DEMOCRITUS UNIVERSITY OF THRACE

DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

Folding studies of the A31P ROP mutant using molecular dynamics simulations

---

Μελέτες αναδίπλωσης της μετάλλαξης A31P της πρωτεΐνης ROP με προσομοιώσεις μοριακής δυναμικής

Vouzina Olympia Dialekti

Supervisor: Dr. Nicholas M. Glykos
Associate Professor of Structural and Computational Biology

Alexandroupolis, September 2022

Acknowledgments

I would like to express my gratitude to my supervisor Dr. Nicholas M. Glykos for his guidance, feedback, and unending patience. Additionally, I would like to thank him for helping me find a way to answer my questions and for introducing me to such a fascinating field of Biology.

I am also grateful for my friends who encouraged and supported me all this time.

Table of Contents

Abstract

ROP is a protein of great research interest and in its native state is a 4-α-helical bundle, with an antiparallel topology. The A31P is one of its mutants studied a lot over the years, and experimentally it has a unique topology, named *bisecting U*. In an attempt to explain its folding, the mutant was generated computationally, but with a native-like topology. The stability of its structure was investigated to find out if it was unstable enough to justify the change of the topology. More specifically, statistical analysis and molecular dynamics simulations for both native ROP and the native-like A31P mutant were done using various tools. The results indicated some differences between the calculations. However, they were not enough to lead to an established conclusion about the folding of the A31P mutant.

Περίληψη

Η ROP είναι μία πρωτεΐνη μεγάλου επιστημονικού ενδιαφέροντος και φυσιολογικά υιοθετεί τη δομή 4-α-ελικοειδούς δεματίου, με αντιπαράλληλη τοπολογία. Ένα από τα πολλά μεταλλάγματα της που έχουν μελετηθεί εκτενώς είναι το A31P, που πειραματικά εμφανίζει διαφορετική τοπολογία από τη φυσική, ονομαζόμενη *bisecting U*. Σε μία προσπάθεια εξήγησης της αναδίπλωσης του μεταλλάγματος, σχεδιάστηκε υπολογιστικά το μεταλλαγμα, όμως με τοπολογία ίδια με αυτή της φυσικής πρωτεΐνης, και ερευνήθηκε αν η δομή του είναι αρκετά ασταθής, ώστε να εξηγείται η αλλαγή. Έπειτα, πραγματοποιήθηκαν στατιστική ανάλυση και προσομοιώσεις μοριακής δυναμικής της δομής της φυσικής ROP και του υπολογιστικά σχεδιασμένου μεταλλάγματος A31P, με τη χρήση διαφόρων εργαλείων. Τα αποτελέσματα των υπολογισμών υπέδειξαν κάποιες διαφορές μεταξύ των δύο πρωτεϊνών. Ωστόσο, δεν ήταν αρκετές για να δώσουν μία απάντηση και να οδηγήσουν σε ένα εμπεριστατωμένο συμπέρασμα για την αναδίπλωση του μεταλλάγματος A31P.

1. Introduction

1.1 Coiled coils

Coiled coils are structural motifs consisting of two or more α-helices bound together. In retrospect, the structural model of the α-helix was first found by Pauling in 1950 and two years later Crick introduced the structure of coiled coils [1,2]. He propounded the "knobs into holes" model, in which the side chains of an α-helix, i.e., the knobs, can fit in the spaces between the knobs, i.e., the holes, of an adjacent α-helix [2]. Later he pointed out how the knobs fall into the holes and the way the helices tend to turn around one another [3]. The α-helices participating in coiled coils are packed differently changing the number of residues per turn in each helix, approximately, from 3.6 to 3.5 and the pairing between knobs and holes is taking place every seventh knob and hole of the neighboring helices creating a heptad repeat [3]. The positions of the residues are labeled "a-g", and the residues in positions "a" and "d" are hydrophobic [3,4].

1.2 4-α-Helical bundle

The 4-α-helical bundle is the most common and simple α-helical domain made up of 4 α-helices packed together. The hydrophobic residues are in the center of the bundle creating a hydrophobic core while the hydrophilic residues are placed externally surrounding the hydrophobic core [5]. The main part of the 4-α-helical bundle that holds the structure steady is the hydrophobic interactions between the side chains of the hydrophobic residues [6]. The helical bundle is built according to the "grooves into ridges" model, which was described by Chothia, *et.al* in 1977 [7]. In this model, the 4 α-helices are almost parallel or antiparallel with a small angle, of 20° approximately, between their axes [5,7]. Along with this rotation, the side chains of the residues form a spiral shape on the surface of the helix creating ridges intermittent by grooves [5,7]. The residues of the grooves and the ridges create a sequence with a difference of three or four residues between each other, while the ridges of one helix are packed into the grooves of the other and vice versa [5,7]. In more detail, in the

helix-helix interactions, the first helix ridges are in a sequence i±4n, and the second helix ridges are in a sequence j±3n, or vice versa [5,7].

In 4-α-helix bundles, the loops between the α-helices are significantly important as they stabilize the structure. In particular, the interloop and the interhelix interaction energies are smaller than the loop-helix interaction energy [8].

1.3 ROP protein

The repressor of primer (ROP) is a left-handed, antiparallel 4-α-helical coiled-coil protein. It is an indirect negative regulator of plasmid replication, found in some plasmids of *E. coli* bacteria, with a large number of mutants available [9]. It is a symmetrical homodimer consisting of two chains, A and B, having 63 amino acids, each. Every chain creates two antiparallel α-helices separated by a loop (residues 3-28 the first, 32-56 the second helix, and 29-31 the loop) and a disordered C-terminal region (residues 57-63) [9,10]. The four α-helices are bound together by hydrophobic interactions, and they create a 4-α-helix/coiled-coil bundle [9,11].

In each α-helix, the residues form hydrogen bonds to stabilize the helix structure. Thus, from residues 28 to 34 only three hydrogen bonds are observed, of which only one is expected [9]. There is one structurally unusual between 32D and 33E. Also, residue 31A participates in two hydrogen bonds, one with each helix, by giving an H^+ from its NH group to the CO group of 26L, creating an unexpected bond, and accepting an H^+ from the NH group of 35A to its CO group, creating a usual α-helix hydrogen bond [9]. Many mutants of ROP protein have been studied through the years. Some do not differ from the wild type structurally, biochemically, and thermodynamically, while others, like Ala2Ile2-6, RM6, and A31P variants, have more differences with native ROP than similarities (*Figure 1*) [12,13,14,15].

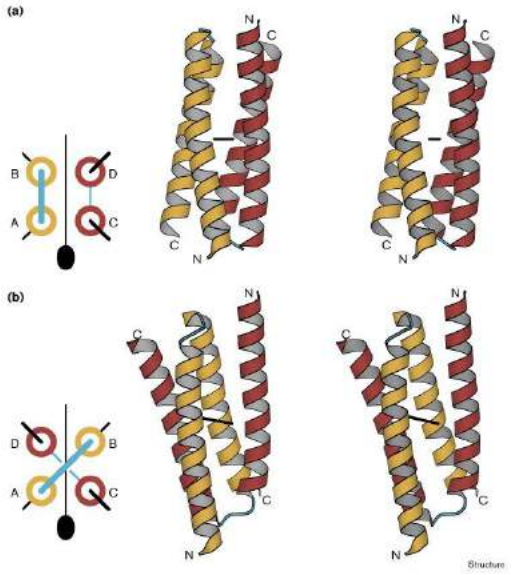

Figure 1 | Structure and topology representation of (a) wild-type ROP, and A31P mutant. Reproduced without permission from Glykos, N. M., Cesareni, G., & Kokkinidis, M. (1999). Protein plasticity to the extreme: changing the topology of a 4-α-helical bundle with a single amino acid substitution. *Structure*, *7*(6), 597–603.

1.4 A31P mutant

The small size and the large number of mutants make ROP a protein easy to study. Some of the mutants do not make a significant difference in the structure and function of the molecule while others are of particular interest to the scientific community. In this thesis, the 31A is the residue of interest and is quite important in forming the loop and the 4-α-helical bundle in general [8,9].

Proline, due to its structure cannot form the dihedral angles of alanine-31 [11]. Furthermore, it cannot give an H^+ to create a hydrogen bond with 26L, causing problems with the stability and the structure of the molecule [15].

In the A31P mutation of ROP protein, alanine is replaced by proline. The replacement resulted in some crucial changes in the molecule, including a major structural differentiation and changes in the packing of the hydrophobic core.

Moreover, the mutation led to an exchange of position of two adjacent helices from different chains, but the connectivity and the polarity remained the same [16].

Thus, the molecule became a mixed parallel and antiparallel helix bundle changing the topology of the protein from *anti* to *bisecting U* topology, [16,17]. Also, the residues forming the loop had changed, in the A31P mutant the loop consists of the residues 27-30, and the distance between the two α-helices of each chain had increased [16,18]. The *bisecting U* topology is probably a result of the removal of one dihedral angle because of the substitution [18]. The packing of the hydrophobic core seems to be looser and a cavity in the middle of the molecule was observed [16,19]. However, the protein did not lose completely its function maintaining some of the binding ability to the RNA target of wild-type ROP [17].

A hypothetical wild-type-like A31P was constructed substituting alanine-31 with proline *in silico*. The hypothetical wild-type-like A31P was used to compare its molecular dynamics results with wild-type results [19]. The results indicated that wild-type-like A31P was more stable than the A31P mutant, but not as stable as the wild-type ROP [19]. The average RMSD in wild-type ROP was 0.62±0.14 Å, in A31P mutant 1.03±0.16 Å, and in the hypothetical wild-type-like A31P 0.70±0.14 Å [19].

1.5 Molecular dynamics simulations

Molecular dynamics (MD) simulations are generally *in silico* experiments that are not prepared in the laboratories to save a large amount of money and time, or because such an experiment is not attainable in the lab. It is a significantly helpful, physics-based tool used to understand, analyze, and predict the atoms' movement, over time [20,21]. The calculation of the total energy of each atom is necessary, such as the calculation of the forces between them and the solution of the classical equation of motion for each one [22]. To specify the total energy of a big molecule, like a protein, empirical methods have been developed, known as force fields [23].

Since the first molecular dynamics simulation by Karplus and his coworkers, who used it to examine the dynamics of a folded protein, there has been a great evolution in the MD field [22,24]. In many laboratories, MD methods were developed

to study protein folding and dynamics. However, force fields are not ideal, and they can definitely be improved over time to become more reliable and useful tools.

During this thesis, a molecular dynamics analysis program and its graphical user interface form, named carma and grcarma, respectively, are used to study the folding of a hypothetical wild-type-like A31P mutant of ROP protein [25,26,27].

1.6 Main question

The aim of the thesis is to explain the difference in the topology noted between the native ROP and the experimental A31P ROP mutant. For this purpose, a native-like A31P mutant, i.e., with the native's protein topology, was designed computationally (in 3D protein structure software) and then compared with the native ROP. To investigate its stability, the energy of the wild-type ROP and the wild-type-like A31P mutant was first checked directly in a web server for protein refinement and the results were statistically analyzed by Student's t-test. They resulted in no statistical difference between the energies of the native and the native-like A31P ROP. Eventually, the study turned to molecular dynamics simulations, and the native ROP and native-like A31P ROP were both analyzed and compared. If the results indicated high instability in the structure of the native-like A31P, this would explain the disparity in the topology of the A31P mutant.

2. Methods

2.1 Mutagenesis

Firstly, in order to answer the main question, three molecular visualization programs, VMD (Visual Molecular Dynamics), Coot (Crystallographic Object-Oriented Toolkit), and PyMol, were utilized for computational mutagenesis [28,29,30]. The ROP protein (PDB ID: 1RPR), and especially model 1 of the file, from the Protein Data Bank (PDB), was used in each tool to create the molecules [31,32,33]. In all software, residue alanine (A) 31 was substituted by proline (P), in both chains, generating A31P mutants. The substitution did not change the protein structure and the new files were saved for further analysis.

2.2 Molecule Refinement

For the refinement of the molecule a web server for protein structure prediction and refinement, named GalaxyWEB, was used, specifically, the GalaxyRefineComplex method [34,35,36,37]. The tool first prepares the refinement, defining the interface residues, followed by structure relaxation, using 2 protocols that differ in the restraint energy. Protocol 1 inserts only distance restraint energy, while protocol 2 distance and position restraints. Finally, as a result, the program displays the 5-lowest-energy models from each protocol. This tool was used for model 1 of ROP protein (PDB ID: 1RPR), which will from now on be referred to as model1_1rpr, and the three new PDB-files generated by the *ab initio* mutagenesis, 1 for PyMol, 1 for VMD, and 1 for Coot.

2.3 Alignment

The MM-align algorithm is specialized in multimeric protein structure alignment, using iterative dynamic programming [38]. It calculates an RMSD value showing the similarities of the comparing structures. The lower the value, the more

alike the structures are. In this case, the program compared model1_1rpr with all the *ab initio* mutant models generated by GalaxyWEB, and then the ones with the lowest values of each group of five were again refined with GalaxyRefineComplex.

2.4 Statistical Analysis

After the second refinement, a Student's t-test was performed to compare the energies of the models (of the 3 mutants generated *ab initio* and the models selected after the alignment) with the new ones of model 1_1rpr [39]. This process was done to figure out if there is a significant statistical difference between the wild-type ROP and the computational mutant's energy. The null hypothesis (H_0) in this analysis was determined that there is no significant statistical relationship between the energies, in contrast, the alternative hypothesis (H_ε) was that there is a statistical relationship. If the calculated p-value was lower than 0.05, the H_0 would be rejected, and vice versa.

2.5 Molecular dynamics

2.5.1 carma & grcarma

The statistical analysis did not provide an answer to the main question, so further investigation was needed. Molecular dynamics simulation seemed quite appealing and could probably lead to an answer. The software used for this process was carma, and grcarma, its automated graphic interface [25,26,27]. It uses two types of files, a DCD and a PSF, containing trajectory data and structure information about the protein. The tools used were "Secondary Structure", "RMSD matrix", "Covariance, average and representative structures", "Cartesian PCA", and "Dihedral PCA" from grcarma, and "torsion-angle RMSD matrix" from straight carma. The cartesian analyses are based on the cartesian coordinates, while the torsion angle and the dihedral on the dihedral torsion angles (φ and ψ), of the atoms. These analyses were done twice, once for the native-like A31P mutant, and once for the native ROP, for the

comparison of the results. The native-like A31P mutant trajectory contained 10092100 frames, while the native ROP, 10000000 frames.

2.5.2 Secondary structure

This program calculates secondary structure assignments for the whole trajectory. The analysis was done three times, the first included all the residues, the second only chain A, and the third only chain B, both excluding the tails, i.e., residues 57-63 that are disordered. Each computation generated a STRIDE plot and a WebLogo graph, shown in the result section [40,41]. The calculations were done for both native ROP and a native-like A31P mutant.

2.5.3 Cartesian RMSD matrix

The cartesian RMSD (Root-Mean-Square-Deviation) matrix calculation compares the frames in all the possible combinations, and the step between the frames can be defined. The procedure, for both native and native-like, was done two times using only the CA atoms of the structure, one including and one excluding tails, with a step of 3364 frames. The excluding tails computation was repeated with a specified maximum RMSD value equal to the maximum value from the one including tails, for comparison purposes.

Two extra RMSD matrices were prepared to compare the results, in order to answer the basic question. The first was generated for native-like A31P mutant, using only the first 10000000 frames, excluding tails. The second one was prepared, for the native ROP, excluding tails, with a specified maximum RMSD value equal to the maximum value of the first. Lastly, these matrices were merged to 1, in which the native ROP appeared in the lower-left corner, while the native-like was in the upper-right one.

2.5.4 Torsion-angle RMSD matrix

The torsion-angle RMSD matrix tool works like the cartesian RMSD matrix, but instead of CA, it uses backbone atoms. Only one calculation was performed for each trajectory, excluding tails, with a step of 3364 frames, the reason is explained in the results section.

2.5.5 Root-Mean-Square-Fluctuation plot

This tool creates an average of the selected atoms that will be used to make an RMSF (Root-Mean-Square-Fluctuation) plot, with xmgrace software, for each atom included in the calculation [42]. The RMSF is a measure of the average deviation of an atom from a reference position over time, in this case between the frames. Practically, it shows the mobility of the atoms, while the lower the RMSF is, the more stable the atom is. This calculation was performed once for each trajectory, excluding tails, using only CA atoms, with a step of 1 frame, generating 3 RMSF plots, each. In the first RMSF plot, all the residues were presented, while in the others only chains' A and B.

2.5.6 Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate technique that analyzes a large amount of data and produces new components, named principal components (PCs), containing only significant information [43]. It generates clusters with similar data and makes the process easier, by decreasing the number of components, while maintaining the data [43].

2.5.6.1 Cartesian PCA

Before the Cartesian PCA (cPCA), the DCD file of each protein was fitted excluding tails. The new ones containing only the CA atoms were used for the analysis.

The number of components produced by cPCA was specified as 5, the ones that participated in the postscript diagrams as 3, the temperature of the simulation as 298 Kelvin, and the maximum number of clusters produced as 50.

The analysis generated a table of clusters and the number of frames included in each, three PC landscapes (PC1-PC2, PC1-PC3, and PC2-PC3) (*Figure 13*), a plot of clusters created by "cPCA.clusters.dat" (*Figure 14*) using plot software, and three kinds of structures for each cluster, an average (not presented), a representative, figures created by PyMol software, and superposition, figures represented by RasMol software (*Figures 15* and *16*) [30,44,45]. The results of the two proteins, native and native-like, were then compared.

2.5.6.2 Dihedral PCA

The dihedral PCA (dPCA) was done using the backbone atoms of the proteins, excluding tails, and the same specifications, as in the cPCA, were used. The dPCA produced the same type of results, but with different data. Finally, the results were again compared in order to find out if there were crucial differences or similarities between them.

3. Results

3.1 GalaxyWEB & MM-align

The results from the mutagenesis and the first run of GalaxyWEB are not presented, but the models produced were aligned. The MM-align results, between the *ab initio* mutants and native ROP/ model1_1rpr, are shown in *Tables 1, 2*, and *3*. It was observed that the values from protocol 2 were, in general, lower than the ones from protocol 1. Therefore, it was indicated that the protocol with distance and position restraints created structures more similar to model1_1rpr than the other one with only distance restraints. The lowest RMSD of each protocol is marked with bold numbers. In *Tables 1,2*, and *3*, models 4 and 9, 1 and 8, and 2 and 10, from *ab initio* mutants generated from PyMol, VMD, and Coot, respectively, were more like model1_1rpr. This is the reason why they were used in the second GalaxyWEB analysis (complete results do not appear), and then in the statistical analysis.

In *Tables 4, 5, 6*, and *7* are featured the energy values of all the models of the model1_1rpr and *ab initio* mutants from the first run of GalaxyWeb, and of each model selected after the alignment. Also, the mean, the standard deviation, and the p-value of the Students' t-tests are presented in the tables. The p-value calculated from the statistical analyses is way lower than 0.05 in all the computations. This indicated that the H_0 hypothesis is rejected and the H_ε is accepted, thus, the energy between the compared data had no significant statistical difference. Overall, the statistical analysis did not lead to a conclusion about the odd folding of *ab initio* A31P mutants.

Table 1 | Models from PyMol *ab initio* mutagenesis and RMSD value, from the comparison with model1_1rpr, calculated by MM-align.

Protocol	Model	RMSD value
Protocol 1	Model 1	1.80
	Model 2	1.51
	Model 3	1.73
	Model 4	**1.44**
	Model 5	1.56
Protocol 2	Model 6	0.72
	Model 7	0.67
	Model 8	0.69
	Model 9	**0.63**
	Model 10	0.65

Table 2 | Models from VMD *ab initio* mutagenesis and RMSD value, from the comparison with model1_1rpr, calculated by MM-align.

Protocol	Model	RMSD value
Protocol 1	**Model 1**	**1.39**
	Model 2	1.79
	Model 3	1.78
	Model 4	1.71
	Model 5	1.74
Protocol 2	Model 6	0.71
	Model 7	0.68
	Model 8	**0.67**
	Model 9	0.69
	Model 10	0.73

Table 3 | Models from Coot *ab initio* mutagenesis and RMSD value, from the comparison with model1_1rpr, calculated by MM-align.

Protocol	Model	RMSD value
Protocol 1	Model 1	1.70
	Model 2	**1.56**
	Model 3	1.66
	Model 4	1.87
	Model 5	1.69
Protocol 2	Model 6	0.70
	Model 7	0.66
	Model 8	0.67
	Model 9	0.68
	Model 10	**0.61**

Table 4 | Statistical analysis of the energies of *ab initio* created mutants with model1_1rpr, generated by protocol 1 of GalaxyWeb, before the alignment.

Protocol 1	model1_1rpr	A31P_pymol	A31P_vmd	A31P_coot
Model 1	-5991.647	-5886.21	-5867.372	-5884.784
Model 2	-5974.395	-5877.609	-5861.14	-5871.263
Model 3	-5962.846	-5870.56	-5854.16	-5865.596
Model 4	-5960.305	-5853.275	-5852.395	-5864.9
Model 5	-5952.57	-5852.151	-5848.602	-5861.095
Mean	-5968.3526	-5867.961	-5856.7338	-5869.5276
StDev	15.19	14.99	7.49	9.27
p-value		**0.0000058**	**0.0000077**	**0.0000079**

Table 5 | Statistical analysis of the energies of *ab initio* created mutants with model1_1rpr, generated by protocol 2 of GalaxyWeb, before the alignment.

Protocol 2	model1_1rpr	A31P_pymol	A31P_vmd	A31P_coot
Model 1	-5902.808	-5803.018	-5821.765	-5818.31
Model 2	-5893.646	-5793.686	-5801.415	-5812.441
Model 3	-5893.615	-5783.327	-5787.831	-5792.686
Model 4	-5891.952	-5781.227	-5782.645	-5789.838
Model 5	-5891.836	-5768.623	-5782.513	-5767.521
Mean	-5894.7714	-5785.9762	-5795.234	-5796.159
StDev	4.58	13.04	16.71	20.17
p-value		**0.0000114**	**0.0000892**	**0.0002588**

Table 6 | Statistical analysis of the energies of the models selected after the MM-align calculation with model1_1rpr. The energies were generated by protocol 1 of GalaxyWeb.

Protocol 1	model1_1rpr	A31P_pymol model4	A31P_pymol model9	A31P_vmd model1	A31P_vmd model8	A31P_coot model2	A31P_coot model10
Model 1	-5991.647	-6029.547	-6027.903	-6029.393	-6017.207	-6033.744	-6018.111
Model 2	-5974.395	-6009.437	-6018.213	-6026.86	-6015.377	-6025.626	-6017.424
Model 3	-5962.846	-6006.041	-6016.038	-6026.649	-5991.809	-6021.955	-6012.742
Model 4	-5960.305	-6000.172	-6009.225	-6022.672	-5988.697	-6021.836	-6001.228
Model 5	-5952.57	-5989.774	-5996.313	-6019.903	-5987.884	-6020.388	-6001.213
Mean	-5968.3526	-6006.9942	-6013.5384	-6025.0954	-6000.1948	-6024.7098	-6010.1436
StDev	15.19	14.65	11.72	3.77	14.78	5.41	8.40
p-value		**0.0034770**	**0.0009292**	**0.0007601**	**0.0099578**	**0.0005514**	**0.0014943**

Table 7 | Statistical analysis of the energies of the models selected after the MM-align calculation with model1_1rpr. The energies were generated by protocol 2 of GalaxyWeb.

Protocol 2	model1_1rpr	A31P_pymol model4	A31P_pymol model9	A31P_vmd model1	A31P_vmd model8	A31P_coot model2	A31P_coot model10
Model 1	-5902.808	-6002.808	-5953.576	-6013.585	-5940.611	-6032.653	-5959.406
Model 2	-5893.646	-6002.626	-5953.227	-6011.912	-5936.729	-6020.916	-5957.7
Model 3	-5893.615	-5995.395	-5947.509	-6006.508	-5933.633	-6017.91	-5956.174
Model 4	-5891.952	-5981.071	-5946.861	-5996.415	-5933.564	-6008.112	-5955.686
Model 5	-5891.836	-5980.462	-5945.467	-5995.995	-5926.319	-6005.012	-5948.22
Mean	-5894.7714	-5992.4724	-5949.328	-6004.883	-5934.1712	-6016.9206	-5955.4372
StDev	4.58	11.10	3.79	8.34	5.25	11.00	4.29
p-value		**0.0000054**	**0.0000001**	**0.0000001**	**0.0000017**	**0.0000015**	**0.00000002**

3.2 Native-like A31P mutant

3.2.1 Secondary structure

The secondary structure analysis revealed that the prevalent secondary structure of the molecule was an α-helix. However, with the use of STRIDE and WebLogo, random coils and turns appeared in some regions (*Figures 2,3,4,* and *5*). Due to the mobility of the turns, variability was observed mainly at the ends of each chain, specifically in residues 1-6 and 55-63 in chain A (*Figure 2a*), and in chain B (*Figure 2b*). Similarly, variability became visible in the residues between the α-helices (27-33 in both chains). In *Figures 4* and *5,* the diversity of the loop was more evident because the tails are disordered, and in that calculation, they were excluded. Also, in some frames, the N-terminal region of both chains seemed to have the structure of α-helix, as suggested by the pink color in *Figure 5*. So, the computation including tails was not used for further analysis.

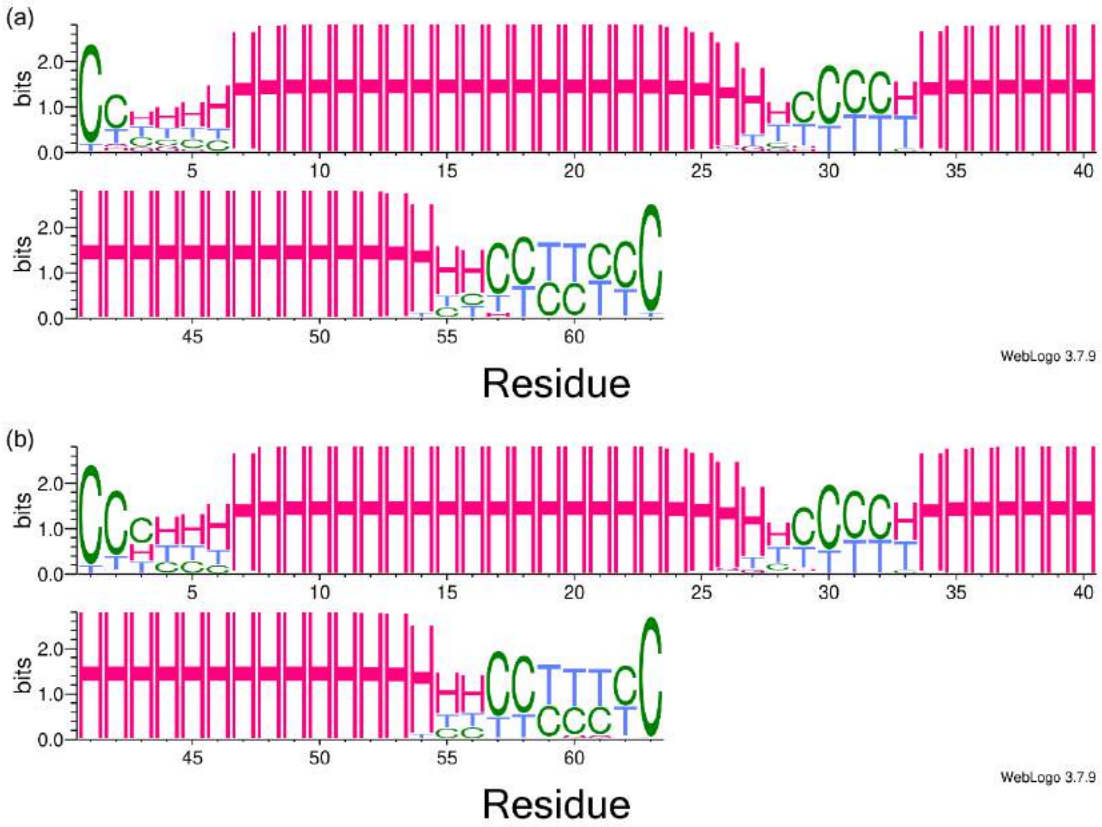

Figure 2 | Secondary structure analysis of native-like A31P mutant, including tails, using WebLogo. Representation of (a) chain A, and (b) chain B. Coding: H for α-helices, G for 3_{10}–helices, T for turns, and C for random coils.

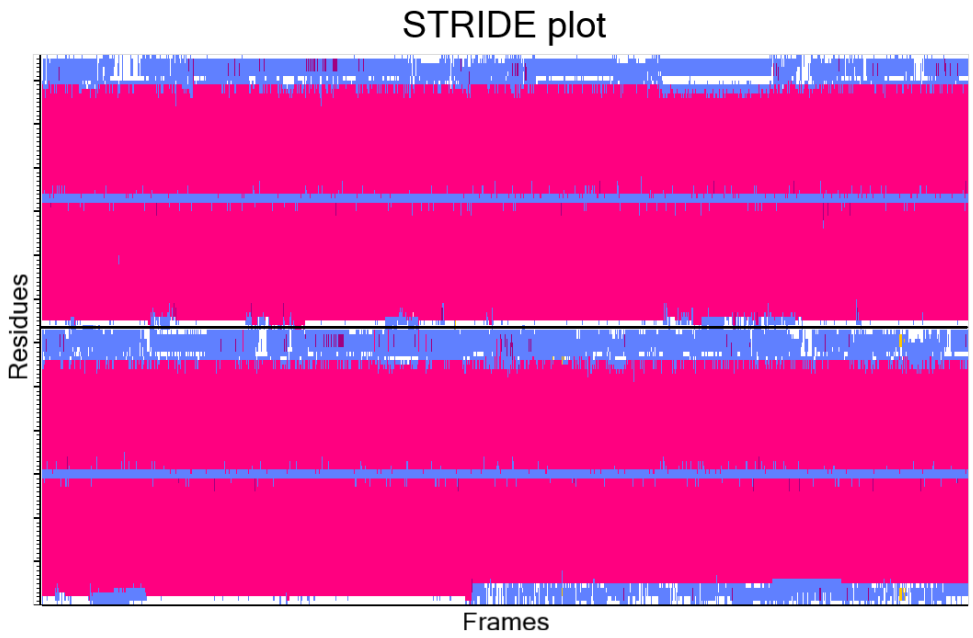

Figure 3 | Secondary structure analysis generated by the program STRIDE, using the native-like A31P mutant, including tails. Coding: pink for α-helices, purple for 3_{10} –helices, dark purple for Pi-helices, yellow for β-sheets, blue for β/G-turns, and white for coils/unassigned. The black line shows the limits of the chains, in the lower half chain A and the upper half chain B.

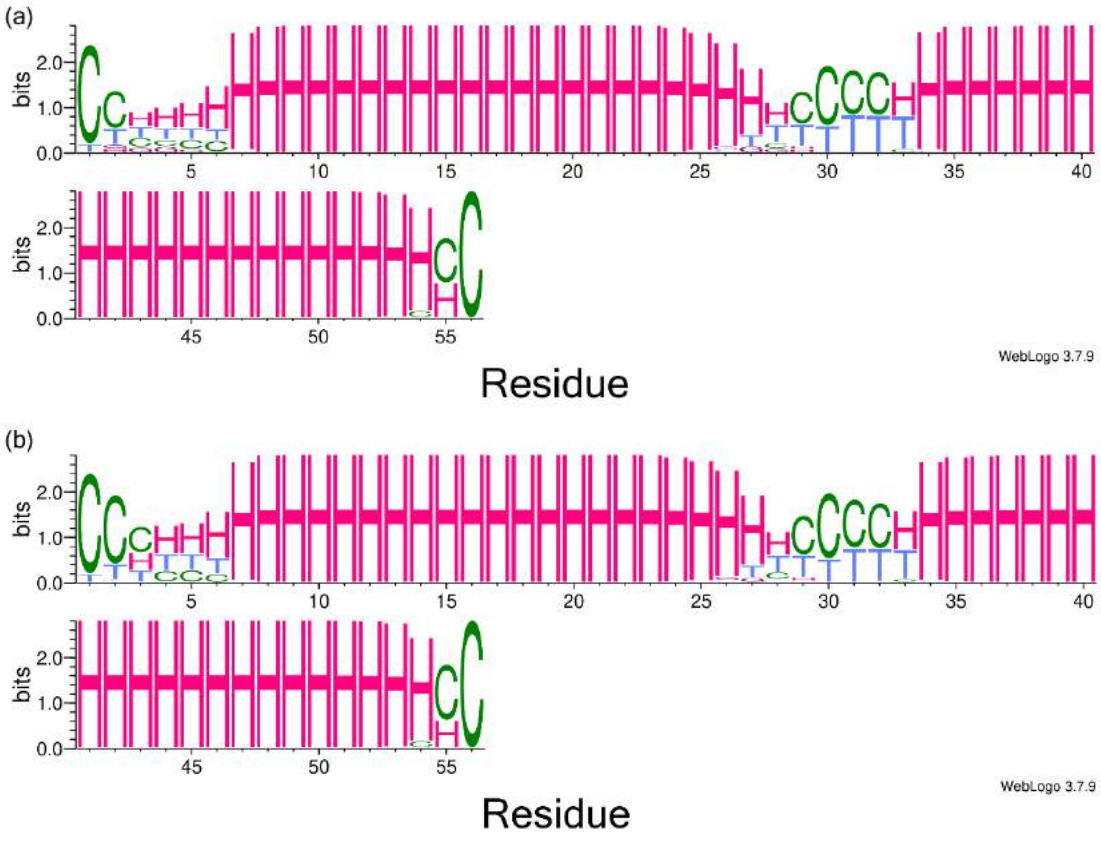

Figure 4 | Secondary structure analysis of native-like A31P mutant, excluding tails, using WebLogo. Representation of (a) chain A, and (b) chain B. Coding: H for α-helices, G for 3_{10}–helices, T for turns, and C for random coils.

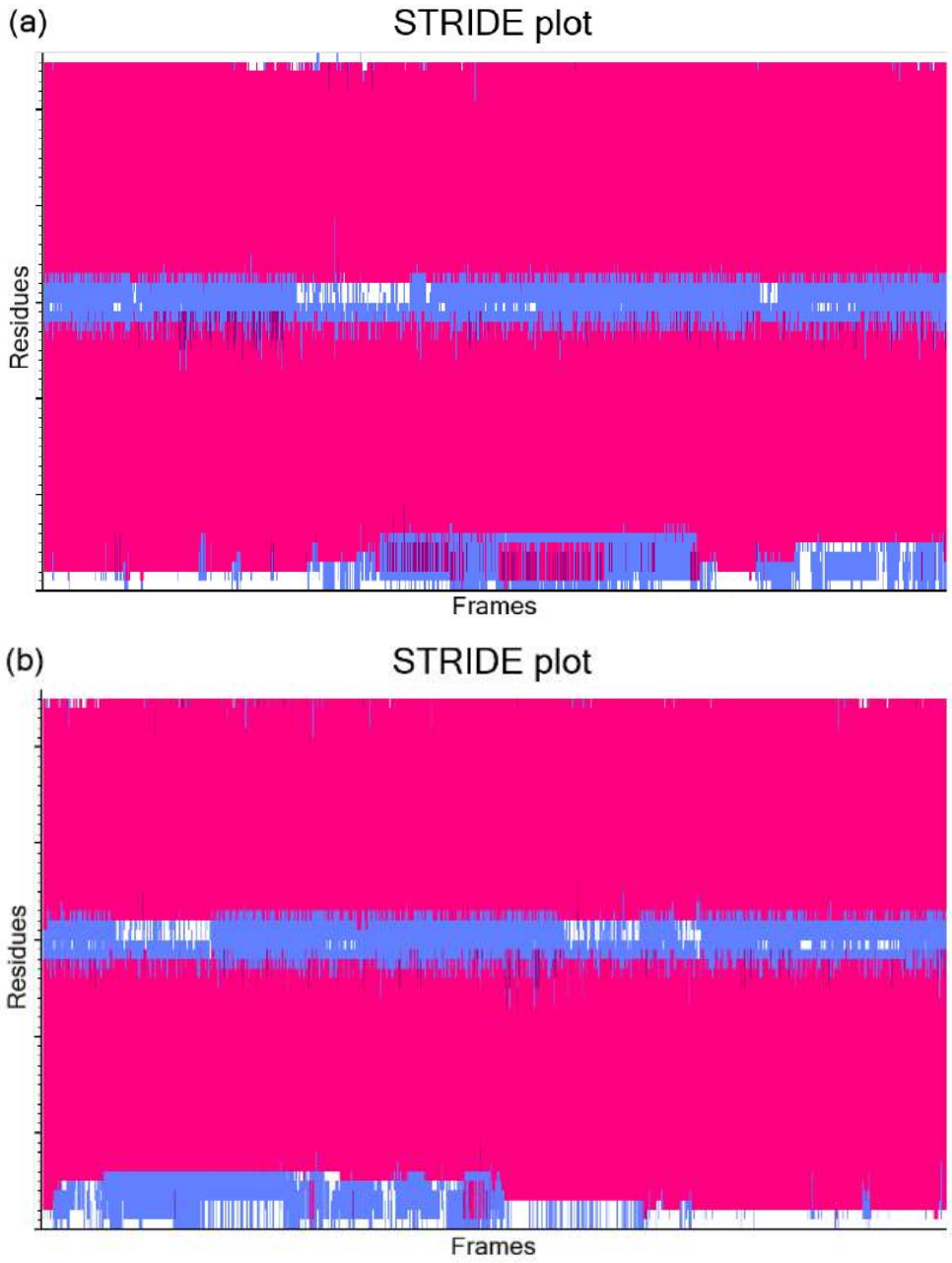

Figure 5 | Secondary structure analysis generated by the program STRIDE, using the native-like A31P mutant, excluding tails. Representation of (a) chain A, and (b) chain B. Coding: pink for α-helices, purple for 3_{10} –helices, dark purple for Pi-helices, yellow for β-sheets, blue for β/G-turns, and white for coils/unassigned.

3.2.2 Cartesian RMSD matrix

The RMSD matrices are symmetrical, and the line of symmetry is diagonal (observed in dark blue). The maximum RMSD noted in the grcarma window, including tails, was 6.92 (*Figure 6*). By contrast, the exclusion of the tails decreased the total

mobility of the molecule, and the max RMSD value was reduced to 5.63 (*Figure 7*). Therefore, a final RMSD matrix, on a scale of 0.00 to 6.92, like the calculation including tails, was prepared for the comparison of the results with and without tails (*Figure 8*). This showed that excluding tails the frames were more similar than those including tails, which meant that in the first case the molecule was more stable, as noted above. The one including tails was prepared to be compared with the excluding tails one, to confirm that the tails should be excluded from the analyses. So, the RMSD matrices in *Figures 6 and 8* did not have any other helpful information and were not used further.

Additionally, it was reported that in the RMSD matrix excluding tails with a maximum RMSD value of 5.63, there were 8 clusters, some of them consisted of many frames while others of fewer (*Figure 9*). These data will be compared with the PCA results, and later with native ROP results to try to answer if there is a reason explaining the structure of the two proteins.

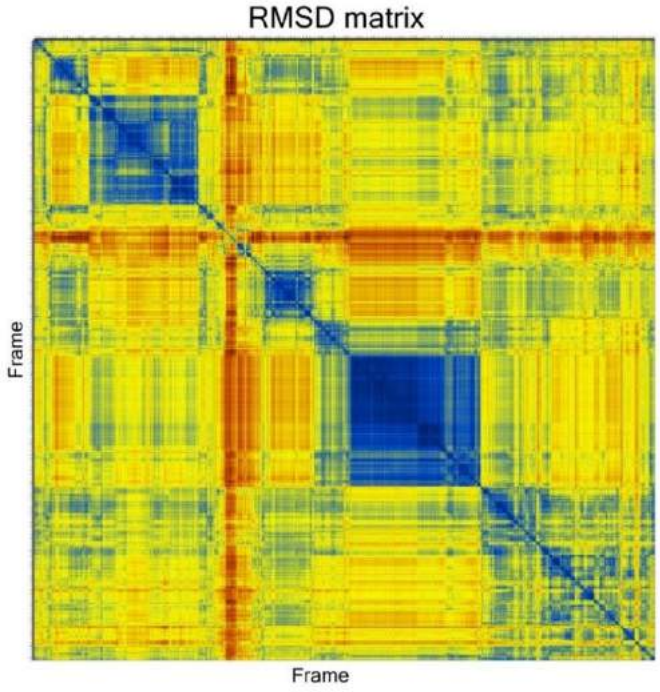

Figure 6 | RMSD matrix using only CA atoms of native-like A31P mutant, including tails. The color ranges from dark blue (minimum RMSD) to dark red (maximum RMSD: 6.92).

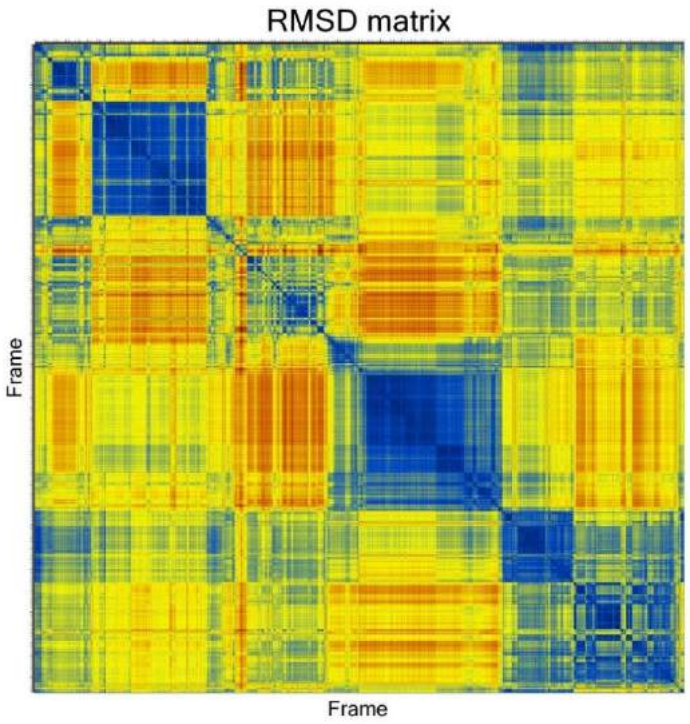

Figure 7 | RMSD matrix using only CA atoms of native-like A31P mutant, excluding tails. The color ranges from dark blue (minimum RMSD) to dark red (maximum RMSD: 5.63).

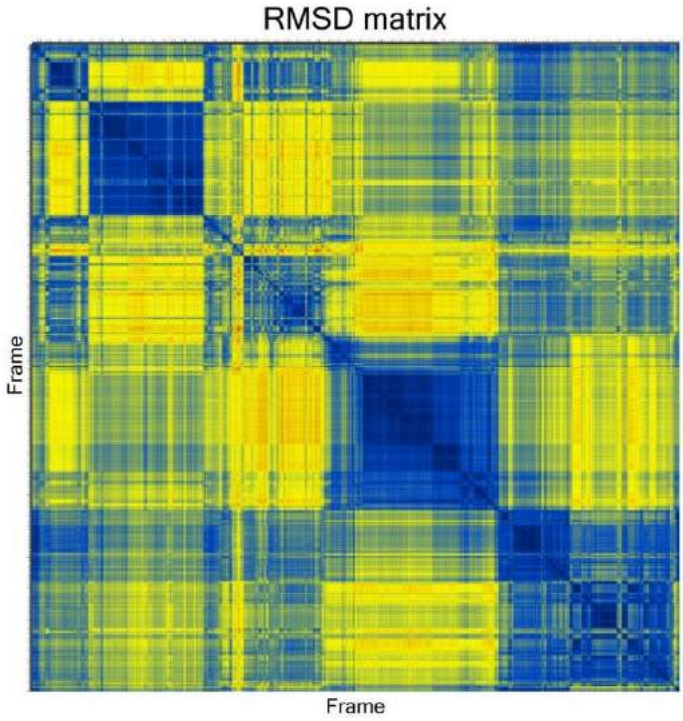

Figure 8 | RMSD matrix using only CA atoms of native-like A31P mutant, excluding tails, with specified maximum RMSD. The color ranges from dark blue (minimum RMSD) to dark red (maximum RMSD: 6.92).

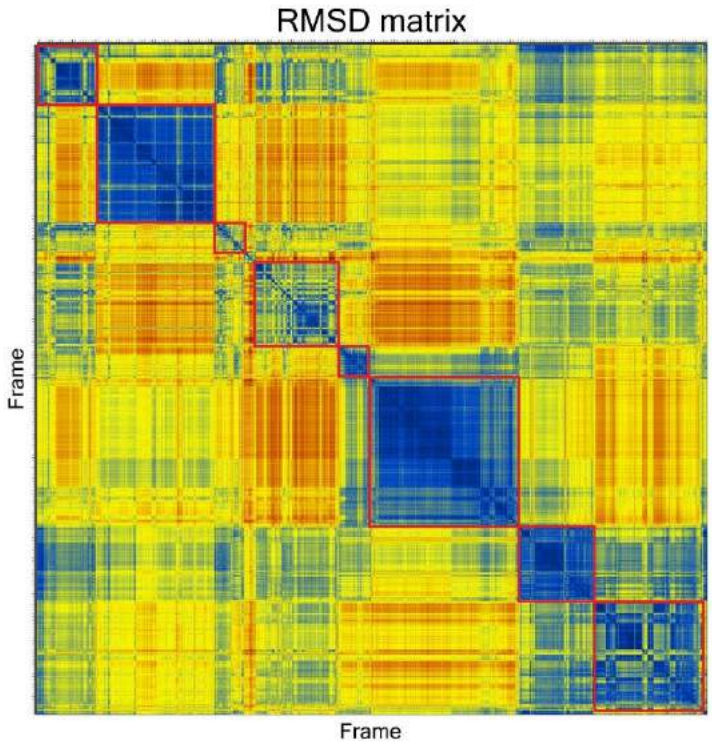

Figure 9 | RMSD matrix using only CA atoms of native-like A31P mutant, excluding tails. The color ranges from dark blue (minimum RMSD) to dark red (maximum RMSD: 5.63). The red lines show the limits of the clusters of frames.

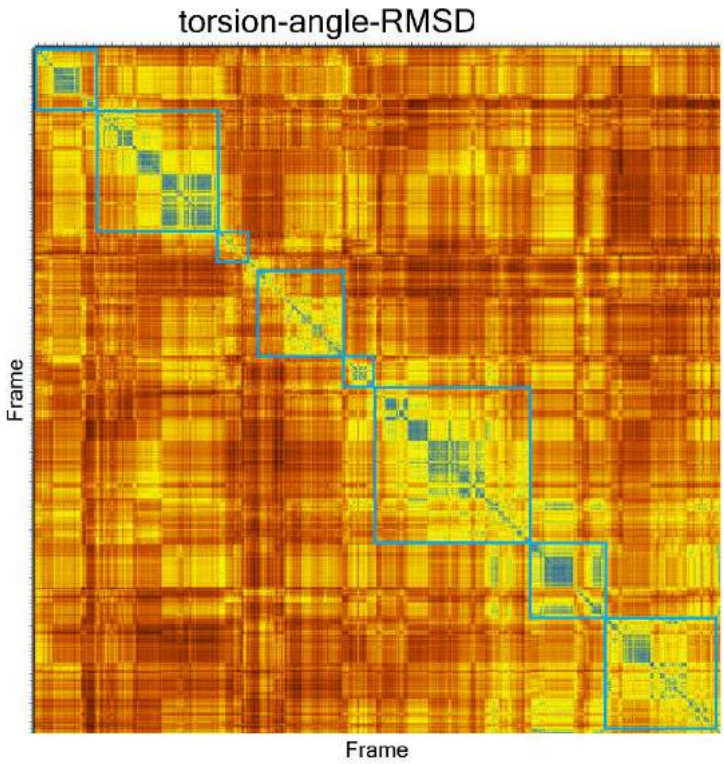

Figure 10 | Torsion-angle-RMSD matrix using backbone atoms of native-like A31P mutant, excluding tails. The color ranges from dark blue (minimum RMSD) to dark red (maximum RMSD). The light blue lines show the limits of the clusters of frames.

3.2.3 Torsion-angle RMSD matrix

The results of the torsion-angle-RMSD matrix calculation did not come in full correspondence with those from the cartesian RMSD matrix. Even though they differed in the scale of RMSD values, the motifs of the clusters observed in the cartesian RMSD matrix were also noticed in the torsion-angle one (*Figure 10*). However, the high noise levels made it hard to understand, and that's why it was not further exploited.

3.2.4 RMSF plot

Moreover, three RMSF plots were created to study the stability of residues in the loop compared to the a-helix. The first plot included residues 1-56 of each chain (*Figure 11*) and the maximum RMSF was 8.5, while the minimum was 0.66. As is depicted in *Figure 11*, the RMSF values of the residues before the first a-helix of each chain were much higher than the others and that is why the region of interest was not presented in detail and no definite conclusions could be drawn about the stability of the loop. Consequently, two separate RMSF plots were created, where, in the first one the maximum value was 2.47 (*Figure 12a*), while in the second one it was 2.28 (*Figure 12b*). Generally, the loop region seemed to have increased values compared to the α-helix.

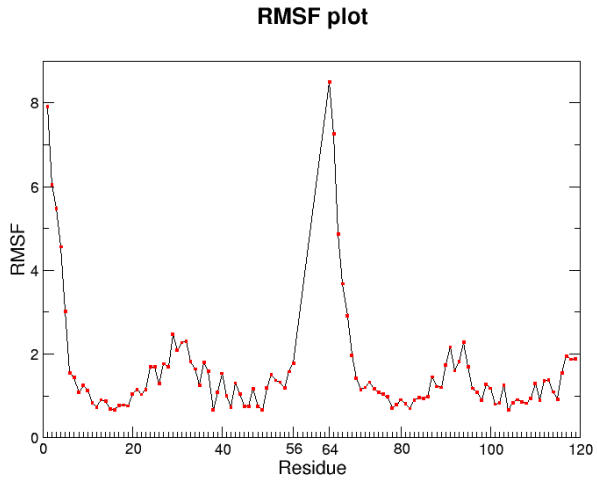

Figure 11 | RMSF plot created by xmgrace plotting tool. The calculation was done using only CA atoms of native-like A31P mutant, excluding tails. The red squares represent the RMSF value of each residue.

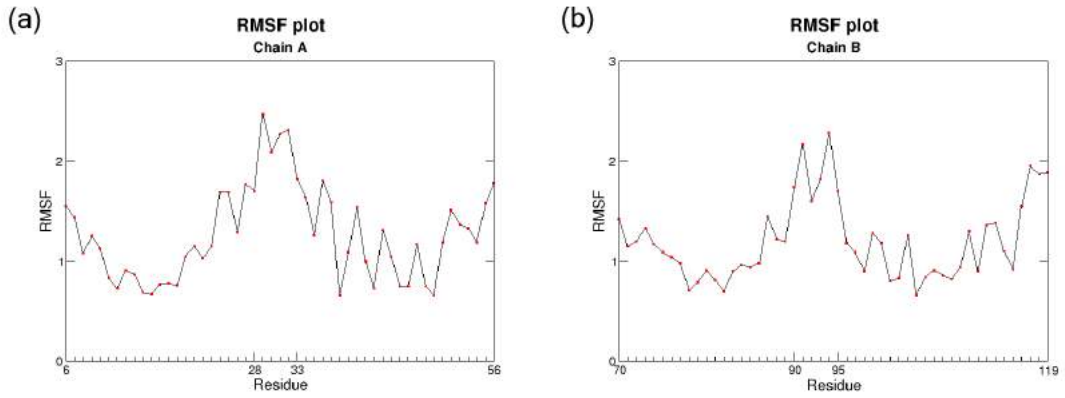

Figure 12 | RMSF plot created by xmgrace plotting tool. The calculation was done using only CA atoms of native-like A31P mutant, excluding tails. (a) residues 6-56 of chain A are presented, and (b) residues 70-119 of the protein, that account for residues 7-56 in chain B. The red squares represent the RMSF value of each residue.

3.2.5 Cartesian PCA

Table 8 and *Figure 14* are a visual representation of the 8 clusters noted in the grcarma window, each depicted in a different color (*Figure 14*), which they varied in terms of the number of frames. Further, after analyzing *Figures 14*, *Table 8*, and the 10 peaks from both PC1-PC2, and PC1-PC3 landscapes (*Figure 13*), it was concluded that some of the clusters (marked with arrows) in the PC landscapes belonged to the same cluster in the plot.

As it has been observed, the size of the clusters varied in the figures due to the different number of frames, as mentioned above. This divergence of the frames could be significant as some clusters were made up of millions of them while others consisted of only a few thousand. Moreover, there was a variance in the similarity of the frames, and thus the color range changed between the clusters. In more detail, in *Figure 13* the ones with darker blue color are structurally more alike than the ones with brighter blue, while in *Figure 14* the colors do not indicate structural features between the frames as the picture shows only the number and size of the clusters.

Table 8 | The clusters and the number of frames in each cluster of cPCA calculated by grcarma, using native-like A31P mutant, excluding tails.

Cluster	Number of frames
Cluster 1	1381639
Cluster 2	1931889
Cluster 3	1081667
Cluster 4	240046
Cluster 5	157611
Cluster 6	65759
Cluster 7	133600
Cluster 8	8902
Total frames in clusters	5001113
Total frames of calculation	10092100

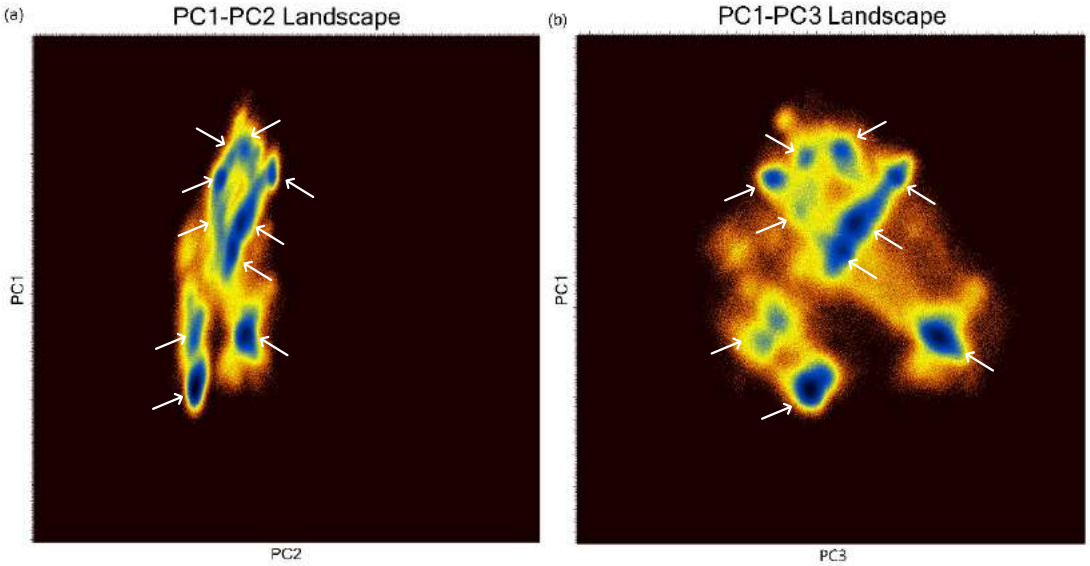

Figure 13 | $PC_{(i)}$-$PC_{(j)}$ landscapes created by cartesian PCA tool of grcarma. The calculation was done using only CA atoms of native-like A31P mutant, excluding tails. The peaks (dark blue), marked with white arrows, represent the clusters calculated by grcarma. (a) PC1-PC2 landscape, and (b) PC1-PC3 landscape.

In conclusion, the data from the cPCA and the RMSD matrix analysis were comparable, as in the second one, excluding tails (*Figures 7*, and *9*) there were 8 clusters, just as many as cartesian PCA isolated. However, there is uncertainty about the similarity between the number of clusters observed in the matrix and those noted in the cPCA, since the count in the matrix was based on the observation.

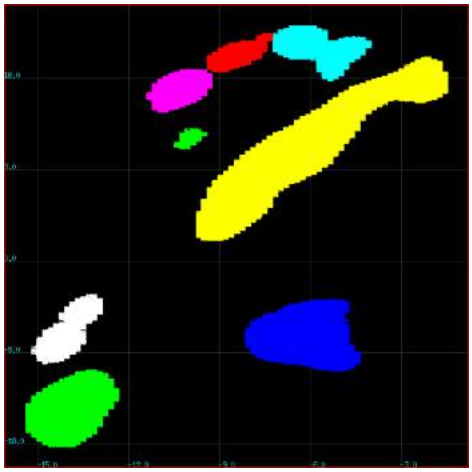

Figure 14 | Plot created from file "cPCA.clusters.dat" performed through grcarma tool, cartesian PCA. The calculation was done using only CA atoms of native-like A31P mutant, excluding tails. The different colors stand for the different clusters.

As for the structures prepared from the cPCA, the representative files contained one structure for each cluster (*Figure 15*), while the superposition files included 500 structures of the cluster prepared through the grcarma tool (*Figure 16*).

The representative structures of the clusters showed some similarities (*Figure 15*), that were mainly localized in the helix regions, due to their stability. Although, the ends were partially diverse in structure and length. The loop regions revealed small structural differences, but they did not have any consequences on the topology of the molecule. The widest variability was observed in the N-terminal regions of each chain because of their high mobility and instability.

In the superposition, the color display used for the images shows the mobility of the atoms and ranges from red to blue (*Figure 16*). In more detail, high mobility atoms were represented with warm colors, while low with colder ones. The helices in all the images were dark blue, meaning they were stable. In some clusters, the ends had significantly higher mobility than in others, as in some images, red regions appeared. Also, there were differences in the loops' stability, proved by the color range in each image. Some of them had only dark blue (*Figure 16b*), while others had cyan (*Figure 16d,* and *f*), green (*Figure 16c, e, g,* and *h*), or even yellow (*Figure 16a*). In conclusion, the region of interest, i.e., the loop, was not stable in most of the native-like A31P mutant frames as demonstrated by this analysis.

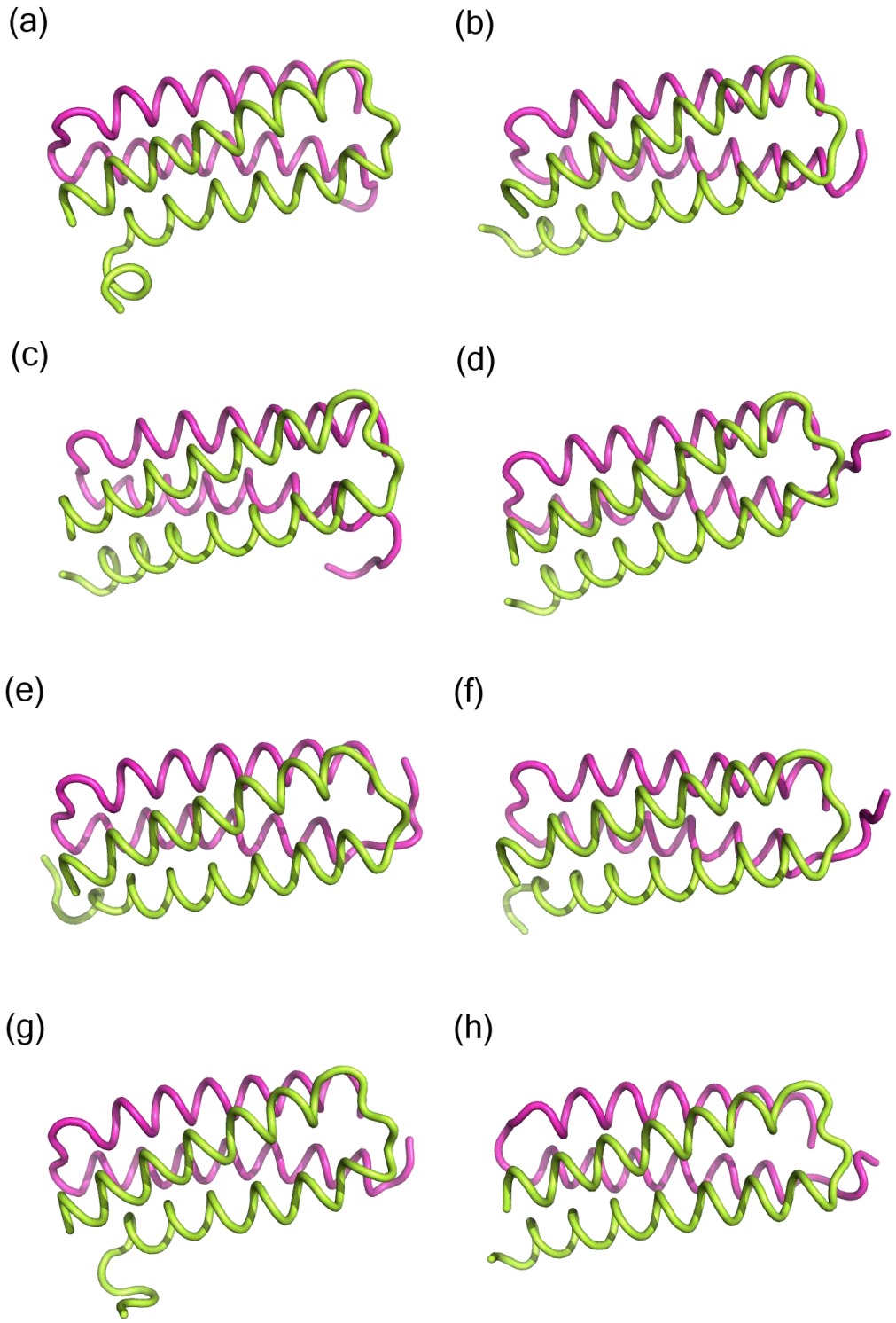

Figure 15 | Representative structures of clusters created by the cartesian PCA, in grcarma, reproduced with PyMol program. The calculation was done using only CA atoms of native-like A31P mutant, excluding tails. Coding: green for chain A, and magenta for chain B. (a) cluster 1, (b) cluster 2, (c) cluster 3, (d) cluster 4, (e) cluster 5, (f) cluster 6, (g) cluster 7, and (h) cluster 8.

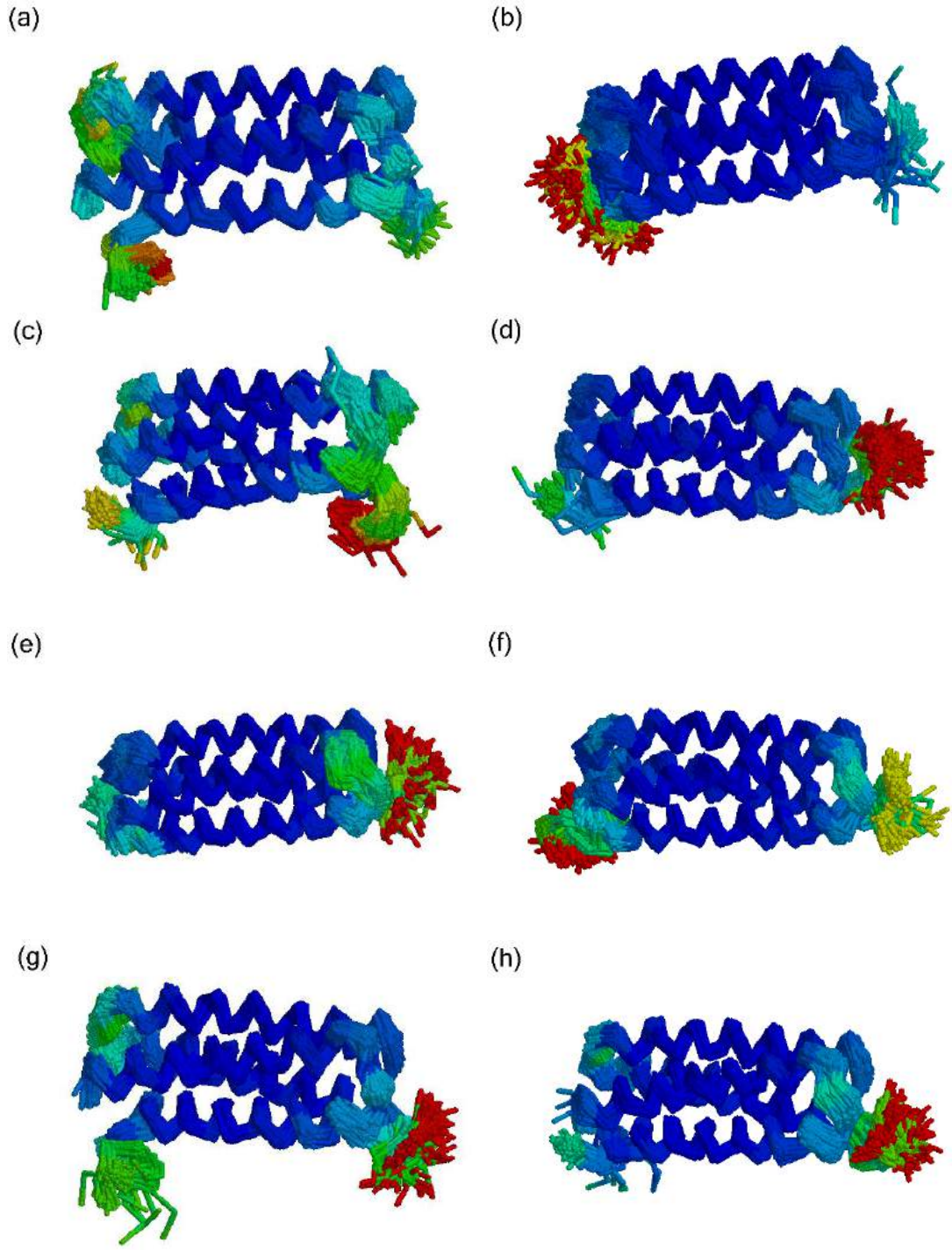

Figure 16 | Superposition structures, containing 500 structures each, of clusters created by the cartesian PCA, in grcarma, reproduced with RasMol program. The calculation was done using only CA atoms of native-like A31P mutant, excluding tails. Color coding: temperature, with a color range from dark blue (low mobility) to red (high mobility). (a) cluster 1, (b) cluster 2, (c) cluster 3, (d) cluster 4, (e) cluster 5, (f) cluster 6, (g) cluster 7, and (h) cluster 8.

3.2.6 Dihedral PCA

After the dPCA, 25 clusters were noted as the output of the grcarma window. The number of frames varied significantly between the clusters, specifically, from hundreds to thousands of frames (*Table 9*, and *Figure 17*). In the PC1-PC2, PC1-PC3, and PC2-PC3 landscapes 22, 18, and 19 peaks were observed, respectively (*Figure 18*). It was noticed that certain clusters had been missed from the landscapes, while some others that were marked with white arrows in the figures belonged to the same cluster in the plot. As explained in the cPCA chapter the size and the color range of the clusters were different, due to the number and the similarity of the frames in each cluster.

Table 9 | The clusters and the number of frames in each cluster of dPCA calculated by grcarma, using native-like A31P mutant, excluding tails.

Cluster	Number of frames
Cluster 1	462486
Cluster 2	593079
Cluster 3	180249
Cluster 4	236263
Cluster 5	207507
Cluster 6	80940
Cluster 7	94445
Cluster 8	74964
Cluster 9	158617
Cluster 10	107376
Cluster 11	137988
Cluster 12	38384
Cluster 13	13086
Cluster 14	29971
Cluster 15	15286
Cluster 16	20264
Cluster 17	11045
Cluster 18	8609
Cluster 19	7246
Cluster 20	13082
Cluster 21	9117
Cluster 22	11241
Cluster 23	4398
Cluster 24	2005
Cluster 25	210
Total frames in clusters	2517858
Total frames of calculation	10092100

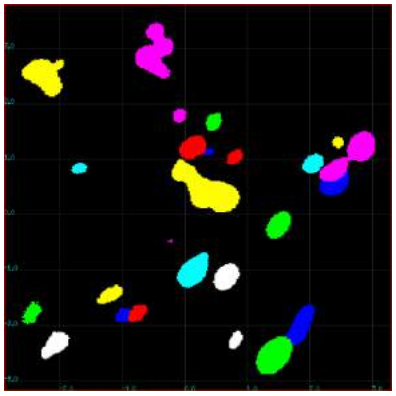

Figure 17 | Plot created from file "dPCA.clusters.dat" performed through grcarma tool, dihedral PCA. The calculation was done using backbone atoms of native-like A31P mutant, excluding tails. The different colors stand for the different clusters.

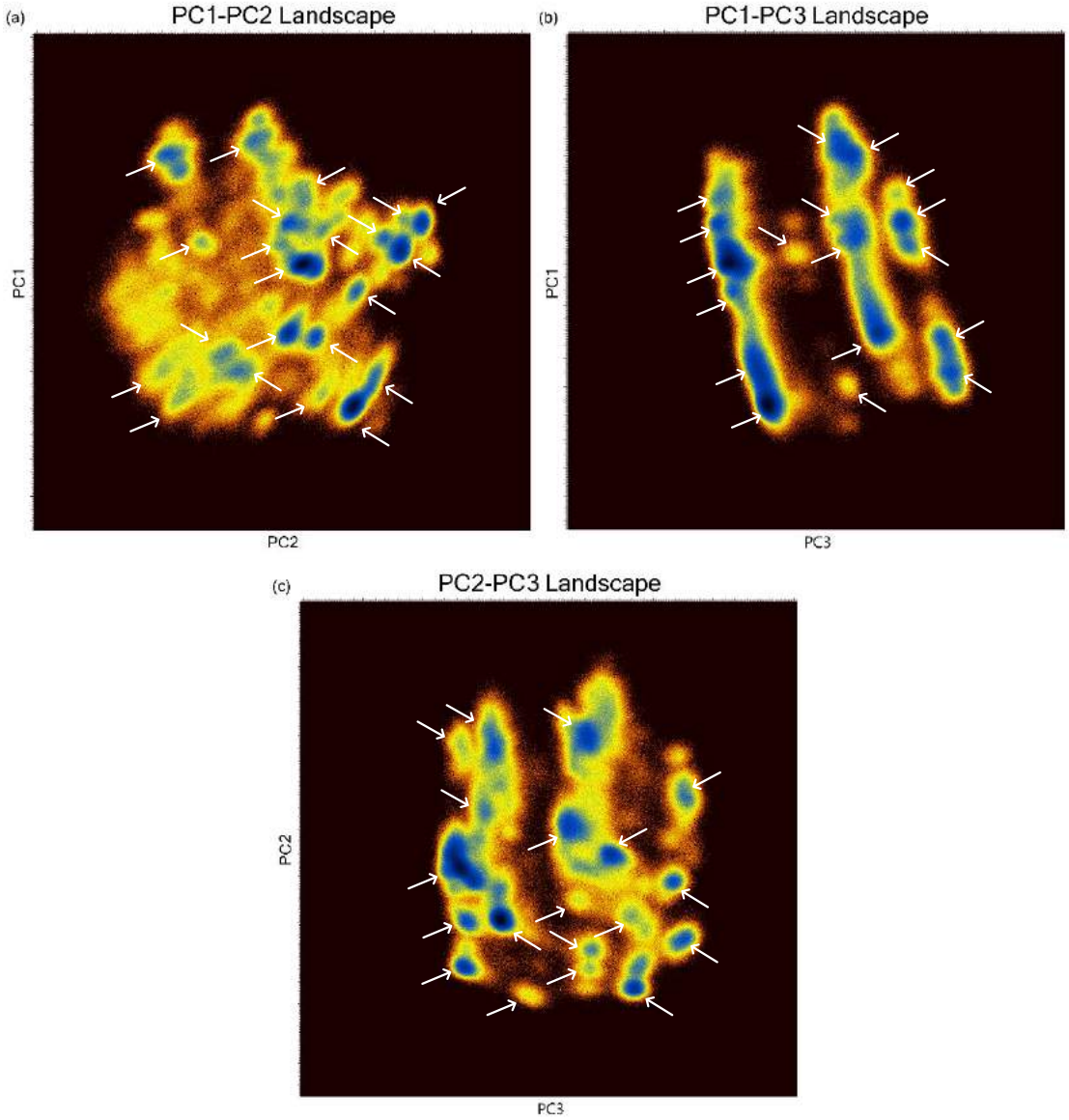

Figure 18 | $PC_{(i)}$-$PC_{(j)}$ landscapes created by dihedral PCA tool of grcarma. The calculation was done using backbone atoms of native-like A31P mutant, excluding tails. The peaks (dark blue), marked with white arrows, represent the clusters calculated by grcarma. (a) PC1-PC2 landscape, (b) PC1-PC3 landscape, and (c) PC2-PC3 landscape.

The results of the dPCA and the RMSD matrix analysis disagreed, as in the RMSD matrix, excluding tails (*Figures 7,* and *9*) there were 8 clusters noticed, in contrast to the 25 clusters noted by grcarma. The reasons why this happened were that in the RMSD matrix calculation the cartesian coordinates of the CA atoms, whereas in the dihedral PCA the dihedral angles of the backbone atoms were used, and the clusters in the matrix were visually selected.

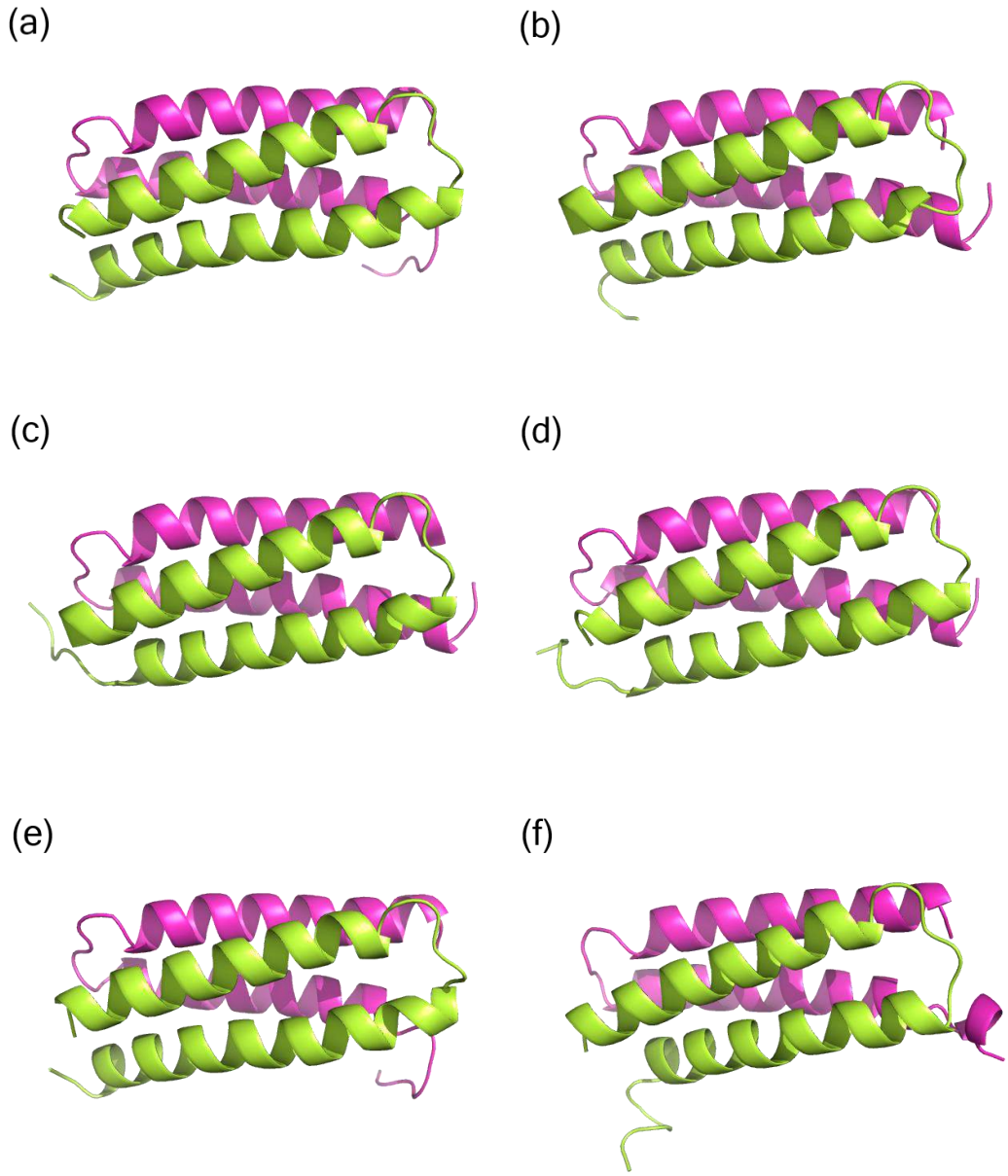

Figure 19 | Representative structures of clusters created by the dihedral PCA, in grcarma, reproduced with PyMol program. The calculation was done using backbone atoms of native-like A31P mutant, excluding tails. Coding: green for chain A, and magenta for chain B. (a) cluster 1, (b) cluster 6, (c) cluster 11, (d) cluster 15, (e) cluster 21, and (f) cluster 25.

The representative structures of the clusters showed differences, localized in the N-terminal regions, due to their instability, and some of the figures are presented indicatively in *Figure 19*. Particularly, a small α-helix was featured in the N-terminal region in chain B of cluster 25 (*Figure 19f*), congruent with the secondary structure results in chapter 3.2.1 (*Figure 5*). Apart from this, the topology of the molecule was maintained. Moreover, the structure of the loop regions was not the same, but it did not cause changes in the topology of the molecule. Though, the α-helices seemed considerably similar between the clusters.

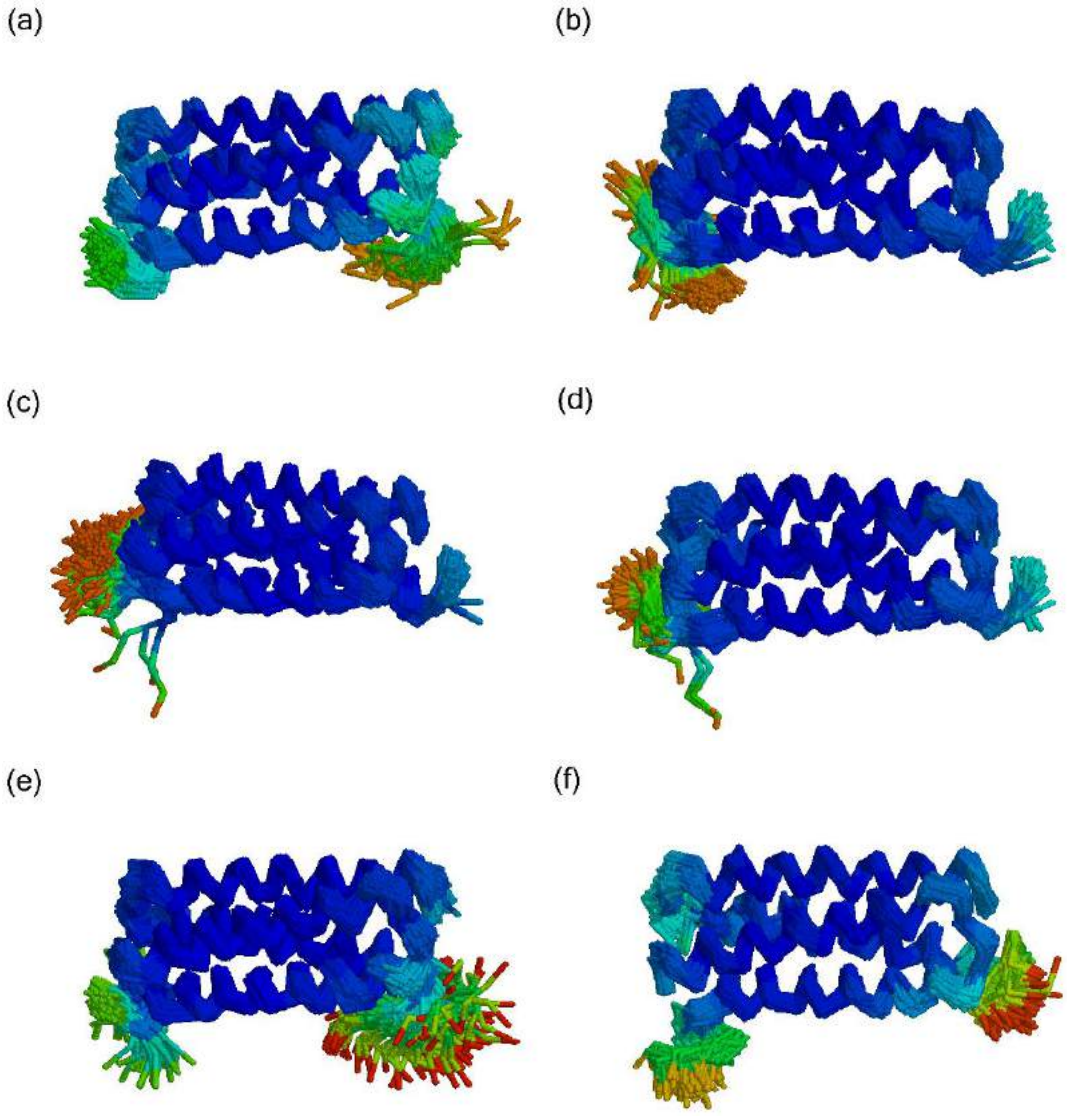

Figure 20 | Superposition structures, containing 500 structures each, of clusters created by the dihedral PCA, in grcarma, reproduced with RasMol program. The calculation was done using backbone atoms of native-like A31P mutant, excluding tails. Color coding: temperature, with a color range from dark blue (low mobility) to red (high mobility). (a) cluster 1, (b) cluster 6, (c) cluster 11, (d) cluster 15, (e) cluster 21, and (f) cluster 25.

In the superposition structures, the color range, from red to blue, was used for the images showing the mobility of the atoms (*Figure 20*). The warm colors represented high mobility atoms, while the colder, low mobility atoms. In all the images, the helices were highly stable, but the N-terminal residues had high mobility, as evidenced by the warmer colors that appeared in the figures. Also, the loops' stability varied, as shown by the presence of blue, cyan (*Figure 20e,* and *f*), and even green colors (*Figure 20a*), in some clusters, whereas in others only blue was noticed (*Figure b, c,* and *d*). Overall, the loop region was slightly unstable in most of the native-like A31P mutant frames as established by the dPCA analysis.

3.2.7 Comparison between cPCA & dPCA

The cPCA was done using the cartesian coordinates of the CA atoms, while the dPCA the dihedral angles of the backbone atoms of the molecule. This caused the different representations in the representative structures of the two calculations, as in the cartesian PCA only the CA atoms were included, providing insufficient information about the secondary structure of the molecule. As explained above, both calculations prepared clusters consisting of a different number of frames. The results, for example, the number of the clusters, may diverge, due to the different parameters used.

More specifically, the cPCA and the dPCA isolated only 8, and 25 clusters, respectively, and the first had a wider range of frames per cluster than the first. Thus, the total number of frames included in the clusters of the cPCA was bigger than that of the dPCA (*Tables 8,* and *9*).

The representative and superposition structures of the two analyses did not present visible, significant differences, except for the extra helix in the N-terminal region of chain B in cluster 25 of the dPCA analysis in the representative structure (*Figure 19f*). Generally, the superposition structures of the dPCA were stabler than of the cPCA, indicated by the color range in *Figures 16* and *20*.

3.3 Native ROP

3.3.1 Secondary structure

The secondary structure analysis confirmed that the main secondary structure of the molecule was α-helix, as known from the bibliography [9,10,11]. Although, in some regions, turns and random coils were observed, as shown in *Figures 21, 22, 23,* and *24*, using WebLogo and STRIDE, respectively. At the beginning and the end of each chain, there was some variability, especially in residues 1-5 and 55-63 in both chains, due to the mobility of the turns.

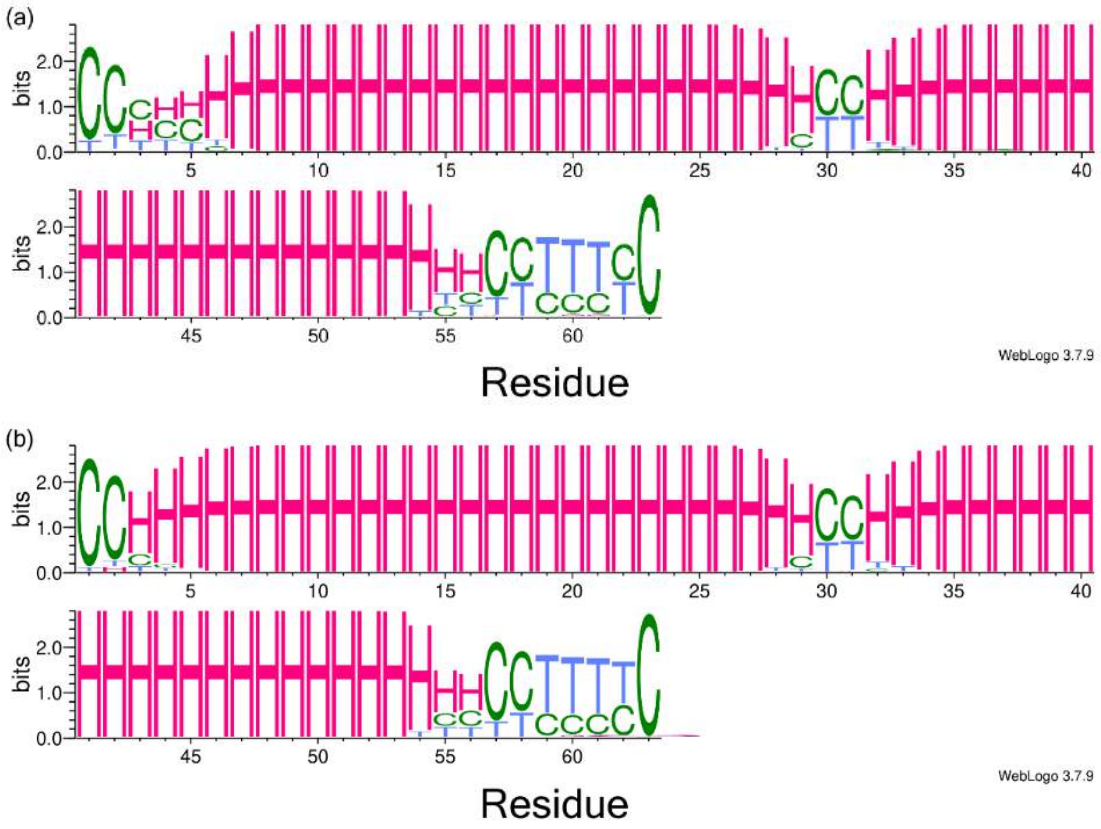

Figure 21 | Secondary structure analysis of native ROP protein, including tails, using WebLogo. Coding: H for α-helices, G for 3_{10}–helices, T for turns, and C for random coils.

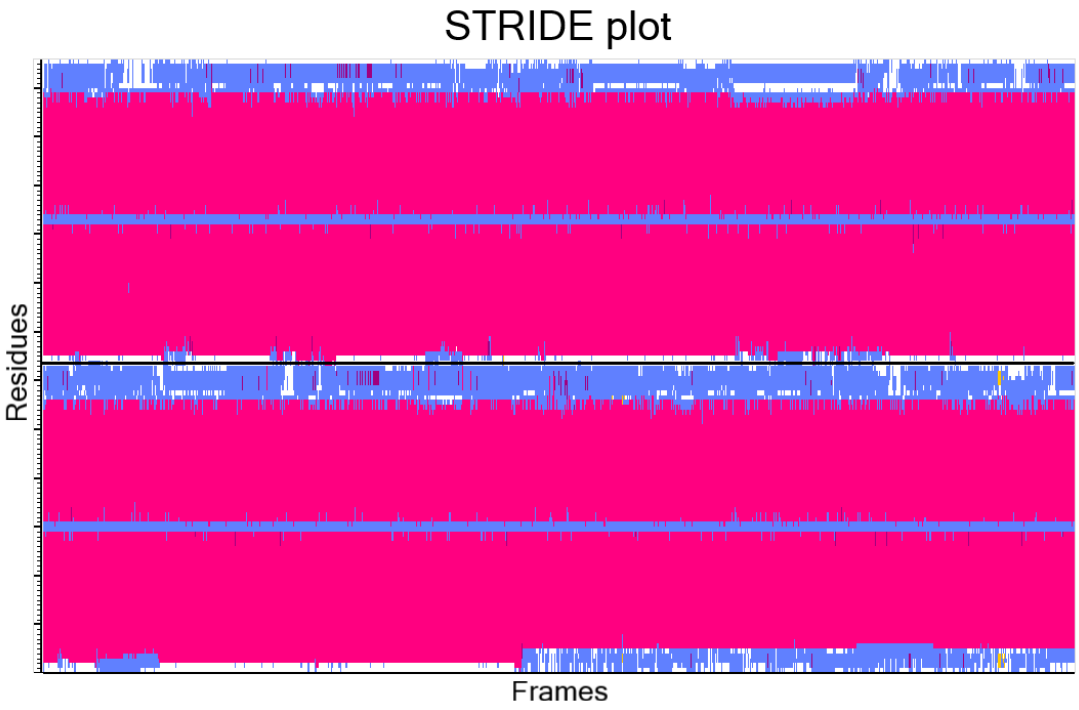

Figure 22 | Secondary structure analysis generated by the program STRIDE, using native ROP, including tails. Coding: pink for α-helices, purple for 3_{10} –helices, dark purple for Pi-helices, yellow for β-sheets, blue for β/G-turns, and white for coils/unassigned. The black line shows the limits of the chains, in the lower half chain A and the upper half chain B.

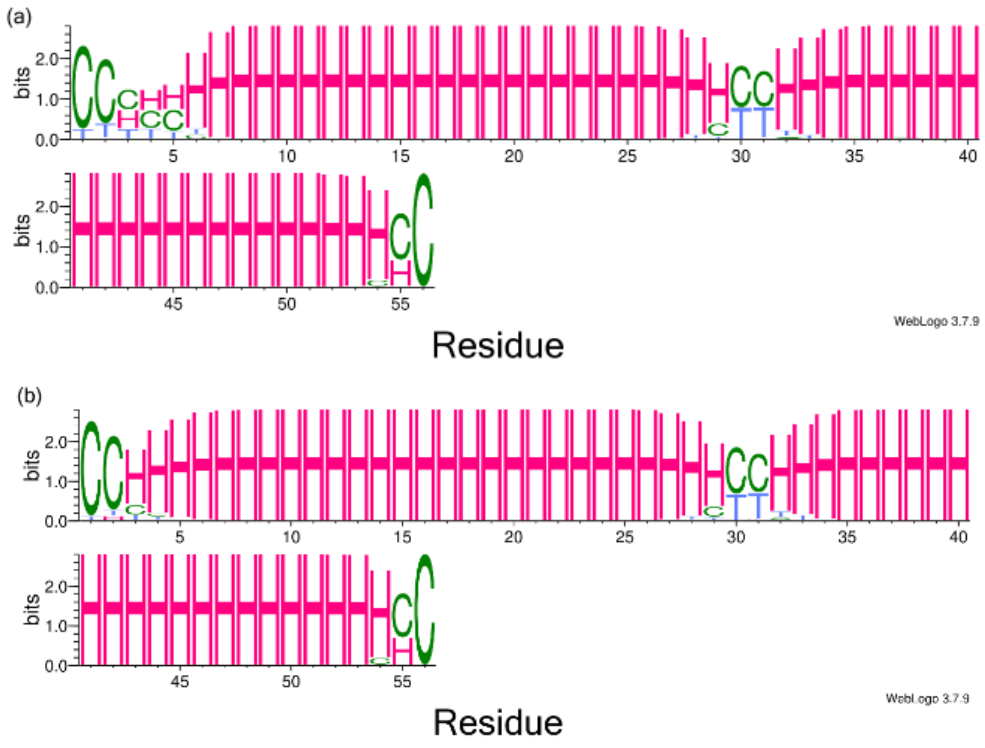

Figure 23 | Secondary structure analysis of native ROP protein, excluding tails, using WebLogo. Representation of (a) chain A, and (b) chain B. Coding: H for α-helices, G for 3_{10}–helices, T for turns, and C for random coils.

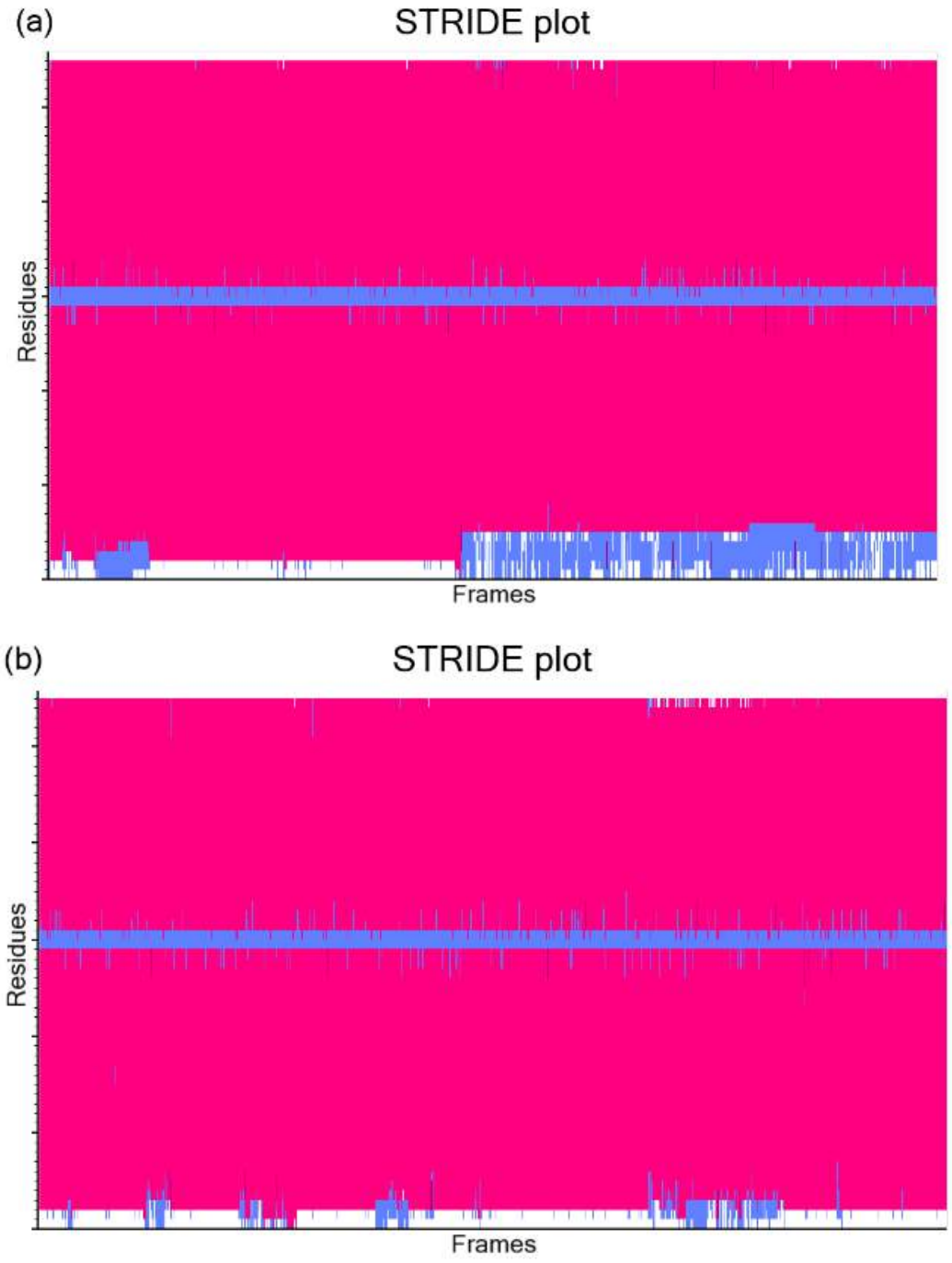

Figure 24 | Secondary structure analysis generated by the program STRIDE, using native ROP, excluding tails. Representation of (a) chain A, and (b) chain B. Coding: pink for α-helices, purple for 3_{10}-helices, dark purple for Pi-helices, yellow for β-sheets, blue for β/G-turns, and white for coils/unassigned.

Moreover, the WebLogo graphs revealed limited diversity in the loop region, specifically, residues 29-32 chain A (*Figure 23a*), and chain B (*Figure 23b*). Although,

according to the STRIDE plots in *Figure 24*, in both chains, the loop region appeared just a little less stable than the helices.

3.3.2 Cartesian RMSD matrix

The RMSD matrix calculation, including tails, resulted in a maximum RMSD value of 6.12 (*Figure 25*). This value was decreased in the excluding tails calculation, due to the disordered tails, to 4.66. Additionally, one more RMSD matrix, excluding tails was generated with a specified maximum value equal to the including tails, to compare their results. It was verified that the tails were disordered and caused some noise to the computations, and thus, they should be eliminated from the analyses. So, the first and the third matrix (*Figures 25,* and 27) were not useful for the rest analyses.

Clustering the frames of the excluding tails RMSD matrix, 3 groups were developed (*Figure 28*). At the upper-left corner, a cluster containing a really large number of frames was observed, while at the center and the lower-right corner two clusters with a decreasing size were.

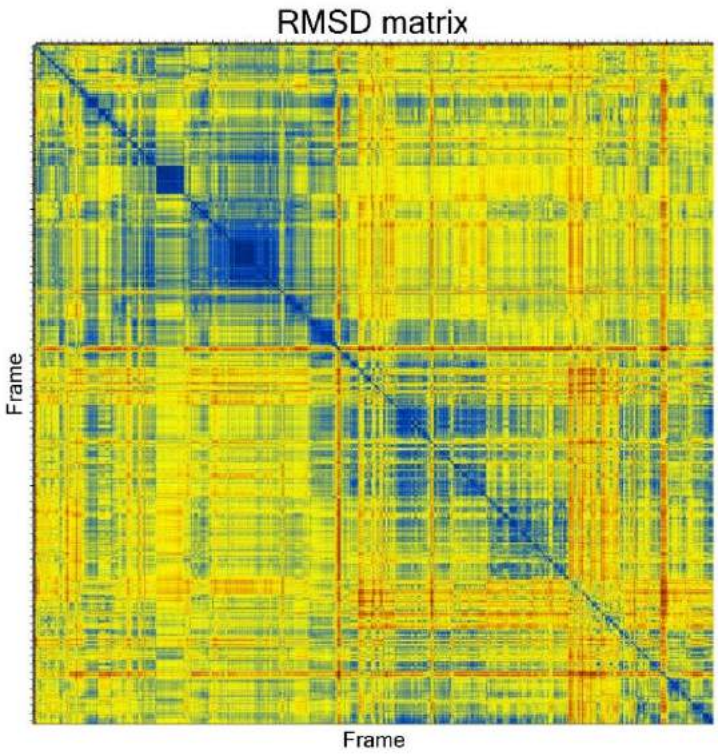

Figure 25 | RMSD matrix using only CA atoms of native ROP, including tails. The color ranges from dark blue (minimum RMSD) to dark red (maximum RMSD: 6.12).

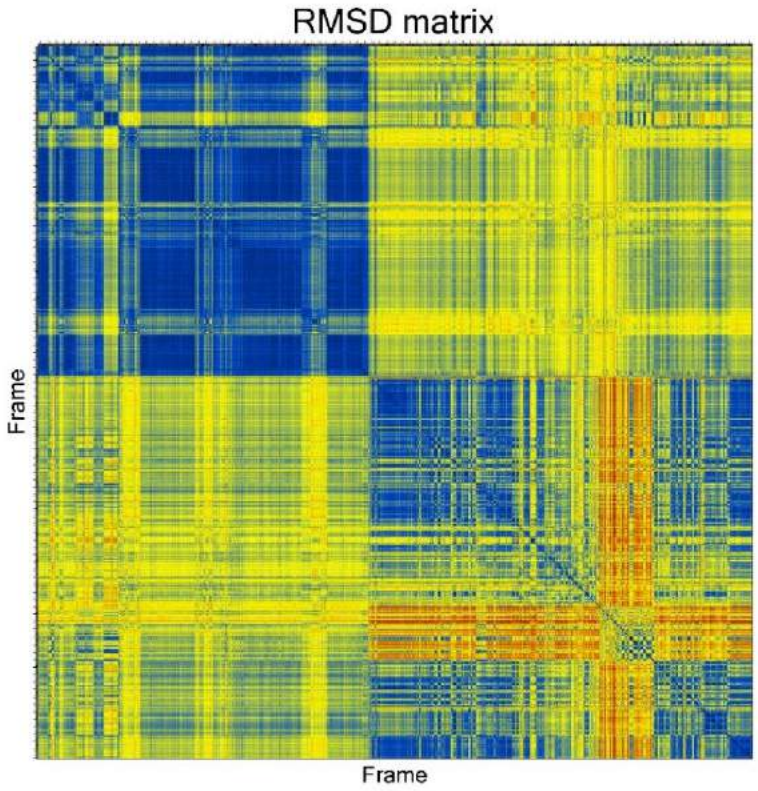

Figure 26 | RMSD matrix using only CA atoms of native ROP, excluding tails. The color ranges from dark blue (minimum RMSD) to dark red (maximum RMSD: 4.66).

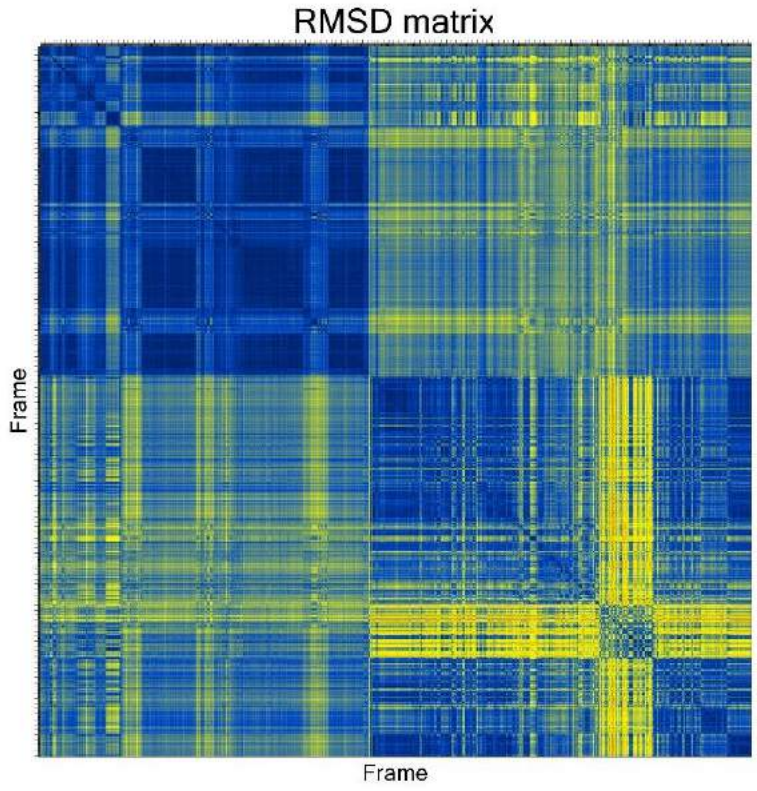

Figure 27 | RMSD matrix using only CA atoms of native ROP, excluding tails, with specified maximum RMSD. The color ranges from dark blue (minimum RMSD) to dark red (maximum RMSD: 6.12).

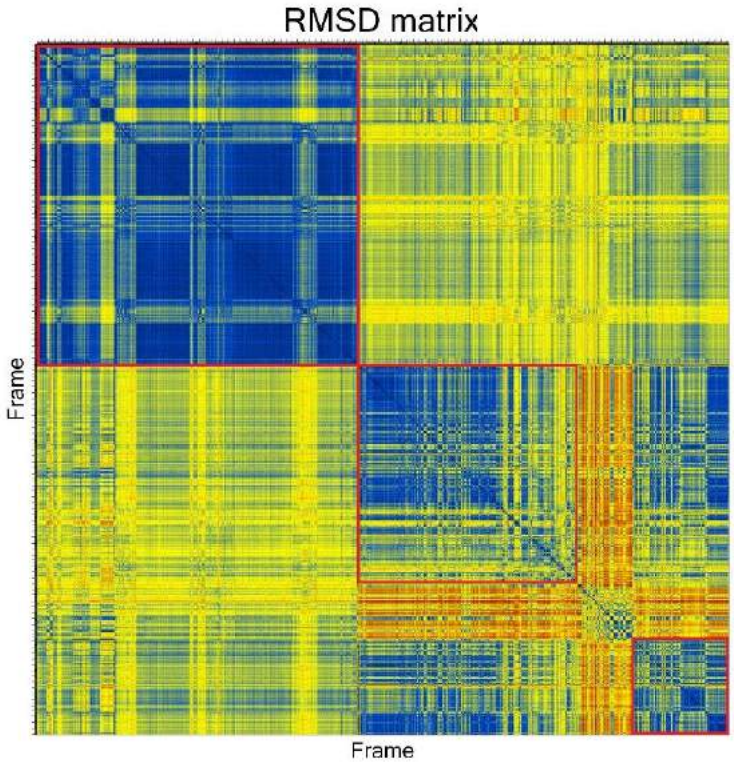

Figure 28 | RMSD matrix using only CA atoms of native ROP, excluding tails. The color ranges from dark blue (minimum RMSD) to dark red (maximum RMSD: 4.66). The red lines show the limits of the clusters of frames.

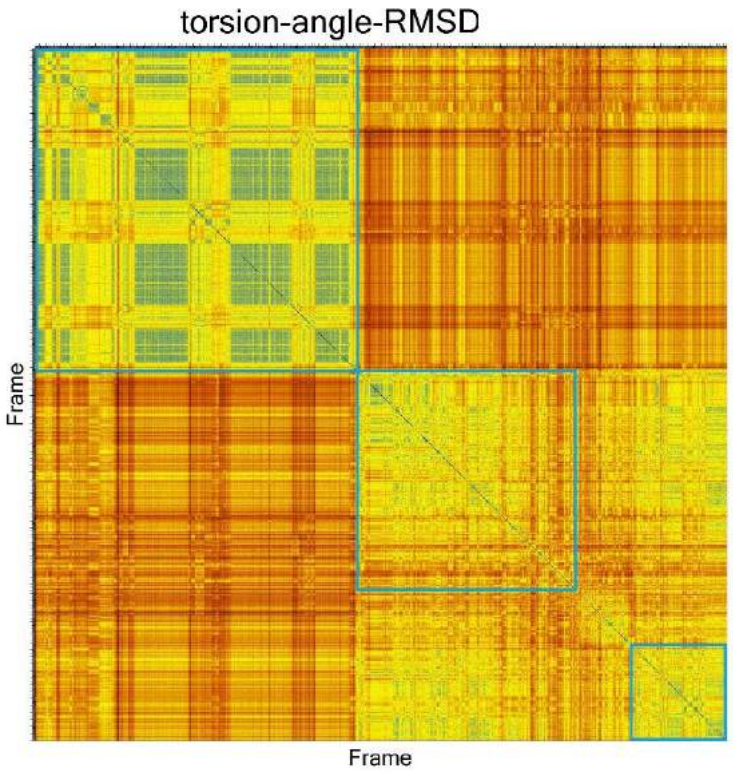

Figure 29 | Torsion-angle-RMSD matrix using backbone atoms of native ROP, excluding tails. The color ranges from dark blue (minimum RMSD) to dark red (maximum RMSD). The light blue lines show the limits of the clusters of frames.

3.3.3 Torsion-angle RMSD matrix

The torsion-angle RMSD matrix (*Figure 29*), featured the same clusters as the cartesian RMSD matrices. However, the scale was much different, and the data generated were noisy. This made the results unclear, so, it was pointless to use them for further analysis.

3.3.4 RMSF plot

The RMSF plots were prepared to estimate the mobility and instability of the loops. In the first plot, with a maximum RMSF value of 7.31 (*Figure 30*), the residues at the N-terminal end of each chain had high mobility, and they should be removed to make the studying of the loop easier. Therefore, two new plots were generated for each chain excluding the N-terminal ends. The maximum RMSF noted in the RMSF plot of chain A was 2.09 (*Figure 31a*), and the maximum value of chain B was 1.63 (*Figure 31b*). As shown in the plots, these values were close to the ones of the residues containing the helices, meaning that the loops are almost as stable as the helices.

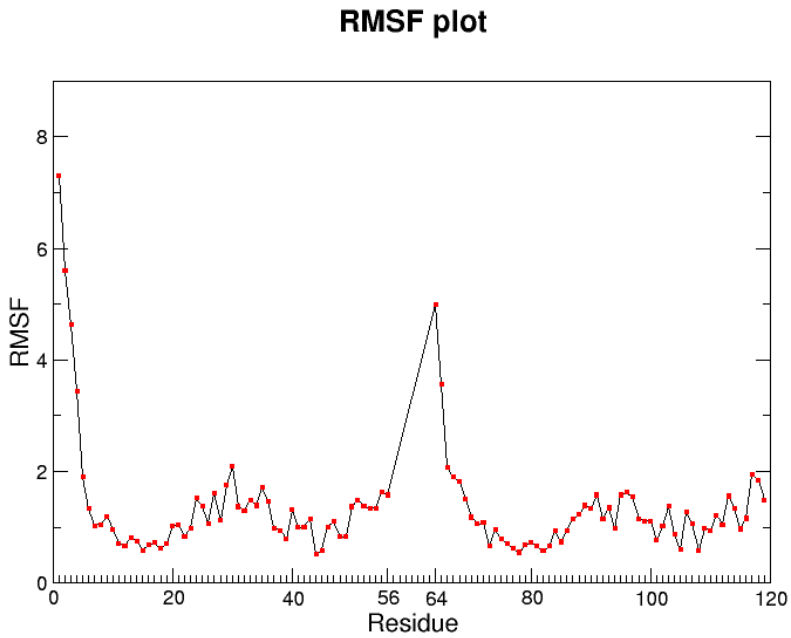

Figure 30 | RMSF plot created by xmgrace plotting tool. The calculation was done using only CA atoms of native ROP, excluding tails. The red squares represent the RMSF value of each residue.

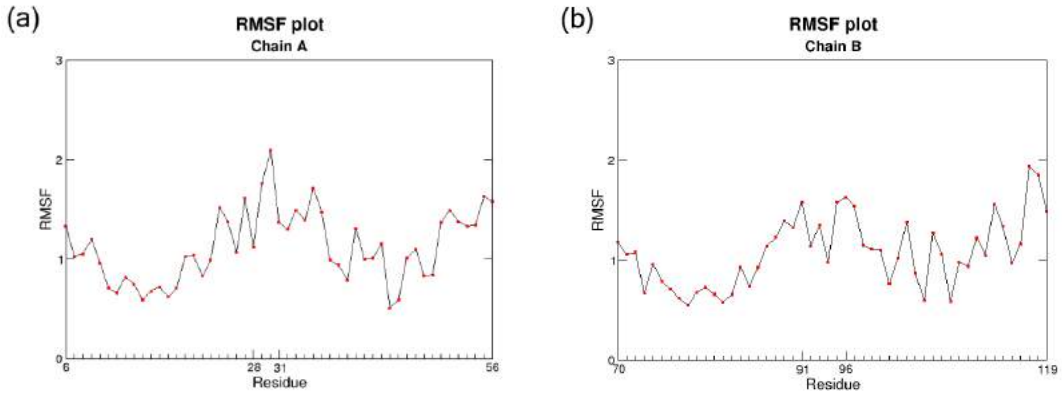

Figure 31 | RMSF plot created by xmgrace plotting tool. The calculation was done using only CA atoms of native ROP, excluding tails. (a) residues 6-56 of chain A are presented, and (b) residues 70-119 of the protein, that account for residues 7-56 in chain B. The red squares represent the RMSF value of each residue.

3.3.5 Cartesian PCA

According to the cPCA, 2 clusters were isolated containing almost 3 million frames the first and 8 hundred thousand the second (*Table 10,* and *Figure 33*). The number of frames in both clusters amplified the fact that the structure is stable. There was a setback during the preparation of the landscapes, that's why they have a slightly different scale than the other PC landscapes (*Figure 32*), while, in this display, the peaks were pointed with dark red color instead of dark blue. Although, 2 peaks were observed in all the PC landscapes, marked with white arrows, as many as those noted in the grcarma window.

Table 10 | The clusters and the number of frames in each cluster of cPCA calculated by grcarma, using native ROP, excluding tails.

Cluster	Number of frames
Cluster 1	2963329
Cluster 2	823041
Total frames in clusters	3786370
Total frames of calculation	10000000

The RMSD matrix analysis resulted in 3 clusters (*Figure 28*), in comparison to the 2 calculated from the cPCA. So, the results of the two computations were not consistent. Although, the cluster observed at the upper-left corner of the matrix could

be equivalent to the first cluster noted in the grcarma window. The reason why there were differences was not detected.

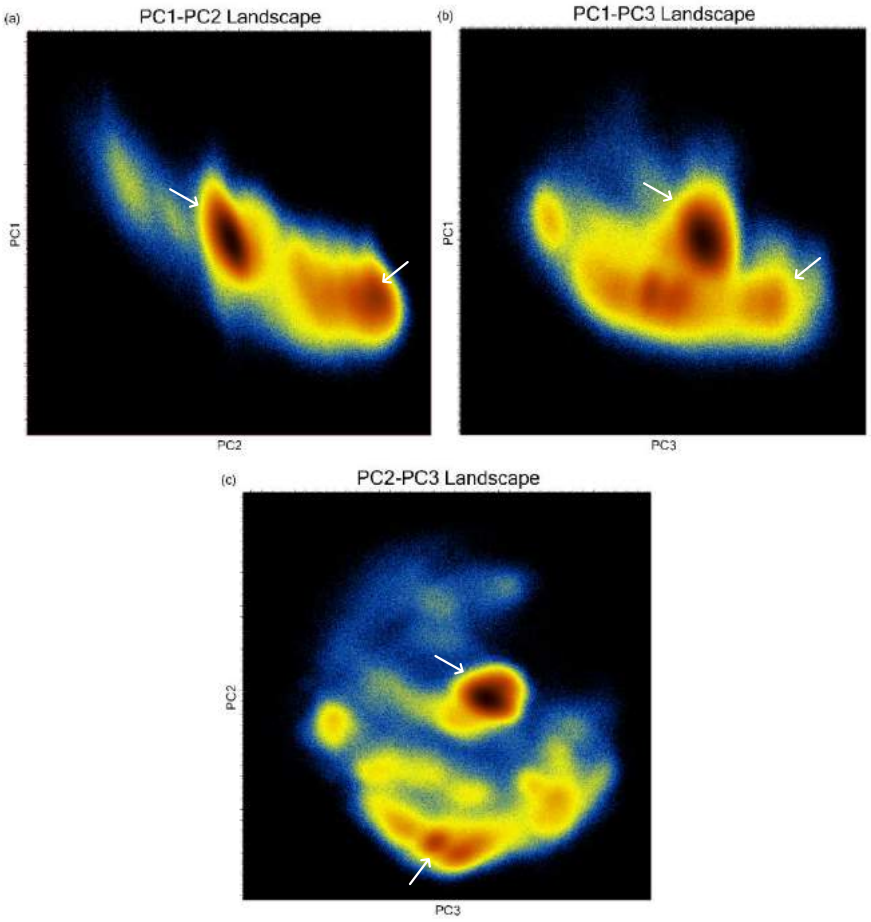

Figure 32 | $PC_{(i)}$-$PC_{(j)}$ landscapes created by cartesian PCA tool of grcarma. The calculation was done using only CA atoms of native ROP, excluding tails. The peaks (dark red), marked with white arrows, represent the clusters calculated by grcarma. (a) PC1-PC2 landscape, (b) PC1-PC3 landscape, and (c) PC2-PC3 landscape.

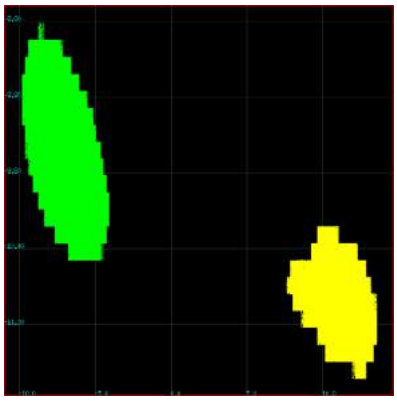

Figure 33 | Plot created from file "cPCA.clusters.dat" performed through grcarma tool, cartesian PCA. The calculation was done using only CA atoms of native ROP, excluding tails. The different colors stand for the different clusters.

To visualize the clusters, representative and superposition structures of the 2 clusters were prepared (*Figures 34,* and *36*). Between the representative ones, no significant diversity was revealed, except for the N-terminal regions. This was confirmed by the superimposed structure (*Figure 35*), in which the two structures did not overlap only in the N-terminal regions of both chains. In the superposition structures (*Figure 36*), the warm colors showed mobility and instability, limited in the terminal regions. In contrast, the loops featured high stability proved by the dark blue color in cluster 1 (*Figure 36a*), which was as dark as the color of the helices, and the lighter blue in cluster 2 (*Figure 36b*), indicating a little lower stability.

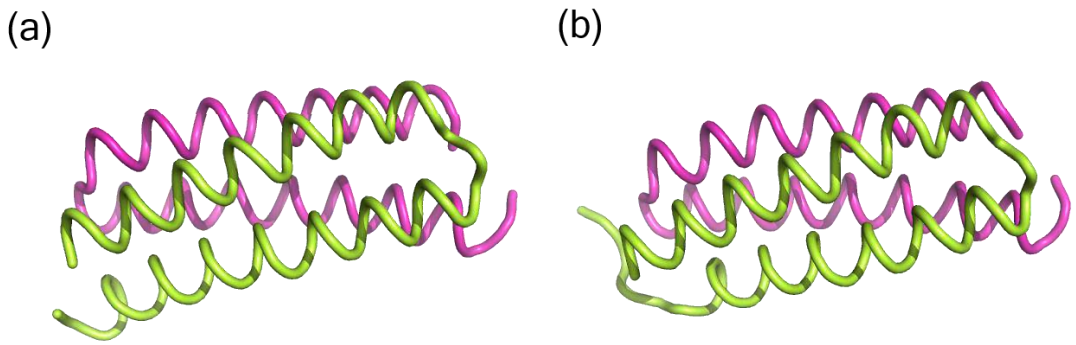

Figure 34 | Representative structures of clusters created by the cartesian PCA, in grcarma, reproduced with PyMol program. The calculation was done using only CA atoms of native ROP, excluding tails. Coding: green for chain A, and magenta for chain B. (a) cluster 1, and (b) cluster 2.

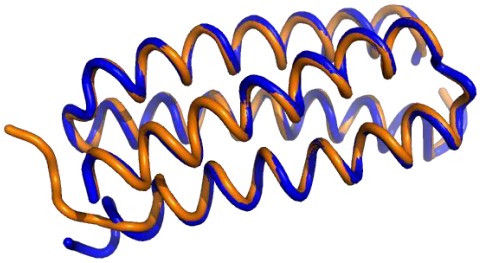

Figure 35 | Superimposed structure prepared with PyMol, comparing the representative structure of cluster 1 (represented in blue), and cluster 2 (represented in orange).

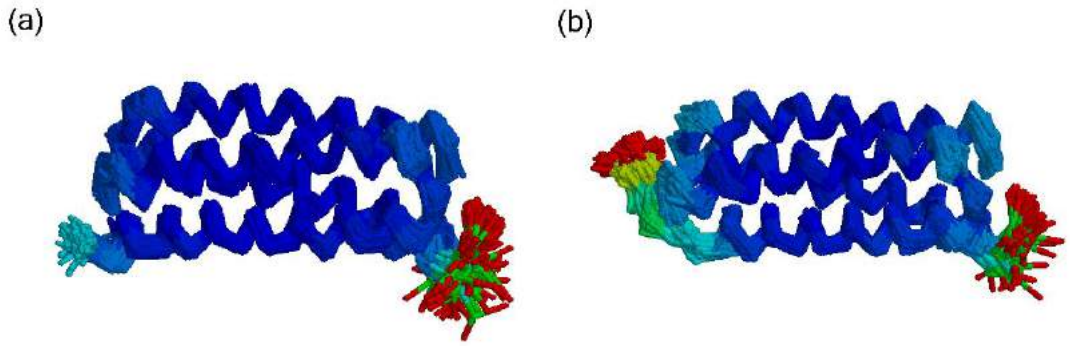

Figure 36 | Superposition structures, containing 500 structures each, of clusters created by the cartesian PCA, in grcarma, reproduced with RasMol program. The calculation was done using only CA atoms of native ROP, excluding tails. Color coding: temperature, with a color range from dark blue (low mobility) to red (high mobility). (a) cluster 1, and (b) cluster 2.

3.3.6 Dihedral PCA

The grcarma calculated 3 clusters (*Table 11,* and *Figure 38*), with a significant variance in the number of frames in each, as proved by their size in *Figure 38*, in which each color stands for a cluster. By contrast in the PC landscapes, more than 3 clusters were noticed (*Figure 37*). A possible reason for this was that in some clusters observed in the PC landscapes, the frames were not similar enough to generate clusters detected by the computation.

Comparing the results from the dPCA and RMSD matrix (*Figure 28*), they both ended up with 3 clusters, 1 of them containing a really large number of frames, and 2 others containing a significantly smaller one. In contrast to the dPCA, which was based on calculation, the RMSD matrix cluster analysis was observationally defined, thus, the clusters marked were not totally certain.

Table 11 | The clusters and the number of frames in each cluster of dPCA calculated by grcarma, using native ROP, excluding tails.

Cluster	Number of frames
Cluster 1	2731800
Cluster 2	71088
Cluster 3	81526
Total frames in clusters	2884414
Total frames of calculation	10000000

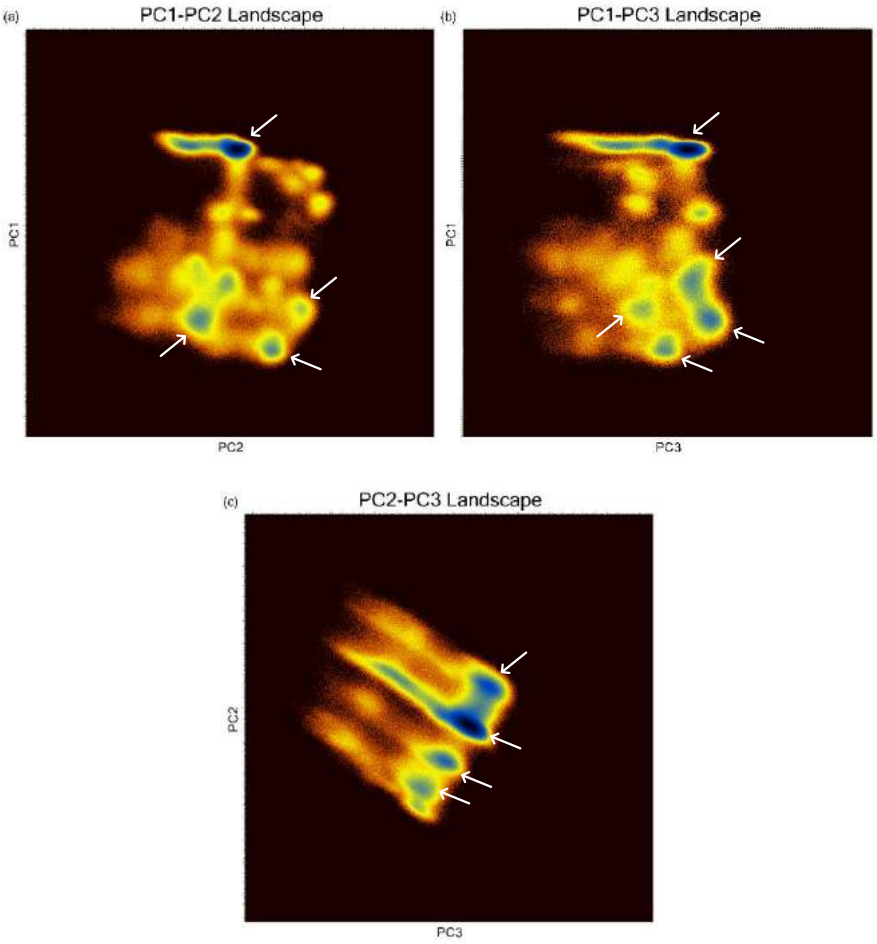

Figure 37 | $PC_{(i)}$-$PC_{(j)}$ landscapes created by dihedral PCA tool of grcarma. The calculation was done using backbone atoms of native ROP, excluding tails. The peaks (dark blue), marked with white arrows, represent the clusters calculated by grcarma. (a) PC1-PC2 landscape, (b) PC1-PC3 landscape, and (c) PC2-PC3 landscape.

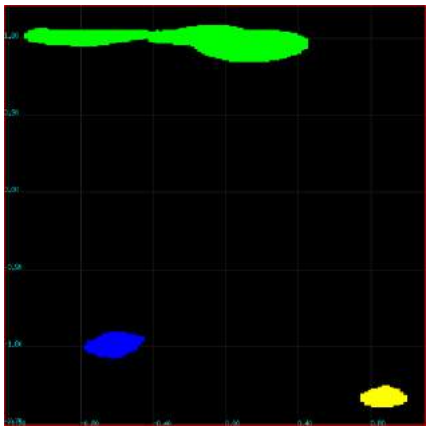

Figure 38 | Plot created from file "dPCA.clusters.dat" performed through grcarma tool, dihedral PCA. The calculation was done using backbone atoms of native ROP, excluding tails. The different colors stand for the different clusters.

In the representative and superposition structures generated from grcarma, the clusters showed some noteworthy similarities and some less significant differences. According to the representative structures, in the helices and the loops no divergence was detected (*Figure 39*). On the other hand, differences were observed in the N-terminal regions of both chains, which were not of particular interest to this study. Additionally, the same similarities and differences were observed in the superposition structures (*Figure 40*). To specify, in cluster 1 the N-terminal region of chain A (*Figure 40a*), was much more stable than of chain B, while in clusters 2 and 3 (*Figures 40b,* and *c*), the exact opposite happened. Overall, the region of interest, was stable in all the clusters, as confirmed by the dark blue color in the figures.

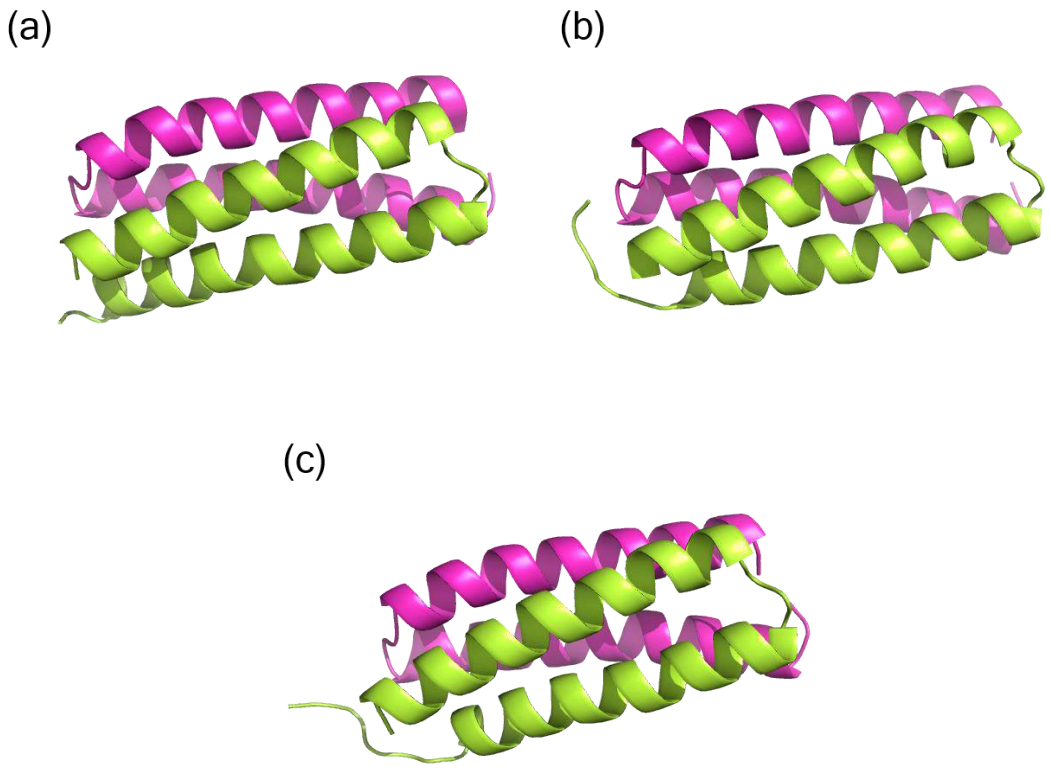

Figure 39 | Representative structures of clusters created by the dihedral PCA, in grcarma, reproduced with PyMol program. The calculation was done using backbone atoms of native ROP, excluding tails. Coding: green for chain A, and magenta for chain B. (a) cluster 1, (b) cluster 2, and (c) cluster 3.

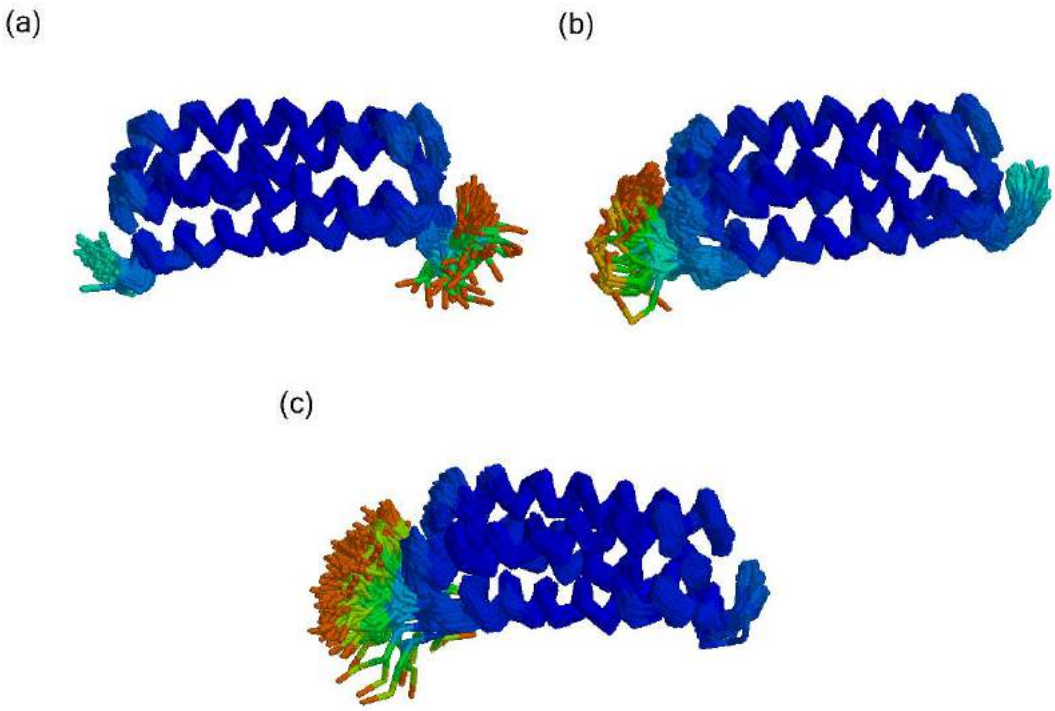

Figure 40 | Superposition structures, containing 500 structures each, of clusters created by the dihedral PCA, in grcarma, reproduced with RasMol program. The calculation was done using backbone atoms of native ROP, excluding tails. Color coding: temperature, with a color range from dark blue (low mobility) to red (high mobility). (a) cluster 1, (b) cluster 2, and (c) cluster 3.

3.3.7 Comparison between cPCA & dPCA

As mentioned in chapter 3.2.7, the two kinds of PCA used different parameters, which might have been the cause of the divergence between their results. In the cPCA calculation, 2 clusters were noted, while in the dPCA 3 clusters were (*Tables 10,* and *11,* and *Figures 33,* and *38*). Furthermore, the number of frames of cluster 1 isolated in both computations had no significant variance, unlike the others. The sum of frames included in clusters was comparable between the 2 analyses, with the sum in the cPCA being greater (*Tables 10,* and *11*).

In addition, the helices and the loops in all computations were stable and helped the molecule keep the same topology (*Figures 34, 36, 39,* and *40*). Despite that, the representative and the superposition structures indicated differences located in the N-terminal region of the chains. This was reasonable, due to the high mobility and instability of this region, as analyzed in a previous chapter.

3.4 Comparison between native ROP & native-like A31P mutant

The comparison was focused on the loop region, which was the region of interest. So, the results with increased noise due to the tails and the N-terminal regions were not useful, and thus they were disregarded.

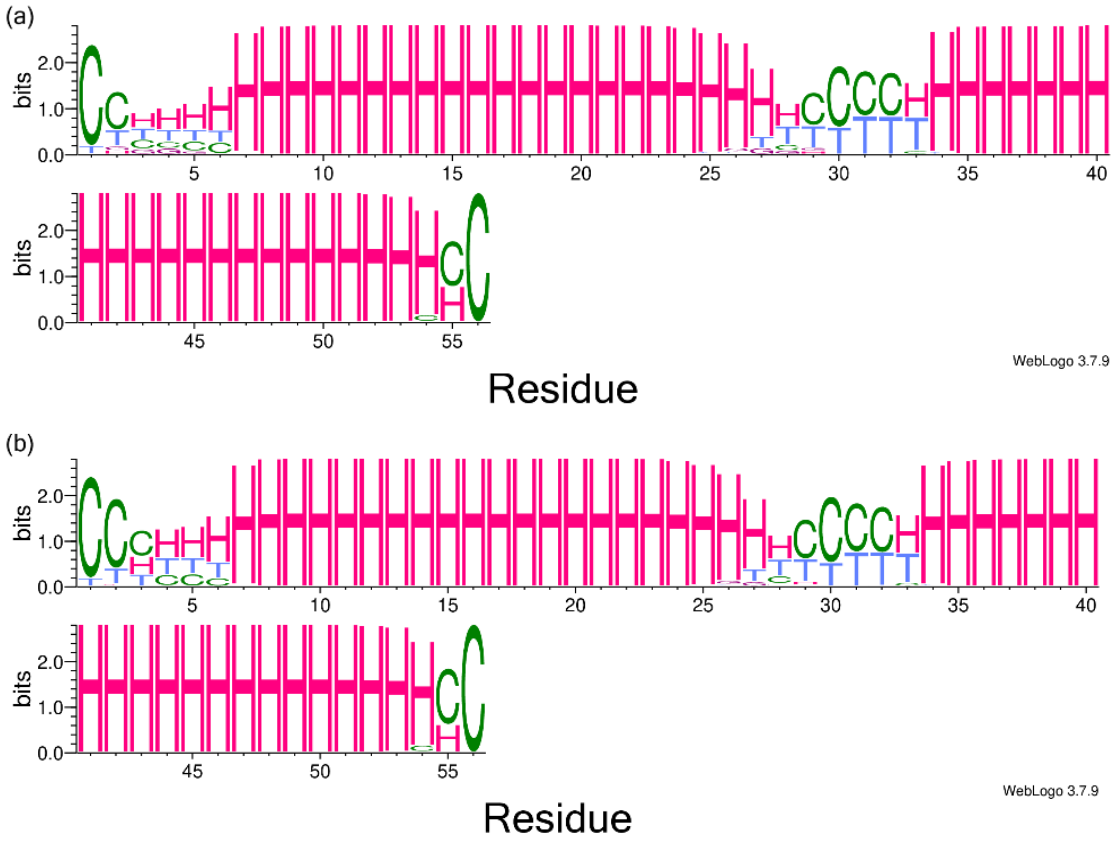

Figure 41 | Secondary structure analysis of native-like A31P mutant, excluding tails, using WebLogo. Representation of (a) chain A, and (b) chain B. Coding: H for α-helices, G for 3_{10}–helices, T for turns, and C for random coils. This figure is the same as Figure 4, page 20.

To start with, the helices in the native ROP were longer and the loop regions were more limited than in the native-like A31P mutant, as shown in *Figures 41, 42, 43,* and *44*. More specifically, the loop in the native-like extended from residue 27 to 33 (*Figure 41*), in both chains, while in the native from 29 to 31 (*Figure 42*). Furthermore, in most of the frames, in both chains of the native-like trajectory, residue 31P seemed to have the structure of a turn, but it was unstable over time, as in some frames it was presented in white color, suggesting a random coil or an unassigned structure (*Figure 43*). In contrast, there were only a few frames that the residue 31A appeared to have

the structure of α-helix, and in all the others it seemed to participate in a turn. In the native it appeared much more stable, confirmed by the low noise observed in *Figure 44*.

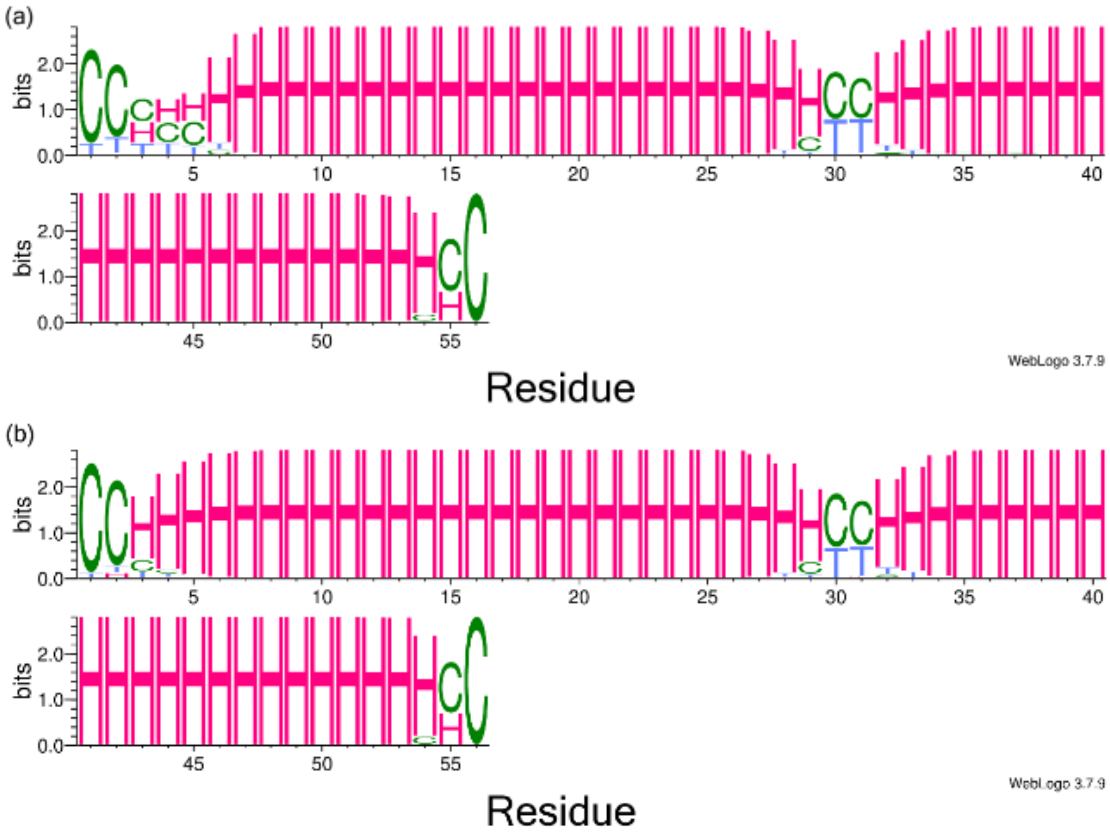

Figure 42 | Secondary structure analysis of native ROP protein, excluding tails, using WebLogo. Representation of (a) chain A, and (b) chain B. Coding: H for α-helices, G for 3_{10}–helices, T for turns, and C for random coils. This figure is the same as Figure 23, page 38.

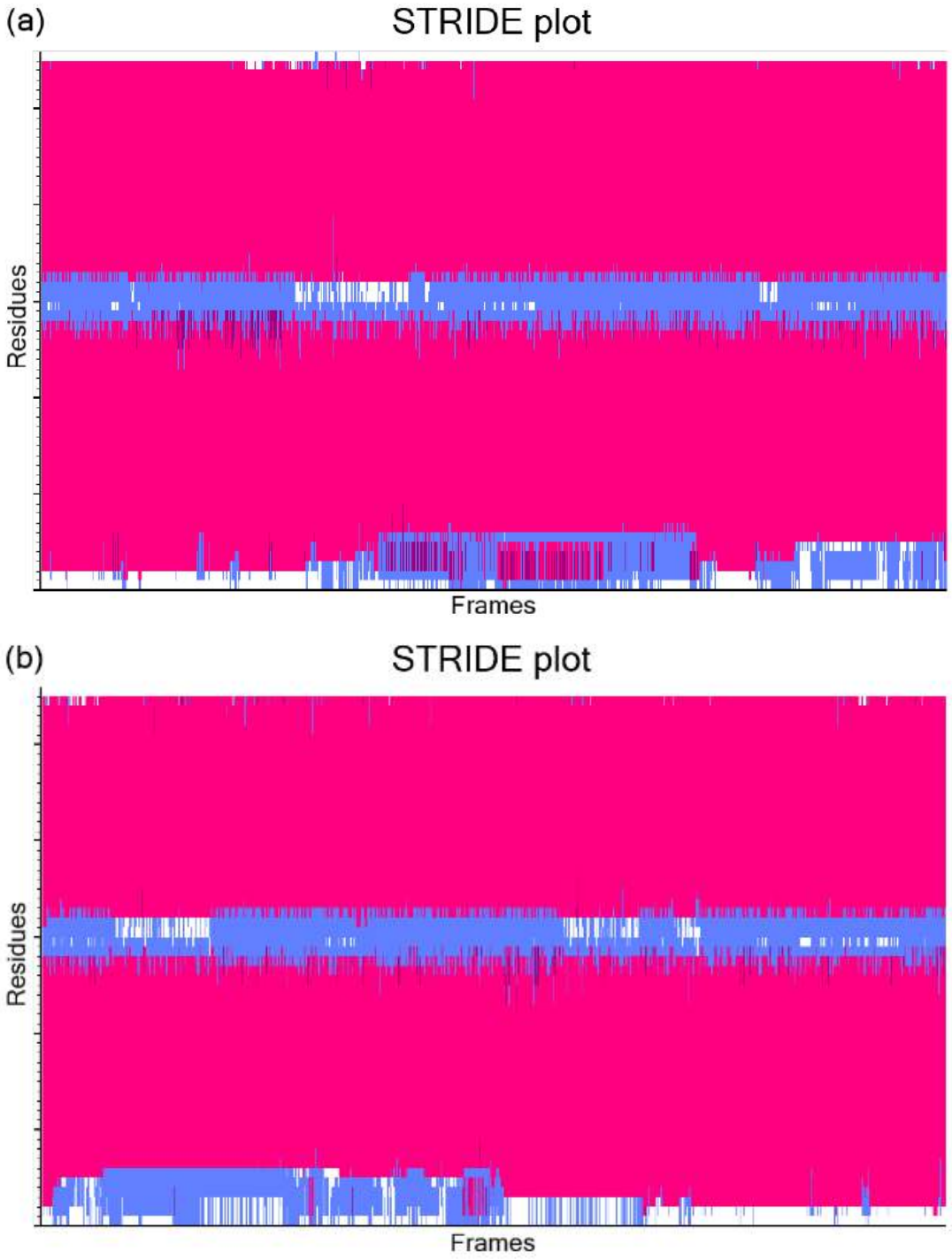

Figure 43 | Secondary structure analysis generated by the program STRIDE, using the native-like A31P mutant, excluding tails. Representation of (a) chain A, and (b) chain B. Coding: pink for α-helices, purple for 3_{10} –helices, dark purple for Pi-helices, yellow for β-sheets, blue for β/G-turns, and white for coils/unassigned. This figure is the same as Figure 5, page 21.

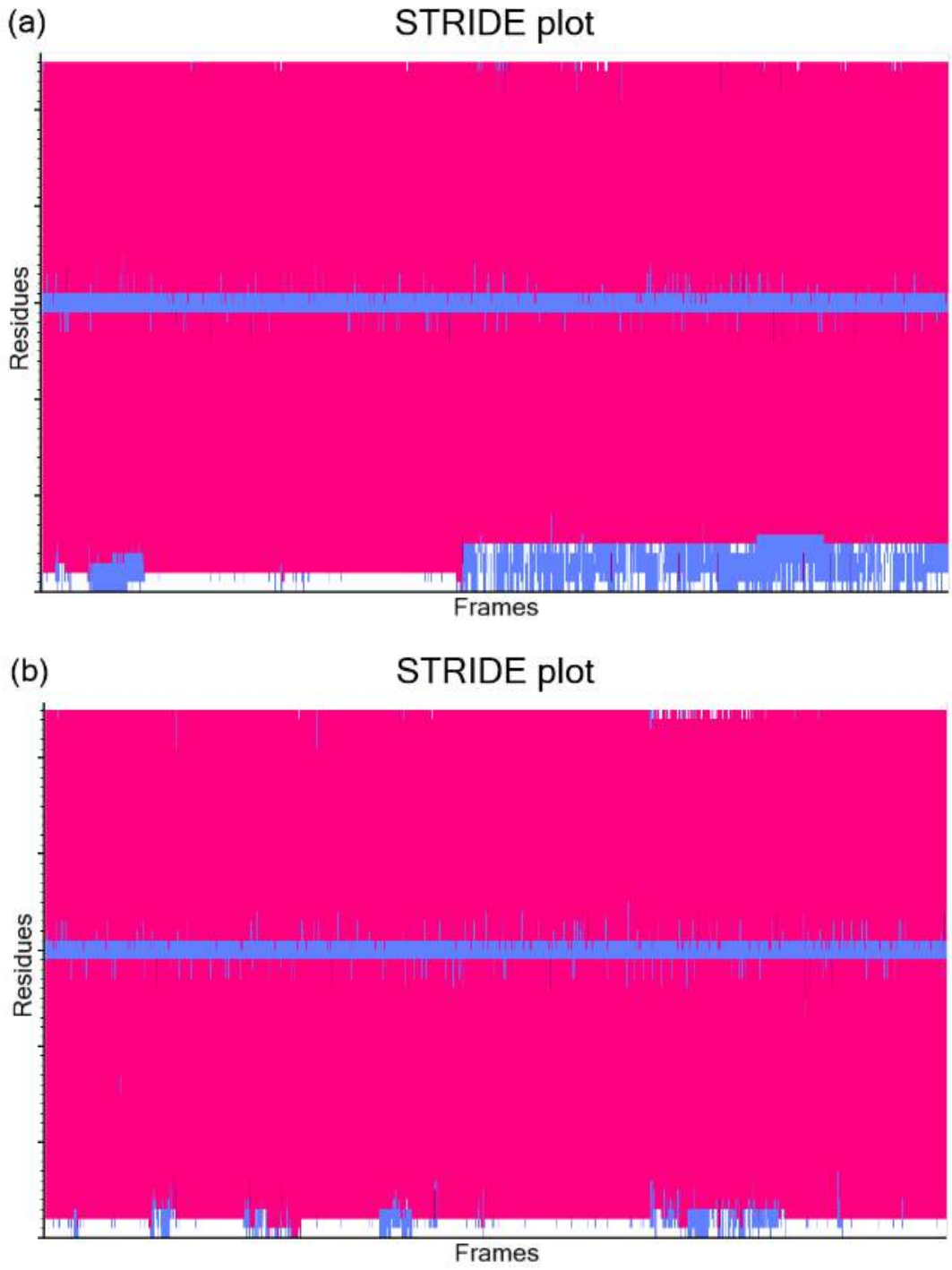

Figure 44 | Secondary structure analysis generated by the program STRIDE, using native ROP, excluding tails. Representation of (a) chain A, and (b) chain B. Coding: pink for α-helices, purple for 3_{10}-helices, dark purple for Pi-helices, yellow for β-sheets, blue for β/G-turns, and white for coils/unassigned. This figure is the same as Figure 24, page 39.

Also, concerning the cartesian RMSD matrix analyses, the structure of the native ROP over time was much more stable than of the native-like A31P mutant, as proved by the colors in *Figure 45*. In the lower-left half of the figure, the predominant

color is blue, whereas, in the upper-right half, a combination of red, yellow, and blue appear. Moreover, the native-like resulted in more clusters with fewer frames, than the ones calculated by the native analysis.

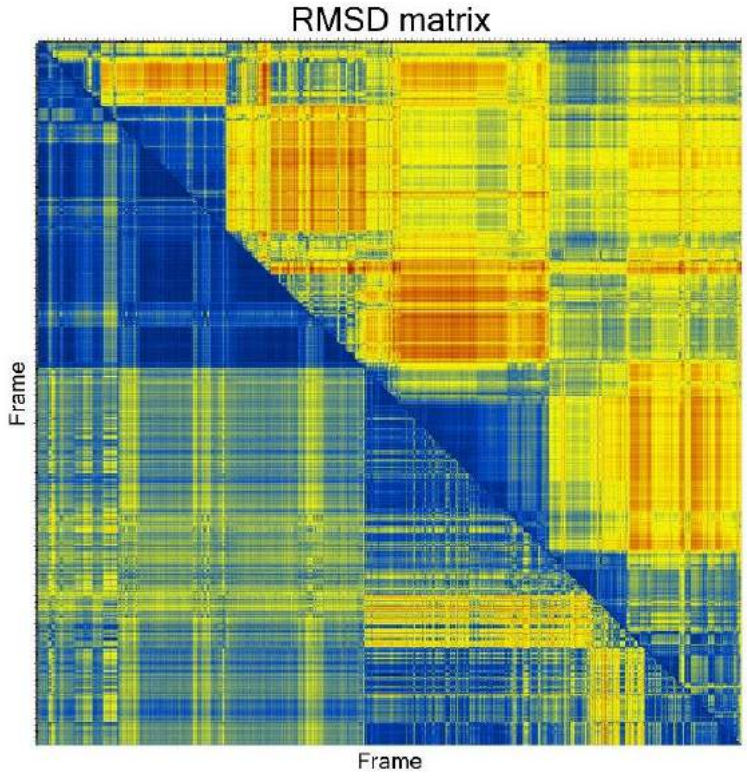

Figure 45 | RMSD matrix using only CA atoms, excluding tails. The color ranges from dark blue (minimum RMSD) to dark red (maximum RMSD: 5.63). The upper-right half represents native-like A31P mutant and the lower-left native ROP.

The results on the stability of the molecules were supported by the RMSF analyses. In general, the residues of the native ROP had lower mobility than those of the native-like A31P mutant (*Figure 46*). In the native-like calculation, the RMSF values of the region of interest were not as low as in the native ROP. Specifically, residue Pro31 (Pro94 in chain B) in the native-like noted RMSF values of 2.27 in chain A, and 2.28 in chain B, while residue Ala31 (Ala94 in chain B) in the native noted RMSF values of 1.37, and 0.98, in chains A, and B, respectively. There was a difference between the values, but it was not particularly big.

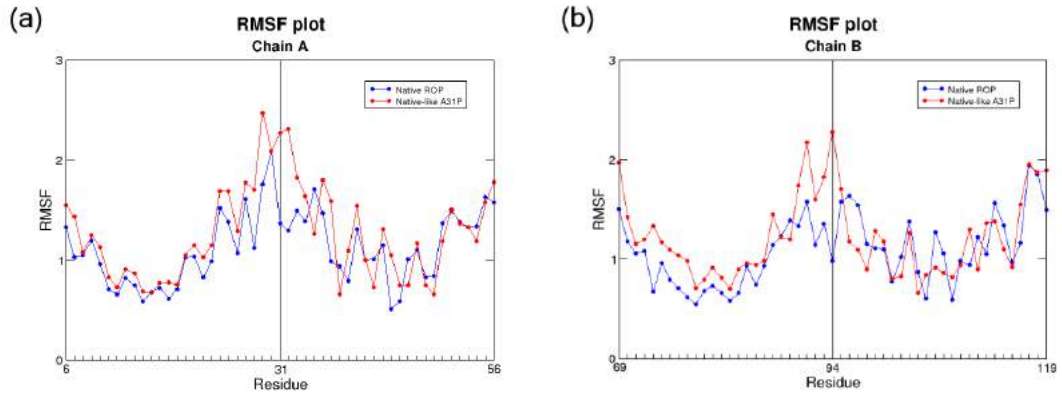

Figure 46 | RMSF plot created by xmgrace plotting tool. The calculation was done using only CA atoms, excluding tails. (a) residues 6-56 of chain A, and (b) residues 69-119 of the protein, that account for residues 6-56 in chain B. The red color stands for the native-like A31P mutant and the blue for the native ROP.

The number of clusters isolated in cPCA and dPCA of the native-like was bigger than the one of the native (*Tables 8, 9, 10,* and *11*). Although, the larger cluster of the native trajectory contained approximately 3 and 2.7 million frames, in the first and the second analysis, respectively, instead of approximately 1.9 and 0.6 million frames included in the results of the native-like trajectory, meaning that the native-like did not have as high stability as the native.

For the comparison of the representative and the superposition structures produced, only the first cluster of each calculation was used as a representation of the computation.

The representative structures agreed with the secondary structure comparison about the difference in the length of the loop. In the representative structures (*Figures 47,* and *48*), and especially in the ones generated by the dPCA (providing more information about the secondary structure), the loop was shorter in the native than in the native-like.

According to the superposition structures, the N-terminal regions had high mobility, as indicated by the warm colors, while the helix regions were highly stable (*Figures 49,* and *50*). Although, the stability of the loops was quite low in the clusters of the native-like analyses, in contrast to the native's which was as high as the helices'. Generally, the region of interest had lower mobility in native than in native-like.

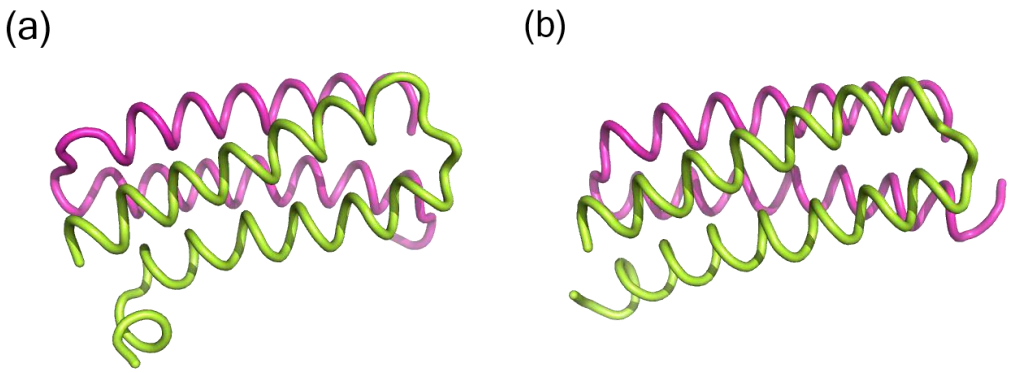

Figure 47 | Representative structures of clusters created by the cartesian PCA, in grcarma, reproduced with PyMol program. The calculation was done using only CA atoms, excluding tails. Coding: green for chain A, and magenta for chain B. Representation of cluster 1 (a) native-like A31P mutant, and (b) native ROP. These figures are the same as Figures 15a, and 34a.

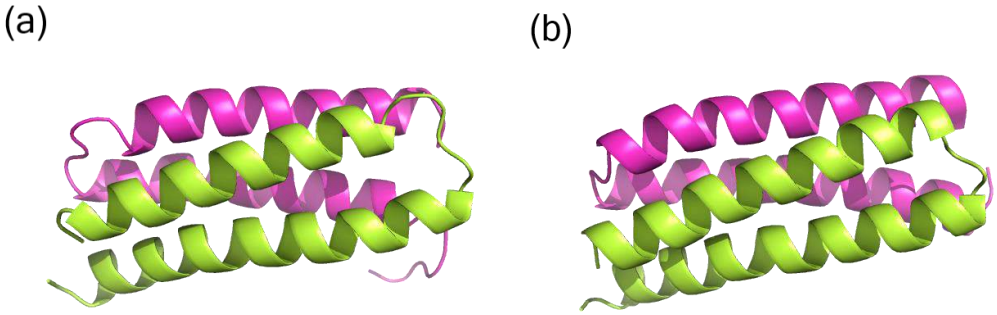

Figure 48 | Representative structures of clusters created by the dihedral PCA, in grcarma, reproduced with PyMol program. The calculation was done using backbone atoms, excluding tails. Coding: green for chain A, and magenta for chain B. Representation of cluster 1 (a) native-like A31P mutant, and (b) native ROP. These figures are the same as Figures 19a, and 39a.

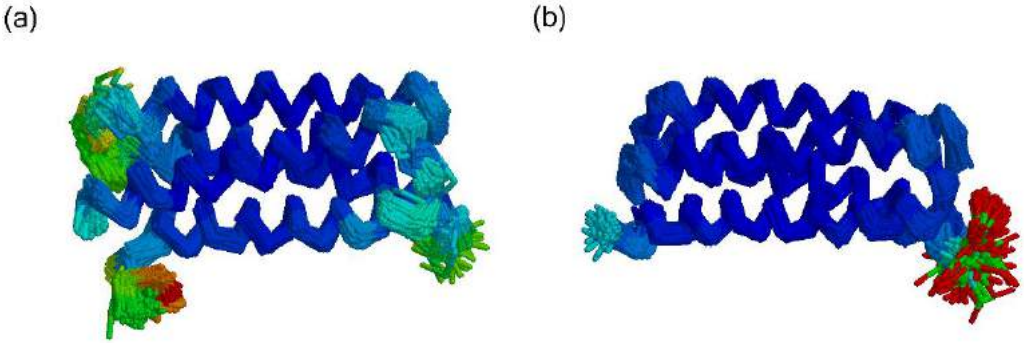

Figure 49 | Superposition structures, containing 500 structures each, of clusters created by the cartesian PCA, in grcarma, reproduced with RasMol program. The calculation was done using only CA atoms, excluding tails. Color coding: temperature, with a color range from dark blue (low mobility) to red (high mobility). Representation of cluster 1 (a) native-like A31P mutant, and (b) native ROP. These figures are the same as Figures 16a, and 36a.

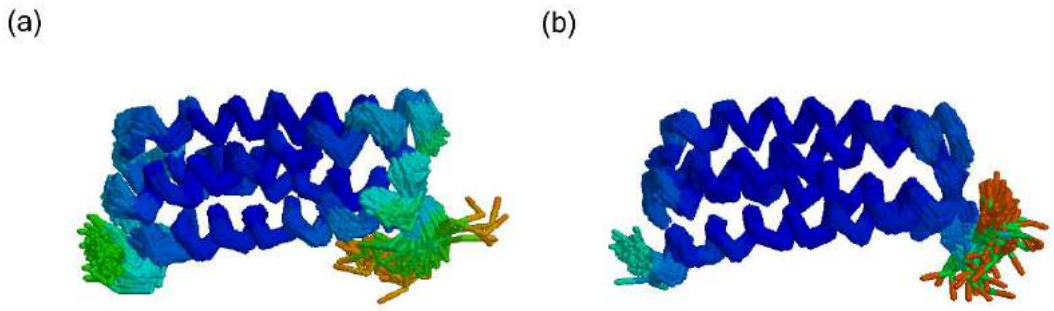

Figure 50 | Superposition structures, containing 500 structures each, of clusters created by the dihedral PCA, in grcarma, reproduced with RasMol program. The calculation was done using backbone atoms, excluding tails. Color coding: temperature, with a color range from dark blue (low mobility) to red (high mobility). Representation of cluster 1 (a) native-like A31P mutant, and (b) native ROP. These figures are the same as Figures 20a, and 40a.

Overall, the differences between the native ROP and the computational native-like A31P mutant were not sufficient, and its structure appeared stable enough to exist. So, there was no answer to the question: “Why does the experimental A31P mutant have such a different topology from the wild-type ROP?”.

4. Discussion & Conclusion

The purpose of this study was to find out if the computationally created mutant is unstable, in an attempt to explain the different topology of the experimental A31P mutant, as shown by crystallographic studies [16, 17]. In the beginning, the statistical analysis did not give any answer, due to the low p-value and the statistical insignificance between the energy of the wild-type and the A31P mutants generated by PyMol, VMD, and Coot. The next move was molecular dynamics simulations for a native ROP and a native-like A31P mutant trajectory. Every calculation indicated that the first was a little stabler than the second, but the structure of the *in silico* created mutant appeared stable enough, once again.

In the process of this study, no reasonable cause was found to justify the folding disparity. The levels of mobility of the molecules, especially of the native-like mutant, might be responsible for that. Another reason could be the number of frames and the time of the simulation, while the molecule might need more time to reach its final folding state. And lastly, the fact that in this study only the native ROP and the native-like A31P mutant were examined and not the A31P mutant with the *bisecting U* topology.

The research could show progress with further investigation and lead to a solution to the problem. First, as for the mobility, the N-terminal regions of both chains appeared unstable, and thus they could be excluded from the calculations to improve the results and minimize the noise. An analysis only for the loop regions could give significant information about the folding, too. Furthermore, a simulation with more frames could provide the molecule more time to fall into its final state, and this could deliver a different outcome. For this purpose, another force field or software that analyzes frames with a higher speed would be helpful in order to perform a longer simulation in less time.

Moreover, the A31P mutant could be involved in the molecular dynamics simulations, to compare the results with both native ROP and native-like A31P mutant. This could reveal some similarities in the folding of native-like and the *bisecting U* mutant, over time. Additionally, to justify the folding of the A31P mutant, more

parameters and analyses could be included in further studies. Overall, finding similar cases, in which the change of just one residue causes such a big difference in the topology of the protein, and assigning them into groups could help scientists perform an extended study, and as the available software improves every day, provide an answer to the folding of A31P mutant in the future.

References

1. PAULING, L., COREY, R. B., & BRANSON, H. R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, *37* (4), 205–211. https://doi.org/10.1073/pnas.37.4.205
2. CRICK, F. H. C. (1952). Is α-Keratin a Coiled Coil? *Nature*, *170*(4334), 882–883. https://doi.org/10.1038/170882b0
3. Crick, F. H. C. (1953). The packing of α-helices: simple coiled-coils. *Acta Crystallographica*, *6*(8), 689–697. https://doi.org/10.1107/s0365110x53001964
4. Lupas, A. N., & Gruber, M. (2005). The Structure of α-Helical Coiled Coils. *Fibrous Proteins: Coiled-Coils, Collagen and Elastomers*, 37–38. https://doi.org/10.1016/s0065-3233(05)70003-6
5. Branden, C., & Tooze, J. (2019). *Εισαγωγή στη Δομή των Πρωτεϊνών* (2nd edition). Ακαδημαϊκές Εκδόσεις.
6. Ho, S. P., & DeGrado, W. F. (1987). Design of a 4-helix bundle protein: synthesis of peptides which self-associate into a helical protein. *Journal of the American Chemical Society*, *109*(22), 6751–6758. https://doi.org/10.1021/ja00256a032
7. Chothia, C., Levitt, M., & Richardson, D. (1977). Structure of proteins: packing of alpha-helices and pleated sheets. *Proceedings of the National Academy of Sciences*, *74*(10), 4130–4134. https://doi.org/10.1073/pnas.74.10.4130
8. Chou, K. C., Maggiora, G. M., & Scheraga, H. A. (1992). Role of loop-helix interactions in stabilizing four-helix bundle proteins. *Proceedings of the National Academy of Sciences*, *89*(16), 7315–7319. https://doi.org/10.1073/pnas.89.16.7315
9. Banner, D. W., Kokkinidis, M., & Tsernoglou, D. (1987). Structure of the ColE1 Rop protein at 1.7 Å resolution. *Journal of Molecular Biology*, *196*(3), 657–675. https://doi.org/10.1016/0022-2836(87)90039-8
10. Eberle, W., Pastore, A., Sander, C., & Rösch, P. (1991). The structure of ColE1 rop in solution. *Journal of Biomolecular NMR*, *1*(1), 71–82. https://doi.org/10.1007/bf01874570
11. 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
12. Willis, M. A., Bishop, B., Regan, L., & Brunger, A. T. (2000). Dramatic Structural and Thermodynamic Consequences of Repacking a Protein's Hydrophobic Core. *Structure*, *8*(12), 1319–1328. https://doi.org/10.1016/s0969-2126(00)00544-x
13. Glykos, N. M., Papanikolau, Y., Vlassi, M., Kotsifaki, D., Cesareni, G., & Kokkinidis, M. (2006). Loopless Rop: Structure and Dynamics of an Engineered Homotetrameric Variant of the Repressor of Primer Protein. *Biochemistry*, *45*(36), 10905–10919. https://doi.org/10.1021/bi060833n
14. Gambin, Y., Schug, A., Lemke, E. A., Lavinder, J. J., Ferreon, A. C. M., Magliery, T. J., Onuchic, J. N., & Deniz, A. A. (2009). Direct single-molecule observation of a protein living in two opposed native structures. *Proceedings of the National Academy of Sciences*, *106*(25), 10153–10158. https://doi.org/10.1073/pnas.0904461106

15. Peters, K., Hinz, H. J., & Cesareni, G. (1997). Introduction of a Proline Residue into Position 31 of the Loop of the Dimeric 4-α-Helical Protein ROP Causes a Drastic Destabilization. *Biological Chemistry*, *378*(10). https://doi.org/10.1515/bchm.1997.378.10.1141
16. Glykos, N. M., Cesareni, G., & Kokkinidis, M. (1999). Protein plasticity to the extreme: changing the topology of a 4-α-helical bundle with a single amino acid substitution. *Structure*, *7*(6), 597–603. https://doi.org/10.1016/s0969-2126(99)80081-1
17. Levy, Y., Cho, S. S., Shen, T., Onuchic, J. N., & Wolynes, P. G. (2005). Symmetry and frustration in protein energy landscapes: A near degeneracy resolves the Rop dimer-folding mystery. *Proceedings of the National Academy of Sciences*, *102*(7), 2373–2378. https://doi.org/10.1073/pnas.0409572102
18. Schug, A., Whitford, P. C., Levy, Y., & Onuchic, J. N. (2007). Mutations as trapdoors to two competing native conformations of the Rop-dimer. *Proceedings of the National Academy of Sciences*, *104*(45), 17674–17679. https://doi.org/10.1073/pnas.0706077104
19. Glykos, N. M., & Kokkindis, M. (2004). Structural polymorphism of a marginally stable 4-α-helical bundle. Images of a trapped molten globule? *Proteins: Structure, Function, and Bioinformatics*, *56*(3), 420–425. https://doi.org/10.1002/prot.20167
20. Karplus, M., & Petsko, G. A. (1990). Molecular dynamics simulations in biology. *Nature*, *347*(6294), 631–639. https://doi.org/10.1038/347631a0
21. Karplus, M., & McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, *9*(9), 646–652. https://doi.org/10.1038/nsb0902-646
22. Scheraga, H. A., Khalili, M., & Liwo, A. (2007). Protein-Folding Dynamics: Overview of Molecular Simulation Techniques. *Annual Review of Physical Chemistry*, *58*(1), 57–83. https://doi.org/10.1146/annurev.physchem.58.032806.104614
23. Hansson, T., Oostenbrink, C., & van Gunsteren, W. (2002). Molecular dynamics simulations. *Current Opinion in Structural Biology*, *12*(2), 190–196. https://doi.org/10.1016/s0959-440x(02)00308-1
24. McCammon, J. A., Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. *Nature*, *267*(5612), 585–590. https://doi.org/10.1038/267585a0
25. Glykos, N. M. (2006). Software news and updates carma: A molecular dynamics analysis program. *Journal of Computational Chemistry*, *27*(14), 1765–1768. https://doi.org/10.1002/jcc.20482
26. Koukos, P. I., & Glykos, N. M. (2013). Grcarma: A fully automated task-oriented interface for the analysis of molecular dynamics trajectories. *Journal of Computational Chemistry*, *34*(26), 2310–2312. https://doi.org/10.1002/jcc.23381
27. Koukos P.I. (2013) "grcarma: A Fully Automated Task-Oriented Interface for the Analysis of Molecular Dynamics Trajectories", Diploma thesis
28. Humphrey, W., Dalke, A., & Schulten, K. (1996, February). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, *14*(1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5

29. Emsley, P., & Cowtan, K. (2004, November 26). *Coot*: model-building tools for molecular graphics. *Acta Crystallographica Section D Biological Crystallography*, *60*(12), 2126–2132. https://doi.org/10.1107/s0907444904019158
30. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
31. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank *Nucleic Acids Research*, 28: 235-242.
32. Eberle, W., Pastore, A., Klaus, W., Sander, C., Roesch, P. (1994) THE STRUCTURE OF COLE1 ROP IN SOLUTION doi: 10.2210/pdb1rpr/pdb
33. Eberle, W., Pastore, A., Sander, C., & Rösch, P. (1991c, May). The structure of ColE1 rop in solution. *Journal of Biomolecular NMR*, *1*(1), 71–82. https://doi.org/10.1007/bf01874570
34. Ko, J., Park, H., Heo, L., & Seok, C. (2012, May 30). GalaxyWEB server for protein structure prediction and refinement. *Nucleic Acids Research*, *40*(W1), W294–W297. https://doi.org/10.1093/nar/gks493
35. Heo, L., Lee, H., & Seok, C. (2016, August). GalaxyRefineComplex: Refinement of protein-protein complex model structures driven by interface repacking. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep32153
36. Heo, L., Park, H., & Seok, C. (2013, May 21). GalaxyRefine: protein structure refinement driven by side-chain repacking. *Nucleic Acids Research*, *41*(W1), W384–W388. https://doi.org/10.1093/nar/gkt458
37. *GalaxyWEB*. (n.d.). Retrieved May 12, 2022, from https://galaxy.seoklab.org/cgi-bin/help.cgi?key=METHOD&type=REFINE
38. Mukherjee, S., & Zhang, Y. (2009, May 14). MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Research*, *37*(11), e83–e83. https://doi.org/10.1093/nar/gkp318
39. Britannica, T. Editors of Encyclopaedia (2022, September). Student's t-test. Encyclopedia Britannica. https://www.britannica.com/science/Students-t-test
40. Crooks, Gavin E., & USDOE. (2003, January 8). WebLogo [Computer software]. https://www.osti.gov//servlets/purl/1230916. https://doi.org/10.11578/dc.20210416.8
41. Frishman, D. (2004, July 1). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research*, *32*(Web Server), W500–W502. https://doi.org/10.1093/nar/gkh429
42. Turner, P. J. (2005). XMGRACE, Version 5.1. 19. *Center for Coastal and Land-Margin Research, Oregon Graduate Institute of Science and Technology, Beaverton, OR*, *2*.
43. Νικολάου, Χ., & Χουβαρδάς, Π. (2015). *Υπολογιστική βιολογία* [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. http://hdl.handle.net/11419/1577
44. Glykos, N. (2017) Home – plot – Utopia. Available at: https://utopia.duth.gr/glykos/plot/
45. Herbert J. Bernstein. (2009). *RasMol*. Retrieved from http://www.rasmol.org/