DEMOCRITUS UNIVERSITY OF THRACE

SCHOOL OF HEALTH SCIENCES

DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

BACHELOR'S THESIS

Quantifying the effect of mutations on protein stability using molecular dynamics simulations: The case of the D30G ROP protein

Ioanna Prigkou, 2517

Supervisor: Nikolaos M. Glykos

Laboratory of Structural and Computational Biology

Alexandroupolis, September 2025

DEMOCRITUS UNIVERSITY OF THRACE

SCHOOL OF HEALTH SCIENCES

DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

BACHELOR'S THESIS

Quantifying the effect of mutations on protein stability using molecular dynamics simulations: The case of the D30G ROP protein

Ioanna Prigkou, 2517

Supervisor: Nikolaos M. Glykos

Laboratory of Structural and Computational Biology

I declare that the present thesis entitled ' Quantifying the effect of mutations on protein stability using molecular dynamics simulations: The case of the D30G ROP protein' is original and was carried out by me personally, as an undergraduate student of the Department of Molecular Biology and Genetics, with Registration Number [2517]. I certify that during the preparation and writing of the thesis, all legal requirements were followed, and that the principles of academic ethics and integrity were fully adhered to, which prohibit the falsification of results, the misuse of others' intellectual property, and plagiarism.

Alexandroupolis, September 2025

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ

ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ & ΓΕΝΕΤΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ποσοτικοποίηση της επίδρασης μεταλλαγών στην σταθερότητα των πρωτεϊνών με την χρήση προσομοιώσεων μοριακής δυναμικής: Η περίπτωση της μεταλλαγής D30G της ROP

Ιωάννα Πρίγκου, 2517

Επιβλέπων Καθηγητής: Νικόλαος Μ. Γλυκός

Εργαστήριο Δομικής και Υπολογιστικής Βιολογίας

Δηλώνω ότι η παρούσα εργασία με τίτλο "Ποσοτικοποίηση της επίδρασης μεταλλαγών στην σταθερότητα των πρωτεϊνών με την χρήση προσομοιώσεων μοριακής δυναμικής: Η περίπτωση της μεταλλαγής D30G της ROP " είναι πρωτότυπη και πραγματοποιήθηκε από εμένα προσωπικά, προπτυχιακό φοιτητή του Τμήματος Μοριακής Βιολογίας και Γενετικής, με Αρ. Μητρώου [2517]. Βεβαιώνω ότι κατά την εκπόνηση της εργασίας και τη συγγραφή της τηρήθηκαν τα προβλεπόμενα από το νόμο, καθώς και ότι ακολουθήθηκαν πλήρως οι αρχές της ακαδημαϊκής ηθικής και δεοντολογίας, οι οποίες απαγορεύουν την παραποίηση των αποτελεσμάτων, την κατάχρηση της διανοητικής ιδιοκτησίας άλλων και τη λογοκλοπή

Αλεξανδρούπολη, Σεπτέμβριος 2025

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisor, Nikolaos M. Glykos, for his help and support during this period. His patience was invaluable for the completion of this project.

I am also deeply thankful to my beloved friends and my family, who were always there for me, reminding me that I belong and I'm never truly alone.

Lastly, I would like to pay my respects to Billie Eilish Pirate Baird O'Connell for the emotional support she unknowingly provided me throughout this difficult year. Her music and presence helped me more than words can express.

Table of Contents

Abstract – Key Words

In the past few years, a great number of computational tools have been developed, for the in silico study of various biomolecules. Some characteristic examples include Alpha Fold (3D Structure Prediction), GROMACS (Folding Simulations) and Grcarma (Trajectory Analysis). These tools can be used to examine the structure and the stability of several proteins, such as ROP. ROP (Repressor of Primer) is a small bacterial protein with a plain secondary structure, for which it is known that some point mutations at the position 30 of the loop region, result in structures with different Tm. This study focuses on three ROP variants (D30G, Native & D30P), and aims to correlate their stability with their Tm values.

Key Words:

AlphaFold, Molecular Dynamics Simulations, ROP, Point Mutation, Loop, Thermal Stability, Tm – Thermal Stability, D30G, D30P

Περίληψη – Λέξεις Κλειδιά

Τα τελευταία χρόνια έχει αναπτυχθεί ένας μεγάλος αριθμός υπολογιστικών εργαλείων με σκοπό την in silico μελέτη διαφόρων βιομορίων. Χαρακτηριστικά παραδείγματα αποτελούν το AlphaFold (πρόβλεψη τρισδιάστατης δομής πρωτεϊνών), το GROMACS (προσομοιώσεις αναδίπλωσης) και το Grcarma (ανάλυση τροχιακών). Αυτά τα εργαλεία μπορούν να χρησιμοποιηθούν για τη μελέτη της δομής και της σταθερότητας πρωτεϊνών, όπως η ROP. Η ROP αποτελεί μία μικρή βακτηριακή πρωτεΐνη με απλή δευτεροταγή δομή, για την οποία έχει βρεθεί ότι διάφορες σημειακές μεταλλάξεις στη θέση 30 του βρόχο, οδηγούν σε δομές με διαφορετικά Tm. Η συγκεκριμένη εργασία αφορά τη μελέτη τριών μεταλλαγμάτων της ROP (D30G, Native και D30P), με σκοπό τη συσχέτιση της σταθερότητάς τους με τα αντίστοιχα Tm.

Λέξεις Κλειδιά:

AlphaFold, Προσομοίωση Μοριακής Δυναμικής, ROP, Σημειακή Μετάλλαξη, Βρόχος, Σταθερότητα, Tm – Θερμική Σταθερότητα, D30G, D30P

1. Introduction

1.1 Folding Problem & AlphaFold 3

For many decades, one of the greatest challenges that concerned the scientific community was the "Protein folding problem". This problem refers to the difficulty of predicting the three dimensional structure of a protein, only by knowing their amino acid sequence. [1].

AlphaFold was first developed and released in 2018 by Deep Mind (subsidiary of Google) and aims to provide a solution to this problem. More specifically, it's a computational method based on artificial intelligence designed to predict the three dimensional structure of a protein based on their primary amino acid sequence. Additionally, it uses the prior knowledge about the physical, chemical and biological background behind the mechanisms of protein folding, while it also incorporates information about their evolutionary history [2].

The most recent version of AlphaFold (AlphaFold 3/ AF3 – 2024) predicts with even greater accuracy the 3D structure and can also identify various interactions between a given protein with other molecules such as DNA, RNA, ligands and ions. Moreover, it can predict interactions among two or more different chains which function as a complex. Lastly, the structures derived from AlphaFold seem to be accurate enough, even in cases where there is no other similar, homologous or closely related structure in any database [3].

1.2 Molecular Dynamics

The knowledge of how a protein moves and folds in the three dimensional space – or any other biomolecule – is essential in order to determine various properties of the system, one of which is its stability. For that reason, it seems extremely important to calculate the trajectory (total positions through which each atom moves) of the system, in a certain period of time. This defines the general folding/movement of the system. For that reason, there has been developed and used molecular dynamic simulations and force fields.

Molecular dynamic simulations are computational tools, used to simulate the folding process of a protein as well as to evaluate the dynamic behavior of a system over time. Among the calculations performed by molecular dynamic simulation software are:

1. The calculation of the positions and velocities of each atom at specific time points.
2. The estimation of the forces between the atoms of a system (potential energy). These forces determine the position of the atoms and consequently of the entire protein.
3. The updating of the new velocities and positions which are calculated during the simulation, using specific algorithms.
4. The creations of trajectories files that can be used in order to analyze and visualize the structure [4].

One of the most important steps during a folding simulation is the calculation of the total energy of the system for each time step. The total energy is the sum of the kinetic and potential energy.

To perform this kind of calculations is essential to use:

1. Equations of classical mechanics

During a molecular dynamic simulation, Newton's Second Law equation of motion needs to be solved. This equation is based on the Cartesian coordinates of the atoms in a molecule. One algorithm commonly used this purpose is the ''leap-frog algorithm'', which calculates the position of each atom at a specific time step (dt) and the corresponding velocity at every intermediate time (dt + dt/2). This is the default algorithm used in "GROMACS", one of the most widely used and established molecular dynamic simulations software. The "Velocity Verlet" algorithm can also be used, which calculates both the position and the velocity at the same point of time [5].

2. Equations of Statistical mechanics

Statistical mechanics are used in order to describe the relationship between the microscopic properties of the atoms of a system, such as their velocity and position, and the macroscopic properties that characterize the whole system, such as temperature and pressure. In other words, they connect the macroscopic behavior of the protein with the microscopic properties of the atoms of which it consists [6,7].

3. Force Fields (potential energy)

These are discussed in the next section.

1.3 Force Fields

Although the kinetic energy of a system can be easily determined, calculating the potential energy seems to be more challenging. This happens because potential energy depends in every kind of interatomic interactions including both bonding and non-bonding interactions [5]. In order to determine the new positions and velocities during the simulation process and to integrate them with the previous results, it's important to know the forces acting upon the atoms of the system [8].

More specifically, these interactions can be described mathematically using the following potential energy functions:

$$V = E_{bonding} + E_{non\text{-}bonding}$$

$$E_{bonding} = V_{strech/lenght} + V_{angles/bending} + V_{dihedral/torsional}$$

$$E_{non\text{-}bonding} = V_{elecrostatic} + V_{Lennard\text{-}Jones}$$

Where V_{strech} refers to the energy associated with the stretch of a chemical bond, $V_{angle,}$ describes the energy between three different atoms that form an angle and $V_{torsional}$ represents the energy of the rotation around the bond.

As for the non-bonding interactions, the Lennard Jones potential describes the Van Der Waals interactions, which can be both attractive and repulsive force that developed between two atoms

These kinds of forces can arise between:

- Two permanent dipoles
- One permanent-dipole and one induced-dipole
- Dispersion Forces (London) [9].

The electrostatic forces, on the other hand, are calculated using Coulomb's Equation. Both type of interactions, depend on the interatomic distance and are essential for the determination of the potential energy.

These forces must be calculated for every time step of the simulation [8].

One possible approach is to calculate the potential energy, using quantum mechanical equations such as Schrodinger' equation. However, this approach was not as efficient as initially expected, especially when the simulation focuses on a large system. Additionally, their solution is extremely computationally demanding [8, 10]

The solution to this problem arises from the development of force fields. Force fields are mathematical models which take into consideration the total of the interatomic forces and compute the potential energy which is responsible for the new positions and velocities that each atom takes during the simulation.

1.4 ROP (Structure & Function)

ROP protein – Repressor of Primer – is a small bacterial RNA binding protein consisting of only 63 amino-acids and derives from the bacterium *Escherichia coli* (ColE1). Its main role is to regulate the replication of the bacterium's plasmids, by controlling RNA-RNA interactions. Its structure is well-known from X-ray crystallography and H NMR experiments [11].

For plasmid's replication, the bacterium uses DNA Pol I and a RNA molecule named RNAII which functions as a primer. In vivo, ROP protein interacts with another RNA molecule (RNAI), forming a complex which prevents the initiation of the replication by interacting with RNAII. ROP does not work like a conventional repressor, but it enhances the negative regulatory role of RNAI. This mechanism is responsible for the preservation of a constant plasmid copy number in the cell. Mutations on ROP or RNAI increase the plasmid's replication frequency and leads to a higher number of copies [12].

More specifically, ROP is a homodimer, consisting of two chains, A and B, each of them forms a pattern of an a-Helix – Loop – a-Helix (HLH). Each a-helix from one of the chains packs with the other with the «knobs-into-holes» model creating a secondary structure element, known as a coiled coil. The two chains also pack together according to the "ridges-in-grooves" model, resulting in a left-handed four-helix-bundle (3D structure) [13].

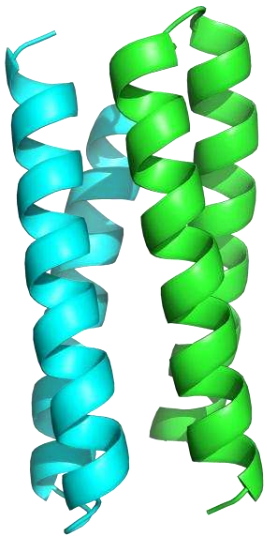


Figure 1: Native ROP, derived from Pymol. Picture of the 4-helix-bundle of ROP where A chain colored green and B chain colored blue.

In ROP's primary amino-acid sequence, there is a periodicity of seven residues which is characterized by the presence of two hydrophobic residues at positions a and d (1 and 4 of each heptad) and two charged residues at the positions e and g (5 and 7) which are forming salt bridges. This results in the formation of a hydrophobic-core at the center of the 4-helix-bundle which consists of eight layers. These layers are placed symmetrically to an intramolecular 2-fold axis which is responsible for the stability of ROP's three-dimensional structure.

This heptad periodicity ends at the loop residues and that is the reason that these residues do not adopt an a-helical conformation due to lack of interactions. Instead they appear to be less stable and more flexible [14].

1.5 Point Mutation & Tm

In Native's ROP primary amino acid sequence, an aspartate (D) is located at position 30 (inside the turn region). Experiments, carried out in the early 90's, have indicated that substitutions of aspartate with any other of the remaining 19 amino acids, does not affect ROP'S RNA-binding ability. However, differences in their thermal stability were observed, as reflected in their Tm (Melting Temperature) values. Tm is the temperature in which half of the proteins are unfolded – lose their secondary and 3D structure. Tm values of all these 20 variants, range from 58.9 ^{o}C to 80.3 ^{o}C, while Native ROP has a Tm value of 68.7^{o}C. The variant with the smallest Tm corresponds to D30P (proline in position 30), whereas the variant with the higher Tm corresponds to D30G (glycine in position 30) [15].

1.6 Main Question

This study focuses on the analysis and comparsion of the stability of Native ROP structure and of two other variants, D30P and D30G. This became possible thought the use of certain computational tools, such as trajectory analysis software and other statistical tools. The main purpose is to determine whether there is any correlation between their structural stability and their melting temperature. More specifically, we aim to examine if an increase in Tm value is quantitatively propotional to an increase in their general stability, as observed in three independent molecular dynamic simulations.

2. Method

2.1 Alpha Fold

For the creation of the primary structure of D30G, wild type (wt) sequence of ROP (PDB ID: 1RPR) was used, from residue 1 to residue 57, with a substitution at position 30 from aspartate (D) to glycine (G).

Figure 2: Amino acid sequence of D30G used for the prediction of the structure from AlphaFold, with the point mutation at position 30. Two copies were chosen due to ROP's homodimeric nature.

AlphaFold 3, among other parameters, calculates, for every structure prediction, their ipTM and pTM scores. These scores evaluate the precision of every three-dimensional structure prediction model. In this case:

1. pTM – Predicted Template Modeling score - shows how similar the structure generated from Alpha Fold, is to the real one.

2. ipTM score - inter-chain predicted TM score - indicates how precise is the prediction regarding to the interaction of two or more different chains – subunits [16].

The greater the score is, the more accurate the prediction is. In the case of D30G, ipTM score is 0.86 and pTM score equals 0.88. Both values indicate that the prediction is very close to reality.

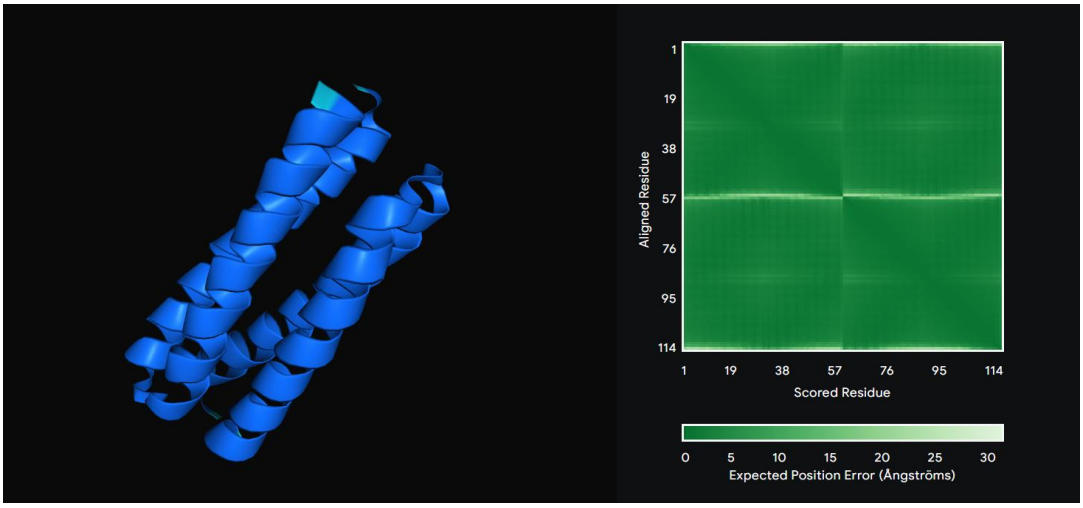


Figure 3: D30G structure and PAE plot generated by AlphaFold.

The "Expected Position Error", is a parameter used to quantify the confidence of the prediction from AlphaFold, for each residue of the structure. Dark green color (like in this case) shows that D30G approaches the real structure, with high confidence.

2.2 GROMACS

GROMACS is well-established and widely-used software, especially in structural biology, and it is designed to perform molecular dynamic simulations.

In order for a molecular dynamics simulation to begin, we need to determine the initial positions – coordinates of all the atoms of the system - and their velocities at t=0 [17]. These initial coordinates can be obtained from X-ray crystallography experiments, H NMR methods or, in this case, from three-dimensional structure prediction models such as AlphaFold.

Therefore, it is essential to add a water solvent model. This step is important in order to simulate the normal environment in which a protein moves and folds. The water molecules that overlap the protein need to be removed and get rearranged in space, in the presence of the protein. That is why an energy minimization step is nessecery. The default water model used is TIP3P (TIP 3-point). It describes the water as a molecule of three atoms (2 H and 1 O) with standard calculated Lennard-Jones forces. Moreover, the length of bonds and angles is fixed [18].

The next step, includes the selection of an appropriate force field. In this case, AMBER99SB*_ILDN was chosen. AMBER99SB*_ILDN is one of the several improved version of Amber ff99SB force field. This improvement relates to the atoms that constituting the backbone of the protein (*) as well as the torsion angle of the side chain (R Group) of some specific amino acids – Isoleucine, Leucine, Aspartic Acid and Asparagine - (ILDN) [19].

Before the production phase of the molecular dynamics simulation, an energy minimization step has to be performed in order to remove any intense Van Der Waals interactions. Skipping this step will lead to inaccurate and unreliable results as well as, in a non-typical folding pattern which does not correspond to reality [17].

The following step is the equilibration of the system. Firstly, equilibration takes place under constant volume with the NVT ensemble (fixed number of atoms – N, constant volume – V and temperature - T). This ensemble is used in order to stabilize the temperature of the system. The second equilibration step is

performed under NPT ensemble (fixed number of atoms – N, constant pressure – P and temperature –T). This ensemble enables the system to stabilize its pressure and adapt its volume. Both of them allow the system to move and fold under conditions that resemble its natural environment. Consequently, the results of the simulation are more representative of the actual dynamics of a protein [20].

Lastly, the production phase of the simulation takes place. The temperature of the simulation was 320K. The files generated from this procedure, were used for the analysis of proteins' stability.

2.3 Xmgrace & xmgr

Xmgr is software used to create two dimensional plots from the data derived from molecular dynamics simulations. Xmgrace is a more recent, improved and optimized version of xmgr [21]. The first one was used for the construction of RMSD and RMSF plots as well as for histograms showing the distribution of RMSD values.

2.4 Carma & Grcarma (RMSD - RMSF - PDB)

Carma is a trajectory analysis software, used to analyze data obtained from molecular dynamics simulations [22]. Grcarma is a fully automated and a more simplified version of Carma software. That makes it easier in use even for more inexperienced users [23]. Various types of analysis can be conducted using both Carma and Grcarma.

For the purpose of this study, several analyses were performed, including:

1. Calculation of RMSF and RMSD values and creation of the corresponding plots
2. Determination of the residues forming the turn's region.
3. Principal Component Analysis (Dihedral & Cartesian PCA) and eigenvalues comparison
4. Extraction of PDB files in order to visualize the structure.

RMSF plot

RMSF (Root Mean Square Fluctuation) indicates the average distance of each each selected atom (for example Ca) of the protein compared to their position in the average structure. For the construction of RMSF plots, the average structure from all the frames of the simulation was initially created. Ca atoms of all residues for both chains (A and B) were used for the calculations. The step between frames was 1.

The same procedure was followed for all of the three structures (Native, D30P and D30G). This resulted in a file named ''average.protein.pdb'', which includes, in different columns, the number of Ca of each residue, the name of the residue and their chain identifier. The last column contains the coresponding RMSF value.

After isolating the last column, this new files containing the RMSF values, were used as an input for xmgr. Different plots were made comparing RMSF values of residues from chain A and B, separately, of Native ROP, D30G and D30P. Three more plots were constructed, from which some of the N-terminal and

C-terminal residues were removed, in order to determine the loop's residues and to observe their stability.

RMSD plot

RMSD (Root Mean Square Deviation) indicates the average distance (in Ångstroms) of each frame of the simulation, from frame 1. The higher the RMSD value is the more the structure differs from the first one. That's why RMSD can be used to evaluate the stability of the system.

For the construction of RMSD plots, ''fitting'' option from grcarma was used. Among the generated files, there was a file named "rms_from_frame_1_protein.dat" which contained the RMSD values for each frame. Using these files as an input for xmgr, three plots were produced. The first plot contained the full length (all residues) Native ROP, D30G and D30P. For the second plot, N-Terminal and C-Terminal residues of every structure, have been removed. The last plot, contained only the residues of the loops/turns which were determined based on their RMSF vlaues.

Extract PDB

Using the "extract pdb" option from grcarma and with step = 100.000 for the Ca atoms, twenty frames were saved for each protein. These structures were aligned based on the residues of the a-helices in order to identify the residues or some specific regions of the loop that appeared to have higher mobility. These results were also compared with their corresponding RMSF values.

2.5 Statistical Analysis with R – Library Sn

In order to calculate mean value, standard deviation and log-likelihood of the RMSD values distribution, a statistical analysis was performed using R. These statistical parameters are essential for quantifying the effect of each mutation.

The analysis performed for both the full-length protein (excluding tail residues) and for only the residues forming the loop region of each structure.

Two statistical models were used for the same data set for all of three structures. Initially, a skewed distribution model was used. This model is more efficient when data follow an asymmetric distribution around their mean value.

The first step was fitting the data, using the command:

"mod <- selm(V1~1, data=testdata)"

After fitting the data, the following values were extracted using the command *"summary (mod)"*:

- ➔ Mean
- ➔ Standard Deviation
- ➔ Log-likelihood
- ➔ Gamma value (only for skewed distribution model). This value corresponds to the skewness showing whether the data tend to lean to the right or to the left.

Lastly, histograms and Q_Q plots were constructed for each structure.

The same procedure was repeated for a second time using the symmetrical distribution model. The only difference was during the fitting step, where the command was modified to:

"mod <- selm(V1~1, data=testdata, fixed.param=list(alpha=0))"

Setting “alpha” parameter to 0, leads to the creation of plots that follow the symmetrical distribution.

In the symmetrical distribution model, “summary (mod)” command does not display a gamma value because skewness in that case equals zero, reflecting the data’s symmetry around their mean value.

2.6 Dihedral PCA & Cartesian PCA

Principal Component Analysis or most commonly known PCA is a computational method that aims to reduce the complexity in the dimensional space. Molecular dynamic simulations generate a large amount of data which are computationally demanding and challenging to analyze, making it difficult to extract valid results. PCA, from all the information, keeps and presents these data that are responsible for the greatest variability while it removes data that do not affect the results significantly.

When referring to the protein folding simulation data, PCA preserves the information responsible for increasing the mobility of the system, thus affecting its stability. PCA can be Cartesian PCA and Dihedral PCA.

- In *Cartesian PCA (cPCA),* Cartesian coordinates (x, y and z), are used because they define the positions of each atom in space, for every time step of the simulation. However, this approach presents some difficulties when separating the internal motions of a system’s atoms from its overall movement [24].

- In order to analyze, with even the smallest detail the internal movement of the atoms, *Dihedral PCA (dPCA)* can be performed. In dPCA, dihedral angles are used. These angles refer to φ (phi) torsion angle which is formed between the N atom and the central carbon (Ca) atom of an amino acid and y (psi) which is formed between C atom and the central carbon (Ca) atom of an amino acid.

Other internal parameters, such as bond length or bond angles, do not change significantly during the simulation and consequently they can not be used to indicate small alternations in the system's stability and mobility [25].

Initially, a dPCA was performed for the full-length structures excluding tail's residues (from 1 to 5 and from 56 to 57). Then, another analysis took place, using only residues of each turn (A and B) separately (from residue 27 to 34).

Following dPCA, two separate analyses were performed using cPCA. The first analysis included residues 27 to 34, which form the loop of chain A, and residues 21-24 and 39-42 which are parts of the a-helices of chain B and locates across turn A. The second analysis, included residues 27 to 34 from the loop of chain B and residues 21-24 and 39-42 from a-helices of chain A. The purpose of these analyses was to observe any minor change in the motion/interaction between each turn and the opposing residues from the a-helices as well as to determine if this movement is related to their Tm.

Initially a fitting step was performed, using grcarma, in order to remove the global and local rotation that could affect and distort the results.

Then, the cPCA option was selected and the following files were obtained (the same type of files was also generated after dPCA):

- Files containing eigenvalues and eigenvectors of every principal component
- Files including information about the clusters that emerged from the analysis
- PDB files for the average and the representative structure of each cluster.

For cPCA, the representative structures of each cluster were visualized, and the a-helical residues of each structure were aligned using PyMOL, for each variant.

2.7 PyMOL

PyMOL is a visualization tool for proteins and other biomolecules. In that case, it was used to visualize the representative structures of each cluster from the PCA analyses as well as the structures derived from .pdb files generated from grcarma after using the “exctract pdb” command.

3. Results

3.1 RMSF plot

RMSF values, are important to determine which residues are less stable and which of them belong to the loop region of the protein. Firstly, a comparsion was perfomed among A and B chains, seperately, of every structure. The following plots obtained from xmgr:

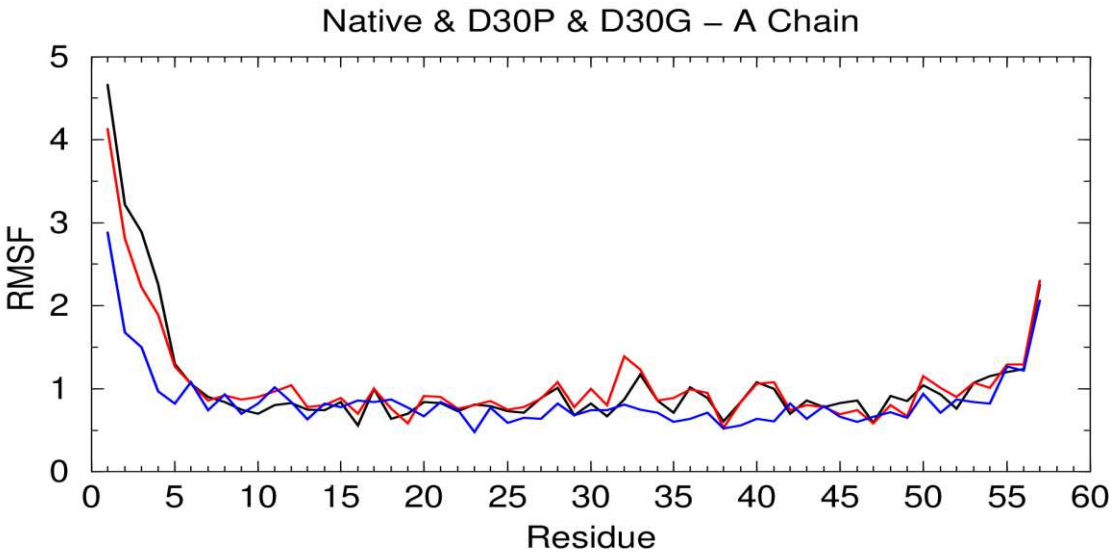


Figure 4: RMSF plot for chain A of Native ROP (black), D30P (red) and D30G (blue)

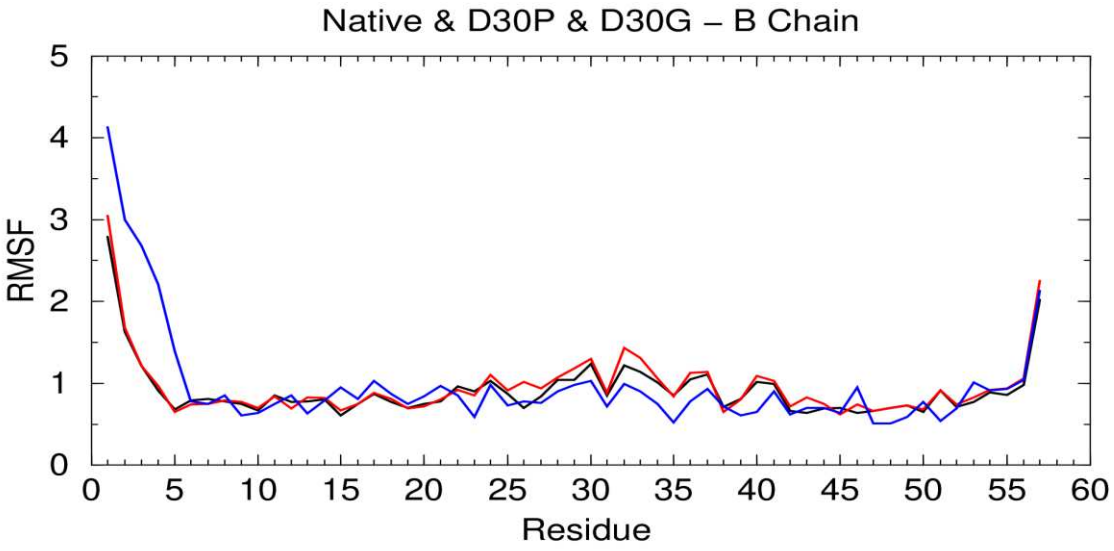


Figure 5: RMSF plot for chain B of Native ROP (black), D30P (red) and D30G (blue)

The increased mobility of the three stuctures is caused by the presence of the residues of the two terminals (figure 4 and 5) which, according to their RMSF values, extend from residues 1 to 5 (for N-terminal) and from 56 to 57 (for C-terminal). More specifically, residue 1 from chain A of Native ROP (Figure 4) has an RMSF value approaching 4.5 Å, while the corresponding value for D30P is almost 4 Å and for D30G is close to 3 Å. The same happens with the residues of the C-terminal for chain A, where residue 57, of all of the three structures, is close to 2 Å.

In order to locate and identify turn's residues, and for that reason alone, residues 1 to 7 and 53 to 57 were removed and new plots were created.

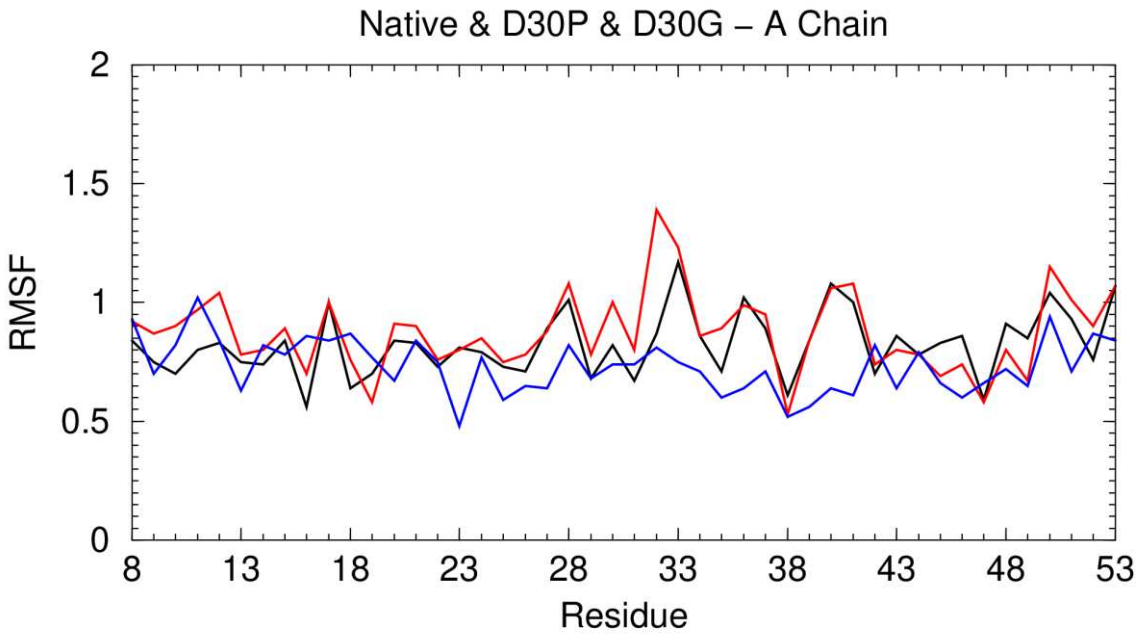


Figure 6: RMSF plot for A Chain of Native ROP (black), D30P (red) and D30G (blue), without residues 1 to 7 and 53 to 57.

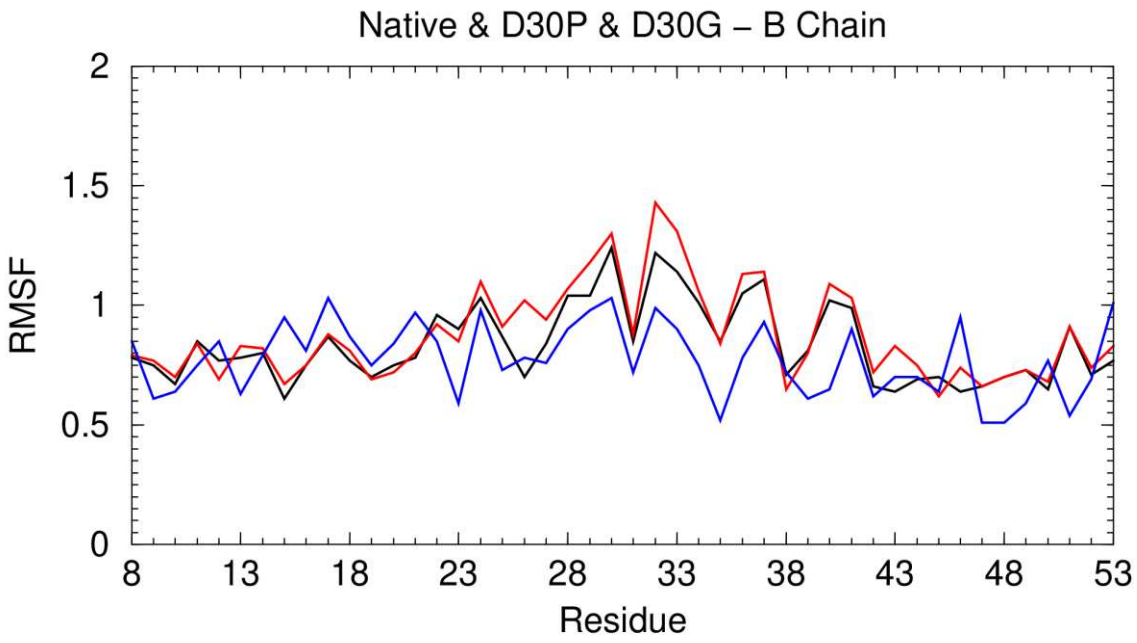


Figure 7: RMSF plot for B Chain of Native ROP (black), D30P (red) and D30G (blue), without residues 1 to 7 and 53 to 57.

After taking into consideration the fact that loop residues have increased mobility [26] and consequently they are taking higher RMSF values, and according to the diagrams above (figures 6 and 7), we can conclude that the loop-turn residues of all of the three structure extend somewhere between 27 to 34.

3.2 Loop Residues

Then, a comparsion of RMSD values, was performed for residues with high RMSF values (27-34) from plots (figures 6 and 7) and for residues that forms the loop, after visualizing the representative structure (representative.protein.pdb file) with Pymol. Additionally, RMSD values for some residues within the a-helix were added in the comparison. These residues are more stable than those in the loop, because they have a well-defined secondary structure due to the presence of a large number of

intermolecular interactions [14]. This results in lower RMSD values and more stabilized structures.

From "representative.protein.pdb" file:

- Residues of the loop from chain A for Native ROP are L29 to D32 and for chain B are L29 to D33.
- Residues of the loops for both chains of D30P are E28 to D32
- Residues of the loops for both chains of D30G are E29 to D32

Then RMSD plots were created, with xmgr, for loop' residues only of Native ROP, according to:

- Structure from PyMOL (29-32)
- Residues with high RMSF values (27-34)
- Seven residues located in the a-helix (20-26)

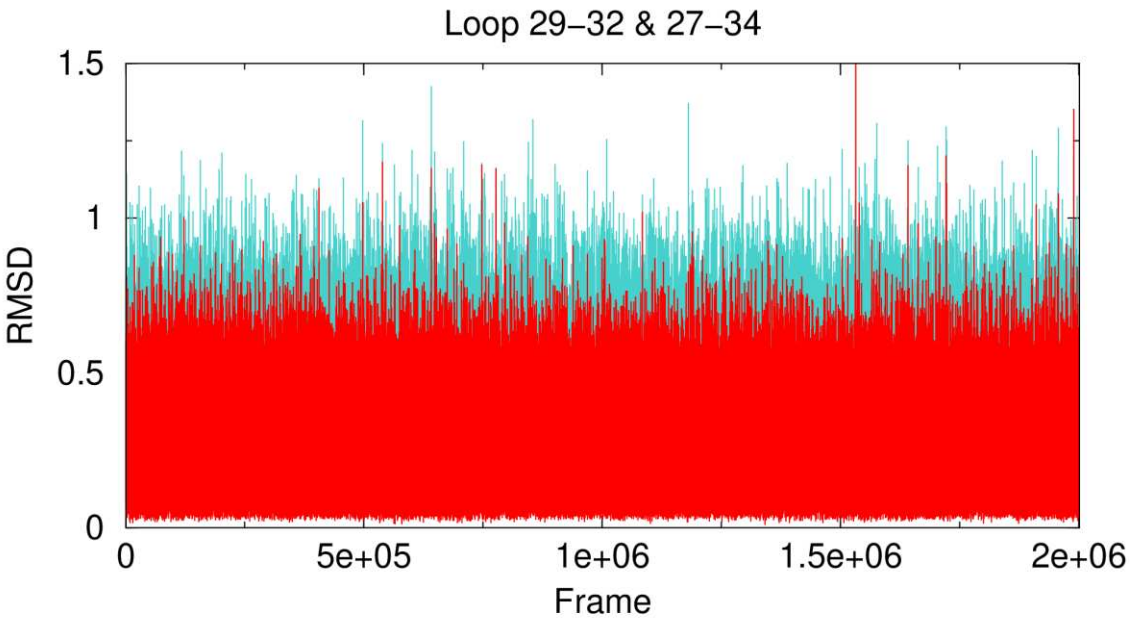


Figure 8: RMSD plot for residues 29 to 32 (red) from pymol, and residues 27 to 34 (turquoise) with the lower RMSF values.

The first plot compared RMSD for residues 29-32 and 27-34

The second plot included RMSD values of residues 20 to 26:

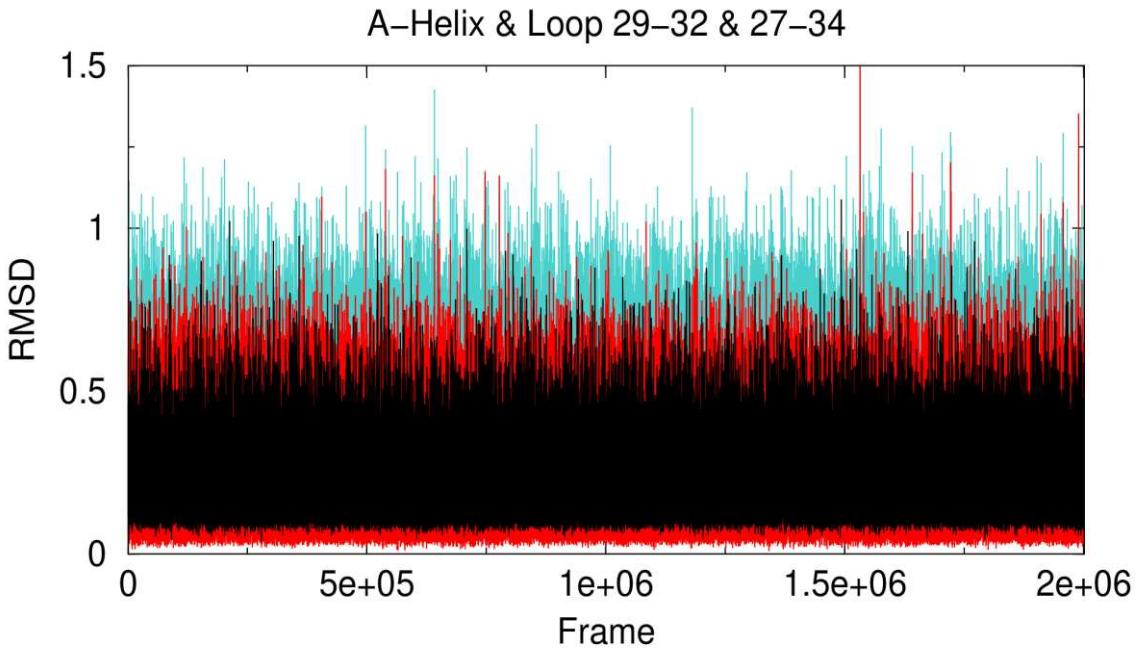


Figure 9: RMSD plot for residues 29 to 32 (red) from pymol, residues 27 to 34 (turquoise) with the lower RMSF value and residues located in the A-helix (black).

Residues 20 to 26 from a-helix (figure 9) seem to have lower RMSD values and smaller fluctuations compared to residues 29 to 32 and 27 to 34. Additionally, residues 29-32, do not only take higher RMSD values but their values are also unstable and more variable compared to those of a-helix. This confirms the fact that loop's residues present higher mobility.

However, RMSD values for 29-32, are lower than the respective values of residues 27-34 (figure 8), leading to the conclusion that turn region extends from residue 27 to 34.

Additionally, in order to correlate each variant's, turn-residues RMSF values, with their Tm, the average RMSF of residues 27-34 was calculated and compared.

Two tables were obtained, one for each turn:

A turn

Chain A	D30G	Native	D30P
Residue 27	0.64	0.89	0.88
Residue 28	0.82	1.01	1.08
Residue 29	0.68	0.68	0.78
Residue 30	0.74	0.82	1.00
Residue 31	0.74	0.67	0.80
Residue 32	0.81	0.87	1.39
Residue 33	0.75	1.17	1.23
Residue 34	0.71	0.86	0.86
Mean	**0.74**	**0.87**	**1.00**
Tm	**80.3 °C**	**68.7 °C**	**58.9 °C**

Table 1: RMSF values for residues 27-34 for each variant and their mean value, Turn A

B Turn

Chain B	D30G	Native	D30P
Residue 27	0.76	0.84	0.94
Residue 28	0.90	1.04	1.07
Residue 29	0.98	1.04	1.18
Residue 30	1.03	1.24	1.30
Residue 31	0.72	0.85	0.88
Residue 32	0.99	1.22	1.43
Residue 33	0.90	1.14	1.31
Residue 34	0.75	1.01	1.06
Mean	**0.88**	**1.05**	**1.15**
Tm	**80.3 °C**	**68.7 °C**	**58.9 °C**

Table 2: RMSF values for residues 27-34 for each variant and their mean value, Turn B

From the tables above, it is observed that RMSF values increase progressively from the D30G variant to the D30P variant. The same results can also be verified from their corresponding mean values. For D30G, the mean RMSF value equals to 0.74 and 0.88 for turn A and B, respectively. These values are lower than those of Native ROP (0.87 and 1.05), while D30P variant exhibits higher RMSF values than both of the other variants. D30P's mean values are 1.00 and 1.15.

3.3 RMSD plot & Histograms

The next step was the creation of RMSD plots and their corresponding histograms showing the distribution of RMSD values for:

(i) the full-length protein,

(ii) the full-length protein excluding tail residues and

(iii) the loop residues both combined and separately for each chain

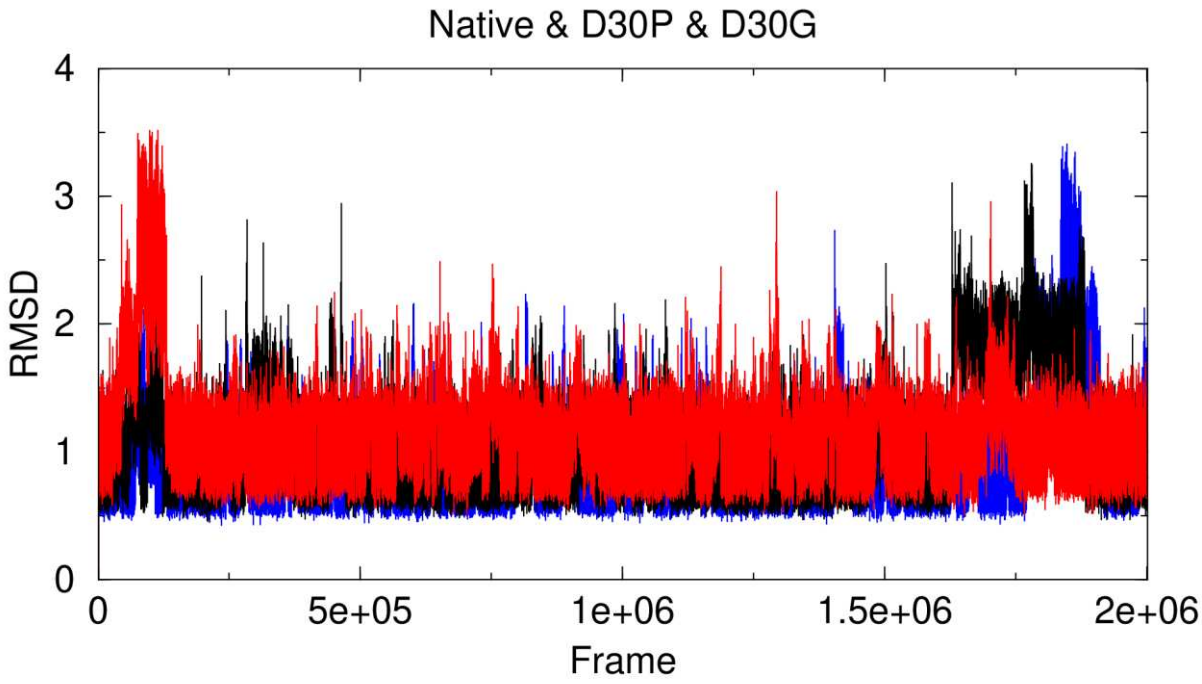


Figure 10: RMSD plot for the full-length Native ROP (black), D30P (red) and D30G (blue)

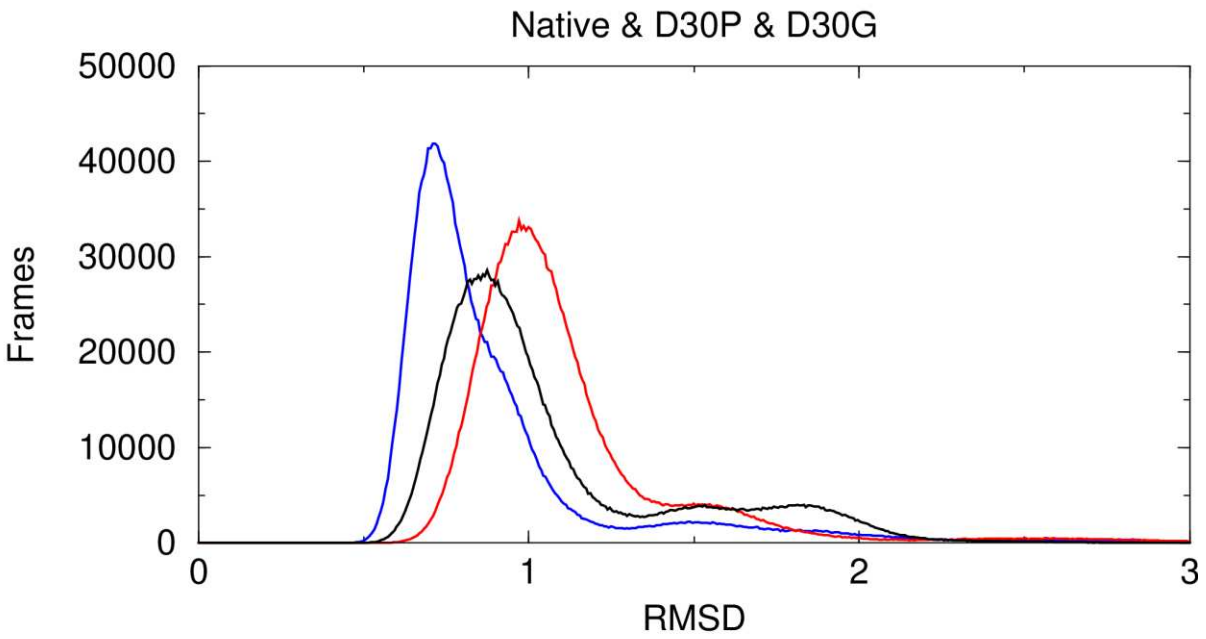


Figure 11: Histogram of RMSD values distribution for all residues of both chains of Native ROP (black), D30P (red) and D30G (blue), excluding residues 1 to 5 and 56 to 57.

For the following plot, several residues have been excluded. These are the residues that form the two terminals of the protein (1 to 5 for the N-terminal and 56 to 57 for the C-terminal). These residues take higher RMSF values according to figures 4 and 5, and they subsequently affect the mobility of the whole protein.

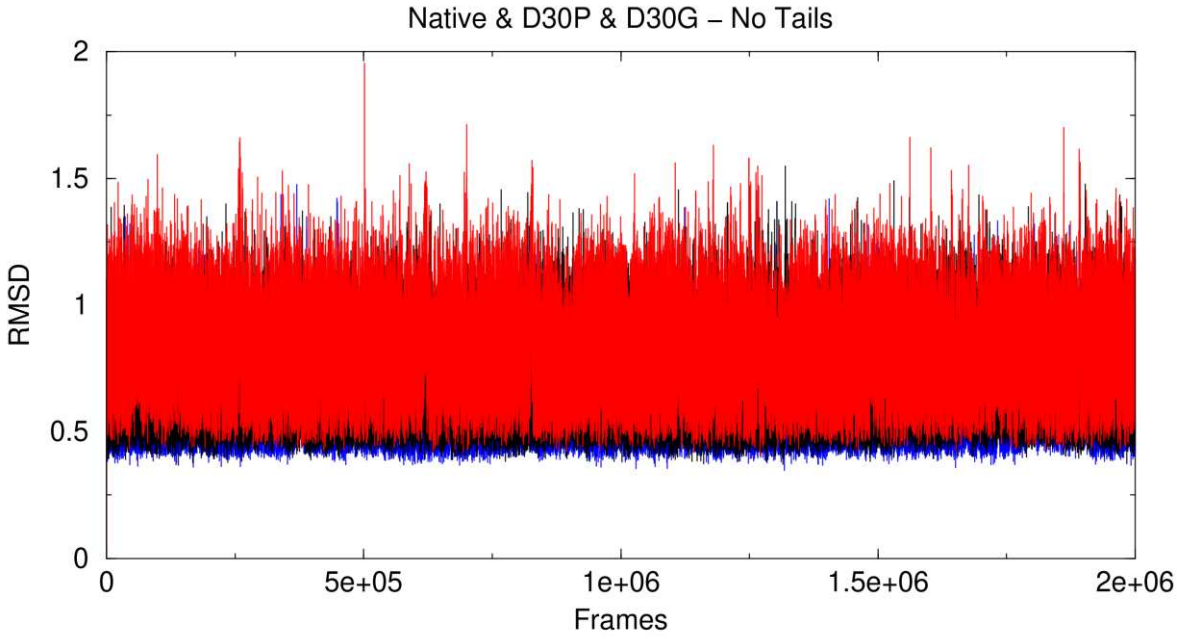


Figure 12: RMSD plot of native ROP (black), D30P (red) and D30G (blue), for all residues of both chains excluding residues 1 to 5 and 56 to 57.

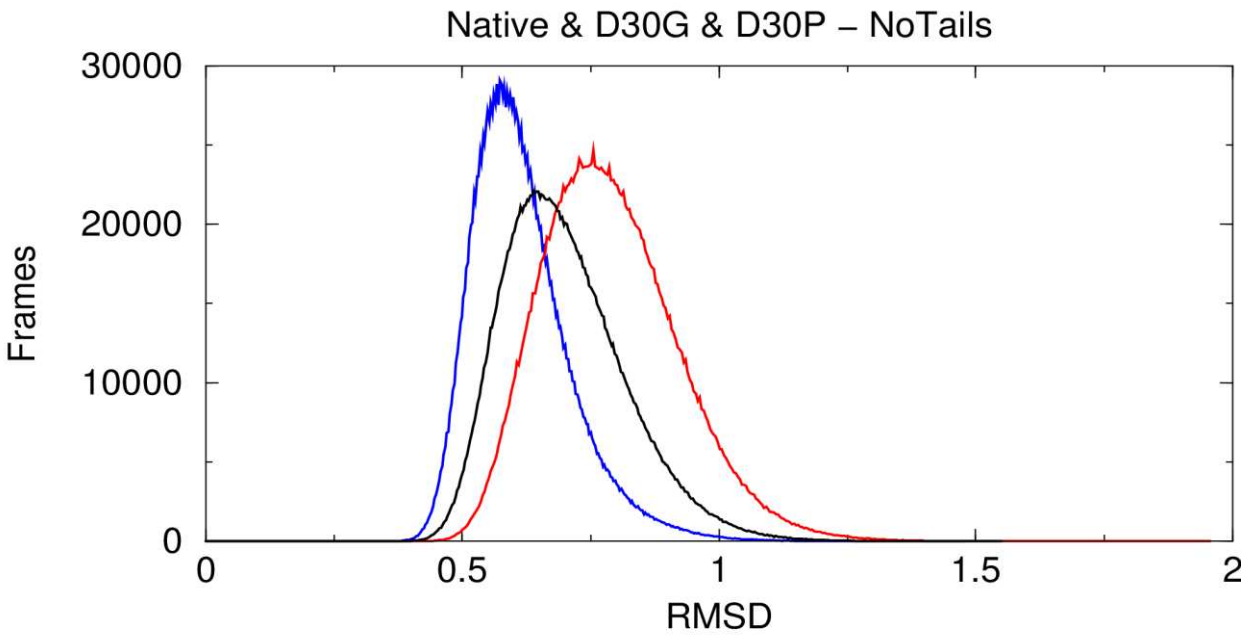


Figure 13: Histogram of RMSD values distribution for all residues of both chains of Native ROP (black), D30P (red) and D30G (blue), excluding residues 1 to 5 and 56 to 57.

Three more plots and histograms were created, for turn A and B (residues 27-34) combined and for both turns separately.

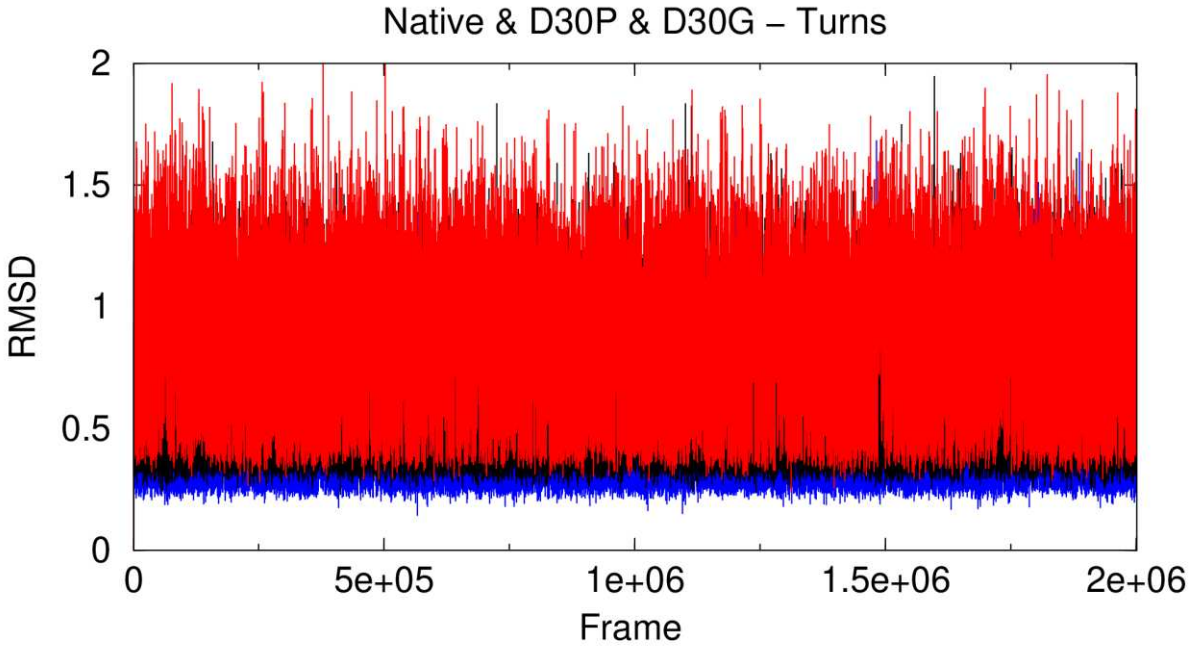


Figure 14: RMSD plot of native ROP (black), D30P (red) and D30G (blue), only for residues 27-34 of both chains.

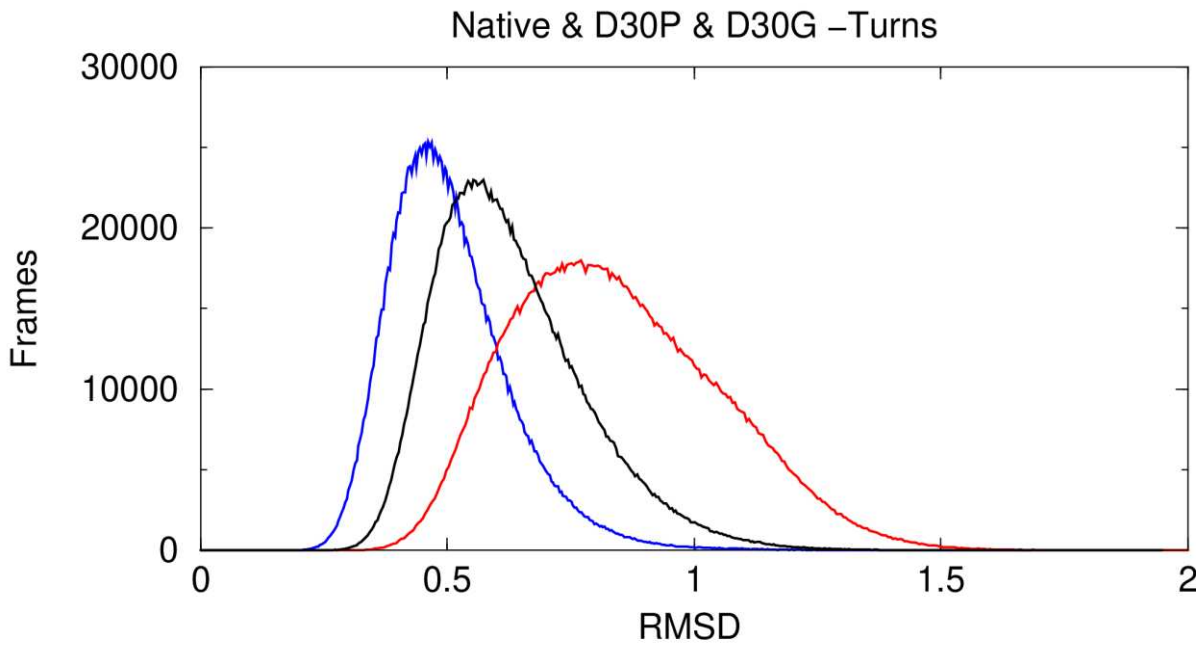


Figure 15: Histogram of RMSD values distribution for both turns (A and B) of Native ROP (black), D30P (red) and D30G (blue).

Residues 27-34 of A chain – A turn

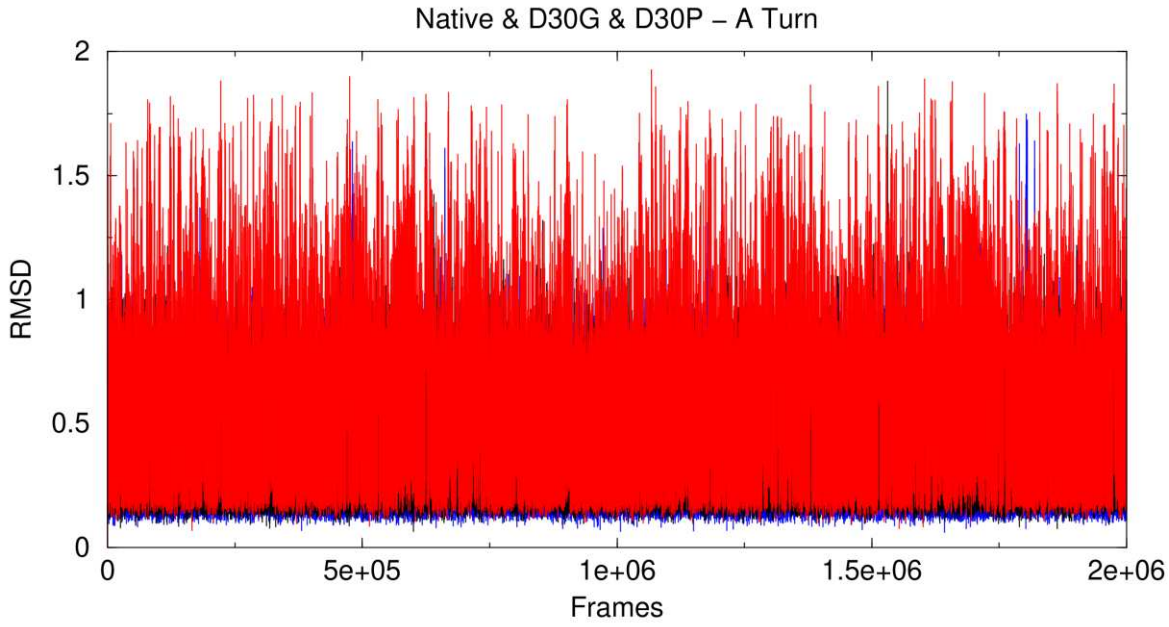


Figure 16: RMSD plot of native ROP (black), D30P (red) and D30G (blue), only for residues of the A turn (27-34).

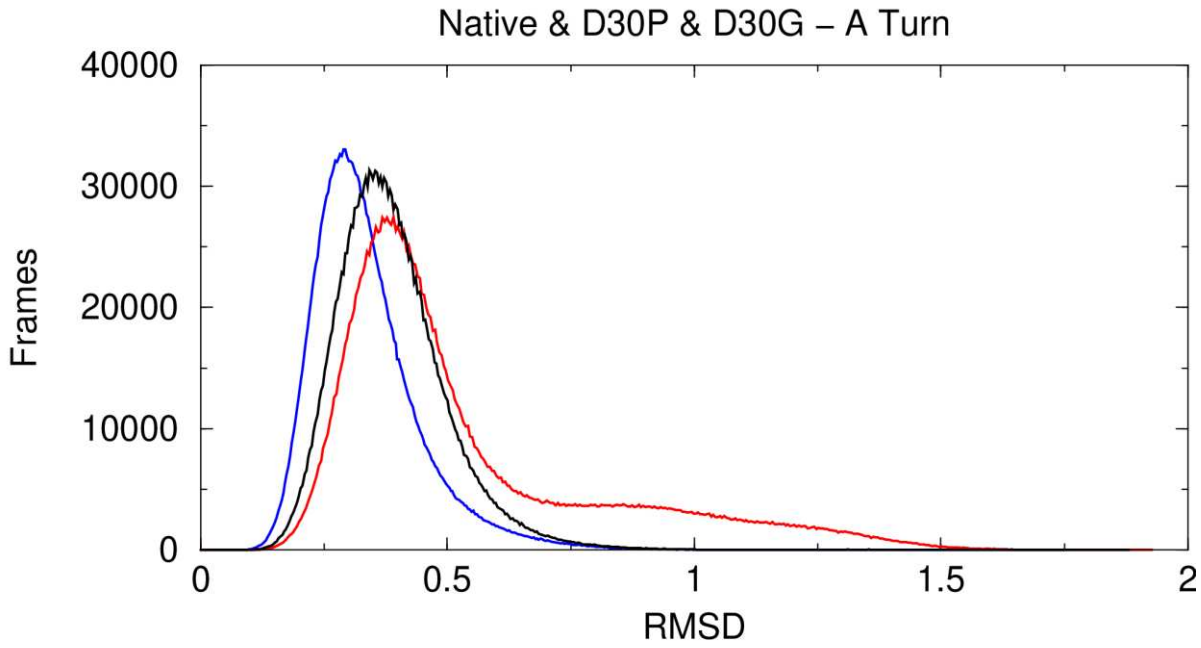


Figure 17: Histogram of RMSD values distribution for A turn of Native ROP (black), D30P (red) and D30G (blue).

Residues 27-34 of B chain – B turn

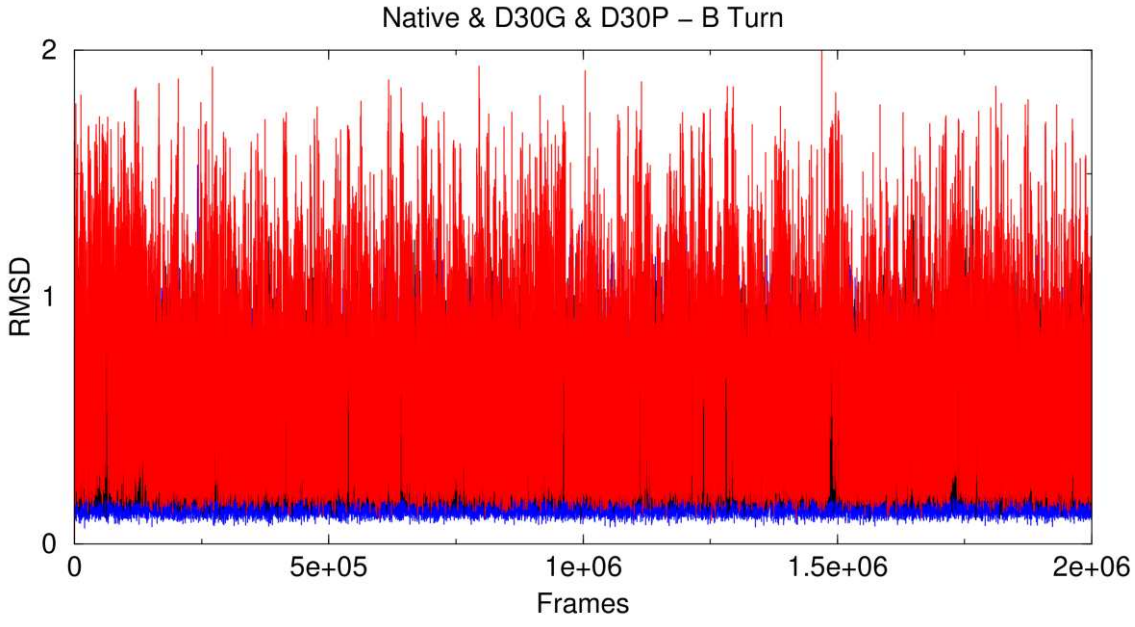


Figure 18: RMSD plot of native ROP (black), D30P (red) and D30G (blue), only for residues of the B turn (27-34).

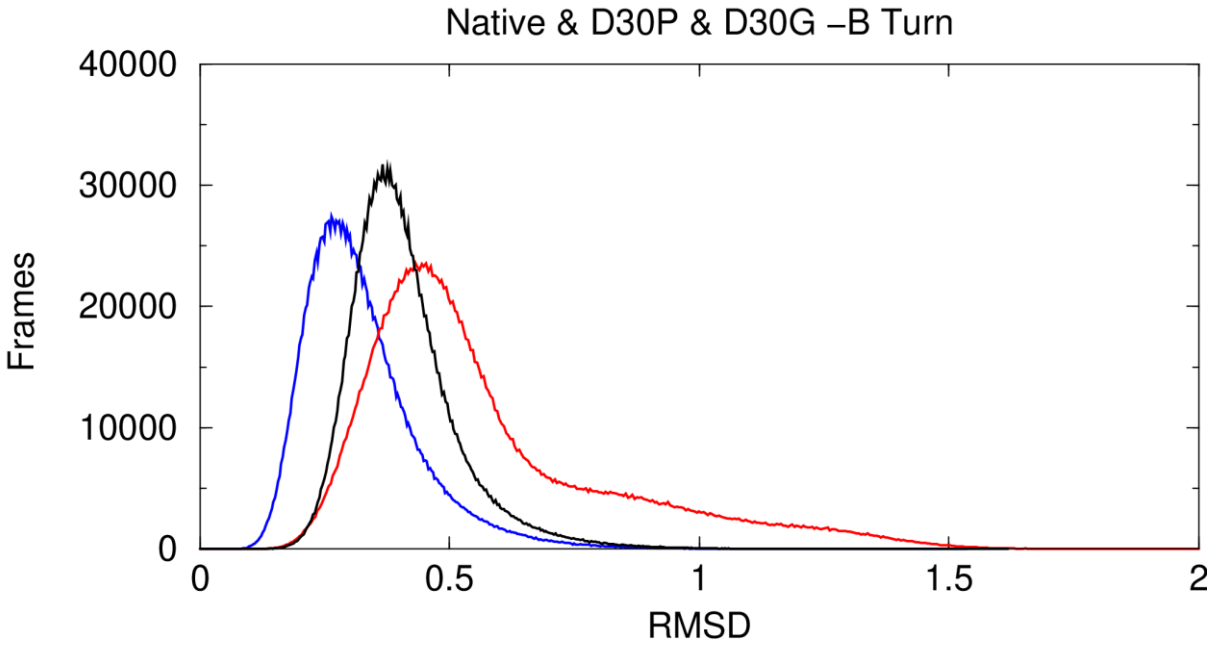


Figure 19: Histogram of RMSD values distribution for B turn of Native ROP (black), D30P (red) and D30G (blue).

In order to show with greater detail, any minor difference in their RMSD values, two additional plots were created for Native and D30G. The first included residues from A turn and the second these residues forming B turn.

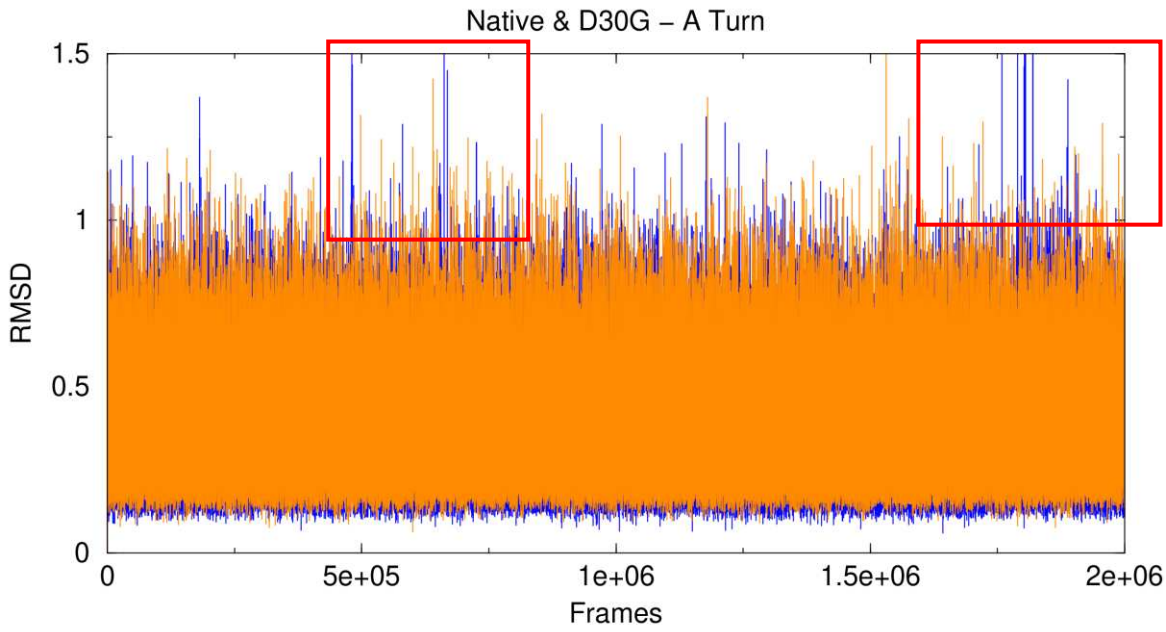


Figure 20: RMSD plot of Native ROP (orange) and D30G (blue) for residues 27 to 34 from loop A.

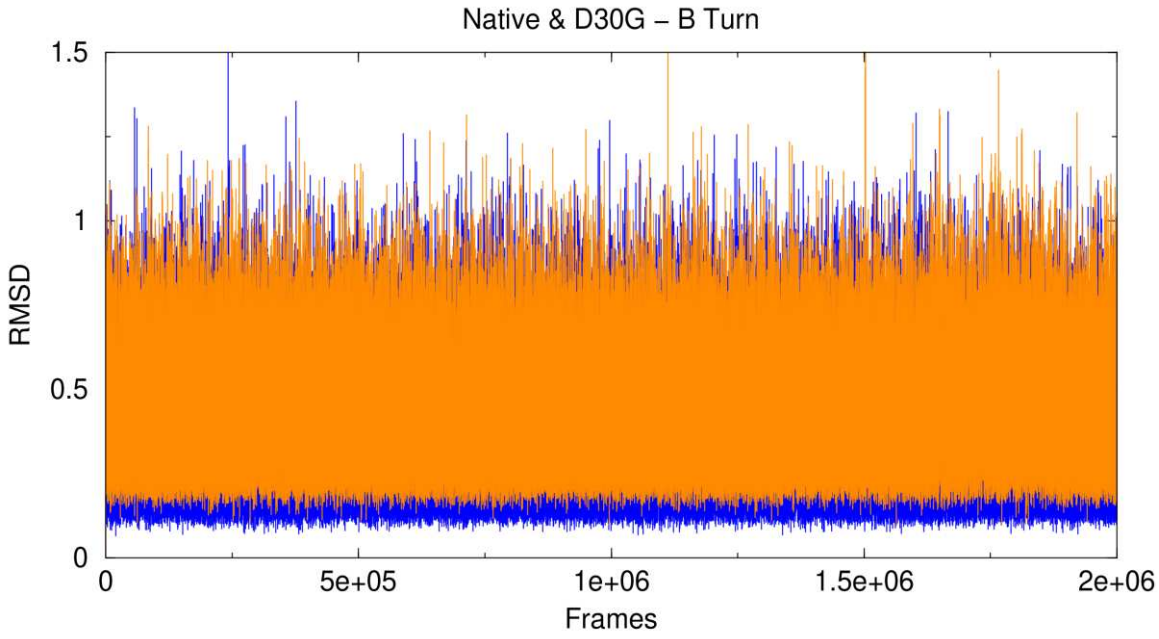


Figure 21: RMSD plot of Native ROP (orange) and D30G (blue) for residues 27 to 34 from loop B.

According to figures 20 and 21, RMSD values of D30G, are generally lower than those of Native ROP. Although, there are some values (table 2), for D30G's A turn marked with red outline, that deviate significantly, and possible corresponding to frames whose conformation differs from their average.

3.4 Statistical Analysis with R

In order to interpreter the results from the statistical analysis, it's important to understand the meaning of some main parameters.

The mean value refers to the average RMSD value of all frames of each structure, while the standard deviation shows the range within these values are observed and how much they diverge from the mean [27]. These two parameters are essential for understanding the dynamic of the system. The

smaller these values are, the more stable the data (RMSD values), and consequently, the system, appeared to be.

On the other hand, skeweness indicates the tendency of some values to deviate asymmetrically from the mean. That means that higher values imply that some RMSD deviate significantly from the mean and the corresponding frame have slightly different conformation. This suggests that the system is less stable.

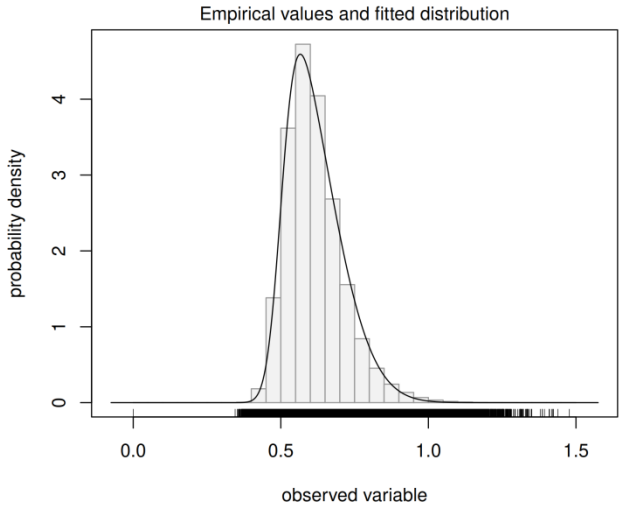
After the analysis with R, using library sn, the following tables and diagrams were obtained:

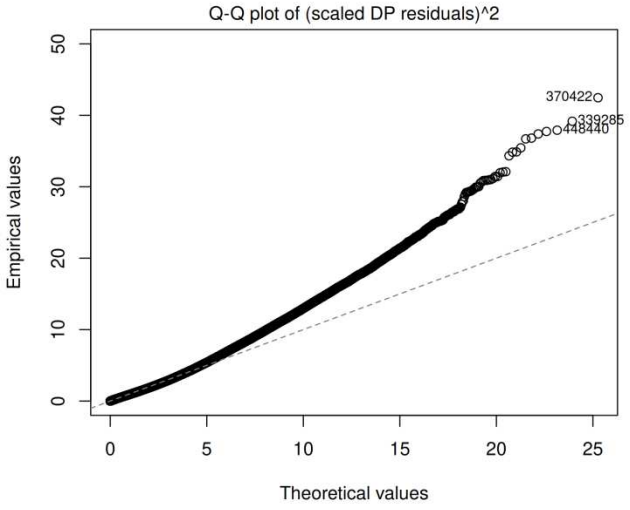
Full-length (excluding tails) – Skewed Distribution

	D30G	**Native**	**D30P**
Mean	6.170 Å	6.984 Å	7.887 Å
Standard Deviation	0.09544 Å	0.1192 Å	0.1327 Å
Skewness	0.75949	0.6948	0.5728
Log-likelihood	1976937	1505558	1258896
Tm	80.3 ^{o}C	68.7 ^{o}C	58.9 ^{o}C

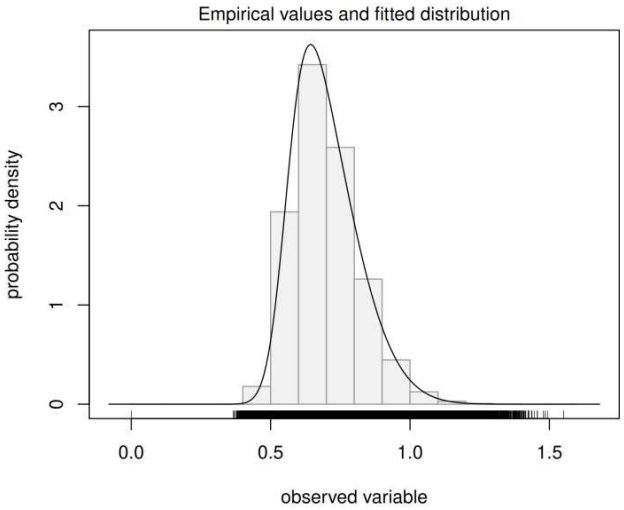
Table 3: Mean, standard deviation, skewness (gamma value) and log-likelihood values for D30G, Native ROP and D30P (full-length excluding tails) –Skewed Distribution.

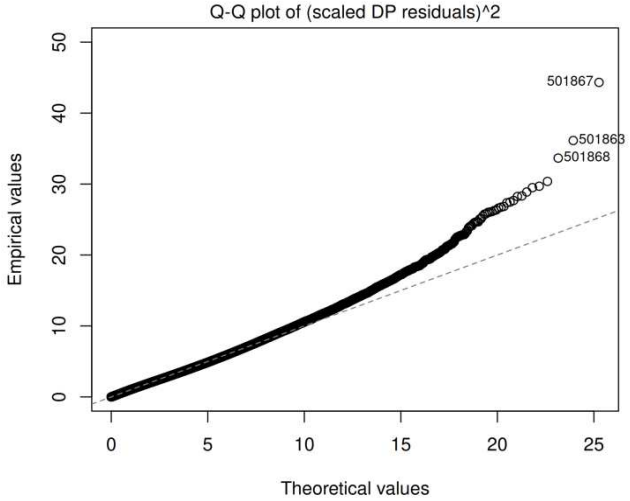
(a) D30G



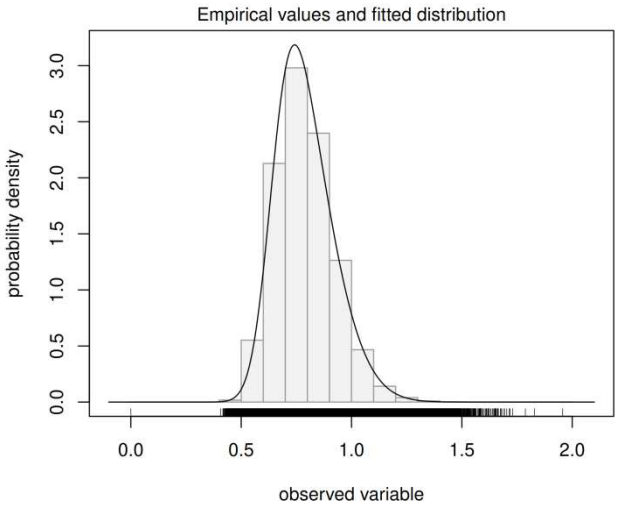


(b) Native





(c) D30P



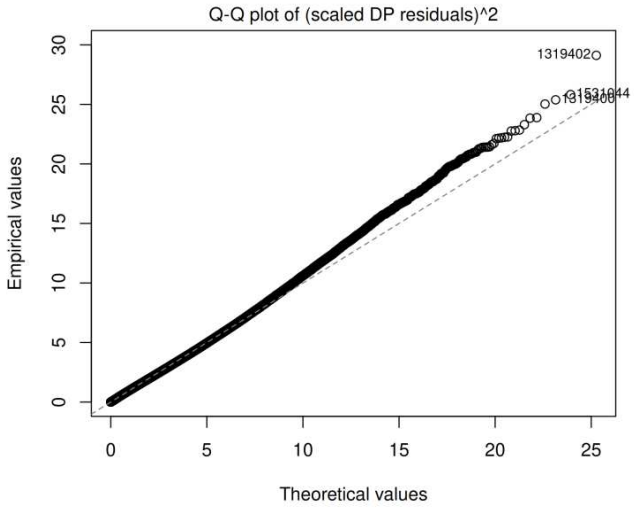


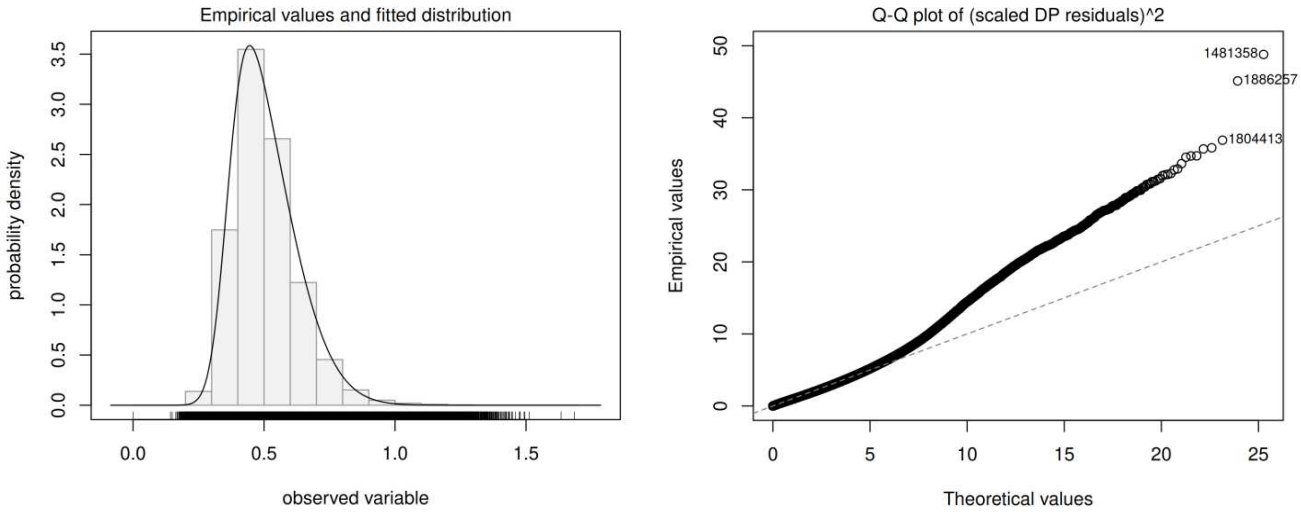
Figure 22: Histograms and Q_Q plots for (a) D30G, (b) Native, and (c) D30P created using a skewed distribution model for all proteins' residues, excluding the N- & C- terminal residues.

Loops Residues – Skewed Distribution

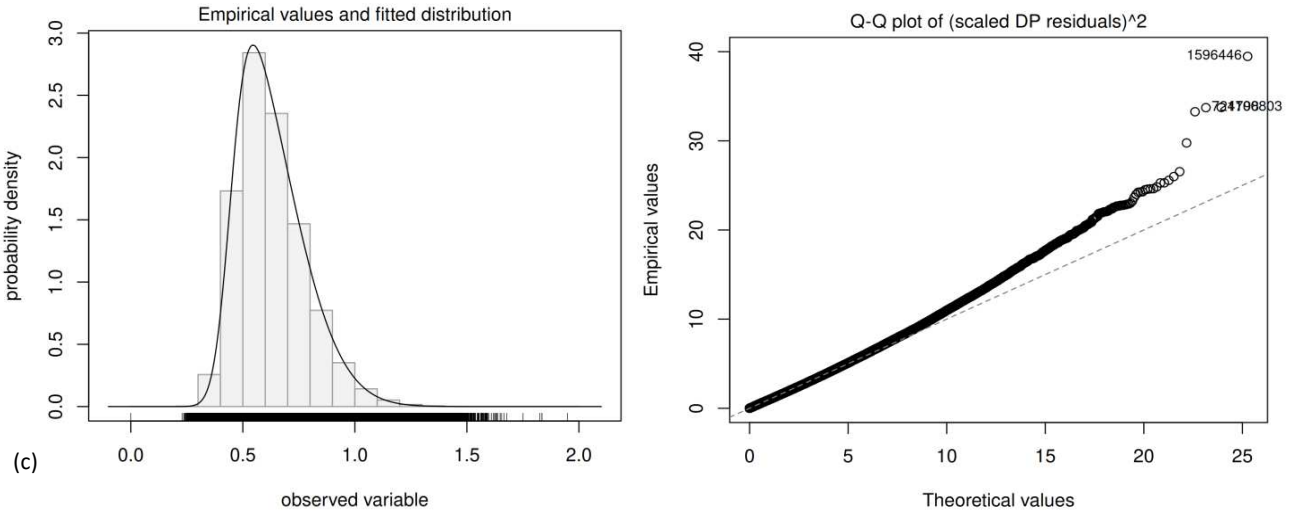
	D30G	Native	D30P
Mean	5.064 Å	6.302 Å	8.411 Å
Standard Deviation	0.1216 Å	0.1517 Å	0.2113 Å
Skewness	0.7340	0.7793	0.6286
Log-likelihood	1483214	1055923	339393
Tm	80.3 ^{o}C	68.7 ^{o}C	58.9 ^{o}C

Table 4: Mean, standard deviation, skeweness (gamma value) and log-likelihood values for D30G, Native ROP and D30P (loop residues) –Skewed Distribution

(a) D30G



(b) Native



(c)

(c) D30P

(b)

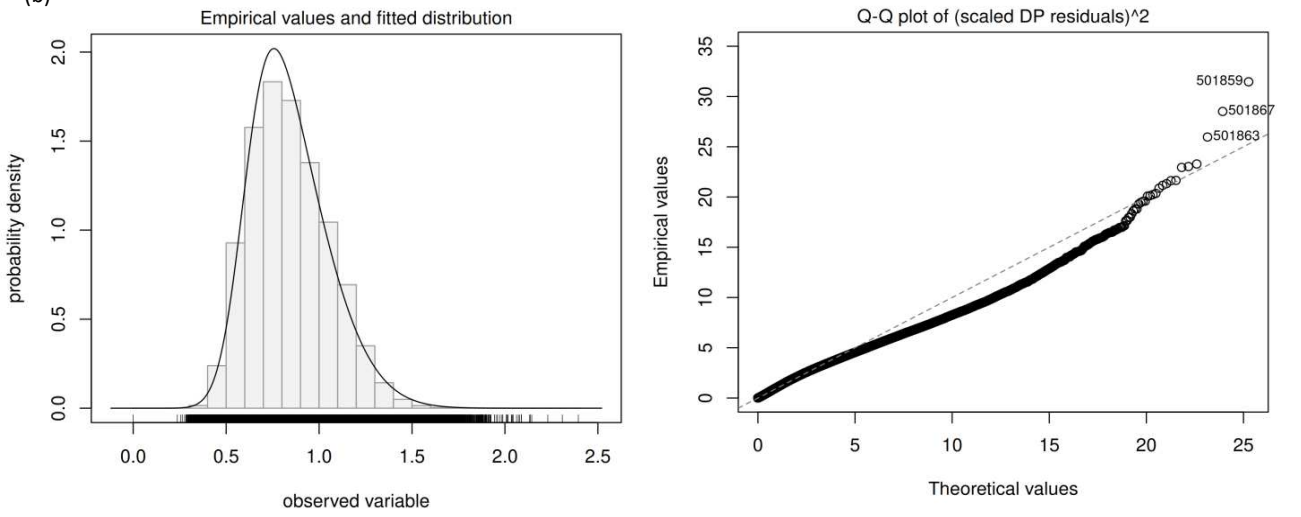


Figure 23: : Histograms and Q_Q plots for (a) D30G, (b) Native, and (c) D30P created using a skewed distribution model for residues of both loops.

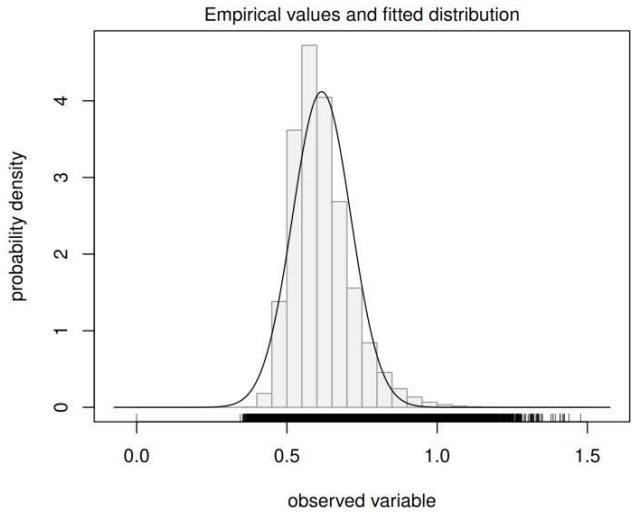
The second statistical model that was used, was the symmetrical distribution model. The results derived from this second analysis include, Q_Q plots and histograms for both the full-length structures and for the turns residues only, as well as tables containing the mean value, standard deviation and log-likelihood.

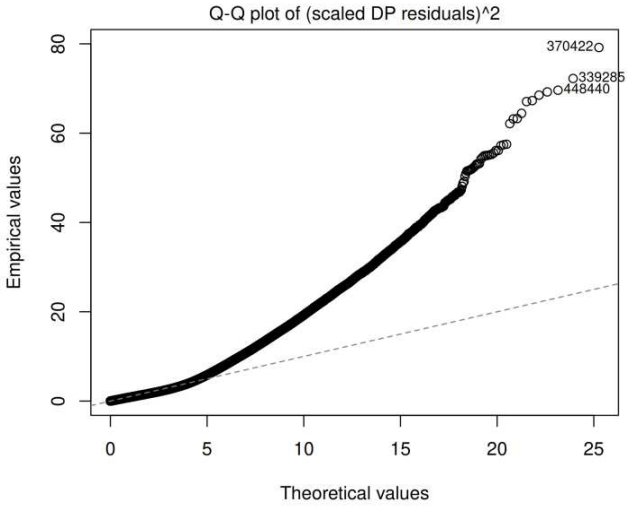
Full-length structure (excluding tails) – Symmetrical Distribution

	D30G	**Native**	**D30P**
Mean	6.158 Å	6.985 Å	7.889 Å
Standard Deviation	0.09684 Å	0.1191 Å	0.1324 Å
Log-likelihood	1831509	1417555	1205300
Tm	80.3 ^{o}C	68.7 ^{o}C	58.9 ^{o}C

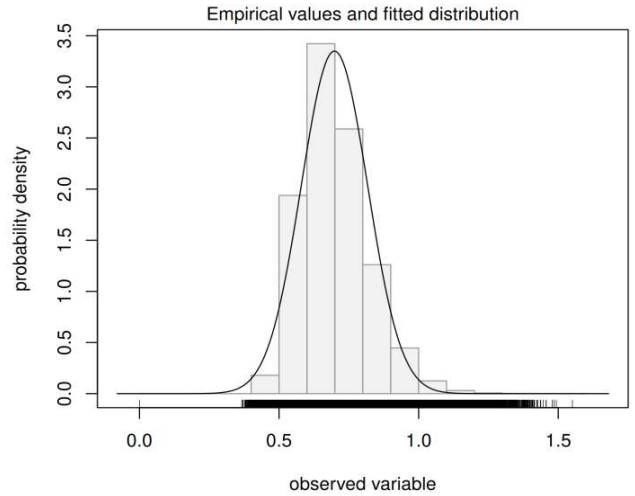
Table 5: Mean, standard deviation and log-likelihood values for D30G, Native ROP and D30P (full-length excluding tails) –Symmetrical Distribution

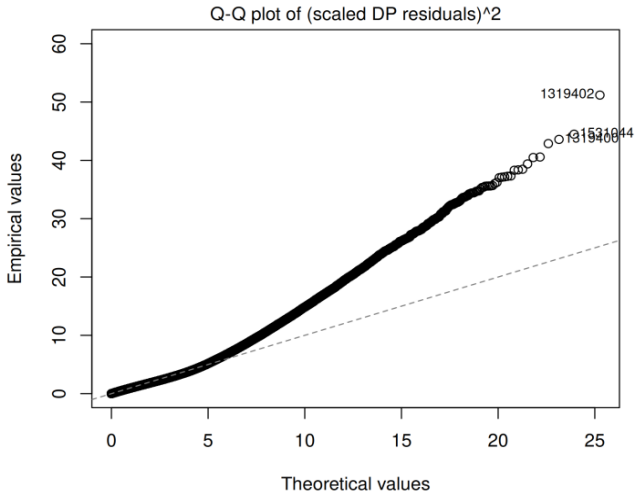
(a) D30G



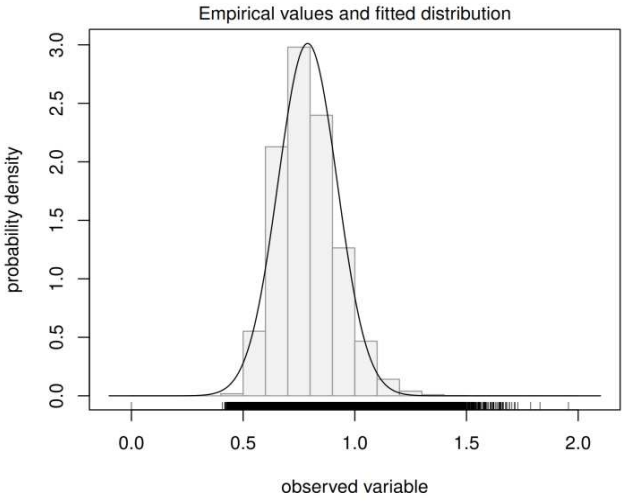


(b) Native





(c) D30P



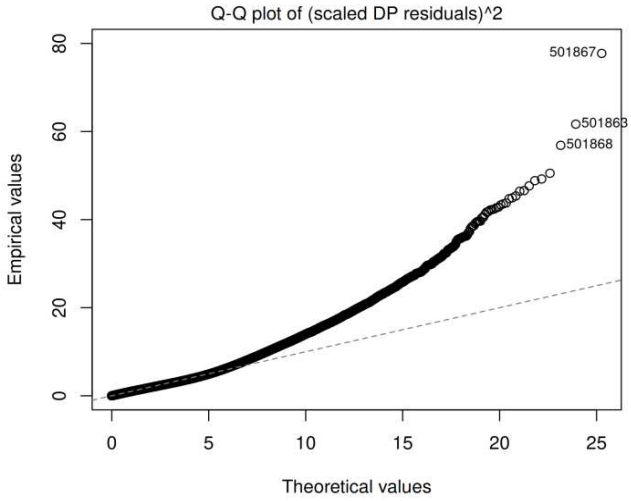


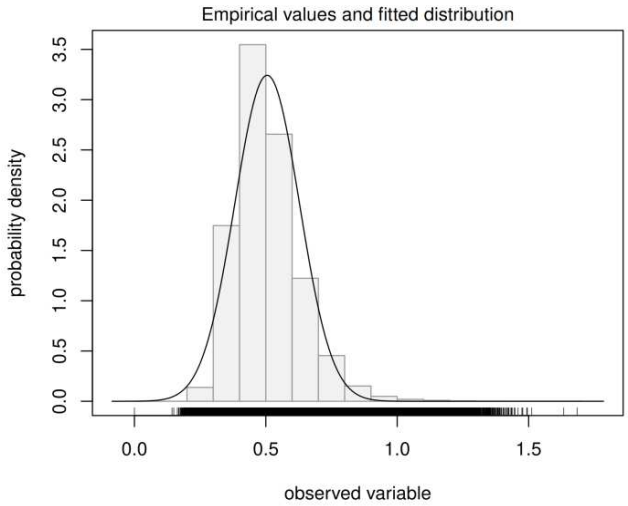
Figure 24: Histograms and Q_Q plots for (a) D30G, (b) Native, and (c) D30P created using a symmetrical distribution model for all proteins' residues, excluding the N- & C- terminal residues.

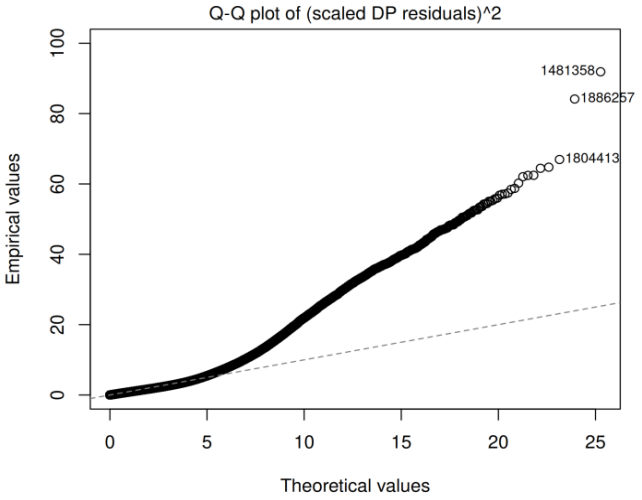
Loop Residues – Symmetrical Distribution

	D30G	Native	D30P
Mean	5.051 Å	6.297 Å	8.431 Å
Standard Deviation	0.123 Å	0.1523 Å	0.209 Å
Log-likelihood	1352562	926421	292745
Tm	80.3 ^{o}C	68.7 ^{o}C	58.9 ^{o}C

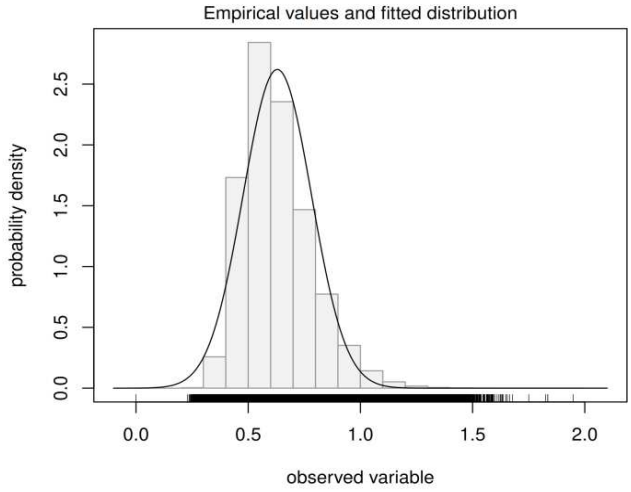
Table 6: Mean, standard deviation and log-likelihood values for D30G, Native ROP and D30P (loop residues) – Symmetrical Distribution

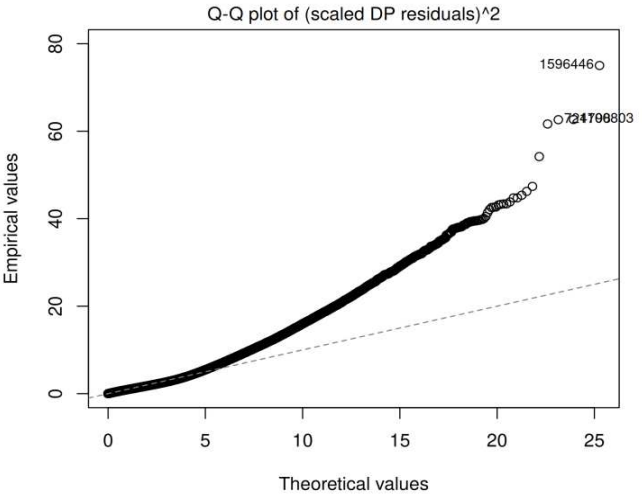
(a) D30G



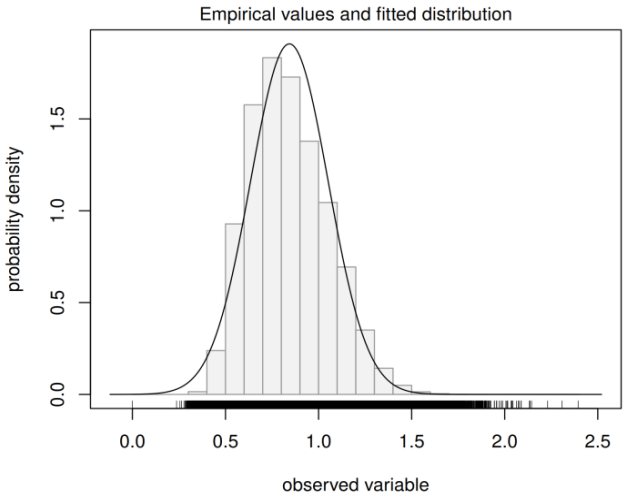


(b) Native





(c) D30P



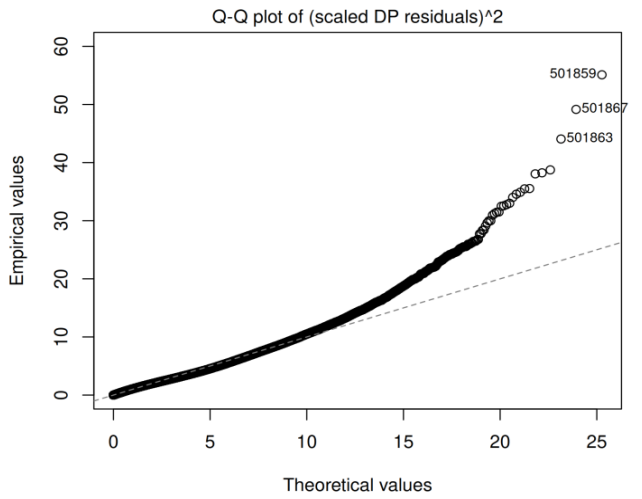


Figure 25: Histograms and Q_Q plots for (a) D30G, (b) Native, and (c) D30P created using a symmetrical distribution model for residues of both loops.

Log-likelihood (LL) is an important statistical parameter used to determine the most representative statistical model. As the LL value increases, the data fit better with model being used.

The variation between log-likelihood values from the symmetrical distribution model to the skewed distribution model, are presented below:

Full-length (excluding tail)	Log-Likelihood (Symmetrical)	Log-Likelihood (Skewed)
D30G	1831509	1976937
Native	1417555	1505558
D30P	1205300	1258896

Table 7: Log-likelihood values from the symmetrical to the skewed distribution model, for the full-length structures.

Loops/Turns	Log-Likelihood (Symmetrical)	Log-Likelihood (Skewed)
D30G	1352562	1483214
Native	926421	1055923
D30P	292745	339393

Table 8: Log-likelihood values from the symmetrical to the skewed distribution model, for the residues of the loops.

The increase in the Log-Likelihood value from symmetrical to skewed distribution (tables 7 and 8) demonstrates that the skewed model interprets the data, derived from molecular dynamic simulations, with greater accuracy. This leads to the conclusion that all of the three proteins, take RMSD values that tend to deviate from the mean, forming a distribution curve that leans to one side.

According to the skewed distribution model, both mean and standard deviation values increase from D30G to D30P (tables 3 and 4). This indicates that D30P is less stable and its RMSD values spreads across a wider area on the plot. On the contrary, D30G's mean and standard deviation values are lower indicating a smaller range. Therefore D30G seems to be more stable.

Skewness (gamma value) also increases from D30G to D30P, meaning that there are some RMSD values, and consequently some frames, which are distant from their structure's average conformation. This suggests that despite the fact that D30G appears to be more stable, there are some time points during the simulation when it changes its conformation and the corresponding frames deviate significantly from the average (higher RMSD).

The only exception in that case, is gamma value for Native ROP loop residues (table 2), which is slightly higher from the corresponding value of D30G but it's still lower from D30P's.

3.5 Dihedral PCA (dPCA)

Eigenvalues & Eigenvectors

Eigenvectors and eigenvalues are concepts from linear algebra that describe the direction and the magnitude, respectively, of each principal component derived from PCA analysis.

In molecular dynamic simulations and protein stability analysis, an eigenvector represents the direction of the movement of every principal component in space. The corresponding eigenvalue indicates the size of this movement and reflects its contribution to the system's general stability. The first principal component (PC1) is responsible for the greatest variability on the data and it also accounts for the biggest impact on the mobility and stability of the protein [28].

Eigenvalues for the full-length structures present below:

Eigenvalues (excluding tail residues)

	D30G (80.3^{o}C)	**Native** (68.7^{o}C)	**D30P** (58.9^{o}C)
PC1	0.2147460729	0.2221473008	0.6239398122
PC2	0.2141319215	0.2183327079	0.5769296288
PC3	0.1289790869	0.1271845251	0.4446907341

Table 9: Eigenvalues derived from Dihedral PCA for each one of the first three principal components, for D30G, Native ROP and D30P for full-length proteins excluding residues 1-5 and 56-57 ((step = 10).

Tables including eigenvalues for each turn separately are presented below:

Eigenvalues (A Turn):

	D30G (80.3°C)	**Native** (68.7°C)	**D30P** (58.9°C)
PC1	0.1287021637	0.1265375614	0.5635805726
PC2	0.1165211126	0.1021214575	0.4116775393
PC3	0.0878125727	0.0845472142	0.2762321830

Table 10: Eigenvalues derived from Dihedral PCA for each one of the first three principal components, for D30G, Native ROP and D30P for A turn (step = 1).

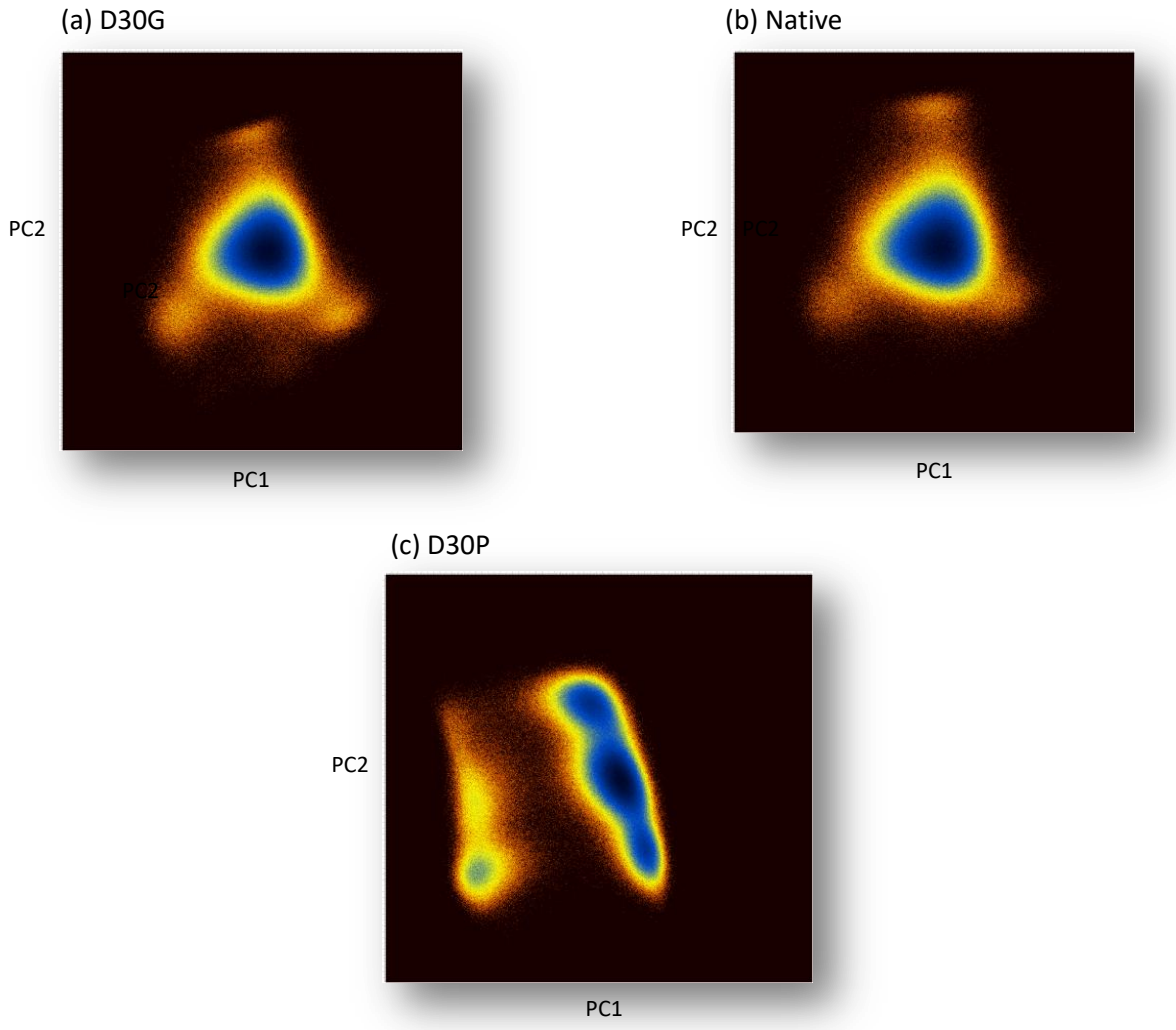


Figure 26: PC1-PC2 plot for A turn's residues of (a) D30G, (b) Native and (c) D30P.

Eigenvalues (B Turn):

	D30G (80.3°C)	**Native** (68.7°C)	**D30P** (58.9°C)
PC1	0.1253450662	0.1233271882	0.5808151364
PC2	0.1150820628	0.1109051555	0.4396424890
PC3	0.0876375511	0.0841571540	0.2766299248

Table 11: Eigenvalues derived from Dihedral PCA for each one of the first three principal components, for D30G, Native ROP and D30P for B turn (step = 1).

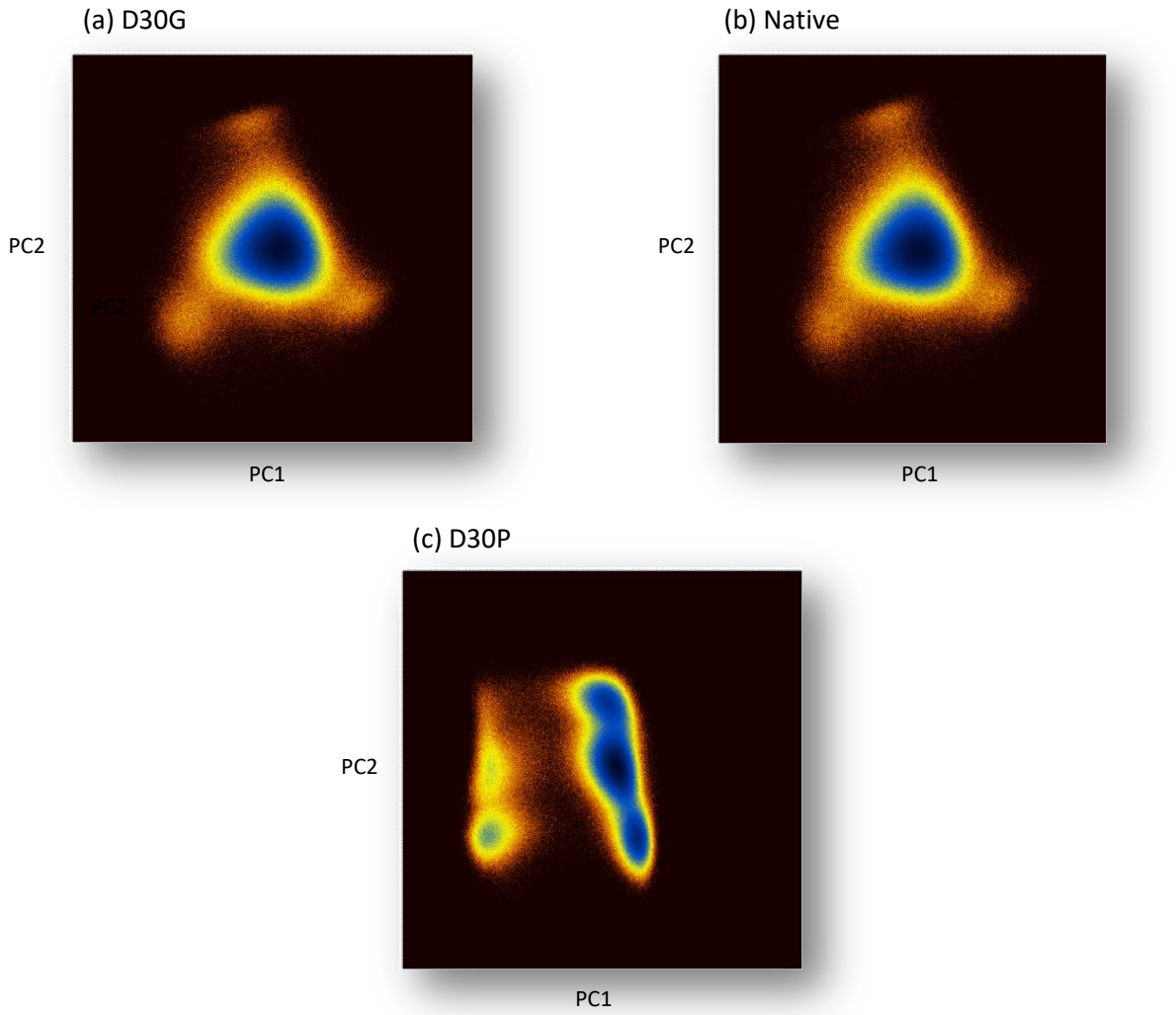


Figure 27: PC1-PC2 plot for B turn's residues of (a) D30G, (b) Native and (c) D30P.

According to tables 10 and 11, the eigenvalues for PC1, PC2 and PC3 of D30P are consistently higher, which confirms its increased mobility. On the other hand, minor differences can be observed between D30G's and Native ROP's eigenvalues. The value for the first principal component (PC1) of D30G equals 0.1287021637, and it's slightly higher than the respective eigenvalue of Native ROP which is 0.1265375614. The eigenvalues for PC2 and PC3 follow the same pattern for both A and B turn residues (table 10 and 11), suggesting a greater flexibility of D30G's loop, considering only dihedral angles.

However, when dPCA performed on the full-length proteins excluding tails residues (table 9), a different conclusion emerged. The PC1 eigenvalue for D30P (0.6239398122) remains higher from those of the other two variants. But in this case, the PC1 eigenvalue of Native ROP (0.2221473008) is higher than the respective value of D30G (0.2147460729). The same results apply for PC2 eigenvalues. Thus, although D30G shows higher internal mobility, it appears generally more stable.

Regarding the PC1-PC2 plots (figures 26 and 27) turns A and B of Native ROP and D30G appear to be more stable during the simulation. The differences between these two variants can be barely observed in the plots. Additionally, the plots confirm that both structures are very stable. On the other hand, D30P again proves that it's less stable, as its plot reveals that multiple distinct conformations exist.

In order to compare the main conformation of the loop region of each variant, the representative structures of the first clusters (both loop A and B) were visualized, as well as their aligned versions, where each turn was superimposed with the corresponding turns from the other two variants.

D30G

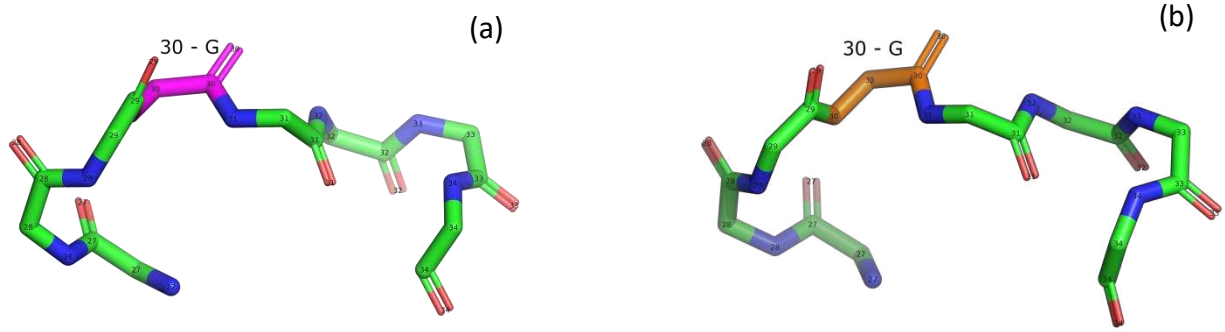


Figure 28: Representative structure of the first cluster generated by Pymol, from data obtained through dPCA analysis, for (a) turn A and (b) turn B (residues 27-34) of D30G.

Native

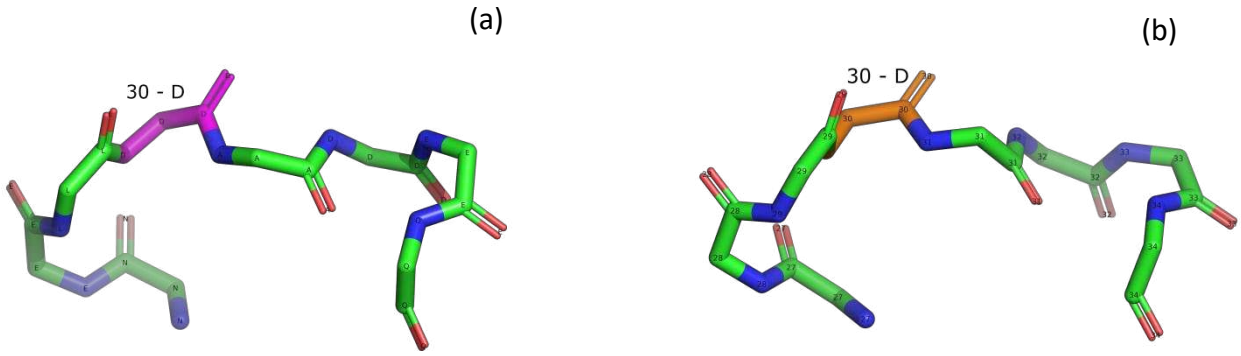


Figure 29: Representative structure of the first cluster generated by Pymol, from data obtained through dPCA analysis, for (a) turn A and (b) turn B (residues 27-34) of Native ROP.

D30P

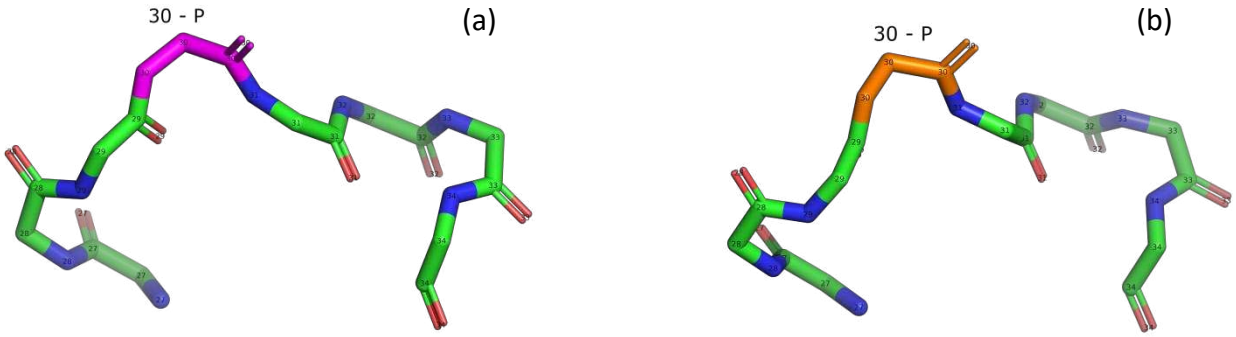


Figure 30: Representative structure of the first cluster generated by Pymol, from data obtained through dPCA analysis, for (a) turn A and (b) turn B (residues 27-34) of D30P.

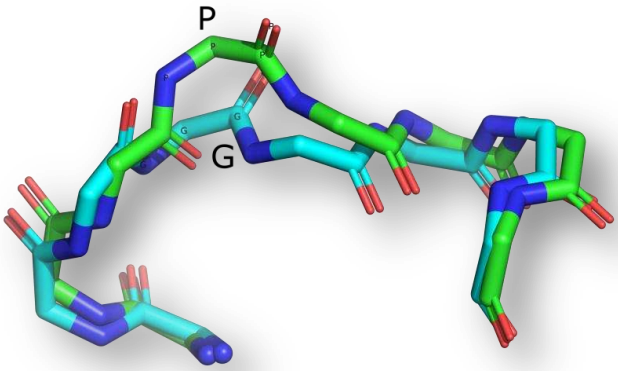


Figure 31: Structures of A-turn (residues 27 to 34) from the first cluster of D30P (green) and D30G (blue), generated using Pymol. Residues 27 and 34 were aligned and additionally residue 30 is shown using the one letter code.

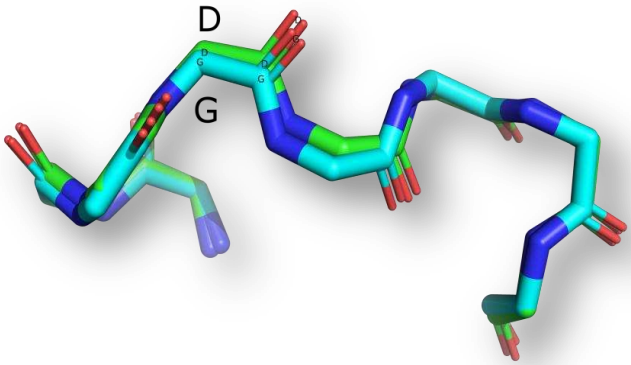


Figure 32: Structures of A-turn (residues 27 to 34) from the first cluster of Native ROP (green) and D30G (blue), generated using Pymol. Residues 27 and 34 were aligned and additionally residue 30 is shown using the one letter code.

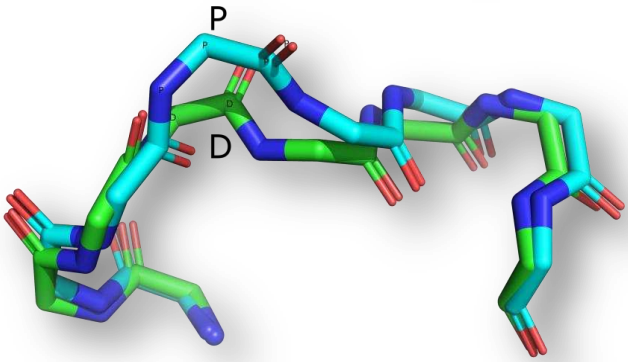


Figure 33: Structures of A-turn (residues 27 to 34) from the first cluster of Native ROP (green) and D30P (blue), generated using Pymol. Residues 27 and 34 were aligned and additionally residue 30 is shown using the one letter code.

From figures 28 to 33, it seems that D30P adopts a noticeably different and a more "open" conformation at loop's region. This may be explained by the presence of proline at position 30. On the other hand, D30G and Native ROP adopt similar and more compact conformations considering the loop region. This observation is also supported by the visualization of the loop region as well as from the extremely higher RMSF and RMSD values and their eigenvalues.

3.6 Cartesian PCA (cPCA)

In order to detect any minor movement/interaction between each turn and the opposing residues from the a-helices, two cPCA analyses were performed. More specifically residues 27-34 were used for each turn and residues 6 to 13 and 48 to 55 for the opposing a-helix.

The tables below contain the eigenvalues for the first five principal components of each variant and the number frames of each cluster. Additionally, the PC1-PC2 and PC2-PC3 plots were also obtained.

Turn (A Chain) + a-helix (B Chain)

Eigenvalues

	D30G	NATIVE	D30P
PC1	5.9572129250	6.6352500916	6.3596129417
PC2	0.9737140536	0.9198949933	1.7323147058
PC3	0.5975292921	0.6345487833	0.9223011732
PC4	0.4683012068	0.4514805079	0.7208867669
PC5	0.3745175004	0.3854705691	0.4575453103

Table 12: Eigenvalues derived from Cartesian PCA for each one of the first 5 principal components of D30G, Native and D30P, for residues 27-34 of A loops and residues 6-13 and 48-55 of the a – helix of B chain (step = 1).

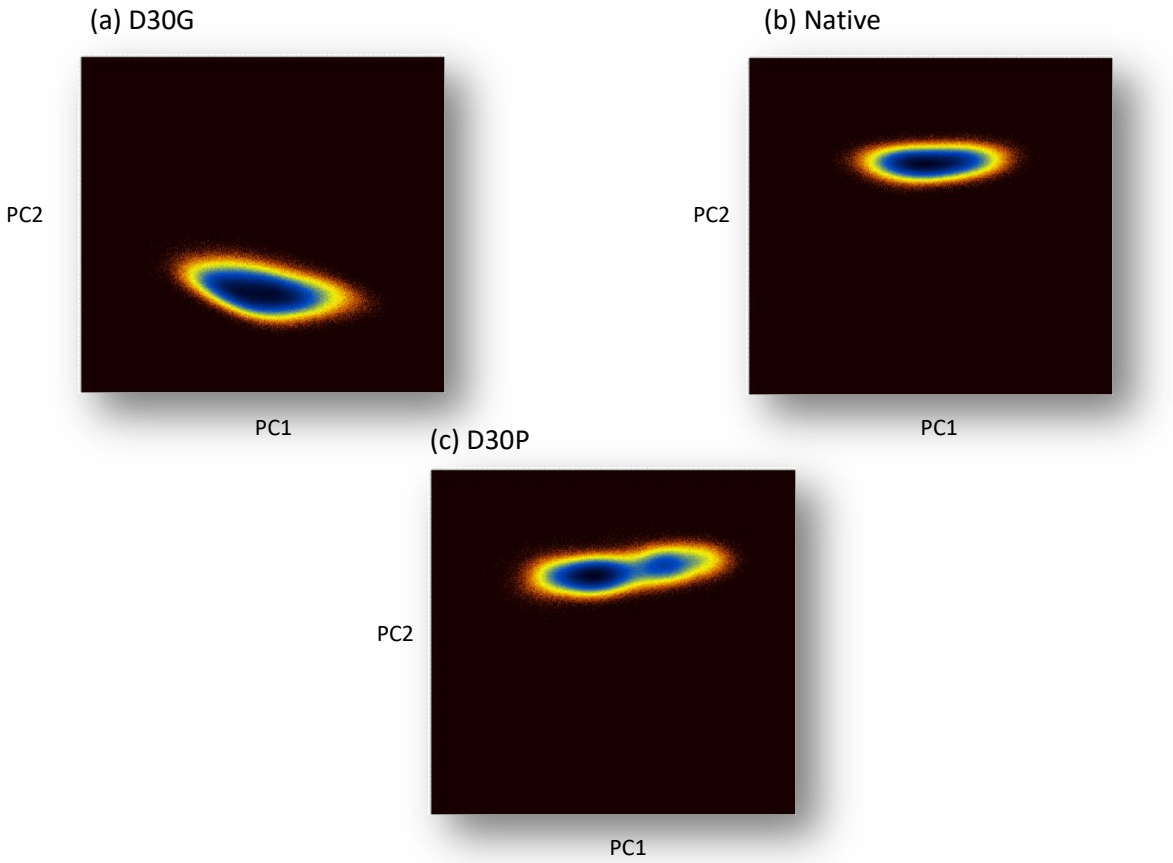


Figure 34: PC1-PC2 plots for residues 27-34 from the A turns and residues 6-13 and 48-55 from a-helix of B chain of (a) D30G, (b) Native and (c) D30P.

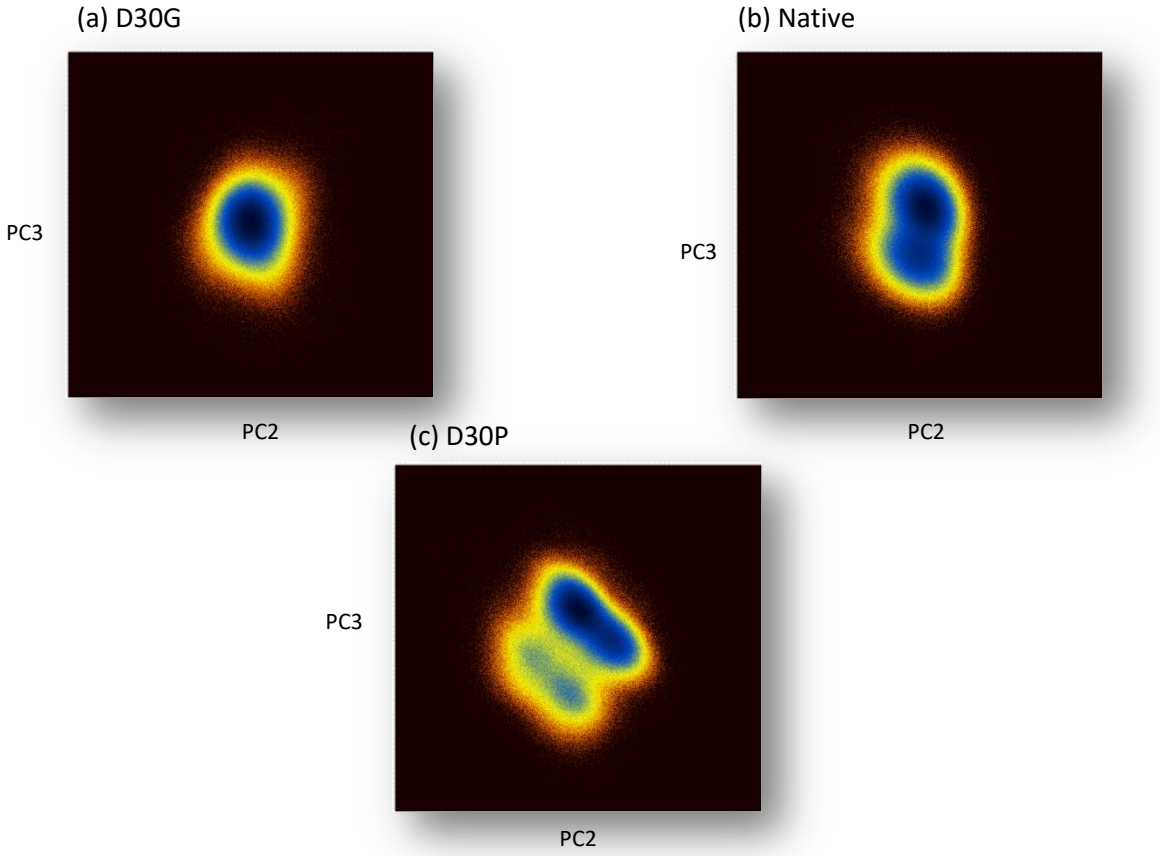


Figure 35: PC2-PC3 plots for residues 27-34 from the A turns and residues 6-13 and 48-55 from a-helix of B chain of (a) D30G, (b) Native and (c) D30P.

Frames per Cluster

	D30G	NATIVE	D30P
Cluster 1	137154	489130	870222
Cluster 2	1473	54	87117
Cluster 3	49		

Table 13: Number of frames per cluster derived from cPCA analysis

Then, using PyMOL, an alignment of all residues of the a-helix was performed:

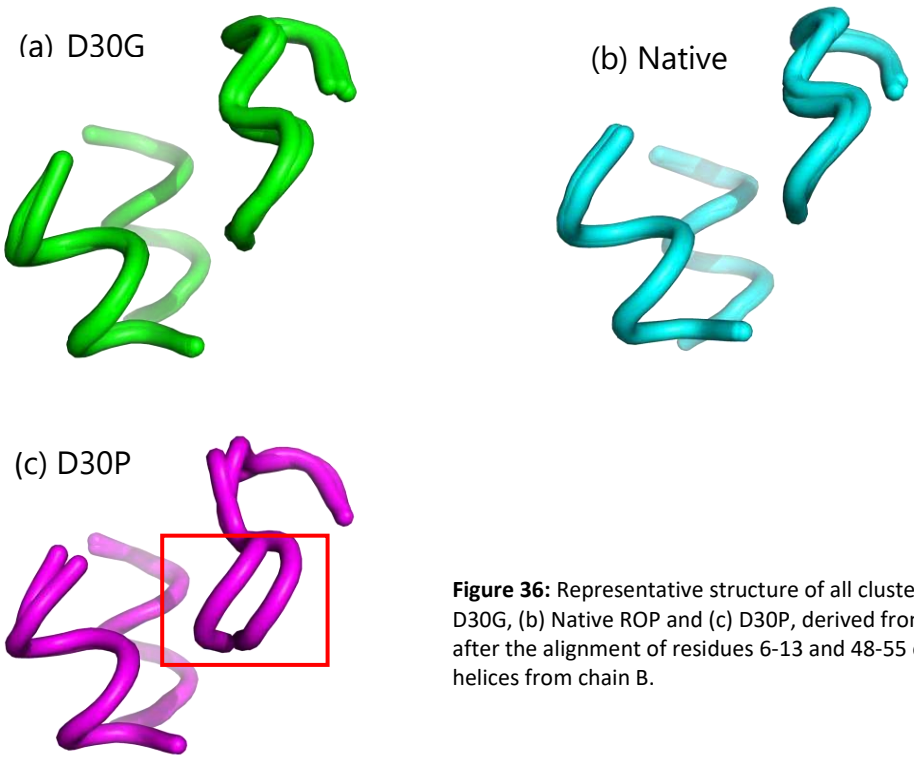


Figure 36: Representative structure of all clusters of (a) D30G, (b) Native ROP and (c) D30P, derived from PyMOL, after the alignment of residues 6-13 and 48-55 of the A-helices from chain B.

From figure 36, it becomes evident that D30P adopts two distinct conformations and consequently its loop interacts more extensively with the opposing a-helices compared to the other two variants.

The same calculations were performed for residues 27-34 of chain B and residues 6 to 13 and 48 to 55 of chain A for each protein

Turn (B Chain) + a-helix (A Chain)

Eigenvalues

	D30G	NATIVE	D30P
PC1	11.3222541809	11.4305343628	12.9947099686
PC2	0.9237968326	1.0249109268	1.8343153000
PC3	0.6409819126	0.6646002531	0.9235197306
PC4	0.4621220231	0.4745688438	0.7192844748
PC5	0.3547840416	0.4421654642	0.4706155062

Table 14: Eigenvalues derived from Cartesian PCA for each one of the first 5 principal components of D30G, Native and D30P, for residues 27-34 of B loops and residues 6-13 and 48-55 of the a–helix of chain A (step = 1).

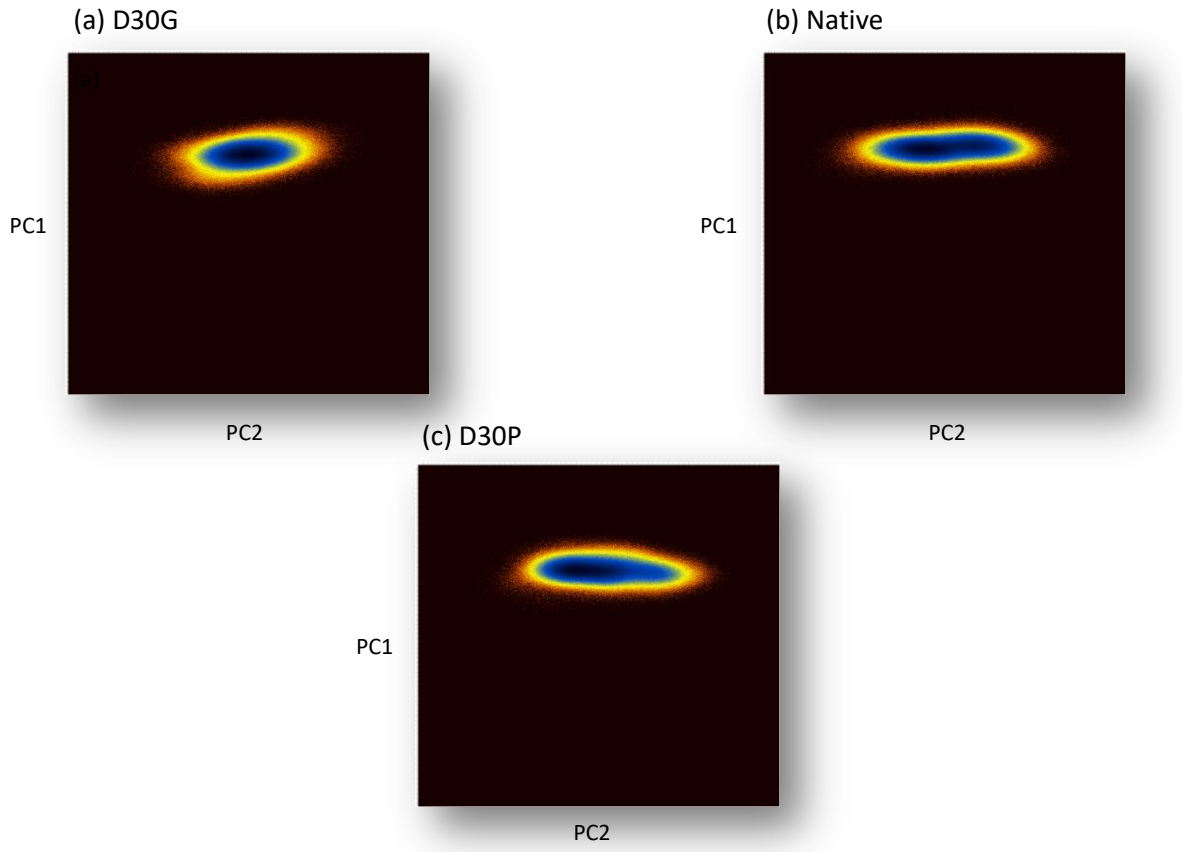


Figure 37: PC1-PC2 plots for residues 27-34 from the B turns and residues 6-13 and 48-55 from a-helix of A chain for (a) D30G, (b) Native and (c) D30P.

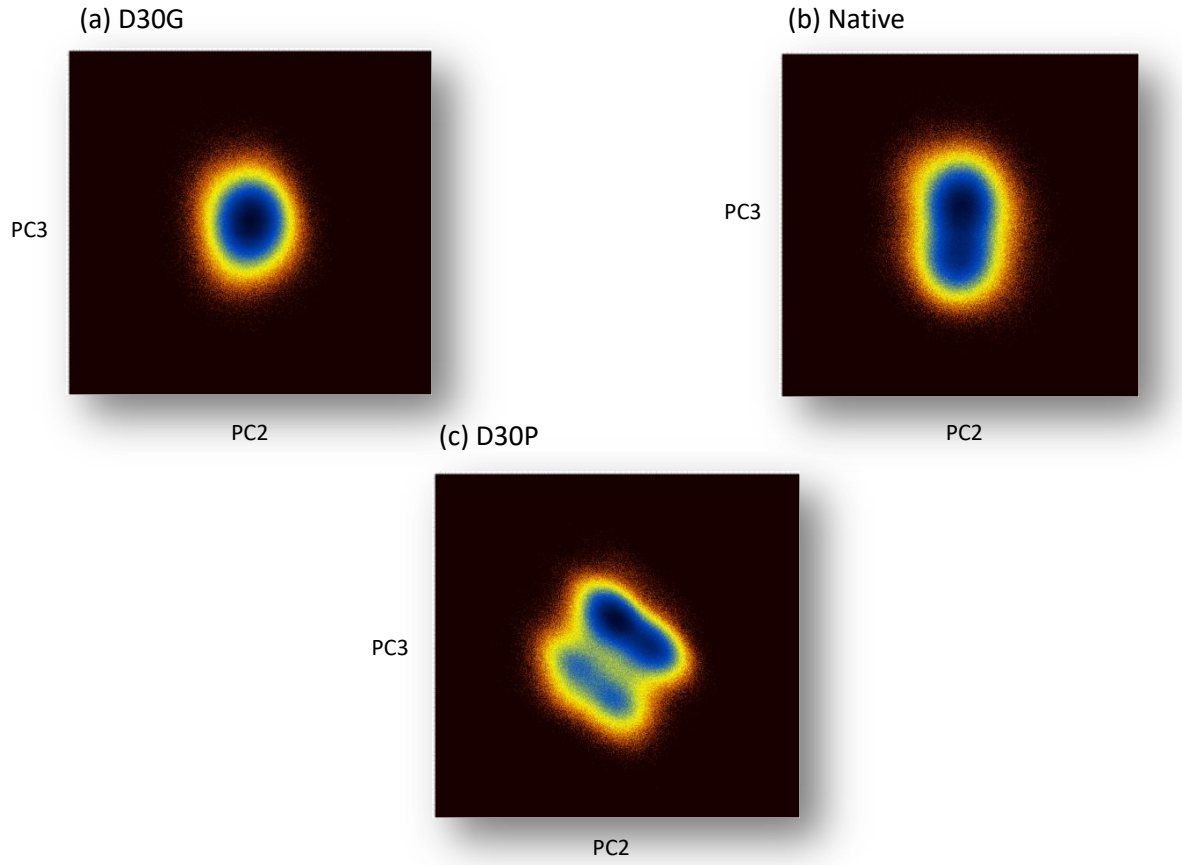


Figure 38: PC2-PC3 plots for residues 27-34 from the B turns and residues 6-13 and 48-55 from a-helix of A chain for (a) D30G, (b) Native and (c) D30P.

Frames per Cluster

	D30G	NATIVE	D30P
Cluster 1	243464	150316	827837
Cluster 2	127	92	158583
Cluster 3		182	
Cluster 4		88	
Cluster 5		50	
Cluster 6		43	

Table 15: Number of frames per cluster derived from cPCA analysis

From the number of frames per cluster (tables 13 and 15), it is observed that the majority of frames for D30P (870222 and 827837, respectively), are concentrated in the first cluster in contrast to the other two variant. This may

be explained, by the fact that D30P tends to remain longer in a single conformation and whereas D30G does not. However, this does not necessarily indicate that this conformation is the most stable or that it corresponds to the lowest energy state.

On the other hand, D30G shows the lowest number of frames in the first cluster (table 13). However, Native ROP has 150316 frames in the first cluster (table 15), which is lower than both D30G and D30P.

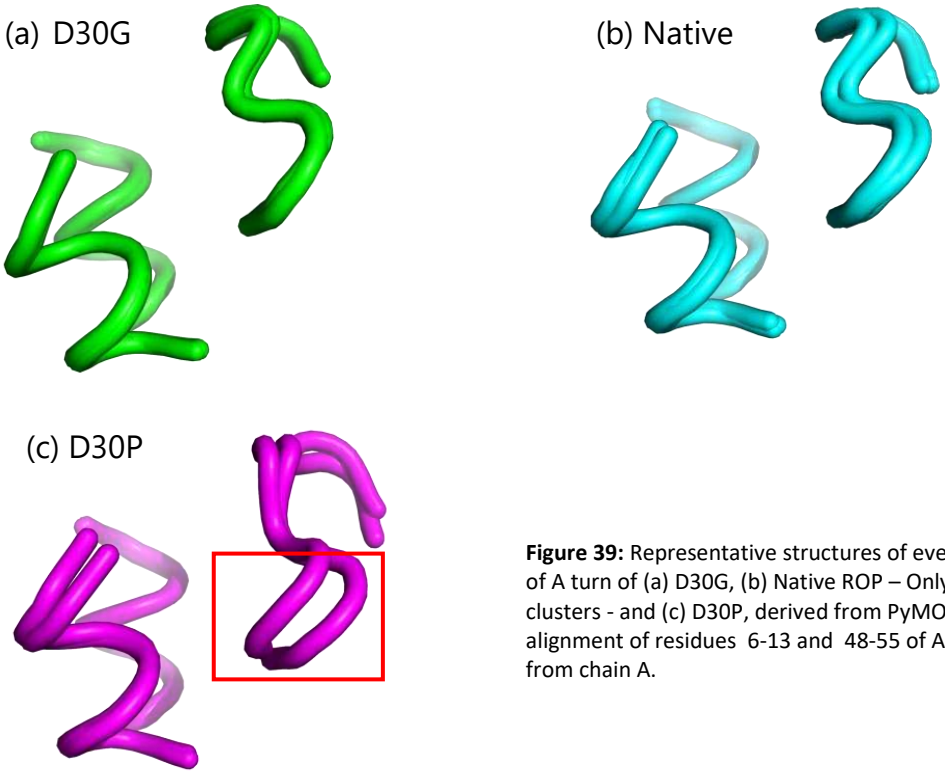


Figure 39: Representative structures of every cluster of A turn of (a) D30G, (b) Native ROP – Only the first 3 clusters - and (c) D30P, derived from PyMOL, after the alignment of residues 6-13 and 48-55 of A-helices from chain A.

Cartesian PCA analysis also revealed that the eigenvalues for PC1, PC2 and PC3 increase progressively from D30G to D30P (tables 12 and 14) although the differences remain relatively small. D30P's eigenvalues are significantly higher, confirming its overall increased structural mobility. In general, D30G displays lower eigenvalues than Native ROP. However, in some specific cases (PC2 and PC4, table 12), slightly higher eigenvalues are observed for D30G.

According to the PC1-PC2 and PC2-PC3 plots (figures 34, 35, 37 and 38), D30P seems to adopt more than one distinct conformation, while the two other variants occupy a more confined region in the plot, indicating greater stability.

Furthermore, structural analysis (figure 39) reveals an increased mobility of D30P's loop towards the opposing a-helices. In contrast, D30G and Native ROP adopt a more stable conformation, and their differences are barely distinguishable in the aligned structures from PyMOL.

3.7 PDB Structures

The final step in evaluating the stability of the loops was their visualization. Twenty frames of each variant generated using grcarma (step=100.00). Then these structures were aligned to the residues forming the a-helices, excluding the loop residues (27-34) in order to identify any significant movement in the turn region.

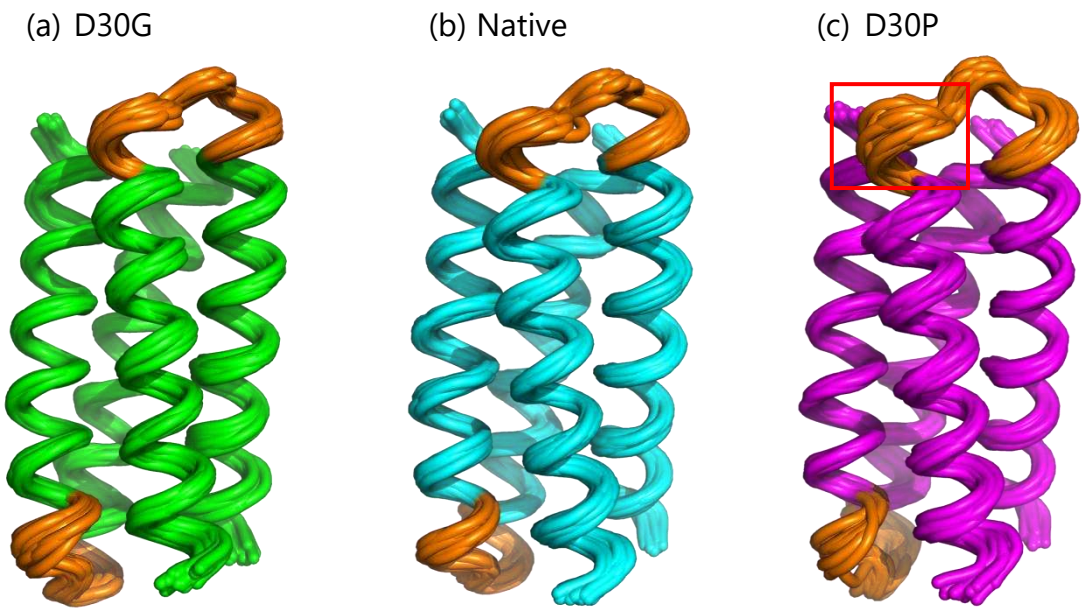


Figure 40: PDB structures of (a) D30G, (b) Native ROP and (c) D30P, derived from PyMOL, after the alignment of all residues excluding residues 27 to 34. In the front, chain B is visible, while loop residues are highlighted in orange.

From figures 40 and 41, D30P's loop appears less stable particularly at residue 28 and 32 (red outlines), which is supported by their corresponding RMSF value (figure 7). The differences between D30G and Native ROP are once again less visible.

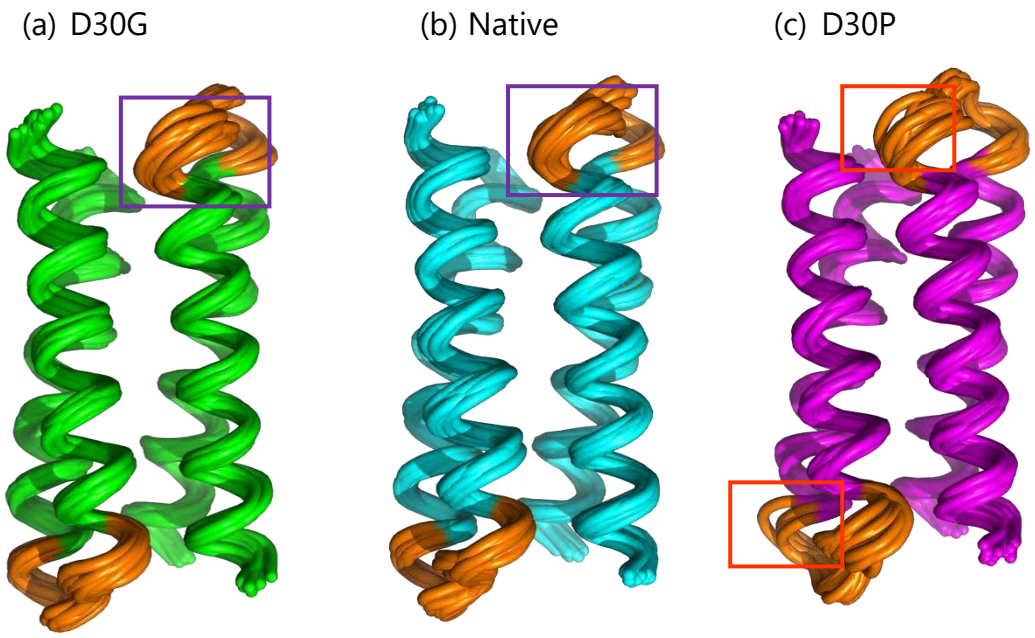


Figure 41: PDB structures of (a) D30G, (b) Native ROP and (c) D30P, derived from PyMOL, after the alignment of all the residues, excluding residues 27 to 34. Loop residues are highlighted in orange.

Although the differences between D30G and Native ROP are minimal, figure 41 reveals an increased mobility in the loop region of D30G (purple outline). This incompatibility observed compared to the previous results, can be attributed to several factors. Firstly, loops are normally regions with higher flexibility compared to a-helices. Additionally, mutations at position 30, may affect the local mobility without nessecery altering the overall structural stability. Moreover, the full alignment of the a-helices may result in the observed motion being localized exclusively to the loop region. Lastly, the frames that were obtained were twenty randomly selected dynamic states, which do not necessarily represent the average or the representative state of each structure.

4. Conclusions – Discussion

This study focuses on the analysis of the stability of three variants of ROP protein – D30G, Native ROP and D30P -. The main purpose was to identify whether there is any correlation between their general structural stability and their thermal stability (Tm). Theoretically, it was expected that while the Tm value increases, the general stability of the protein would also increase.

The results derived from their RMSF values and plots, the histograms of RMSD value distributions and the statistical analysis with R, reveal that there is a correlation between their thermal and their overall stability. The same results can be extracted from the cPCA analysis, where the decrease in their Tm corresponds to increased mobility and interaction between the loop-turn region and the opposing residues from the a-helices. Eigenvalues for most of the principal components of D30P are higher than the respective values of Native ROP, and those of Native ROP are higher than D30G's.

However, dPCA analysis indicates that, when considering only the residues forming each loop separately, there are some differences compared to the previous results. Although D30G's RMSD and RMSF values are generally lower (which indicates an increased overall stability), its dPCA eigenvalues are higher indicating increased mobility of the loop. On the other hand, Native ROP's dPCA eigenvalues are lower, which suggests that its loop region is more stable. These subtle differences in loop dynamics do not necessarily imply that Native ROP is in general less stable than D30G, but they may be attributed to the methodological differences between cPCA and dPCA.

Although dPCA suggests a slightly higher internal mobility in D30G's loop region compared to Native ROP, cPCA indicates that Native ROP exhibits higher rigid body mobility. This is because cPCA takes into account the Cartesian coordinates of every atom in the system, capturing the magnitude of the collective movement of the structure. This means that Native ROP undergoes larger-scale rigid body motion, which can only be observed through Cartesian PCA. On the other hand, D30G shows higher flexibility in terms of phi and psi torsional angles of turn's residues, but this increased internal mobility is not reflected in its general stability. D30G appears overall more stable due to reduced amplitude of collective rigid body movements (lower cPCA eigenvalues).

In the case of D30P, this clear difference in its stability, compared to the other two variants, may be caused by the presence of proline. Proline is a bulky amino acid whose side chain is bonded to the backbone, resulting in limited torsional movement [29]. That likely affects the dynamic behavior of the loop and consequently affects the overall stability of the protein.

The general conclusion that arises from all these analyses is that, the presence of a different amino acid at position 30 within the loop/turn – such as glycine in D30G or proline in D30P – affects both loop dynamics and the stability of the full-length structure. Consequently, the observed differences in Tm are associated with their thermal stability (Tm). These results provide an initial indication of the link between thermal and general stability, though further studies should be conducted.

References

1. Dill KA, Ozkan SB, Shell MS, Weikl TR., (2008) ''The protein folding problem'', Annual Review of Biophysics, 37:289-316. doi: https://doi.org/10.1146/annurev.biophys.37.092707.153558
2. Jumper J, Evans R, Pritzel A, et al., (2021), "Highly accurate protein structure prediction with AlphaFold", *Nature*, 596(7873):583-589. doi: https://doi.org/10.1038/s41586-021-03819-2
3. Abramson J, Adler J, Dunger J, et al., (2024), "Accurate structure prediction of biomolecular interactions with AlphaFold 3", *Nature*, 630(8016):493-500. doi: https://doi.org/10.1038/s41586-024-07487-w
4. Kumar G., Mishra R.R., Verma A., ''Forcefields for Atomistic-Scale Simulations: Materials and Applications'', Lecture Notes in Applied and Computational Mechanics, vol 99. Introduction to Molecular Dynamics Simulations. In: Verma, A., Mavinkere Rangappa, S., Ogata, S., Siengchin, S. (eds) Springer, Singapore. https://doi.org/10.1007/978-981-19-3092-8_1
5. H.J.C. Berendsen, D. van der Spoel, R. van Drunen, (1995), "GROMACS: A message-passing parallel molecular dynamics implementation", Computer Physics Communications, 91: 43-56, doi: https://doi.org/10.1016/0010-4655(95)00042-E
6. Manias E., (1995), "Nanorheology of strongly confide molecular fluids: a computer simulation study", University of Groningen, https://zeus.plmsc.psu.edu/~manias/PDFs/thesis.pdf
7. Baldovin M., Gradenigo G., Vulpiani A., Zanghì N.,(2025), "On the foundations of statistical mechanics", Physics Reports,1132:1-79, doi: https://doi.org/10.1016/j.physrep.2025.05.003
8. Vieira I. H. P., Botelho E. B., de Souza Gomes T. J., Kist, R., Caceres R. A., & Zanchi F. B., (2023), "Visual dynamics: a WEB application for molecular

dynamics simulation using GROMACS.", BMC bioinformatics, 24(1), 107, doi: https://doi.org/10.1186/s12859-023-05234-y

9. LibreTexts, (n.d.), "Lennard-Jones Potential", LibreTexts Chemistry, Retrieved from: https://chem.libretexts.org , (06/08/2025)
10. Meller J., (2001), " Molecular Dynamics", Wiley Online Library, doi: https://doi.org/10.1038/npg.els.0003048
11. Eberle W, Pastore A, Sander C, Rösch P., (1991), "The structure of ColE1 rop in solution", Journal of Biomolecular NMR, 1(1):71-82. doi: https://doi.org/10.1007/bf01874570
12. Castagnoli, L., Scarpa, M., Kokkinidis, M., Banner, D. W., Tsernoglou, D., & Cesareni, G., (1989), ''Genetic and structural analysis of the ColE1 Rop (Rom) protein'', The EMBO journal, 8(2), 621–629, https://doi.org/10.1002/j.1460-2075.1989.tb03417.x
13. Braden C., Tooze J., (2019), Εισαγωγή στην Δομή των Πρωτεϊνών, 2η Έκδοση, Ακαδημαϊκές Εκδόσεις Ι.ΜΠΑΣΔΡΑ & ΣΙΑ Ο.Ε., 53-54
14. Glykos NM, Papanikolau Y, Vlassi M, Kotsifaki D, Cesareni G, Kokkinidis M. Loopless, (2006), ''Rop: structure and dynamics of an engineered homotetrameric variant of the repressor of primer protein'', Biochemistry, 45(36):10905-10919. doi: https://doi.org/10.1021/bi060833n
15. Predki, P. F., Agrawal, V., Brünger, A. T., & Regan, L. (1996), "Amino-acid substitutions in a surface turn modulate protein stability", Nature structural biology, *3*(1): 54–58. https://doi.org/10.1038/nsb0196-54
16. Vouzina O. D., Tafanidis A., & Glykos N. M., (2024), ''The Curious Case of A31P, a Topology-Switching Mutant of the Repressor of Primer Protein: A Molecular Dynamics Study of Its Folding and Misfolding'', Journal of chemical information and modeling, 64(15), 6081–6091, doi: https://doi.org/10.1021/acs.jcim.4c00575

17. EMBnet, (2016), ''Molecular Dynamics Tutorial'', Retrieved from: https://web.archive.org/web/20161230070935/http://www.ch.embnet.org/MD_tutorial , (12/07/2025)
18. LAMMPS Developers, ''TIP3P water model'', Retrieved from: https://docs.lammps.org/Howto_tip3p.html, (12/07/2025)
19. Lindorff-Larsen, K. Maragakis, P. Piana, S., Eastwood M. P., Dror R. O., & Shaw D. E., (2012), "Systematic validation of protein force fields against experimental data", *PloS one*, *7*(2), https://doi.org/10.1371/journal.pone.0032131
20. GROMACS Tutorials, (n.d.), "Lysozyme in Water: Equilibration with NVT and NPT ensembles", Retrieved from: http://www.mdtutorials.com/gmx/lysozyme/07_equil2.html (10/08/2025)
21. AMBER, (2024*),* ''Plotting [Tutorial]'' Retrieved From: https://ambermd.org/tutorials/Plotting.php, (15/07/2025)
22. Glykos N. M., (2006), "Software news and updates. Carma: a molecular dynamics analysis program", Journal of computational chemistry, 27(14), 1765–1768., doi: https://doi.org/10.1002/jcc.20482
23. Koukos PI, Glykos NM., (2013), ''Grcarma: A fully automated task-oriented interface for the analysis of molecular dynamics trajectories'', Journal of Computational Chemistry, 34(26): 2310-2, doi: https://doi.org/10.1002/jcc.23381
24. Sittel F., Jain A., Stock G., (2014), "Principal component analysis of molecular dynamics: on the use of Cartesian vs. internal coordinates", The Journal of Chemical Physics, 141(1), doi: https://doi.org/10.1063/1.4885338
25. Altis A., Nguyen P.H., Hegger R., Stock G., (2007), "Dihedral angle principal component analysis of molecular dynamics simulations", The Journal of Chemical Physics, 126(24), doi: https://doi.org/10.1063/1.2746330

26. Arnittali M., Rissanou A. N., Amprazi M., Kokkinidis M., & Harmandaris V., (2021), ''Structure and Thermal Stability of wtRop and RM6 Proteins through All-Atom Molecular Dynamics Simulations and Experiments''. International journal of molecular sciences, *22*(11), doi: https://doi.org/10.3390/ijms22115931
27. The R Foundation, (2025), ''The R Project for Statistical Computing'', Retrieved From : https://www.r-project.org/, (17/08/2025)
28. David C. C., & Jacobs D. J., (2014), ''Principal component analysis: a method for determining the essential dynamics of proteins'', Methods in molecular biology (Clifton, N.J.), 1084, 193–226, doi: https://doi.org/10.1007/978-1-62703-658-0_11
29. Moradi M., Babin V., Roland C., Darden T. A., & Sagui, C., (2009), "Conformations and free energy landscapes of polyproline peptides", Proceedings of the National Academy of Sciences of the United States of America, 106(49), 20746–20751, doi: https://doi.org/10.1073/pnas.0906500106