4DK3:A
4GRW:H
1U0Q:A
5HVF:B
5HVH:C
1QD0:A
4LAJ:H
1BZQ:K
4I0C:C
3TPK:A

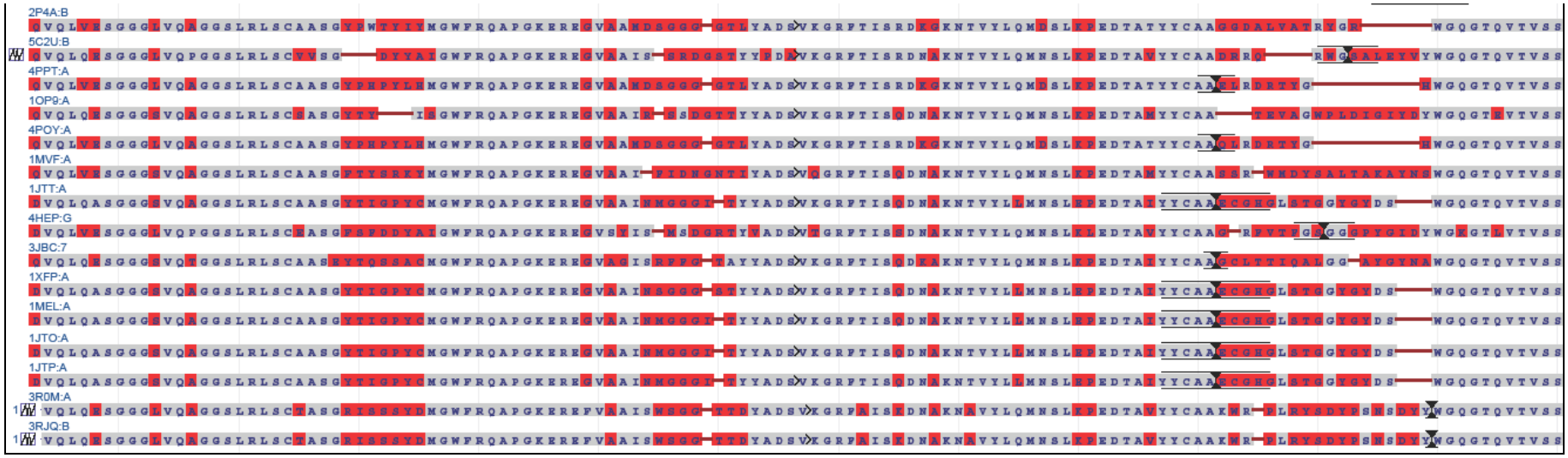



**Figure S1 45 sequences find with BLAST searching in PDB and align with sequence D11.**


**Figure S2 Short list alignment with D11.**

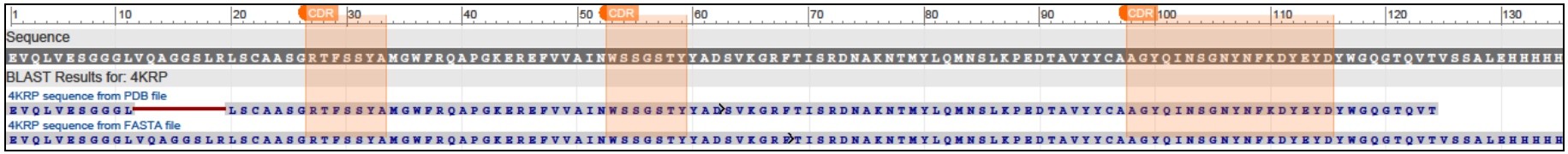
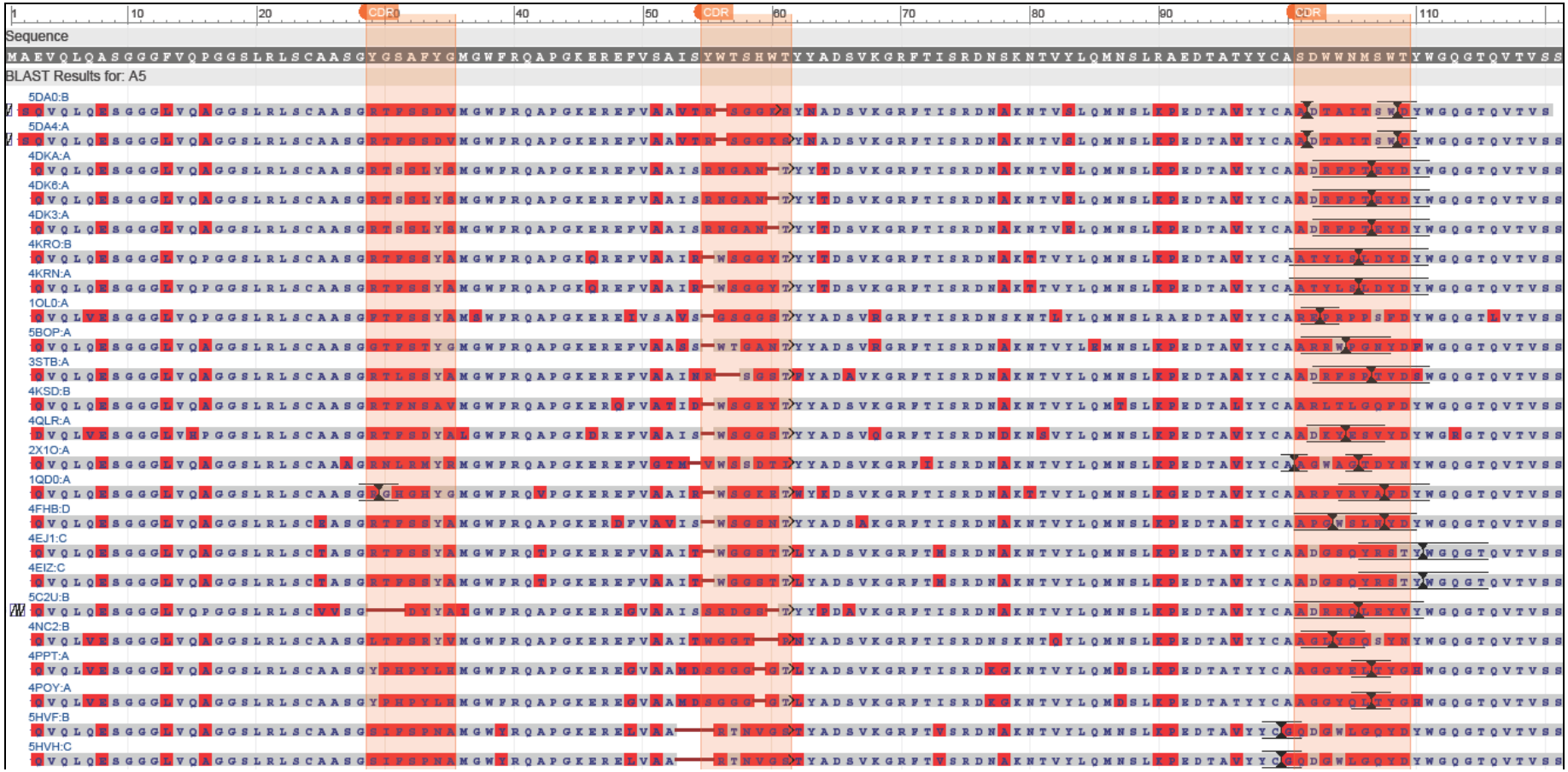



**Figure S3 Alignment of sequence 4KRP that was eliminated.**



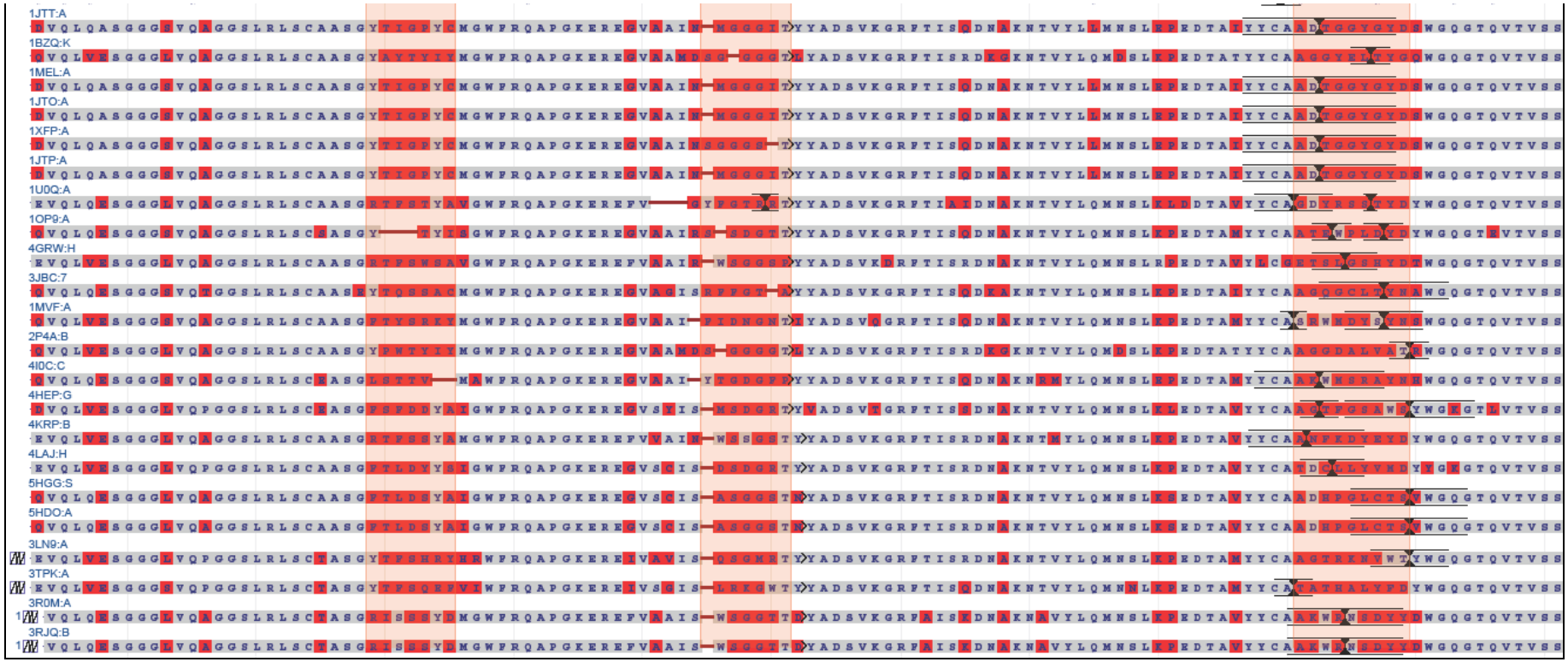



**Figure S4 Sequences find in PDB and align by BLAST with A5.**

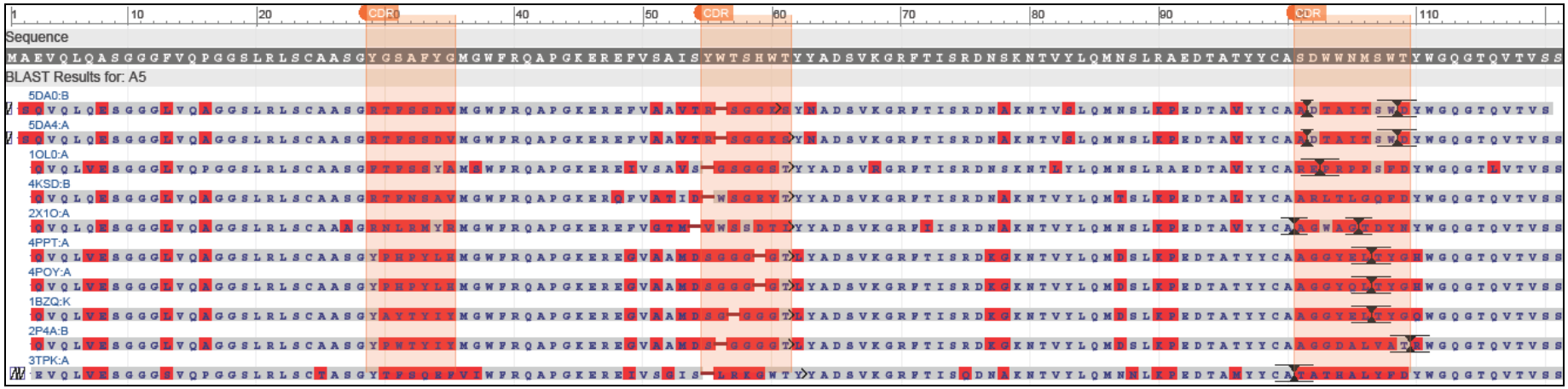



**Figure S5 Short list alignment with A5.**

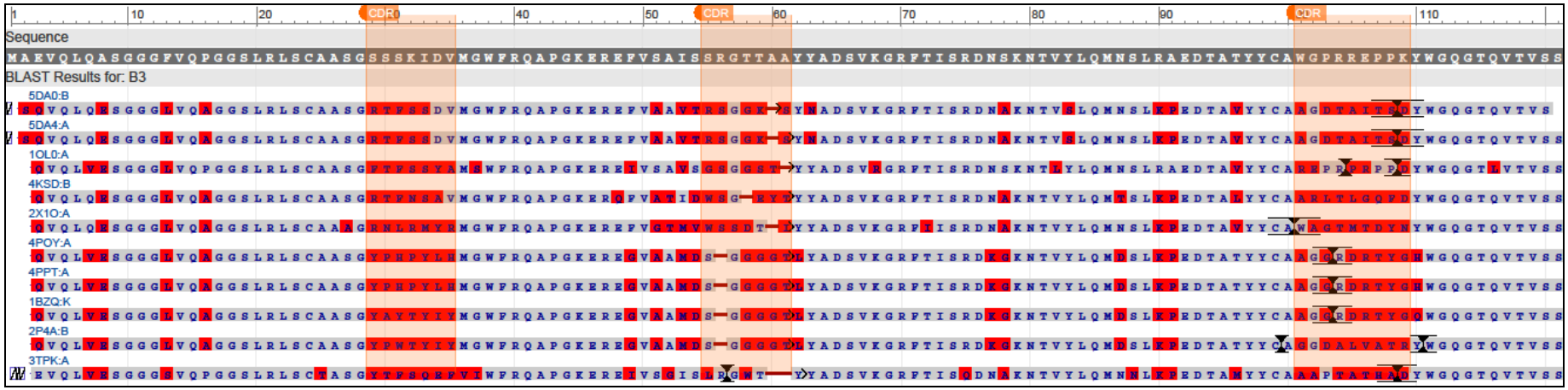



**Figure S6 Short list alignment with B5.**

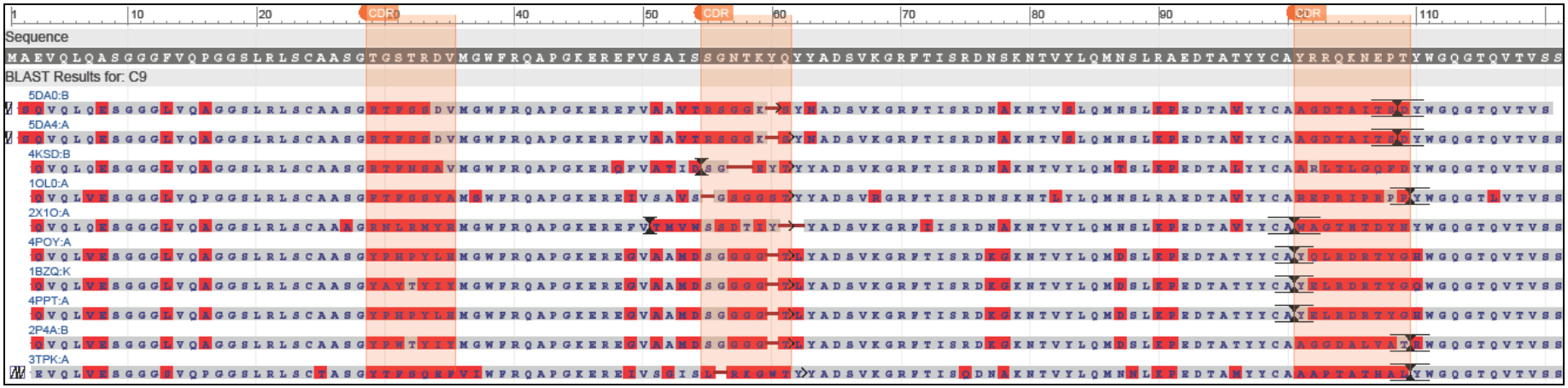



**Figure S7 Short list alignment with C9.**

```
; Parameters describing what to do, when to stop and what to save
integrator  = steep           ; Algorithm (steep = steepest descent minimization)
emtol       = 1.0 (1000)      ; Stop minimization when the maximum force < 1.0 (1000) kJ/mol/nm
emstep      = 0.005 (0.01)    ; Energy step size
nsteps      = 50000           ; Maximum number of (minimization) steps to perform

; Parameters describing how to find the neighbours of each atom and how to calculate the interactions
cutoff-scheme = Verlet
nstlist     = 20 (1)    ; changed to run on GPU cluster
ns_type     = grid      ; Method to determine neighbour list (simple, grid)
rlist       = 1.0 (none); Cut-off for making neighbour list (short range forces)
coulombtype = PME       ; Treatment of long range electrostatic interactions
rcoulomb    = 1.0       ; Short-range electrostatic cut-off
rvdw        = 1.0       ; Short-range Van der Waals cut-off
pbc         = xyz       ; Periodic Boundary Conditions (yes/no)
```

**Figure S8 Content of mdp file. In yellow are highlighted the differences between first and second minimisation. The first number is the parameter used for the second, while the number in parentheses is the one used for the first.**