DEMOCRITUS UNIVERSITY OF THRACE HEALTH SCIENCES SCHOOL
DEPARTMENT OF MOLECULAR BIOLOGY & GENETICS

DEMOCRITUS UNIVERSITY OF THRACE
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS
MSc TRANSLATIONAL RESEARCH IN BIOMEDICINE

MSc Thesis

"Structural computational biology studies of an antimicrobial peptide with a WWW motif"

Author: Georgia Mitraina
Supervisor: Dr. Nicholas M. Glykos, Associate Professor of Structural and Computational Biology

Alexandroupolis, 2023

Acknowledgements

I would like to express my deep gratitude and respect to my supervisor, Dr. Nicholas M. Glykos for his unwavering guidance, patience and the opportunity he provided me to be a part of his research team, allowing me to become familiar with the field of bioinformatics. His insightful comments, contagious enthusiasm, along with his understanding and the sense of security he offered, greatly contributed to the realization of this thesis and to my immersion in this realm of science. I would also like to extend my thanks to my family and all those who stood by me, for their support and encouragement throughout my academic journey.

Table of contents

Abstract:

Molecular Dynamics simulations are a method with extensive usage in various applications that aims to understand protein folding by studying the motion and behavior of various chemical systems. This thesis describes the computational study of a peptide, named Tetra-F2W-RK using Molecular Dynamics simulations. The peptide is rich in tryptophan (Trp) and has the ability to interact with bacterial DNA. The study focuses on the behavior of the peptide in various solutions, namely dimethylsulfoxide (DMSO), 2,2,2-trifluoroethanol (TFE) and water, as well as comparing the results with experimental data carried out by researchers D. Zarena, B. Mishra and co-workers, who studied Tetra-F2W-RK using micelles. The role of the solvent is highlighted since it is an important factor in the structure, behavior and properties of biological systems. Then, the methodology of the present work is presented, where details about the simulation of the peptide are mentioned, specifically the parameters of the simulation and the methods of data analysis in the three under consideration solutions. The results demonstrate that Tetra-F2W-RK does not acquire helical structures in all three solutions, but is more stable for longer time when the solvent is TFE or water. This is followed by a comparison of the experimental values with the results of the simulations, which mostly agree and the need for further study is highlighted in order to elucidate poorly understood aspects of the behavior of the peptide. Overall this work contributes to the understanding of the structure and stability of the Tetra-F2W-RK peptide in various solutions, having important applications in biochemistry and pharmaceuticals.

Keywords: Molecular Dynamics, Tetra-F2W-RK, DMSO, TFE, water, GRCARMA, CARMA, helical structure

Περίληψη:

Οι προσομοιώσεις Μοριακής Δυναμικής αποτελούν μια ευρέως χρησιμοποιούμενη μέθοδο, που στοχεύει μέσω της μελέτης της κίνησης και της συμπεριφοράς διαφόρων χημικών συστημάτων στην κατανόηση της αναδίπλωσης των πρωτεϊνών. Η παρούσα διπλωματική εργασία περιγράφει την υπολογιστική μελέτη ενός πεπτιδίου, που ονομάζεται Tetra-F2W-RK χρησιμοποιώντας προσομοιώσεις Μοριακής Δυναμικής. Το πεπτίδιο είναι πλούσιο σε τρυπτοφάνη (Trp) και έχει την ικανότητα να αλληλεπιδρά με το βακτηριακό DNA. Η μελέτη επικεντρώνεται στη συμπεριφορά του πεπτιδίου σε διάφορα διαλύματα, συγκεκριμένα στο διμεθυλοσουλφοξείδιο (DMSO), την 2,2,2-τριφθοροαιθανόλη (TFE) και το νερό, καθώς και στη σύγκριση των αποτελεσμάτων με πειραματικά δεδομένα που πραγματοποιήθηκαν από τους ερευνητές D. Zarena B. Mishra και των συνεργατών τους, που μελέτησαν το Tetra-F2W-RK χρησιμοποιώντας μικκύλια. Επισημαίνεται ο ρόλος του διαλύτη αφού αποτελεί σημαντικό παράγοντα της δομής, της συμπεριφοράς και των ιδιοτήτων των βιολογικών συστημάτων. Κατόπιν, παρουσιάζεται η μεθοδολογία της παρούσας εργασίας, όπου αναφέρονται λεπτομέρειες σχετικά με την προσομοίωση του πεπτιδίου και συγκεκριμένα οι παράμετροι της προσομοίωσης και οι μέθοδοι ανάλυσης των δεδομένων στα τρία υπό εξέταση διαλύματα. Τα αποτελέσματα αποδεικνύουν ότι το Tetra-F2W-RK δεν αποκτά ελικοειδείς δομές και στα τρία διαλύματα, όμως ενδείκνυται πιο σταθερό για μεγαλύτερη χρονική διάρκεια, όταν διαλύτης είναι το TFE και το νερό. Ακολουθεί σύγκριση των πειραματικών τιμών με τα αποτελέσματα των προσομοιώσεων, τα οποία συμφωνούν σε μεγάλο ποσοστό και επισημαίνεται η ανάγκη για περαιτέρω μελέτες προκειμένου να αποσαφηνιστούν μη κατανοητές πτυχές της συμπεριφοράς του πεπτιδίου. Συνολικά η παρούσα εργασία συνεισφέρει στην κατανόηση της δομής και της σταθερότητας του πεπτιδίου Tetra-F2W-RK σε διάφορα διαλύματα, έχοντας σημαντικές εφαρμογές στη βιοχημεία και τη φαρμακευτική.

Λέξεις-κλειδιά: Μοριακή Δυναμική, Tetra-FW2-RK, DMSO, TFE, νερό, GRCARMA, CARMA, ελικοειδής δομή

CHAPTER 1: Introduction

1.1 Proteins

Proteins are the most pluripotent and abundant biological macromolecules in living organisms, which are composed of amino acids. Since they are involved in almost every biological process, a protein analysis shows how these molecules interact and cooperate to create and maintain a functional biological system. So, their functions encompass catalyzing reactions, facilitating molecule transport and storage, offering structural reinforcement and immune defense, enabling motion, relaying nerve impulses and governing cellular growth and differentiation.

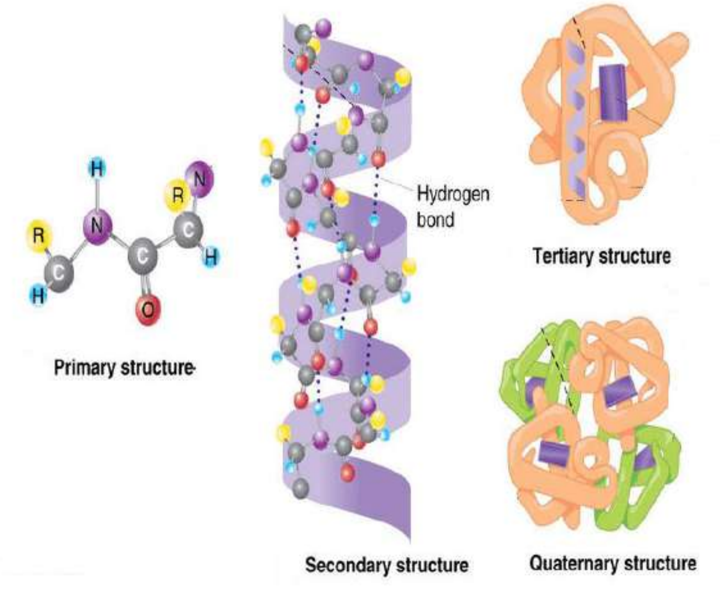


Figure 1.1: Levels of protein structure: primary structure, secondary structure, tertiary structure and quaternary structure. (Adapted without permission from ResearchGate)
https://www.researchgate.net/publication/279193494_A_Study_of_Intelligent_Techniques_for_Protein_Secondary_Structure_Prediction/figures?lo=1

As I mentioned, proteins are polymers constructed from amino acid monomers. This sequence of amino acids within polypeptide chain is called primary structure. Secondary structure is based on the ability of proteins to fold into three-dimensional structures and is stabilized through hydrogen bonds between adjacent amino acids. A-helices and β-sheets are the two most frequent components of secondary structure. On the other hand, interactions between distant amino acids constitute the tertiary structure of proteins. In addition, some proteins also display a quaternary structure, in which the functional protein is formed by more than one polypeptide chains.

(Figure1.1)

For numerous years, researchers have been engaged in the endeavor of predicting the three-dimensional arrangement of proteins based on their sequence of amino acids. A complete understanding of the biological function of proteins, however, presupposes solving the commonly known “protein folding problem”, which is rendered impossible due to its complexity.

Finally, other characteristics of proteins are that the contain functional groups such as alcohols, thioethers etc., as well as that they can interact with each other and with other molecules to create complex assemblies. [1][2][3]

1.2 Protein folding problem

The protein folding problem is a fundamental issue of molecular biology. The main problem, known as Levinthal’s paradox, revolves around the necessity for a functional outcome despite the protein’s initial synthesis in a linear molecular form. For this reason it must reach its native conformation. Nevertheless there are plenty of conformational states for a long protein molecule. Many researchers have tried to solve the Levinthal’s paradox, but the solution is still unsolved.

The history of this topic begins about 50 years ago, with Christian Anfinsen’s experiments. He used the enzyme Ribonuclease-A in order to study denaturation-renaturation issues and he demonstrated that under specific conditions, the tertiary structure of a protein is shaped by both its primary structure and the sequence of amino acids. The “Thermodynamic hypothesis of protein folding”, which he developed until 1962, refers that the lower free energy state is preferred by the protein (Gibbs free energy). From these experiments, Anfinsen concluded that the physical molecule is the most thermodynamically stable configuration. For his discoveries, he received half of the Nobel Prize in Chemistry in 1972. [4][5][6]

In 1968, Cyrus Levinthal referred that the folding time for lengthy protein molecules is typically exponentially prolonged due to the extensive array of available conformational possibilities. In his attempt to interpret the factors, which they are responsible for the speed of protein folding, concluded that it is impossible to form the most stable thermodynamic protein conformation through random displacements. He proposed the existence of defined folding pathways, a sequence of successive events which should be both thermodynamically and kinetically favored. After his experiments, further studies were carried out by many researchers with the aim of finding intermediate states during folding. However the folding process is not fully described by any of them. [7][8][9]

1.3 Models of protein folding

Different models of protein folding have been established mostly based on the recognized protein structures, aiming to reduce the extensive conformational space that needs to be explored and shorten the experimental folding time. However none of these can interpret the folding of the whole of the proteins. Following are some examples of these models:

1.3.1 Diffusion Collision Model

Karplus and Weaver proposed this model in 1976. According to this, the protein consists of many micro domains, which can assume all possible conformations quickly in comparison with the folding time of the entire protein. Micro domains are less stable, so in order to gain stability, they move and collide with each other forming larger structures. As secondary structural stability increases, diffusion collision becomes more likely. The creation of tertiary structure is the result of each step's conformation toward native structure. [10]

1.3.2 Nucleation Condensation Model

This model proposes the formation of a nucleus, characterized by stability resulting from the interplay of secondary and tertiary structure interactions. Subsequently, the nucleus serves as a blueprint for the accelerated assembly of additional structure around it, leading to the folding of the entire protein around it. This reduces the number of configurations. To conclude the primary feature of the “nucleation condensation model” revolves around the development of secondary and tertiary structures that they occur simultaneously and interact with each other. [11][12][13]

1.3.3 Jigsaw Puzzle Model

This model states that all proteins do not follow the same folding pathway; instead proteins can achieve their native conformation through multiple pathways through the analogy of jigsaw puzzle. [14][15][16]

1.3.4 Energy Landscapes Model – folding funnels

The most recent models and simplified representations of the structures and interactions created by statistical engineering techniques and the microscopic interactions that arise between proteins are models of “energy landscapes” or “folding funnels”. This model predicts a funnel-shaped energy landscape that follows a protein during the folding process, as it adopts its native conformation. The depth of the funnel signifies the reinforcement of the native state over the disordered ones while its width represents all possible configurations, i.e. the entropy formed in system which is studied. As the protein approaches its native structure, the entropy decreases. The energy value associated with each point on the funnel’s surface represents a potential configuration. Moreover the schematic representation of energy landscapes consists of two- and three-dimensional diagrams. An energy landscape illustrates the fluctuations in free energy variation across different conformations based on varying degrees of freedom. [17] The following images present the different types of energy landscapes. [18][19][20]

In the Levinthal's "golf course" energy landscape, it takes a while for a ball, which moves unpredictably across a level surface, to locate and enter the hole, in the native structure N. (Figure 1.2)

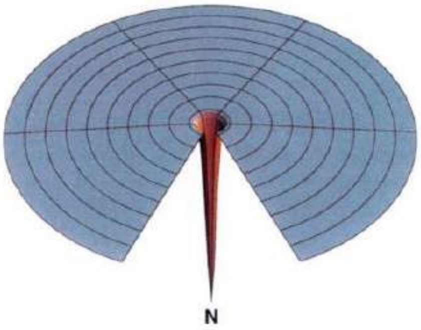


Figure 1.2: The Levinthal's "golf course" landscape (Adapted without permission from Quora) https://www.quora.com/Why-do-prions-violate-Levinthals-paradox

The "grooved golf course" landscape displays a proposed path solution to the random search problem. The folding molecule starting from a configuration A, travels via a tunnel on the landscape, in the physical structure N. (Figure 1.3)

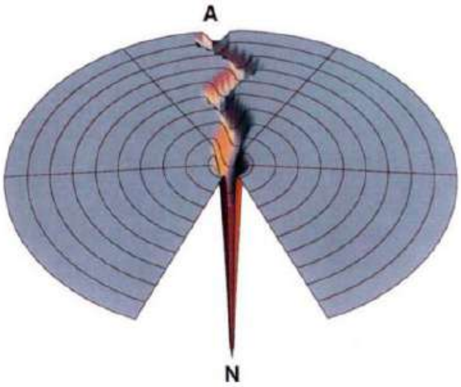


Figure 1.3: The "grooved golf course" landscape (Adapted without permission from PubMed) https://pubmed.ncbi.nlm.nih.gov/8989315/

The "smooth funnel" energy landscape is an idealization that demonstrates that the smooth decline of the protein's free energy results in a decrease in the number of possible conformations. (Figure 1.4)

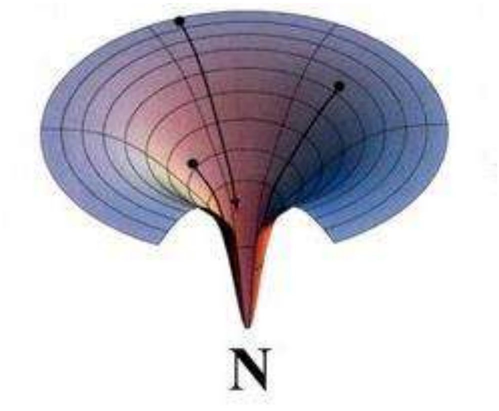


Figure 1.4: The "smooth funnel" landscape (Adapted without permission from ResearchGate) https://www.researchgate.net/publication/258020421_Small_angle_X-ray_scattering_studies_on_proteins_under_extreme_conditions/figures?lo=1

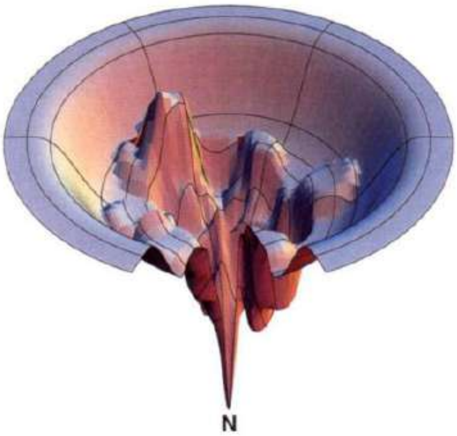


Figure 1.5: The “bumpy bowl” landscape (Adapted without permission from PubMed) https://pubmed.ncbi.nlm.nih.gov/8989315/

The “bumpy bowl” refers to a rugged energy landscape that in the route to the native structure N contains kinetic obstacles, energy barriers and constrained paths (Figure 1.5)

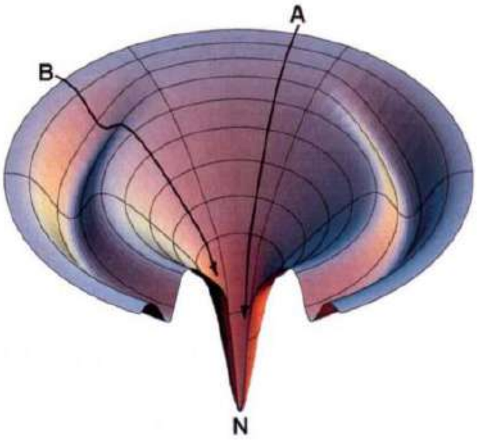
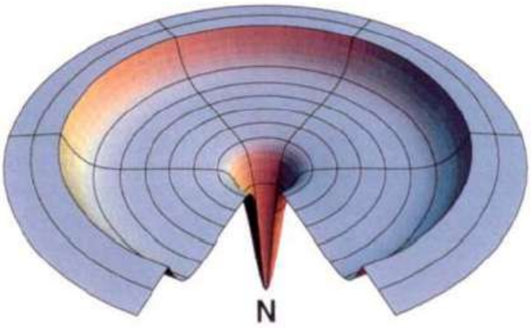


Figure 1.6: The “moat” landscape (Adapted without permission from PubMed) https://pubmed.ncbi.nlm.nih.gov/8989315/

The “moat” energy landscape illustrates that a protein can fold rapidly (path A) or more slowly (path B) due to a kinetic trap (moat) (Figure 1.6)



The “champagne glass” energy landscape shows how the entropy of the conformation can cause barriers in the folding process. Roaming across the flat plateau hinders the chain from quickly reaching its native structure. (Figure 1.7)

Figure 1.7: The “champagne glass” landscape (Adapted without permission from PubMed) https://pubmed.ncbi.nlm.nih.gov/8989315/

1.4 Methods to study protein folding

1.4.1 Experimental approaches

Various experimental methods have been employed over the years to investigate protein folding and structure. Among the most prevalent techniques are X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and Circular Dichroism.

1.4.1.1 X-ray crystallography

X- ray crystallography is the most reliable and widely used method of obtaining information for determining the structure of proteins and biological macromolecules. In order to derive a three-dimensional molecular structure from a crystal, the initial procedure involves crystallizing a purified sample. Following this, the crystalline specimens are subjected to an X-ray beam for analysis. Diffraction patterns are then analyzed, in which X-ray waves experience diffraction, manifesting as distinct directions with well-defined amplitudes and phases. This provides information about the crystal packing symmetry. The spot intensities can be used to produce an electron density map, which locates the electrons in the crystal and serves to determine the molecular structure of the protein. [21][22][23][24]

1.4.1.2 Nuclear Magnetic Resonance (NMR)

With the Nuclear Magnetic Resonance (NMR), the interaction of electromagnetic radiation with material is studied. The common utility of the NMR method is based on the study of the structure of small molecules, its specificity at the level of single atoms and its ability to determine the distribution of the structures of more complex molecules such as proteins. A constraint of the technique is the confined volume of concentrated protein solution within a potent magnetic field. [25][26]

Certain atomic nuclei possess a magnetic moment or spin.

When exposed to applied radio frequency (RF) pulses of electromagnetic radiation, these spins shift to a disoriented state. Upon returning to their aligned state, they emit radio frequency radiation, which can be computed and exhibited as a spectrum. Notably, the excitation of one nucleus influences the absorption and emission of radiant from nearby nuclei. This property of nuclear spin is exploited by the NMR method.[27][28][29]

In summary, the steps of the method include first the preparation of the protein solution, followed by NMR measurements and assignment of NMR signals to individual atoms, conformation constraints such as distances between atoms are collected and finally the 3D structure is determined.

There is one-dimensional, two-dimensional and three-dimensional NMR. Multidimensional NMR spectroscopy offers insights into protein structure, dynamics, stability and interactions at the granularity of individual atoms. By using 2D experiments, NMR has become the first technique applied to determine three-dimensional by biopolymer structures in solution and other non-crystalline states. 2D NMR spectra offer a more comprehensive understanding of a molecule compared to 1D NMR spectra. They hold particular significance in structure determination due to the enhanced information they provide. The proximity of pairs of atoms can be detected by measuring Nuclear Overhauser Effects (NOEs). [30][31][32]

To sum up, NMR is capable of elucidating the structure of relatively small proteins (<~20kDa). However achieving a high-quality protein structure necessitated the use of intricate experimental methodologies and the meticulous analysis of NMR spectra. The primary benefit of using NMR to investigate crowding effects lies in the capability to introduce NMR active isotopes (e.g. 15N, 13C). This facilitates the scrutiny of the protein of interest within an environment containing non-enriched crowding agents, while the main disadvantage of NMR is its insensitivity. However, the limitations in sensitivity are overcome and ultimately this method has great potential for the study of protein folding. [33]

1.4.1.3 Circular Dichroism (CD)

Circular Dichroism (CD) stands as an exceptional, rapid and straightforward spectroscopic technique. It’s utilized for discerning the secondary structure and folding attributes of proteins obtained through recombinant approaches, with its efficacy often influenced by temperature variations. [34][35][112]

The prevalent applications of CD often revolve around assessing whether an expressed and purified protein undergoes proper folding, as well as investigating interactions between different proteins. It is also used to determine the effects of mutations on the stability and conformation of proteins. CD spectra analysis, performed while varying temperature, offers valuable insights into the equilibrium structure and conformation of a molecule in solution. This method can be applied to molecules of diverse sizes under various solvent conditions without the need of crystallization. An advantageous aspect of CD is its ability to swiftly measure multiple samples containing proteins within physiological buffers. However, it’s important to note that CD lacks the capacity to provide residue-specific information achievable through x-ray crystallography or NMR techniques. [36][37][38][25]

In summary, while Circular Dichroism (CD) lacks the atomic resolution provided by x-ray crystallography and NMR spectroscopy, its utility remains significant for analyzing the structures of biomolecules like saccharides, proteins and nucleic acids. Finally, its solution approach for membrane proteins is particularly advantageous.

1.4.2 Computational methods

Since experimentally determining the structure of a protein is expensive, time-consuming and difficult, the contribution of computational methods is exceptionally significant. These methods can be categorized into three main groups for predicting protein structure: ab initio methods, fold recognition methods, comparative modeling and artificial neural networks. [28][39]

1.4.2.1 Ab initio methods

The term ab initio was coined to refer methods that solely rely on fundamental physical principles to deduce the folded structure of a protein.[40] Specifically, it pertains to structure prediction techniques that do not rely on experimentally established structures, nor do they compare a target and a known protein. Instead they exclusively compare and analyze fragments, specifically short amino acids subsequences from the target, with fragments of known structures retrieved from the Protein Data Bank. Once the appropriate fragments are identified, they are assembled into a sequence that predicts the structure of its native state, often with the help of evaluation through scoring functions derived from conformational statistics of known proteins. Scoring functions are increasingly improved by introducing information from independent secondary structure predictions. [41][42][43] To conclude, the ab initio methods are preferred in cases where the target has no homologue already present in biological databases. By using a homolog, it would be relatively easy to anticipate the configuration of a desired protein structure. [39][44]

1.4.2.2 Fold recognition methods

Fold recognition methods are designed to forecast the structure of amino acid sequences. A core principle underpinning these methods is that the structural aspect is more conserved in evolution than the sequence itself, i.e. a similar sequence exhibits a similar structure, while a similar structure does not necessarily imply a similar sequence. [39] As a result, the variety of distinct folds is more constrained than what might be inferred from the diversity of sequences. In light of this, fold recognition methods strive to identify a reference fold. This model pertains to a specific target sequence within the context of known folds, even in situations where there is no discernible sequence similarity. [45] Some of the disadvantages of these methods are their slowness, the need for human contribution in order to interpret the results and the inaccuracy of sequence-structure alignments. The lack of automation is perhaps the main problem since it is not practical when the volume of sequences is large. [46]

1.4.2.3. Comparative modeling

Comparative modeling stands as a reliable method for predicting the structure of a protein. This method compares and aligns the amino acid sequence of interest with one or more sequences with an already known native structure. [39] The procedural steps in this method encompass several stages, starting with fold assignment by reference. The initial step assesses the similarity between the target and at least one established template structure. Subsequently, the target-template alignment is established, followed by the generation of a model based in this alignment. Lastly, the model is meticulously evaluated for potential errors.

Fundamental to this method is the assumption that sequence similarity implies structural similarity. If the level of sequence similarity, indicated by the target-template identity, surpasses 50%, it can be inferred that the sequences likely share similar structures; the predictions are of high quality and followed by identification of the results by other methods as well. If the percentage of the degree of similarity is less than 30%, then the prediction will probably be incorrect. [47][48]

1.4.2.4. Artificial Neural Networks-AlphaFold

AlphaFold is an exceptionally advanced machine learning approach used for predicting the three-dimensional structures of proteins. By incorporating new neural network technologies and training procedures based on evolutionary and geometric constraints of protein structures, AlphaFold offers outstanding precision in protein structure prediction. [121]

Chapter 2: Molecular Dynamics Simulation

2.1 Introduction

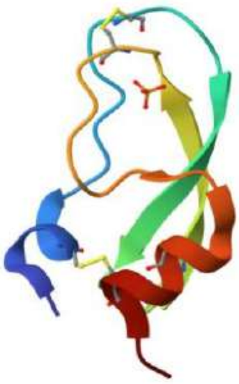


Figure 2.1: Structure of bovine pancreatic trypsin inhibitor (BPTI). (Adapted without permission from Protein Data Bank) https://www.rcsb.org/structure/5PTI

Molecular Dynamics studies the motion of molecules under the influence of interatomic forces. The first studies they were mainly applied to gases, because in tem, the particles move freely and they were easier to study. With the advent of fast computers these studies were extended to both liquids and solids. Molecular dynamics simulations were originally introduced by Alder and Wainwright in the late 1950s, aimed at studying hard-sphere interactions. [54] In 1974 Rahman and Stillinger conducted the pioneering molecular dynamics simulation of a realistic system, namely liquid water. In 1977 the first protein simulation was made, specifically the simulation of bovine pancreatic trypsin inhibitor (BPTI). (Figure 2.1) Subsequently molecular dynamics simulations have expanded their scope beyond solvated proteins to encompass protein-DNA complexes and lipid systems. They have proven instrumental in addressing various inquiries, including the thermodynamics of ligand binding and the folding mechanisms of small proteins. [49][50][51]

The main idea of molecular dynamics simulations is to change and evolve a system (change in position and velocity) as a function of time, so that the system goes through all possible states. In addition, it is possible to simulate systems that evolve at extreme values of variables, such as temperature and pressure, which cannot be studied experimentally. In addition molecular dynamics simulations are used in various experimental procedures such as x-ray crystallography and NMR structure determination. [52][53][106]

Molecular dynamics includes two main categories: MD (Molecular Dynamics Simulations) and MC (Monte Carlo Simulations). Of course, there are also hybrid techniques that combine features from both. MD simulations yield insights into the dynamic characteristics of systems, encompassing factors like transport coefficients, time-dependent perturbation responses, rheological properties and spectra but are also a useful tool for the theoretical study of the behavior of biomolecules over time, their structure and the interactions between of molecules. On the other hand MC simulations rely on statistical and probabilistic methods. [54][55]

In this way, the researchers manage to create an initial idea, for example about the structure and behavior of the molecule of interest. Finally computer simulations serve as a bridge between the microscopic length and time scales of molecular interactions and the macroscopic world, as well as the theory with experiment.

2.2. Classical Mechanics and Integration / molecular interaction

Molecular dynamics simulations rely on the principles of Newton's second law of motion (classical mechanics). Thus once the forces applied to each atom of the system, given its mass, are calculated, its acceleration, velocity and position during the simulation time can be determined. By completing the equations of motion, a trajectory is created with the velocities, accelerations and positions of the system's atoms that change as a function of time. Given therefore, that knowledge of positions and velocities is possible, to predict the state of the system at any instant, the equations are call deterministic or causal. [58][59]

So according to Newton's second law:

$\mathbf{F} = \mathbf{ma}$ ***(1),*** where F represents the overall force exerted on the particle, m signifies the mass of the particle and a denotes the acceleration of the particle.

The force (F) can also be expressed as a function of the change in potential energy:

$\mathbf{F} = -\frac{\boldsymbol{dV}}{\boldsymbol{dr}}$ ***(2),*** where dV is the change in potential energy and dr is the change in position of the particle

Combining equations (1) and (2), we get:

$\mathbf{\frac{dV}{dr} = m\ \frac{d^2r}{dt^2}}$ ***(3),*** where dt is the time duration and

$\mathbf{a = -\frac{1}{m}\frac{dV}{dr}}$ ***(4)***

In conclusion, for a trajectory simulation, you need the initial positions of the atoms, an initial distribution of velocities and information about the acceleration acting on the system. The initial positions of the atoms can be derived by conducting experiments such as X-ray crystallography and NMR spectroscopy. [60] On the other hand the initial velocity distribution can be calculated from the Maxwell-Boltzmann formula: $\boldsymbol{p(v) = \left(\frac{m}{2\pi k_B T}\right)^{1/2} \exp\left[-\frac{1}{2}\frac{mv^2}{k_B T}\right]}$ ***(5),*** where T is the temperature of the system, kB is the Boltzmann constant

It is a fact that the calculation of the acceleration is a complex process, since it is calculated through the potential energy using dynamic fields, which is calculable dependent on the positions of all atoms in the system, which amounts to 3N atomic positions. Given this, the integration of the equations of motion is numerically approximated through algorithms. The most common among them are:

- Verlet algorithm (simple algorithm, offers stability for relatively long time intervals, less accurate than Velocity verlet)
- Leap-frog algorithm (less accuracy in velocity calculation than Velocity verlet)
- Velocity verlet (calculates velocities in phase with positions)
- Beeman's algorithm (more accurate in calculating velocities than Verlet, although they produce the same products)

Most of these algorithms are based on Taylor series expansions. These series are used with the aim of reducing the number of terms of an equation so that its solution is easier. Nevertheless, the integration algorithms can show inaccuracies in their results, so the choice of the algorithm that should be used in each case must be made with prudence, control and criteria taken so that the results produced are as compatible as possible with the reality. [61][62][63]

2.3 Force fields

Force fields are empirical functions utilized in molecular dynamics simulations and aim to calculate the potential energy of a system of atoms and its forces as a function of the nuclear positions, the positions of the atoms and the interactions between them. Most force fields manage to combine computational efficiency and accuracy. [65]

In molecular mechanics, potential energy includes bonding or otherwise internal interactions, i.e. interactions between atoms connected by covalent bonds and non-bonded (non-covalent, external). The potential energy of the system is determined by summing the contributions from both bonded interactions and the non-bonded interactions. [66][67]

$$\boldsymbol{V(R) = E_{bonded} + E_{non\text{-}bonded}} \ \textit{(6)}$$

Bond interactions include bond length, bond angle and dihedral angle rotation. So Ebonded is a sum of these three terms:

$$\mathbf{E_{bonded} = E_{bonded\text{-}stretch} + E_{angle\text{-}bend} + E_{rotate\text{-}along\text{-}bond}} \ \textit{(7)}$$

Ebond-stretch refers to the energy associated with the interaction between two atoms that are covalently bonded. The bond energy depends on the displacement of the atoms from the original bond length, r_o. In formula 8, Kb represents the force constant that dictates the bond strength. Both the initial bond length and the force constant are unique to each pair of bonded entities in a molecular system.

$$\mathbf{E_{bond\text{-}stretch}} = \sum_{\mathbf{1,2\ pairs}} \mathbf{Kb}\ (\boldsymbol{r} - \boldsymbol{r_0})^2 \ \textit{(8)}$$

Eangle-bend refers to the change in the bond angle θ from the initial value θ_0. The values of θ_o and K_θ can vary, based on the specific chemical characteristics of the atoms involved in forming the angle.

$$\boldsymbol{E}_{\mathbf{bond\text{-}bend}} = \sum_{\mathbf{angles}} \mathbf{K_\theta}\ (\boldsymbol{\theta} - \boldsymbol{\theta_0})^2 \ \textit{(9)}$$

The Erotate-along-bond term factors in the system's potential energy arising from rotations of the dihedral angles. This potential displays periodic behavior, representing the steric barriers between atoms separated by three covalent bonds, often described by a cosine function.

The Enon-bonded term, which signifies nonbonded interactions, encompasses both van der Waals and electrostatic interaction energies. The Enon-bonded term is calculated by the following formula: $\mathbf{E_{non\text{-}bonded} = E_{van\text{-}der\text{-}Waals} + E_{electrostatic}}$ ***(10)***

It pertains to individuals separated by three or more bonds or to entities originating from distinct molecules.

The van der Waals interaction between two atoms emerges from a balance between repulsive and attractive forces. This interaction is described by the Lennard Jones potential. $\mathbf{E_{van\text{-}der\text{-}Waals} = \sum\nolimits_{nonbondedpairs} \left(\frac{A_i k}{r_i^{12} k} - \frac{C_i k}{r_i^{6} k}\right)}$***(11),*** where A and C are individual dependent constants. The likelihood of interaction between two atoms rises as their distance decreases.

These exists a specific distance, called the equilibrium distance at which the potential energy attains its lowest achievable value. In non-bonded pairs, if the the distance between the atoms becomes smaller than the equilibrium distance, repulsive forces emerge. Conversely as the distance increases, attractive forces become predominant.

The term Eelectrostatic is the electrostatic interaction energy and is described by Coulomb's law: $\mathbf{E_{electrostatic} = \sum\nolimits_{nonbondedpairs} \frac{q_i q_k}{Dr_{ik}}}$ ***(12),*** where D stands for the effective dielectric constant and r represents the distance between two atoms with charges qi and qk. [50] [63] [68]

However the development of parameter sets is an intricate endeavor, demanding meticulous optimization, parameterization and continuous enhancements to bolster their precision. Some of the most well-known and widely used dynamic fields are: [69] [117] [118]

- Assisted Model Building for Energy Refinement (AMBER)
- Chemistry at Harvard Macromolecular Mechanics (CHARMM)
- Groningen Molecular Simulation (GROMOS)
- Optimized Potential for Liquid Simulation (OPLS)

Although the aforementioned force fields employ similar calculation methods for potential energy, disparities emerge in terms of their parameterization techniques and the computation of both bond and non-bond interactions. The force fields undergo a continuous process of evolution and refinement to enhance the alignment between Molecular Dynamics results and experimental data, striving for improved accuracy. [65][113][114]

2.4 Solvent in Molecular Dynamics Simulation

The role of solvent in Molecular Dynamics Simulation is notably significant due to its impact on the molecule's structure, dynamics and the thermodynamic parameters of biological systems, as well as on the electrostatic interactions between molecules.

In Molecular Dynamics simulation two basic types of solvents can be used, the implicit and the explicit solvent. The implicit solvent model involves replacing the aqueous environment of discrete molecules by a continuous medium, thereby greatly reducing the number of tracked particles. Thus a dielectric constant occupies a position in the potential energy function. This method is particularly fast. [70] On the other hand, the explicit solvent model relies on a comprehensive approach involving the calculation of interactions between all atoms of both the solute and the solvent. This method has a high computational cost and requires the system to be bounded to prevent diffusion of the solvent molecules and to use a certain number of solvent molecules. [71]

Water solution, dimethyl sulfoxide (DMSO) solution and 2,2,2-Trifluoroethanol (TFE) solution were used in the present study. Water is the most common but at the same time important solvent in nature. It has a particularly important role thanks to the ability of its molecules to interact with other water molecules by forming hydrogen bonds. In addition, the residence time of water molecules and their diffusion characteristics deviate from those in the bulk and surface solvent regions. From a thermodynamic perspective, these distinctions can play a role in the formation of protein complexes. The solvent affects the determination of the structure of a molecule, the dynamics as well as the electrostatic, thermodynamic parameters and the functionality of the molecules, for this reason, its presence is considered essential in molecular dynamics simulation. [72][73]

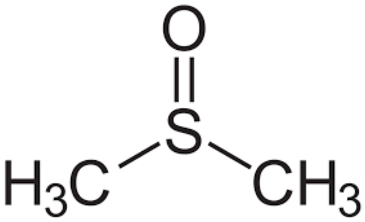


Figure 2.2: Structure of Dimethyl sulfoxide (Adapted without permission from Wikipedia) https://en.wikipedia.org/wiki/Dimethyl_sulfoxide

DMSO on the other hand, is a very common solvent in organic chemistry, chemical engineering and cell biology. It has a polar group S=O and two hydrophobic groups CH3 and offers the possibility of cell fusion, increasing membrane permeability and changing protein properties. (Figure 2.2) It is particularly capable of solubilizing various compounds such as hydrophobic helical peptides thanks to its low dielectric constant (ε=46.8) and its high dipole moment (4.0 D). [74][75][76][105]

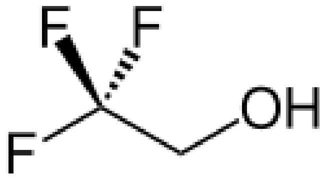


Figure 2.3: Structure of 2,2,2-Trifluoroethanol (Adapted without permission from Wikipedia) https://en.wikipedia.org/wiki/2,2,2-Trifluoroethanol

Finally, TFE is the most commonly used alcohol of the last decades. It has the ability to stabilize well-ordered conformations, either a-helix or β-sheet, and solubilize peptides, thereby promoting the formation of secondary structure in polypeptides and proteins. It possesses a relatively low dielectric constant (ε=26.7) and a small dipole moment (2.52 D). Remarkably, it does not interfere with the hydrogen bonds formed by amide and carbonyl backbone groups. In addition, it preserves the tertiary structure of proteins and does not disrupt van der Waals interactions due to its weak interaction with nonpolar amino acid residues. [76][77][78]

2.5 Tetra-F2W-RK peptide: A structural and functional overview

The Tetra-F2W-RK peptide is a significant antimicrobial peptide that has attached the interest of the scientific community due to its antimicrobial properties for combating a range of bacterial infections and unique structure. Regarding the composition of the peptide, Tetra-F2W-RK consists of a carefully sequence of nine amino acids (WWWLRKIWX), with an emphasis on the presence of four tryptophan (Trp) amino acids at positions W1, W2, W3 and W8. Additionally, it includes arginine (Arg) at position R5 and lysine (Lys) at position K6. The amphipathic nature of this peptide, combined with the complex three-dimensional arrangement of tryptophan residues, which constitutes approximately 50% of its composition, may be crucial for its antimicrobial activity. This structure enables the peptide to interact with bacterial membranes in a distinctive manner. Its amphipathic nature means that the structure contains both hydrophobic (lipophilic) and hydrophilic (water-loving) regions. This allows the peptide to penetrate bacterial membranes, causing increased inhibition of bacterial growth or even their death.

The three-dimensional structure of Tetra-F2W-RK, primarily characterized by an amphipathic alpha-helical configuration. This intricate arrangement, meticulously analyzed through nuclear magnetic resonance (NMR) techniques. (Figure 2.4) Furthermore, it is within this structure that the aromatic rings of Trp residues and other side chains occupy the hydrophobic side of the helical framework. An interesting observation by researchers regarding the peptide's structure is the

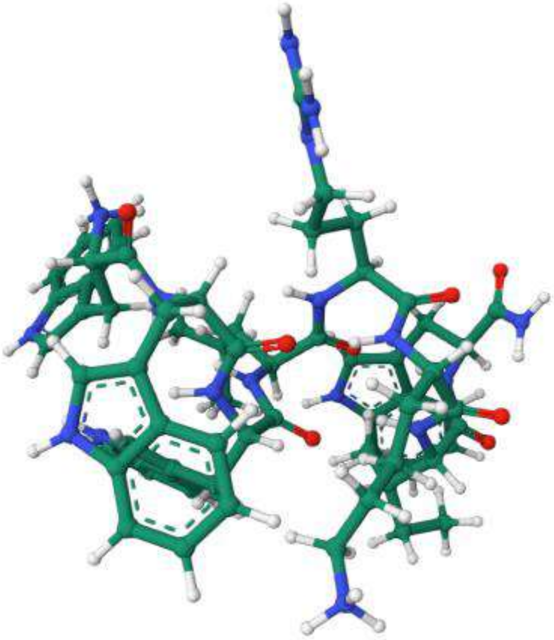


Figure 2.4: NMR structure of WW291 in micelle solvent (Adapted without permission from Protein Data Bank, PDB) https://www.rcsb.org/3d-

identification of a region where three tryptophan residues (W1, W2, W3) form a π arrangement, with W2 acting as the horizontal rod and W1, W3 as the two legs. This observation likely contributes to the stabilization of the structure and the antimicrobial activity of the peptide.

As established through research, Tetra-F2W-RK exhibits strong antimicrobial activity. It is capable of eradicating the human pathogenic bacterium Staphylococcus aureus USA300, known for its antibiotic resistance. Its antimicrobial action has been designed based on the peptide's structure and the interactions of the included amino acids. Despite its antimicrobial action, the peptide exhibits cytotoxicity against human red blood cells at specific concentrations. However, studies have highlighted its selectivity, making it a candidate for the development of antimicrobial agents that are more selective against pathogens and less harmful to the human organism. Moreover amino acids W1, W2, L4 and W8 in Tetra-F2W-RK play a pivotal role in its interaction with bacterial membranes. Alterations in these amino acids, such as substitutions with Arg and Lys, can influence the peptide's action on bacterial membranes.

The Tetra-F2W-RK peptide was deposited in the Protein Data Bank (PDB) with the identifier 6NM2 on January 10, 2019 and it became publicly available on July 15, 2020. The deposition was carried out by Zarena D. and Wang G., with funding from the National Instituted of Health/ National Institute of Allergy and Infectious Diseases (NIH/NIAID). According to the protein database Tetra-F2W-RK has a total structural weight of 1.27 kDa and consists of 93 atoms. It includes 9 modeled and 9 deposited amino acids, forming a unique protein chain. [104][120]

2.6 Purpose of thesis:

The purpose of this thesis is to investigate the folding behavior of the Tetra-F2W-RK peptide through Molecular Dynamics Simulations in common solvents, namely DMSO, TFE and water, and to compare it with the experimental data of researchers Zarena D. and Wang G., who utilized micelles. In conclusion, the primary objective of this study is to determine whether the peptide adopts the same conformation in the selected solvents as it does within micelles and, consequently, whether organic solvents can serve as viable alternatives to micelles.

Chapter 3: Methods

3.1 Introduction

To study through molecular dynamics the folding mechanism of the peptide TetraF2W-RK, the NAMD program was used and the AMBER 99SB-STAR-ILDN force field was used for its simulation. The NAMD program is software for simulating large biomolecular systems with high performance and presents compatibility with AMBER and CHARMM force fields. [79][107][116]

In order to reduce the computational cost and to increase the performance of the simulation, the parallel connection is preferred, i.e. a cluster of computers. The simulation was executed using Norma. [80] Norma is a stateless computing cluster belonging to the Beowulf-class. Norma operates on the Caos NSA GNU/Linux distribution. It is equipped with 40 CPU cores, boasting a total of 46 Gbytes of physical memory and 6 GPGPUs. These cores are distributed among 10 nodes, each powered by Intel's Q6600 Kentsfield 2.4 GHz quad processors. The nodes are interconnected through a dedicated HP ProCurve 1800-24G Gigabit Ethernet switch. Among these nodes, nine are outfitted with four cores, 4 Gbytes of physical memory and 2 gigabit network interfaces each. On the other hand, one node utilizes Intel's i7 965 extreme processor, boasting 6 Gbytes of physical memory. This particular node is also equipped with a GTX-295 card that possesses CUDA capability. Four out of the eight Q6600-basec nodes are enhanced with an nvidia GTX-460 GPU. The head node features four cores, 8 Gbytes of physical memory, 1.5 Tbytes storage in the form of a RAID-5 array of four disks, 3 gigabit network interfaces and an nvidia GTX-260 GPU. Norma serves predominantly for computational biology and crystallography initiatives undertaken by the Structural and Computational Biology group. This cluster is situated within the Department of Molecular Biology and Genetics of the Democritus University of Thrace. [80]

3.2 Simulation with NAMD

NAMD needs the following three kinds of files in order to perform the molecular dynamics simulation:

1. A PDB (Protein Data Bank) file, serves as a repository for the atomic coordinates and/or velocities, or forces pertinent to the system. These files can be created by the user or pulled from the pdb database. [81]
2. A force field parameter file, in which the numerical parameters for analyzing the potential of the system are stored. In addition it determines bond strength, equilibrium lengths, etc. In the present case, the parameter file of the AMBER 99SB-STAR-ILDN dynamic field was used.
3. A configuration file, in which the user specifies how NAMD should run the simulation. [82][83]

3.3 System preparation and simulation steps

The following figure shows the steps required to run a molecular dynamics simulation. Relatively, these steps are coordinate initialization, energy minimization, initial velocities assignment, heating, equilibration, temperature control, production phase and finally trajectory analysis. (Figure 3.1) [84]

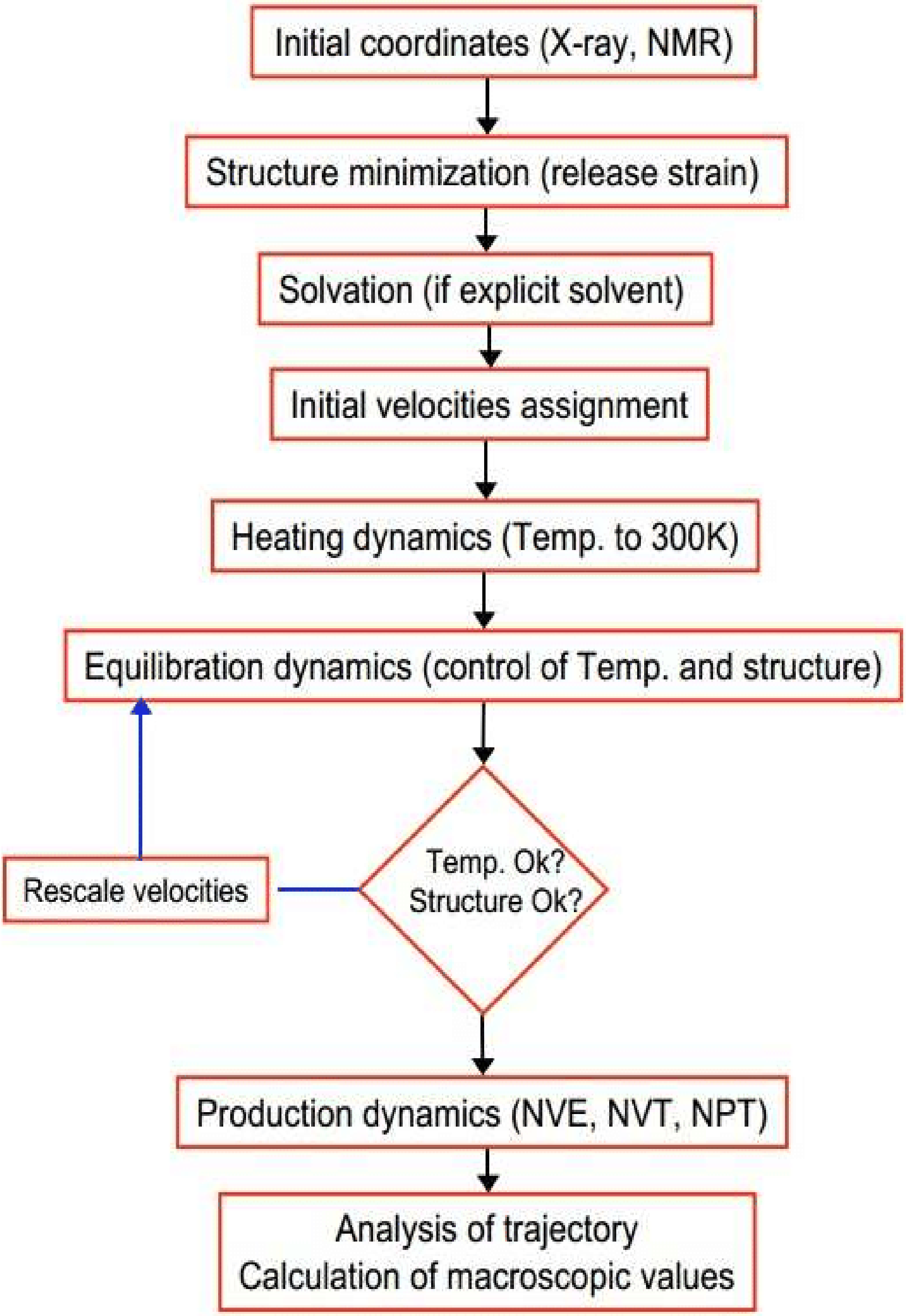

Initial coordinates (X-ray, NMR)
Structure minimization (release strain)
Solvation (if explicit solvent)
Initial velocities assignment
Heating dynamics (Temp. to 300K)
Equilibration dynamics (control of Temp. and structure)
Rescale velocities
Temp. Ok?
Structure Ok?
Production dynamics (NVE, NVT, NPT)
Analysis of trajectory
Calculation of macroscopic values

Figure 3.1: Steps followed during Molecular Dynamics simulation (Adapted without permission from Deep Biswas, A. (2015).) Deep Biswas, A. (2015). MOLECULAR DYNAMICS STUDY OF Anopheles gambiae ODORANT RECEPTOR (AgOR1) AND MOLECULAR DOCKING STUDY WITH 14 PREDICTED INHIBITORS. doi:https://doi.org/10.13140/RG.2.2.25604.94088

In MD simulation cases an initial configuration of the system is required for this reason, often an X-ray or NMR solved structure is used as the initial structure or a theoretical structure derived from homology modeling. In order to remove any van der Waals interactions to avoid structural distortions and by extension unstable simulation that will affect the generated data, before starting the simulation, the energy is minimized. During the minimization phase, the process involves systematically adjusting the positions of the entities while calculating the local energy each time. This is done in order to locate and approach local energy minima. [85]

The next step was the heating phase in which the initial speeds are set to low temperatures. So the simulation begins and during it, periodically, new velocities are assigned to a slightly higher temperature until the desired temperature is reached. In the specific cases that will be analyzed in more detail below, the temperature limits were 280K and 380K and the temperature was elevated in increments of 20K until reaching the final target of 320K over duration of 32 picoseconds (ps).

When this temperature is reached, the simulation moves to the equilibrium phase, i.e. the examination of various properties such as pressure, structure, temperature and the energy, until the properties become constant with respect to time. The equilibrium phase refers to Newton's second law and applies to every atom of the system by determining its orbital. In the event of a significant increase or decrease in temperature, the velocities are adjusted proportionally to ensure that the temperature returns to the desired value again. Temperature and pressure control were achieved using the Nosé-Hoover Langevin dynamics and Langevin piston barostat control methods. [79][86][87][110] In the present cases the pressure was kept at 1 atm. Also the long-range electrostatic interactions were calculated using the Particle Mesh Ewald (PME) method. In the specific cases calculated every two time steps with a grid spacing of approximately 1 Å and tolerance of 10^{-6}. Finally the SHAKE algorithm was employed to uphold constraints on bonds involving hydrogen atoms, adhering to a specific tolerance threshold of 10^{-8}. [79][84][86][88][89]

The last phase of the simulation is the production phase, where the simulation runs for a required amount of time ranging between several hundreds of ps and ns or even more, depending on the individual characteristics of each different simulation. So in the production phase, the coordinates, velocities and energy of the system at different times recorded in the previous phase are used. For the production phase, the Verlet-I multistep integration algorithm was used to calculate the velocities and coordinates of the atoms. [84]

The folding simulation dynamics study of the TetraF2W-RK peptide was performed using the NAMD program and the AMBER99SB-STAR-ILDN force field. This force field was used because it has been repeatedly shown to be able to correctly fold many peptides. DMSO, TFE and water were used as solvents.

Specifically to study the simulated folding dynamics of the peptide using DMSO, the NAMD program was used for a large set of 15 μs. The process of solvating and ionizing the system was executed using the LEAP program, a component of the AMBER tools distribution. Periodic boundary conditions were used, a cubic unit cell large enough that the minimum separation between adjacent cells to be at least 16 Å and adaptive tempering method, which is typically equivalent to simulation single-copy exchange folding with a continuous temperature range. The system underwent an energy minimization process involving 2000 coupled gradient steps. Furthermore the system was equilibrated for duration of 10ps under constant temperature and pressure (NpT conditions) until volume equilibration was achieved. The Langevin damping factor was established to 1 ps^{-1} and the piston oscillation period was set to 400 fs, with a decay time of 200 fs. For the production phase the internal time step was 2.5 fs with non-bonded interactions calculated one step at a time. The cutoff for van der Waals interactions was configured at 8 Å using a transfer function. The trajectory was generated by saving atomic coordinates at intervals of 1.0 ps. The entire simulation spanned duration of 15 μs, generating a total of 15.807.750 frames.119]

To study the dynamics of peptide folding simulations with TFE and water as solvents, the NAMD program was also used but for 10 μs and 3,95 μs respectively using the TIP3P water model. [89][107] Similarly to the case were DMSO was the solvent, an adaptive tempering method was applied. The first energy minimization of the systems was for 1000 coupled gradient steps. Equilibration was done for 10 ps in NpT conditions without limiting boundaries until volume equilibration. The Langevin damping factor was adjusted to 1 ps^{-1} while the piston oscillation period was established at 200 fs, alongside a decay time of 100 fs. In the case of water simulation, the time step was 2 fs and for the TFE simulation was extended to 2.5 fs. The cutoff for van der Waals interactions was applied to 9 Å through a switching function. The trajectories were derived by storing their atomic coordinates system every 0.8 for water and 1.0 ps for TFE simulations. The simulations with solvent TFE and water resulted in 10,000,000 and 3,950,400 frames respectively.

Chapter 4: Results

4.1 Introduction

Two programs were mainly used to study the folding of the Tetra-F2W-RK peptide. CARMA and GRCARMA are computer programs that help with trajectory analysis. [90][91][108] These programs necessitate the utilization of two input files: a dcd file and a psf file. The dcd file encapsulates the simulation trajectory, providing the coordinates of all atoms throughout the simulation. Also a frame corresponds to a distinct set of coordinates. On the other hand, the psf file (protein structure file) includes the structural details such as atoms, angles, bonds and other pertinent information. [92]

To achieve the purpose of this work as already mentioned above, the following analyzes were also used:

- RMSD matrix (Root Mean Square Deviation) analysis from a selected reference structure based on mean deviation. RMSD matrices are frame-to-frame comparison of all peptide conformations observed during the simulation. [93]
- Analysis of secondary structure through the program STRIDE (STRuctural Identification) and the logo-generating program Weblogo
- Analysis based on clustering (cluster analysis)
- Principal Component Analysis (PCA)

4.2 RMSD analysis

The Root Mean Square Deviation (RMSD) is a widely used structural biology technique for the analysis of macromolecular structures and dynamics. It has the ability to calculate the average distance between atoms and serves as a quantitative measure for comparing the structure of a partially folded protein and the structure of the protein when it is in its native conformation. [109] RMSD is calculated by the following equation where values are shown in Å:

$$\mathbf{RMSD= \sqrt{\Sigma(x_i - x_{ref})^2} / N,}$$

Here, x_i, signifies the coordinates of the individuals at a specific time, x_{ref} denotes the coordinates of atoms within the reference molecule and N stands for the total number of atoms.

The lower the RMSD value, the more similar the two structures are. When the RMSD value is equal to 0.0Å, the two structures do not show structural differences and therefore they are identical. Generally two structures show quite large similarities when the RMSD value is less than 2.0Å.

The RMSD result is presented in a table where all the structures obtained from the simulation obtained from the simulation are placed in ascending order on the axes. The RMSD value is shown by color switching. Therefore the blue color corresponds to low RMSD values i.e. to stable structures, the red color to high RMSD values and the yellow color in the intermediate RMSD values. The blue areas shown on the diagonal line of the table denote a structure that remains constant for a time interval that it is proportional to the extent of this area, while the remaining blue areas outside the diagonal correspond to similar structures that appeared during the simulation.

The RMSD matrix produced by GRCARMA with step 3500 for DMSO is shown below:

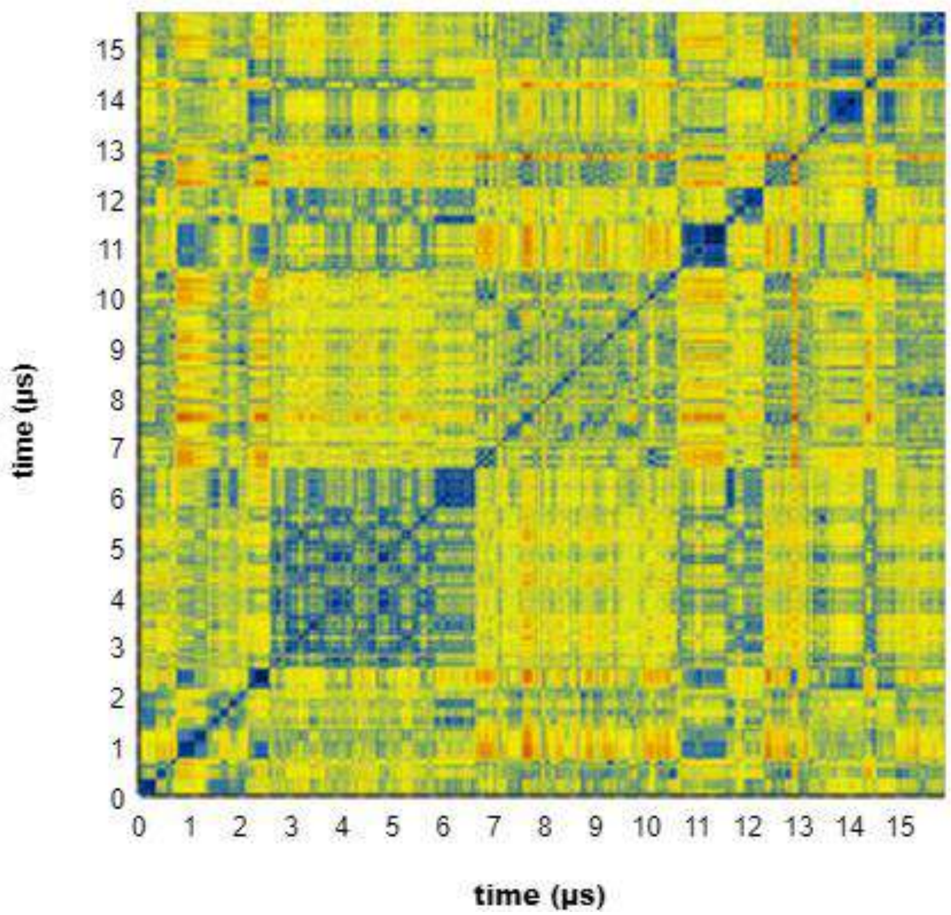


Figure 4.1: RMSD matrix diagram for Tetra-F2W-RK when solvent is DMSO. The number of total frames is 15,807,750. It is produced by GRCARMA.

The peptide appears to show several discrete moments where it displays a stable structure but not for a long time since the blue discrete regions are small in extent. These regions are located at time 0.3 μs, 1 μs, 2.5 μs, 11.2-11.6 μs, 13.5 μs and 13.8 μs.

On the other hand the use of TFE as solvent produced the following RMSD matrix with step 2000 respectively:

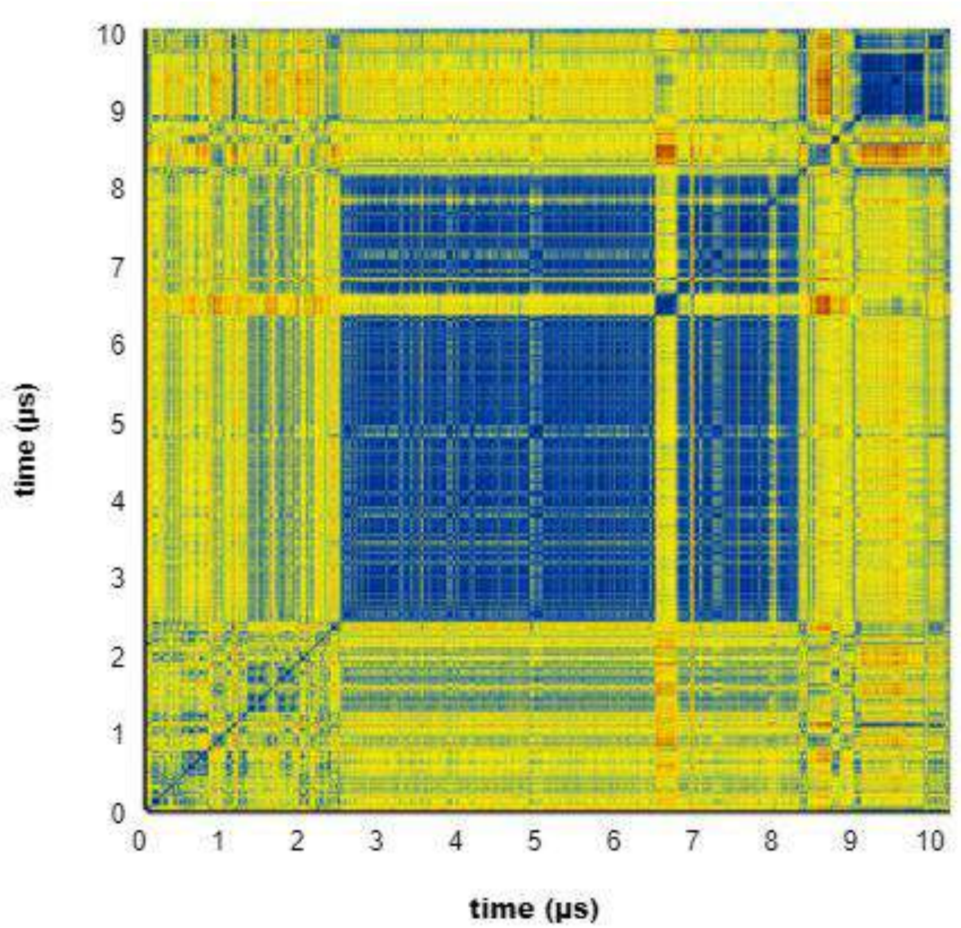


Figure 4.2: RMSD matrix diagram for Tetra-F2W-RK when solvent is TFE. The number of total frames is 10,000,000. It is produced by GRCARMA.

In the present case the peptide seems to be quite stable for a long time. Specifically the time intervals that indicate stable structures (blue areas) are 9.-10 μs etc. However at 2.5-8.3 μs, peptide exhibits a stable structure too.

Finally in the case where water was used as solvent, grcarma with a step of 500 produced the following RMSD matrix:

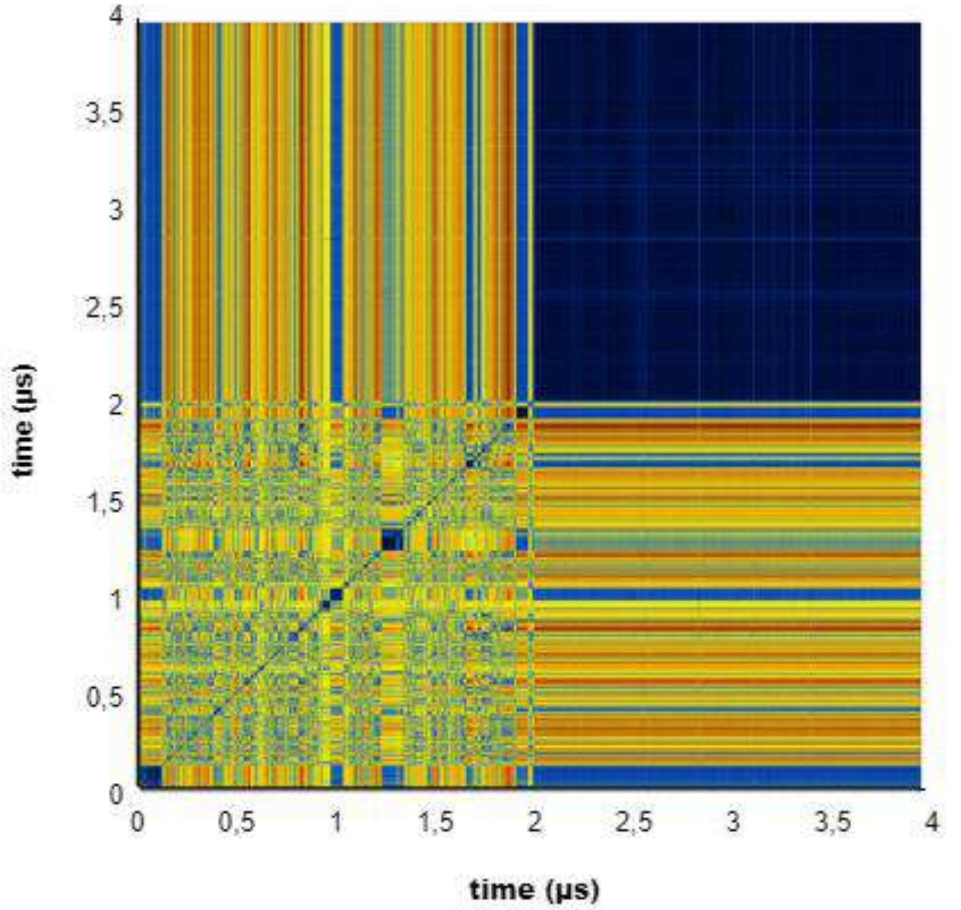


Figure 4.3: RMSD matrix diagram for Tetra-F2W-RK when solvent is water. The number of total frames is 3,950,400. It is produced by GRCARMA.

It is clear that the peptide shows small blue areas and eventually acquires a stable structure for a long time, specifically from 2-3.95 μs. The remaining stable structures appear at times 1μs, 1.3μs, 1.8 etc.

4.3 Secondary structure analysis

Knowledge of secondary structure is an important step in determining 3D protein structures. For this purpose, i.e. the identification of the main structural features adopted by the peptide during the simulation, the program STRIDE was used. STRIDE is a secondary structure assignment software tool based on an automated algorithm that uses a combination of hydrogen bond energy and statistically derived backbone dihedral angle information. [94]

The result is a colored diagram of the simulated trajectory. Thus the colors encode elements of secondary structure, in more detail:

COLOR	SECONDARY STRUCTURE ELEMENT
PINK	A-HELICES
PURPLE	3-10 HELICES
YELLOW	B-SHEET
BLUE	TURNS
WHITE	RANDOM COIL

Table I: Color coding for secondary structure

In addition, for better understanding and more accurate description of the secondary structure of the peptide, WebLogo charts were created. WebLogo is a method that generates sequence logos of sequence representations. Each logo consists of stacks of letters, where each stack refers to one position in the sequence. The mapping of the letters and secondary structure elements is the following:

- H for α-helix
- G for 3-10 helix
- I for π-helix
- E for β-sheet
- B for β-bridge
- T for turn
- C: coil (none of the conformations mentioned above)

Here are the results obtained by STRIDE and WebLogo in the first case considered, where DMSO was used as solvent and the step was 300.

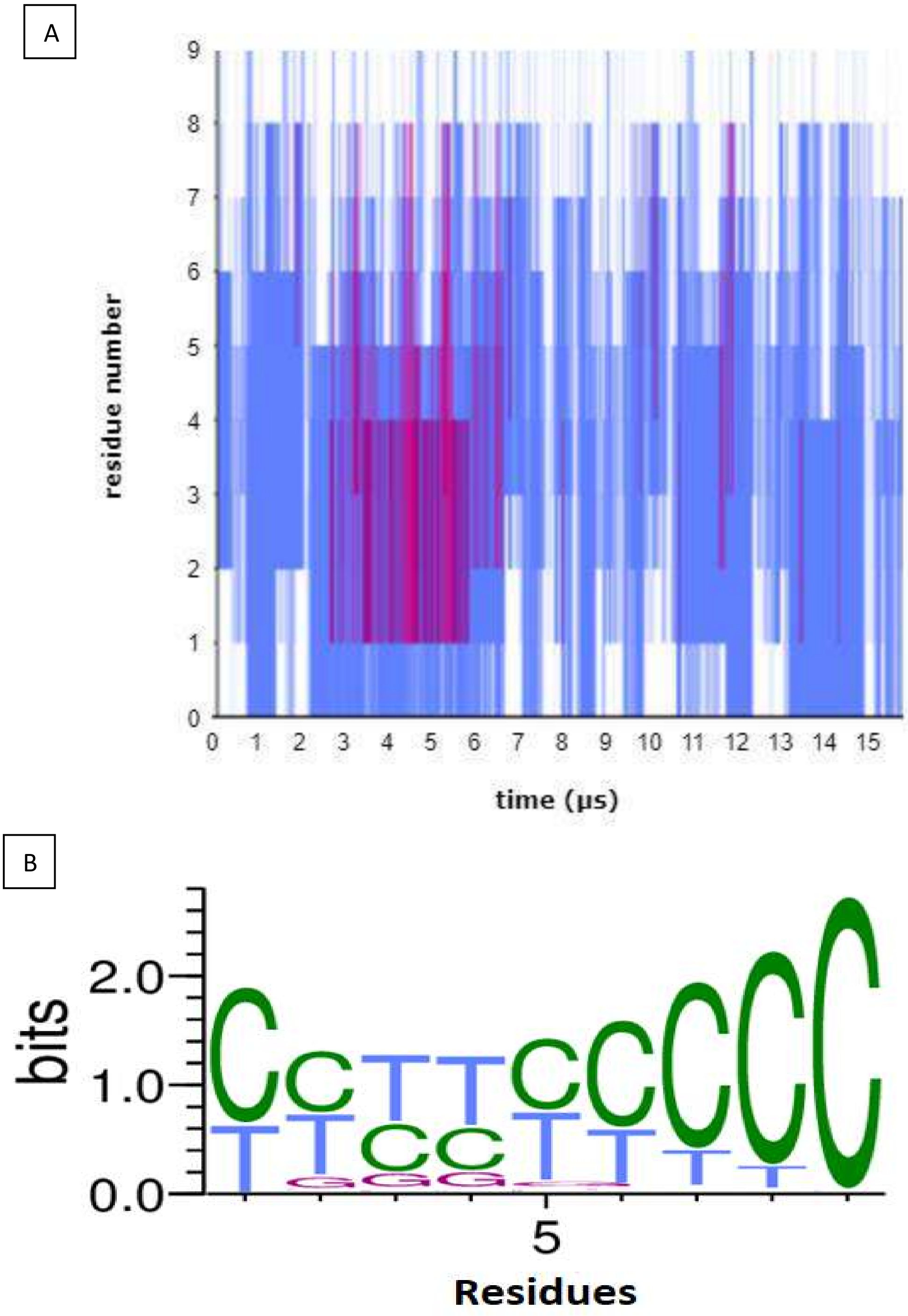


Figure 4.4: Secondary structure analysis for Tetra-F2W-RK when solvent is DMSO. [A] Secondary structure diagram created by STRIDE. [B] The depiction generated from WebLogo, which illustrate the secondary structure assignments based on STRIDE analysis for each residue, across all frames of the simulation.

Based on the STRIDE analysis (Figure 4.4A), we can observe a strong preference for configuration of turns and 3-10 helices since the colors blue and purple are the most noticeable in the shape. The pink color indicating the α-helices can be hardly distinguished so it does not seem to affect the structure of the peptide much. Based on WebLogo graph (Figure 4.4B) we can more precisely observe the same data. In more detail, residues 2-6 show 3-10 helix configurations.

In the case where the solvent was TFE, the results of STRIDE and WebLogo with step 500 are as follows:

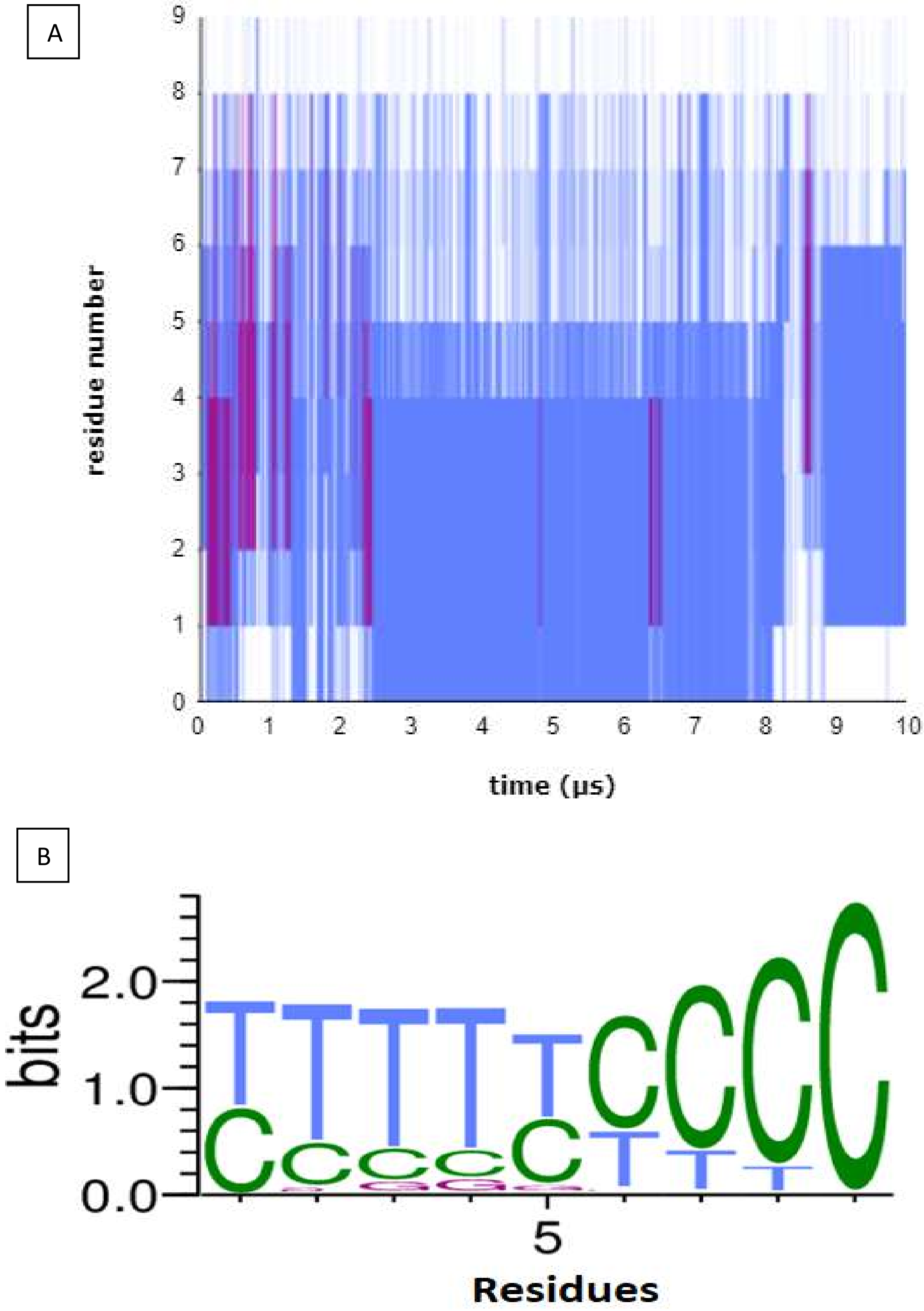


Figure 4.5: Secondary structure analysis for Tetra-F2W-RK when solvent is TFE. [A] Secondary structure diagram created by STRIDE. [B] The depiction generated from WebLogo, which illustrate the secondary structure assignments based on STRIDE analysis for each residue, across all frames of the simulation.

According to STRIDE analysis (Figure 4.5A), the peptide in this case appears to predominantly use the turn and 3-10 helix conformations, with blue and purple colors respectively predominating. Likewise, the WebLogo graph (Figure 4.5B) confirms this data. Residues 2-6 show a 3-10 helix conformation but in this case not with the same frequency (letter height). In addition, in both cases complete identification of the last 4 residues is observed.

In the last case considered where the solvent was water, the programs STRIDE and WebLogo produced the following results with step 500:

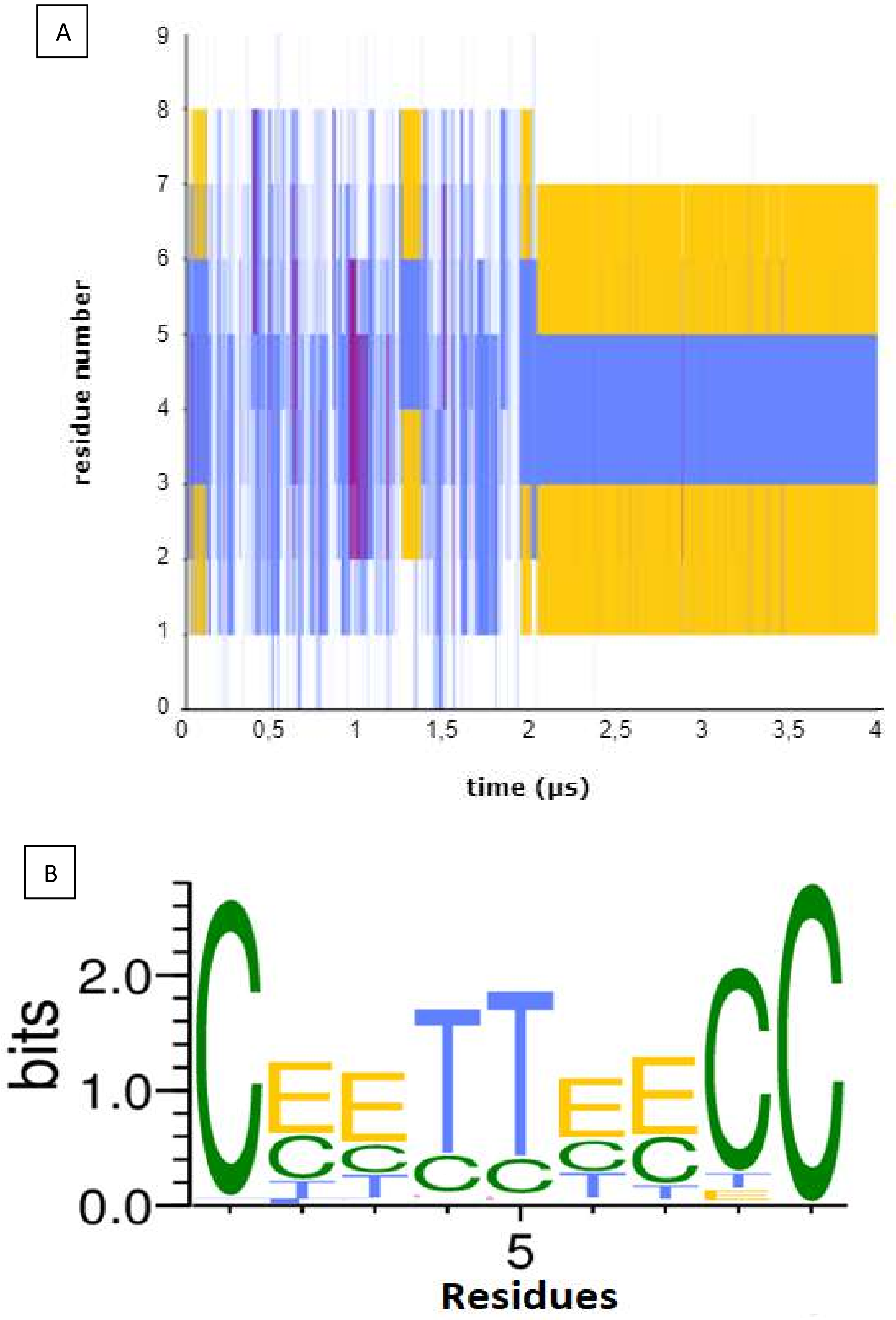


Figure 4.6: Secondary structure analysis for Tetra-F2W-RK when solvent is water. [A] Secondary structure diagram created by STRIDE. [B] The depiction generated from WebLogo, which illustrate the secondary structure assignments based on STRIDE analysis for each residue, across all frames of the simulation.

As we can be seen for the STRIDE analysis (Figure 4.6A), in this case the peptide takes more conformations than in the previous two cases. With predominant blue and yellow colors as is evident, the turns and the β-sheets are in the majority, while the purple color is also subtly observed, i.e. the 3-10 helices. The WebLogo (Figure 4.6B) gives a clearer picture of this analysis. Residues 2-3 and 6-7 show β-sheets and residues 4-5 show a turn, making a β-hairpin evident. Furthermore the presence of the letter “C” at the ends is evident, i.e. the existence of a coil.

4.4 Principal component analysis (PCA) and cluster analysis

Principal component analysis (PCA), also known as quasiharmonic analysis or the essential dynamics method holds a prominent status as one of the most extensively employed statistical methods, especially in the field of Molecular Dynamics. It is a multifactorial technique applied to the systematic reduction of the dimensions required to describe the dynamics of the proteins of a complex system, so that their analysis is possible. An important element of this approach is a covariance matrix, which provides information on the correlations of two system points. Thus, PCA constitutes a linear transformation that diagonalizes the covariance matrix, thereby unveiling the instantaneous linear correlations existing among the variables. [95][96]

There are two categories of Principal Component Analysis used in MD simulation data analysis, Cartesian Principal Component Analysis (cPCA) and Dihedral Principal Component analysis (dPCA). Cartesian PCA involves dimensionality reduction by considering the Cartesian coordinates of the atoms that define the atomic displacements in each conformation. On the other hand Dihedral PCA focuses on dimensionality reduction based on the φ,ψ dihedral angles of the main chain. Although Cartesian PCA is a useful technique for studying protein structure, it has some drawbacks. Mixing internal and global motions can lead to artifacts and, by extension, failure to discriminate configurations. This disadvantage was proven by Yuguang Mu in a study on the reversible folding and unfolding of pentaalanine in explicit water, where the PCA performed with Cartesian coordinates did not work in this particular case. In order to circumvent issues stemming from the circular nature of these variables, a solution known as Dihedral PCA method was introduced. This involves transforming the space of dihedral angles into a linear metric coordinate space (i.e. a vector space with a well-defined distance between two points) using the trigonometric functions like sinφ and cosφ. Dihedral PCA can therefore yield more precise data, as internal coordinates such as bond lengths and bold angles generally experience limited amplitude changes. As a result this analysis directly engages with the relevant part of the dynamics, thereby eliminating superfluous noise. In addition, the use of internal coordinates avoids physical problems related to the mixing of internal and total motion, while at the same time since Dihedral PCA is based on backbone dihedral angles, it can distinguish the main conformational states of peptides. It is worth noting that in PCA analysis of trajectories from a protein simulation, it is important to integrate both. [97][98][99][115]

Cluster analysis

Clustering is the process of dividing a large group of objects (data) into smaller groups. The purpose of this is to gain a deeper understanding of the trajectory information and analyze it effectively. In each group the data share more similarities (i.e., similar molecular configurations) compared to data in other groups. This categorization helps identify patterns of motion. Cluster analysis is typically applied to multivariate data in which numerous measurements are made on a group of objects, but there is no prior knowledge of the group structure of the data, assuming that such structure exists. It is crucial to highlight that cluster analysis can be employed as part of a subset of multidimensional scaling techniques, which may include principal component analysis, principal coordinate analysis, or nonlinear mapping. In conclusion, PCA data can further categorized into clusters, depending on the similarities of the data. [100][101][111]

In the 3 cases of solvents examined (DMSO, TFE, water) dihedral Principal Component Analysis (dPCA) was used, after adjusting the peptide by removing the NME (N-methyl amide capping group) residue. The analyses were done with the grcarma program, with a maximum limit of 10 clusters.

In the *case I* where the solvent is DMSO, 10 clusters were produced. The following Table II shows the populations of each cluster obtained from the dPCA analysis for DMSO. A total of 58,63% were included in clusters

(3376962 out of 5759548). The two main clusters are cluster 1 with 1008479 frames out of 5759548 (17,5%) and cluster 3 with 595975 frames out of 5759548 (10,34%)

Cluster	Frames (out of 5759548)	Percentages
1	1008479	17,5 %
2	269038	4,67%
3	595975	10,34%
4	375033	6,51%
5	172584	2,99%
6	163450	2,83%
7	257851	4,47%
8	227666	3,95%
9	216661	3,76%
10	90225	1,56%

Table II: The populations of the ten clusters obtained from the dPCA analysis for DMSO along with their percentages.

In the *case II* where TFE was the solvent, 7 clusters were produced. The following Table III lists the populations of each cluster from the dPCA analysis for TFE. In this case, 100% were included in clusters, with main clusters among them, cluster 1 with 1183642 frames out of 2315880 (51,1%), cluster 2 with 388467 frames out of 2315880 (16,77%) and cluster 3 with 315275 frames out of 2315880 (13,61%).

Cluster	Frames (out of 2315880)	Percentages
1	1183642	51,1%
2	388467	16,77%
3	315275	13,61%
4	151578	6,54%
5	127413	5,5%
6	88600	3,82%
7	60905	2,62%

Table III: The populations of the seven clusters obtained from the dPCA analysis for TFE along with their percentages.

Finally, in the *case III*, where water was used as solvent, 6 clusters were produced. The populations of each cluster from the dPCA analysis for water are shown in the following Table IV. Similarly to the case of TFE, and in this case 100% were included in clusters, with main cluster, cluster 1 with 1857559 frames out of 1950116 (95,25%)

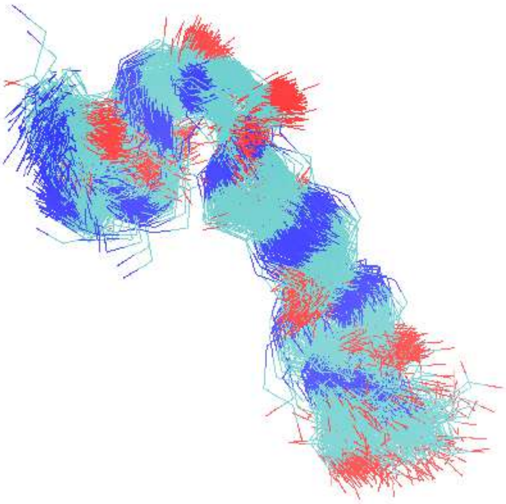
Cluster	Frames (out of 1950116)	Percentages
1	1857559	95,25%
2	36019	1,84%
3	19510	1%
4	15849	0,81%
5	16399	0,84%
6	4780	0,24 %

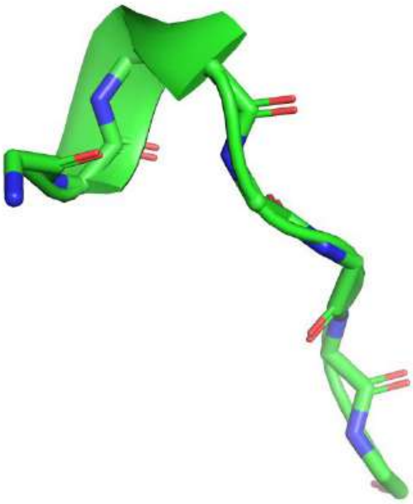
Table IV: The populations of the six clusters obtained from the dPCA analysis for water along with their percentages.

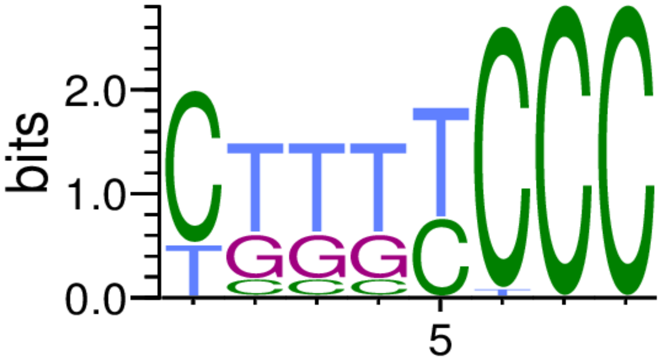
All the resulting clusters in each case are then further analyzed. The 1st image to the left of each cluster shows the 3D model as generated through the VMD (visual molecular dynamics) program, using the superposition files produced by dPCA. VMD is a program for the presentation and study of molecular assemblies, such as biopolymers and supports a large variety of rendering and coloring techniques and may display any number of structures at once. The color coding for backbone structures is cyan for C atoms, blue for N atoms and red for O atoms. [102] To the right of this image, in each cluster is shown the configuration of each of them. These configurations are derivatives of the PYMOL program and were obtained from the representative files produced by dPCA. PYMOL is a molecular visualization program for rendering and animating 3D structures. The colors represent various molecular components of the residues. Green stands for the carbon groups, red for the oxygen groups and blue for the nitrogen groups. [103] The 3rd image following each cluster depicts a WebLogo diagram specific to each cluster. As already mentioned above the letters used represent for α-helices (H) , 3-10 helices (G), π- helices (I), β-sheets (E), turns (T), β-bridge (B), random coils (C).

Case I- Clusters when solvent is DMSO

1st CLUSTER

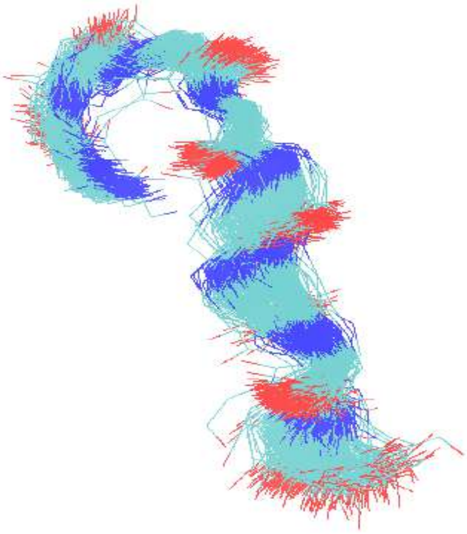


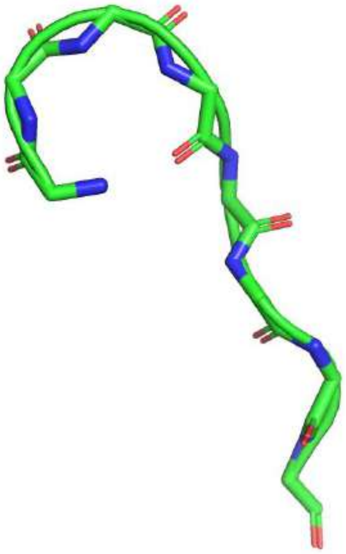


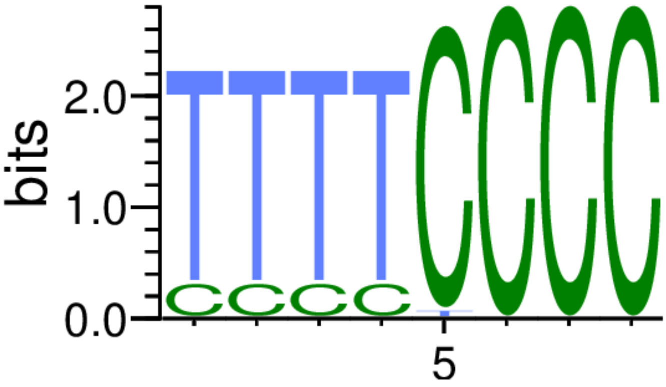


According to the three images, cluster 1 (most populated cluster) presents in residues 2-4, a 3-10 helix conformation, while the rest of the peptide forms turns and coils.

2nd CLUSTER

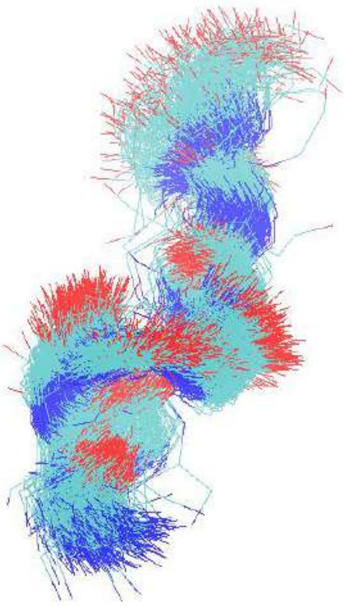


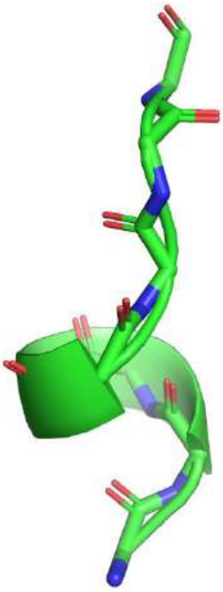


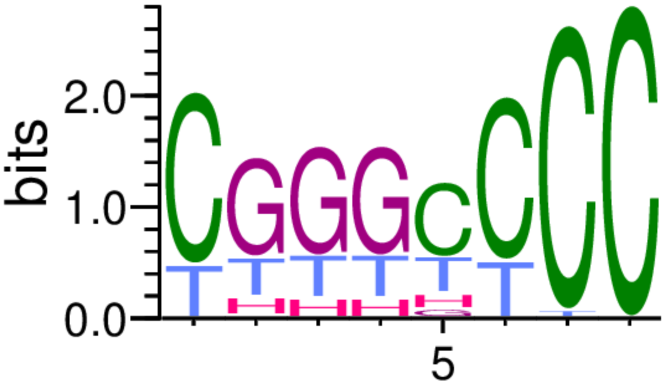


Cluster 2 features only turns and random coils.

3rd CLUSTER

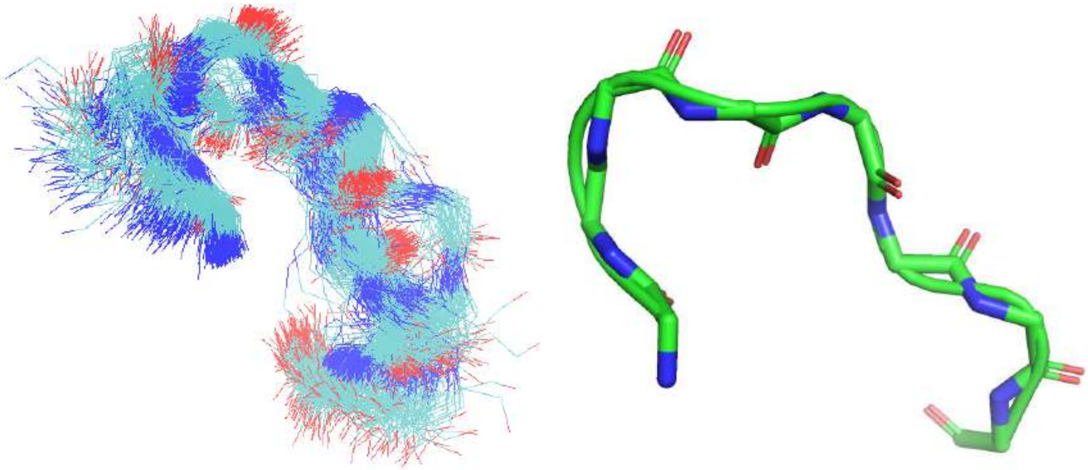


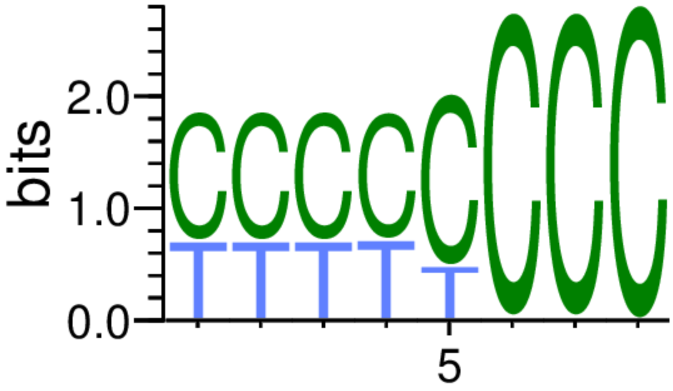




Cluster 3 (second most populated cluster) displays in residues 2-4, 3-10 helix conformation and in a particularly low frequency α-helix conformation in residues 2-5, the rest of the peptide presents turns and coils.

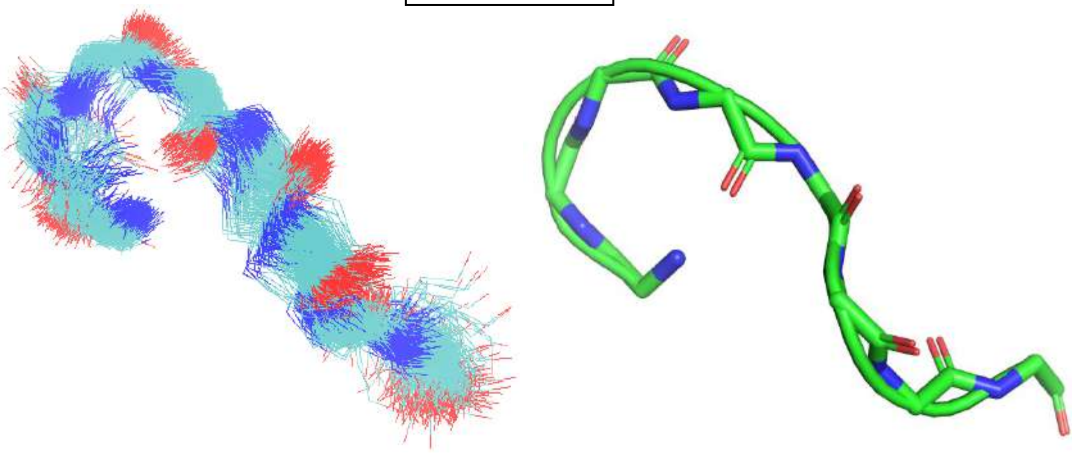
4th CLUSTER

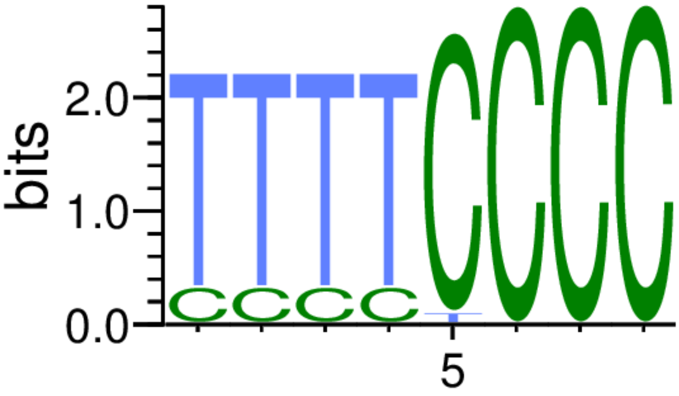




Turns and random coils are the only elements of Cluster 4.

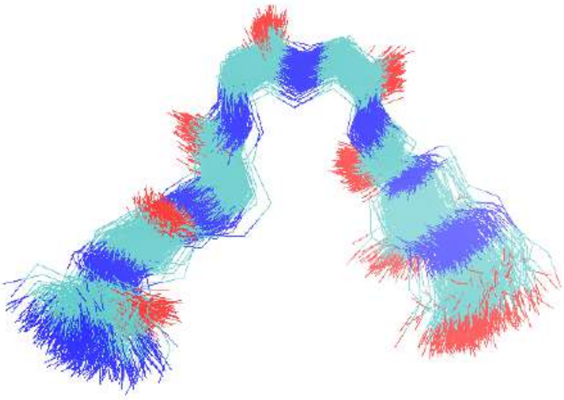
5th CLUSTER

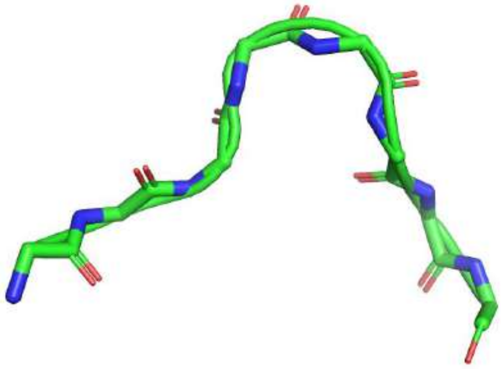


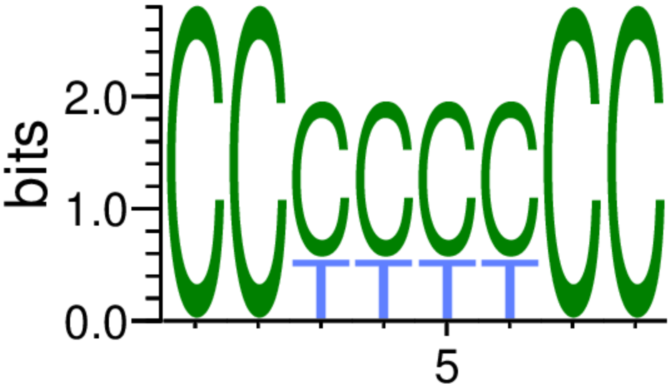


Only turns and random coils are present in cluster 5.

6th CLUSTER

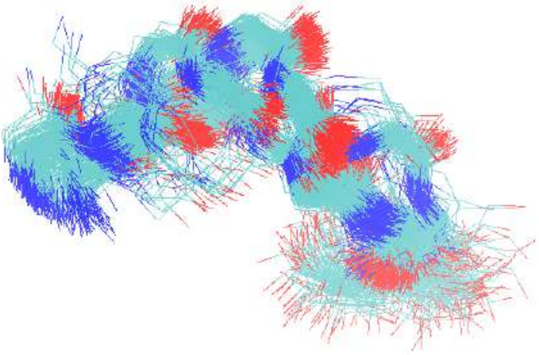


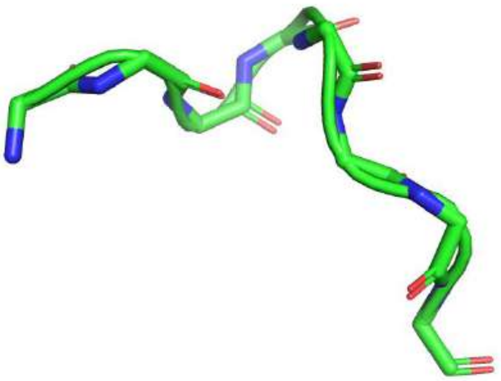


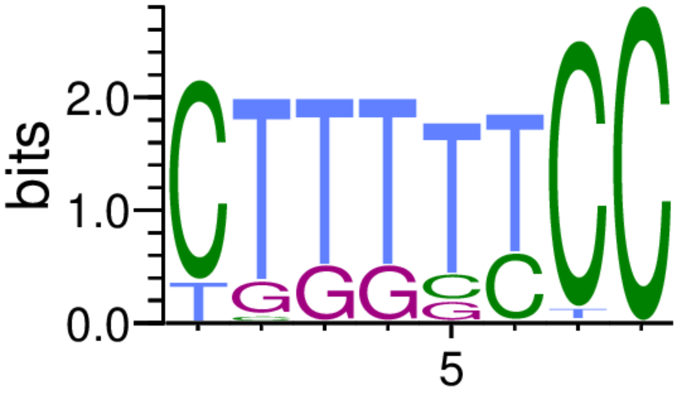


Cluster 6 forms only turns and coils.

7th CLUSTER

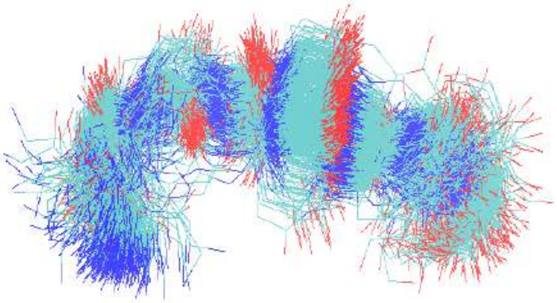


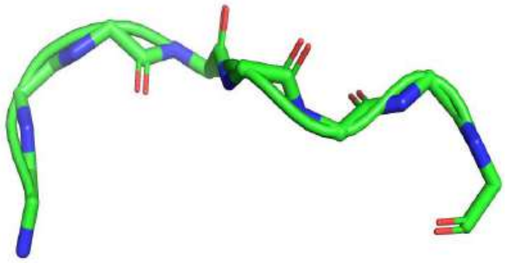


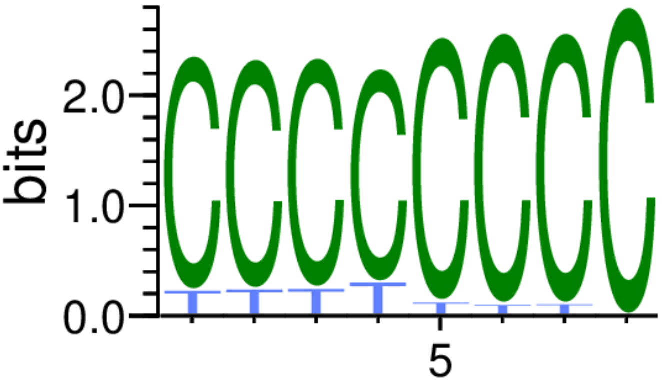


Cluster 7 forms turns and coils, while at low frequency residues 2-5 built a 3-10 helix conformation.

8th CLUSTER

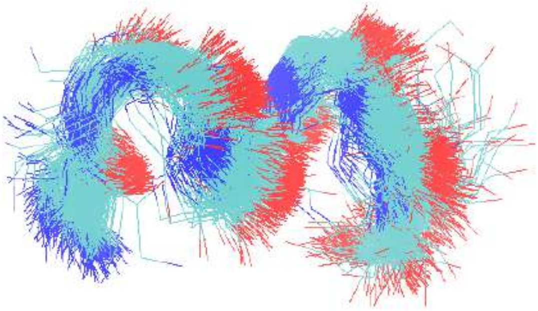


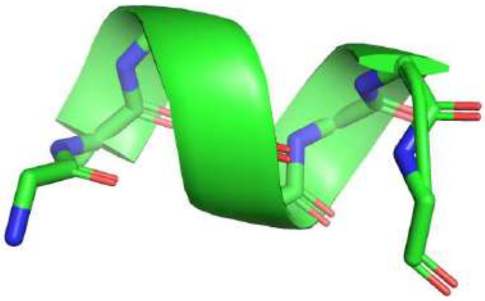


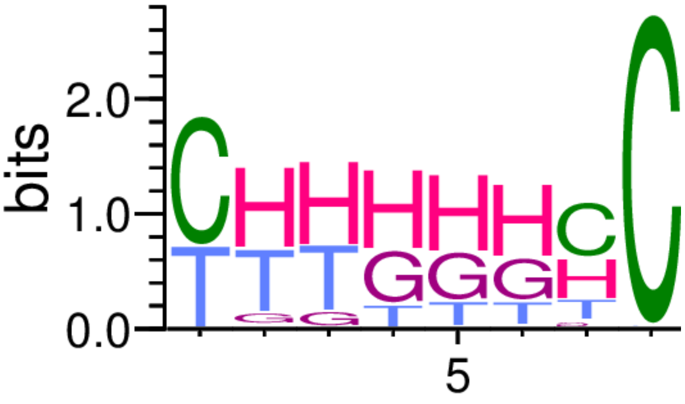


The main structural characteristic in cluster 8 is random coil, while at low frequency turns are also observed.

9th CLUSTER

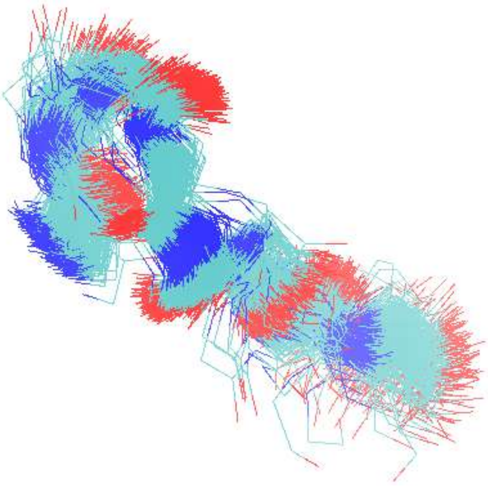


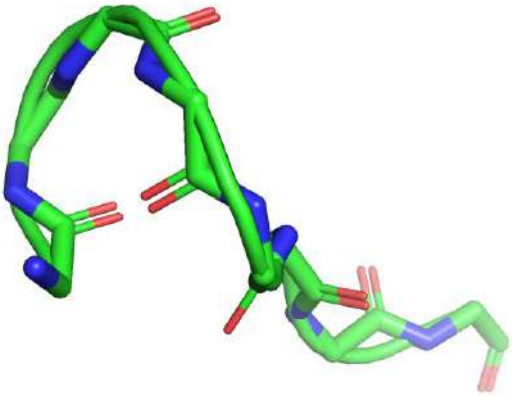


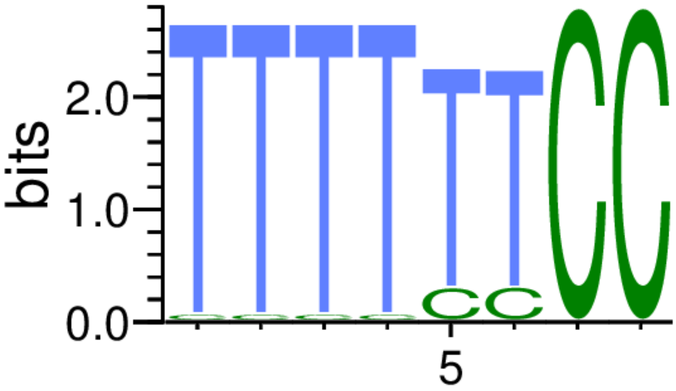


In cluster 9, residues 2-7 build an α-helix conformation, less frequently residues 2-6 forms 3-10 helix conformation and the rest of peptide includes turns and coils.

10th CLUSTER





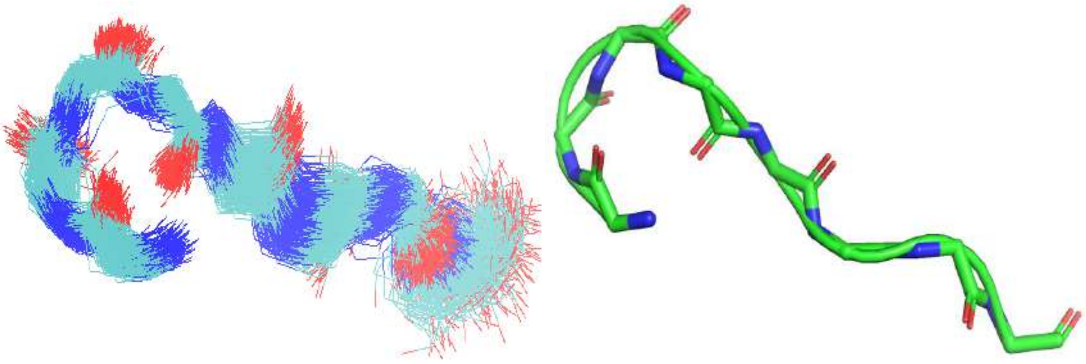


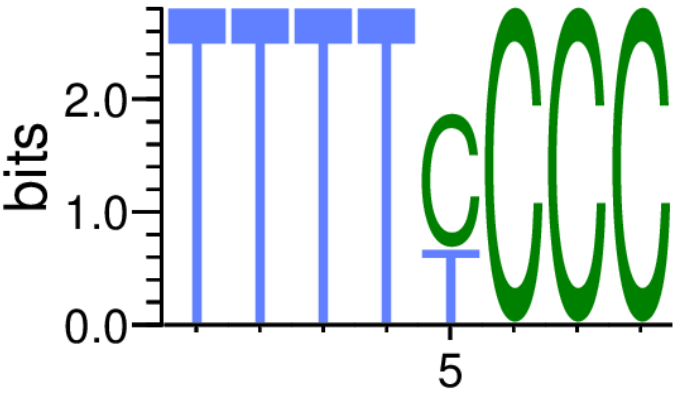
Cluster 10 includes only turns and random coils.

Figure 4.7: Superposition of structures (left) that belong to each cluster derived from the dPCA analysis for DMSO. The superpositions appear complex and noisy. Representative of each cluster, it is produced by Pymol program (right). WebLogo graph (3rd image) of each cluster, as derived from STRIDE analysis.

<u>Case II- Clusters when solvent is TFE</u>

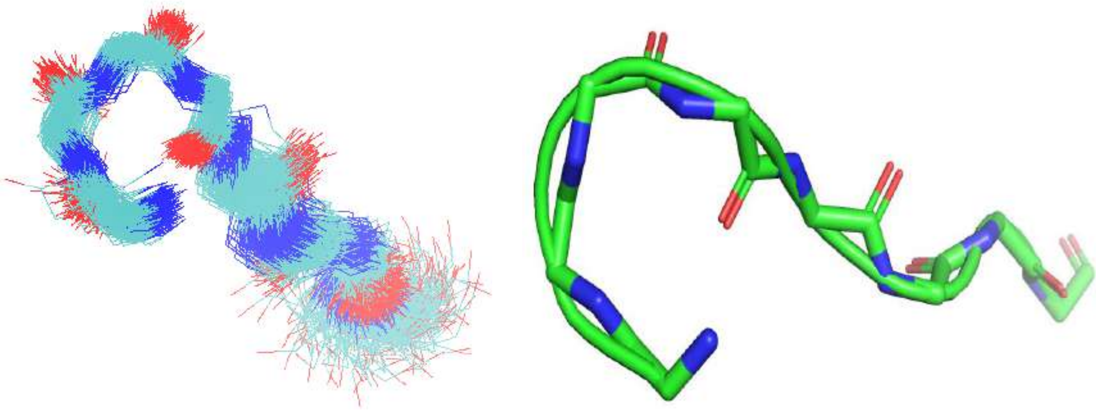
1st CLUSTER

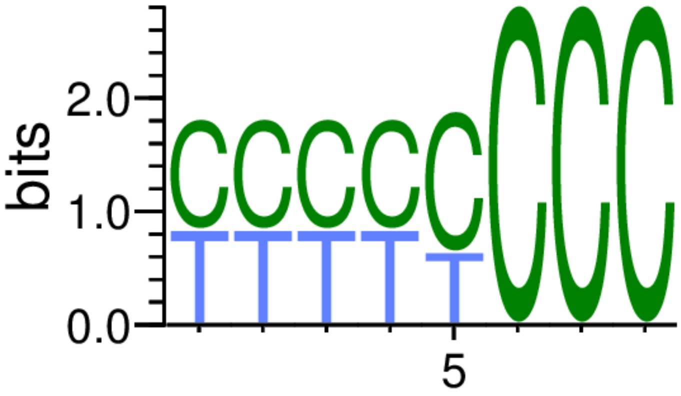




Cluster 1 (the most populated cluster) forms only turns and random coils.

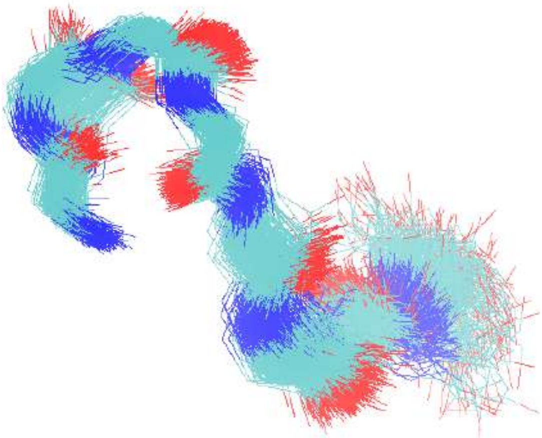
2nd CLUSTER

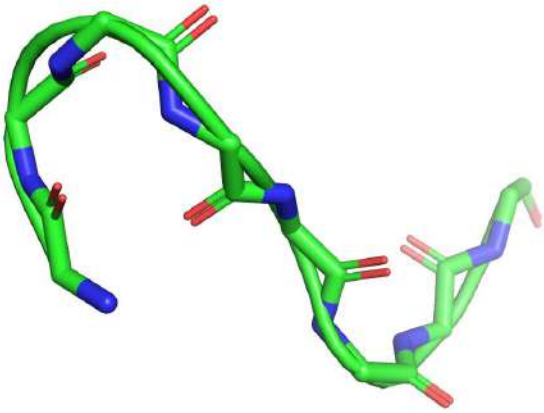


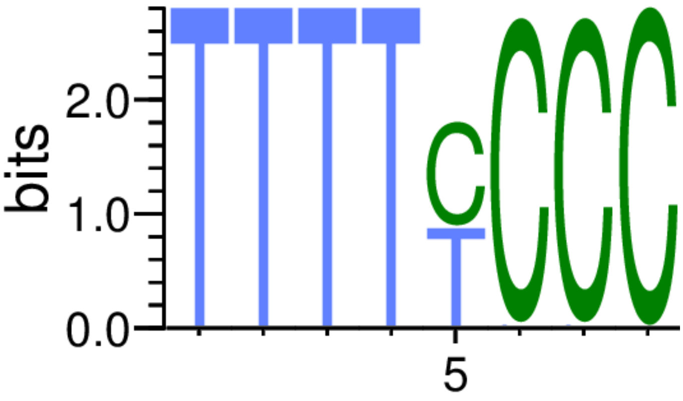


Cluster 2 (the second most populated cluster) includes only turns and coils.

3rd CLUSTER

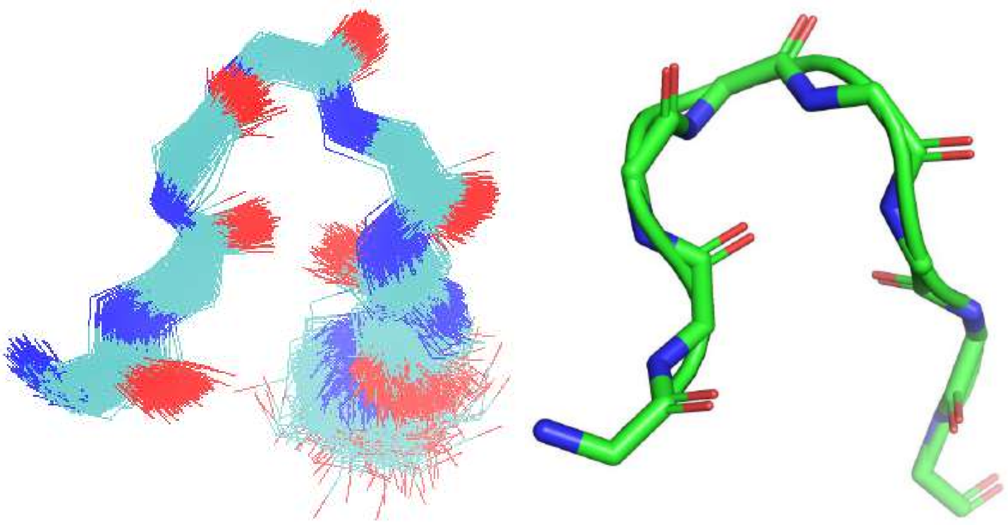


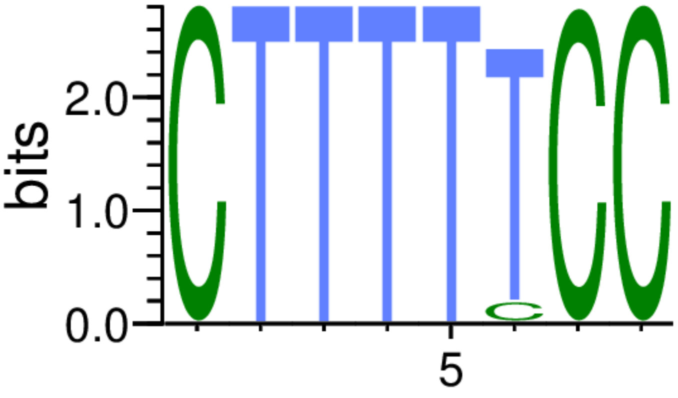




Only coils and turns are included in cluster 3 (the third most populated cluster).

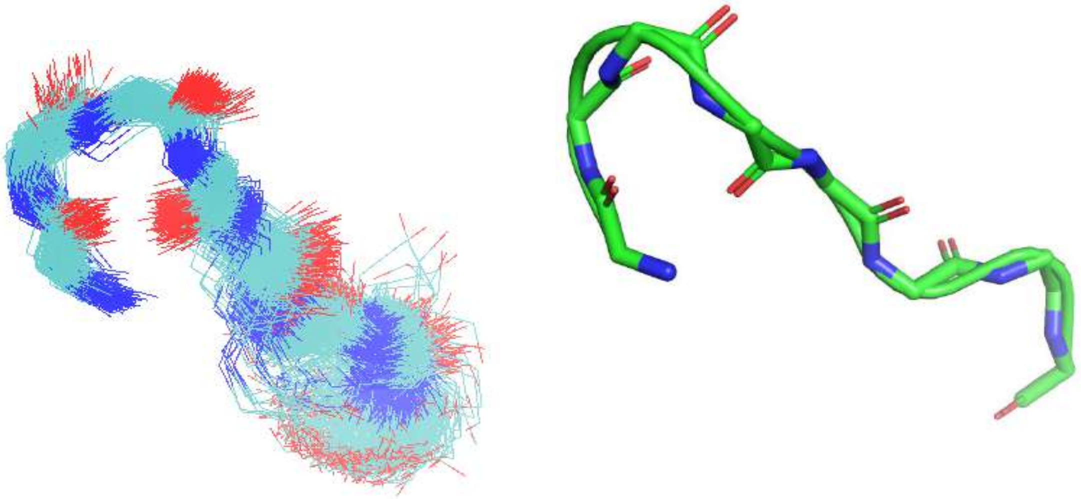
4th CLUSTER

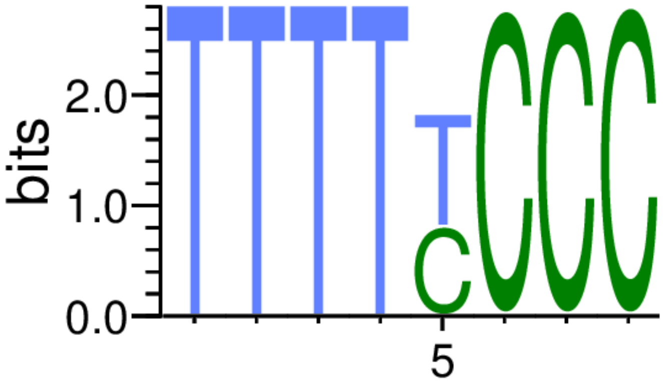




Cluster 4 contains only turns and random coils.

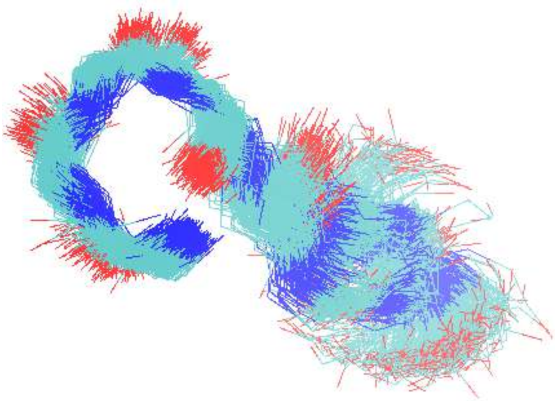
5th CLUSTER

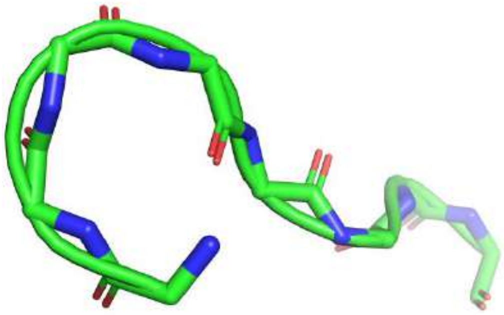


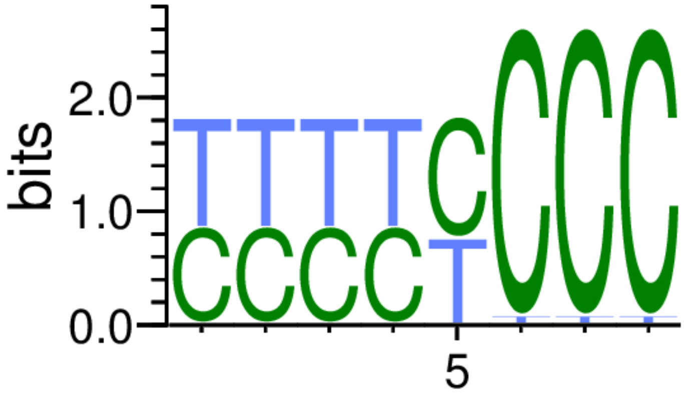


Cluster 5 forms only turns and coils.

6th CLUSTER

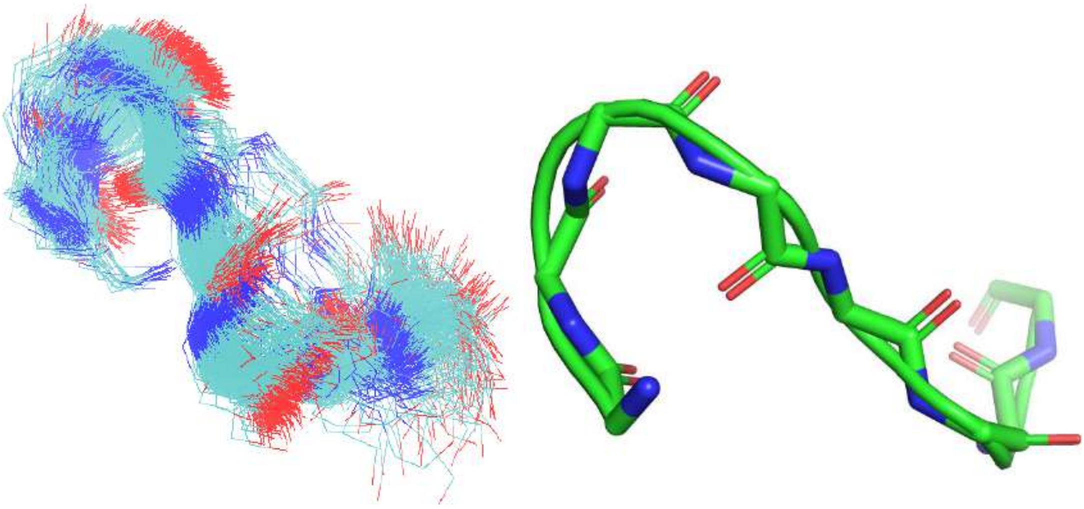


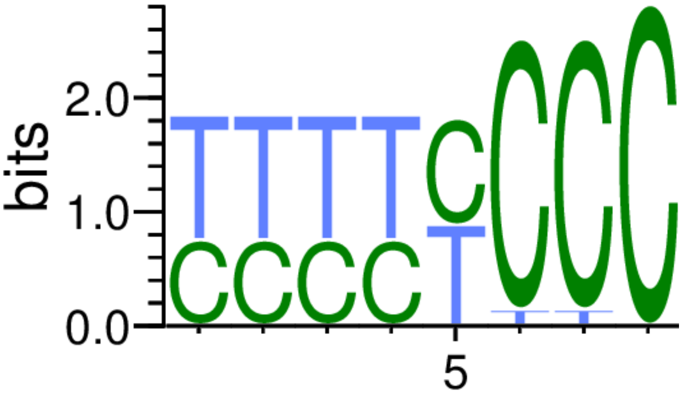




Cluster 6 forms only turns and random coils.

7th CLUSTER



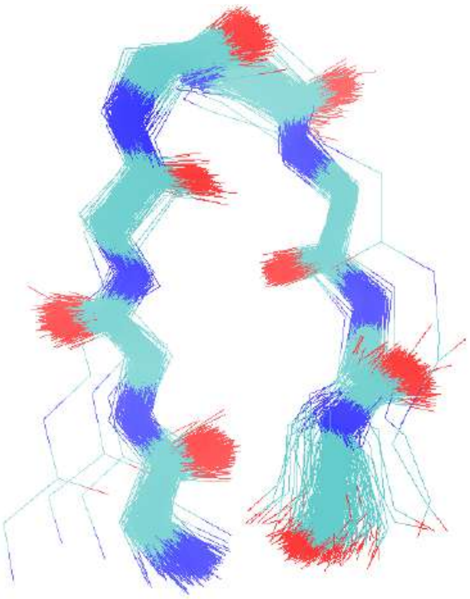


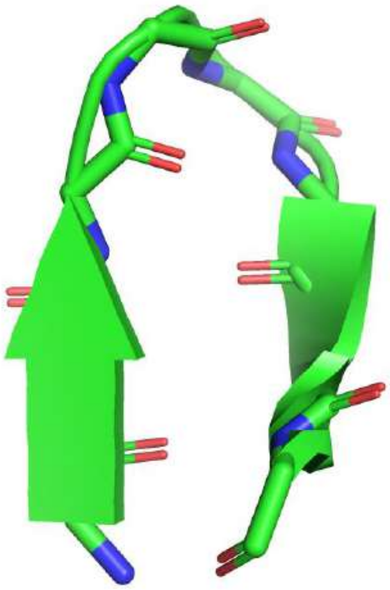
Only turns and coils are present in cluster 7.

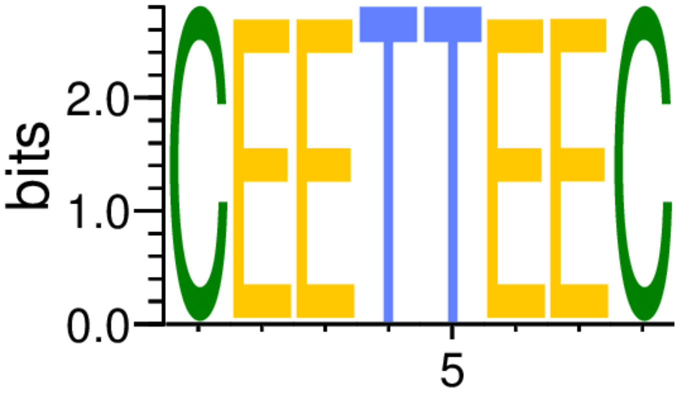
Figure 4.8: Superposition of structures (left) that belong to each cluster derived from the dPCA analysis for TFE. The superpositions are little noisy but, N-terminus has more compact structures than C-terminus. Representative of each cluster, it is produced by Pymol program (right). WebLogo graph (3rd image) of each cluster, as derived from STRIDE analysis.

Case III- Clusters when solvent is WATER

1st CLUSTER

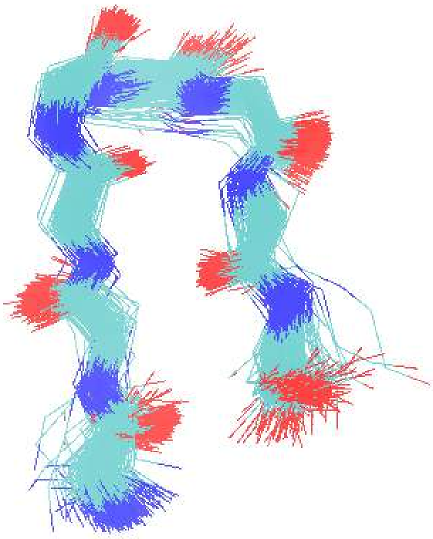


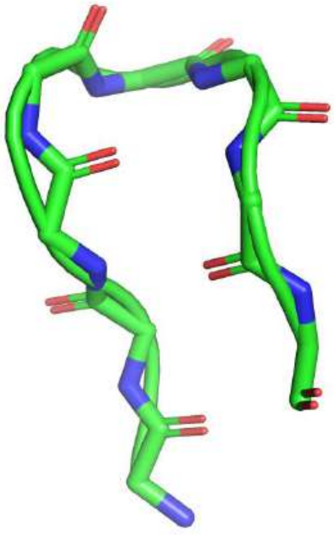


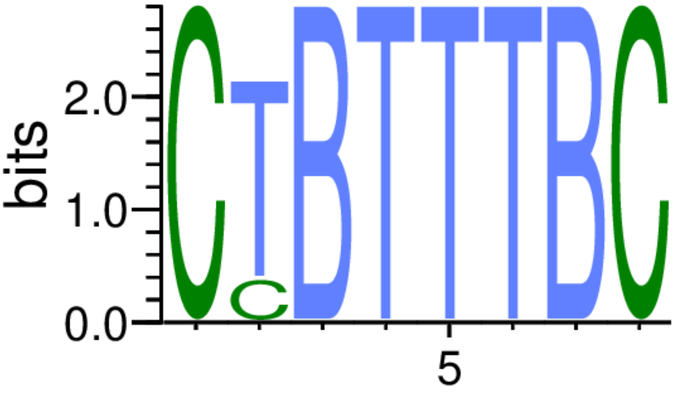


Residues 2-3 and 6-7 in cluster 1 (the most populated cluster) participate in β-sheets conformations. The rest of the peptide contains turns in residues 4-5, (existence of β-hairpin) while including coils at the ends.

2nd CLUSTER

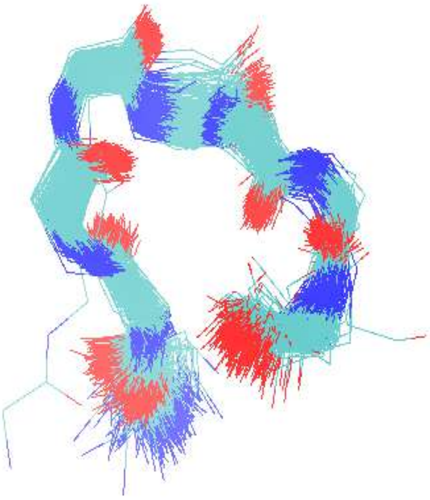


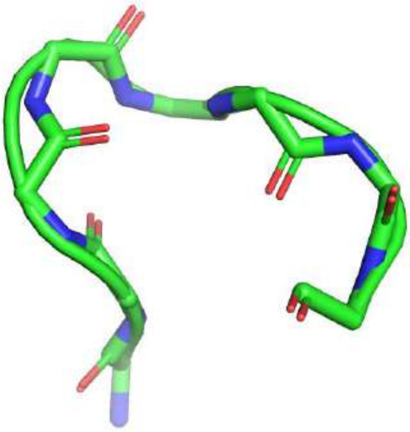


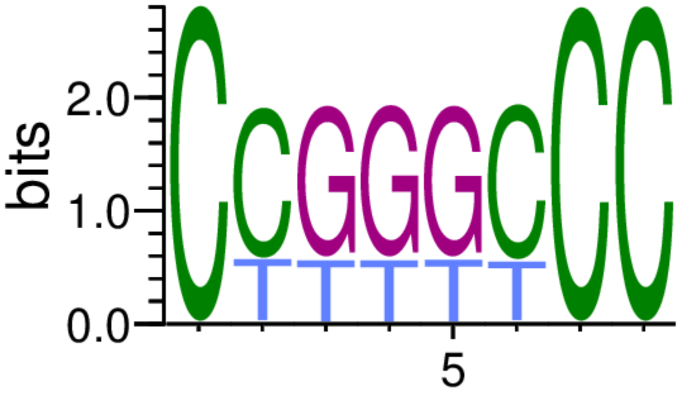


Residues 3 and 7 in cluster 2, build β-bridges. The rest of the peptide includes turns and coils at the ends.

3rd CLUSTER

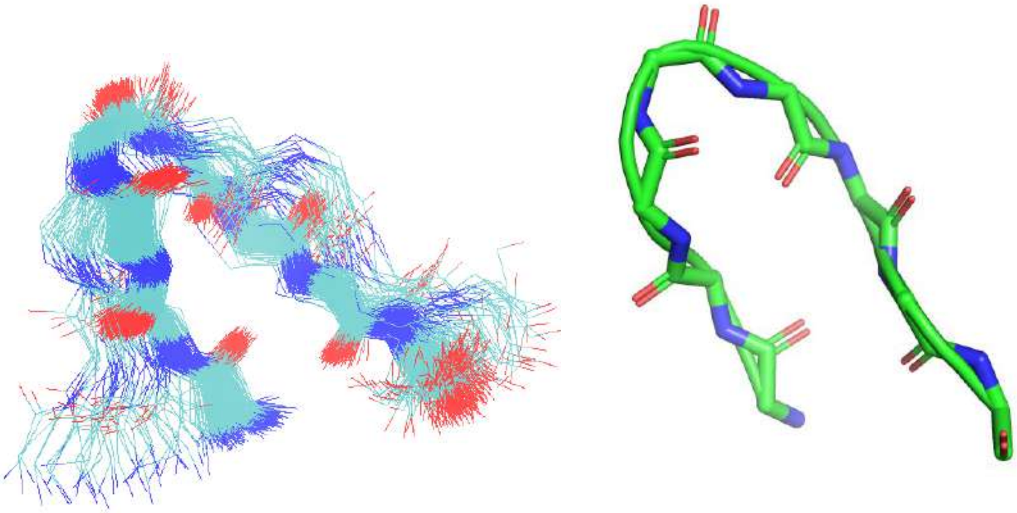


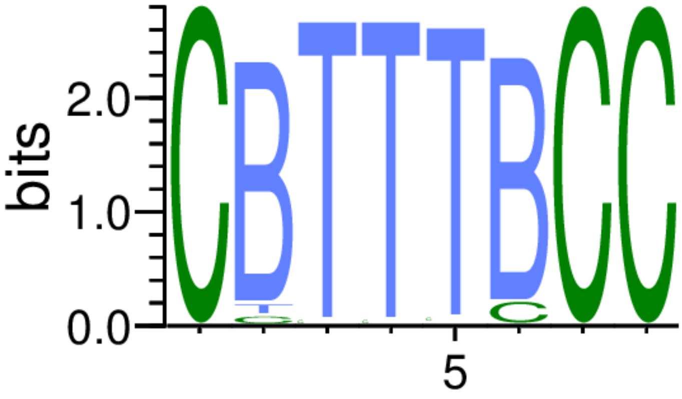




Cluster 3 presents in residues 3-5, a 3-10 helix conformation, while the rest of the peptide forms turns and coils.

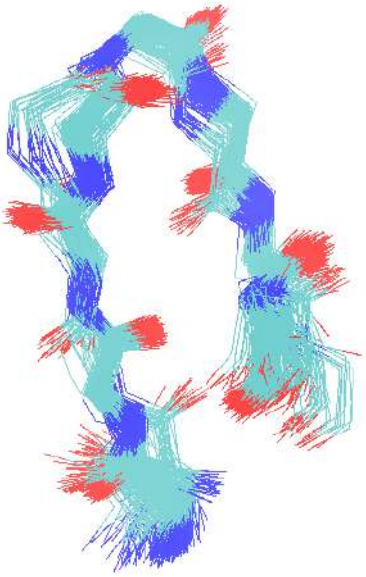
4th CLUSTER

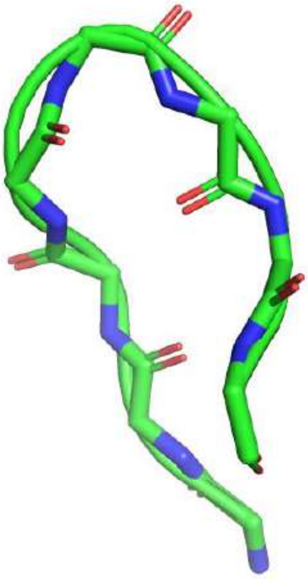


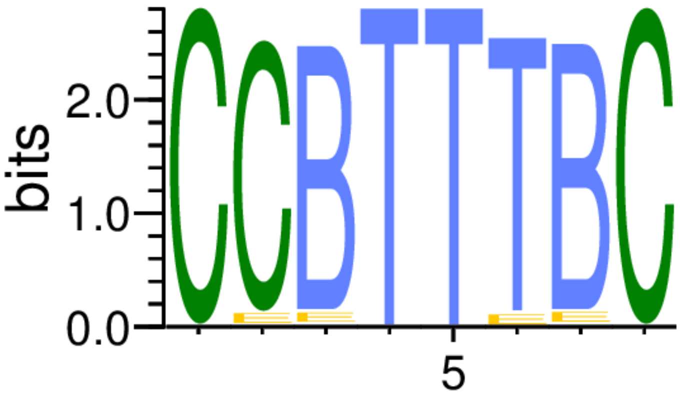


The main structural characteristics in cluster 4 are coils and turns, but it is observed in residues 2 and 6, β-bridges.

5th CLUSTER

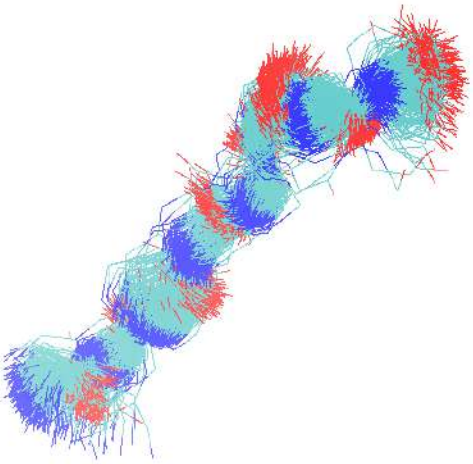


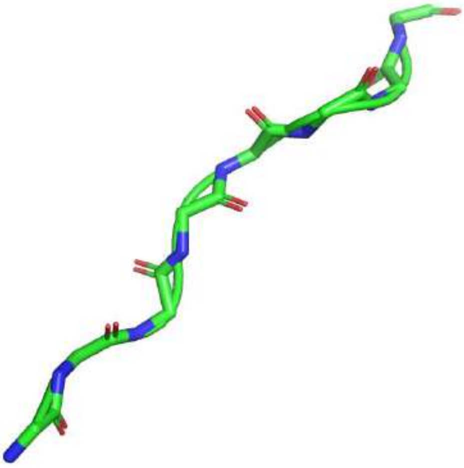


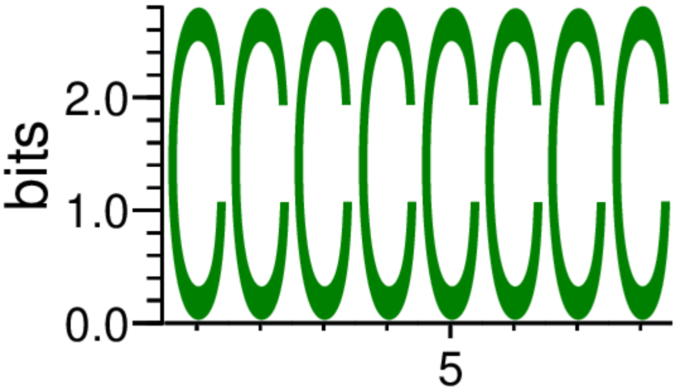


The main structural characteristics in cluster 5 are coils and turns, but it is observed in residues 3 and 7, β-bridges and in very low frequency residues 2-3, 6-7 have β-sheets.

6th CLUSTER







Cluster 6 consists exclusively only random coils.

Figure 4.9: Superposition of structures (left) that belong to each cluster derived from the dPCA analysis for water. The superpositions are not much noisy except of cluster No 4. Representative of each cluster, it is produced by Pymol program (right). WebLogo graph (3rd image) of each cluster, as derived from STRIDE analysis.

Chapter 5: Comparison with experimental data

5.1 DMSO vs. Micelles

Comparing the results obtained from our study with the experimental results conducted by Zarena D. et al., we observe certain similarities and differences. The study of the Tetra-F2W-RK peptide using micelles provides significant information about its structure and interaction with the membrane. The peptide Tetra-F2W-RK is examined within a microscopic environment that includes micelles, modeling the membrane. Based on the results of Zarena D. et al., the peptide appears to adopt a stable structure at specific time points. Also this study shows a strong preference of the peptide for turns and helices.

On the other hand, when DMSO is used as the solvent for studying the peptide, the maintenance of peptide's structure is also observed, but the stable structure is short-lived. The structural regions of the peptide remain stable at specific time intervals. However these stable regions are relatively short-lived. The STRIDE analysis in this case also demonstrates the peptide's preference for turns and 3-10 helices. It is evident that DMSO is not ideal for maintaining the peptide's structure for an extended period. Micelles are preferable for studying the structure and interaction of the Tetra-F2W-RK peptide because they provide a natural environment that closely resembles biological membranes. Additionally, they allow the peptide to adopt stable structures that can be more effectively examined. Micelles consist of lipid molecules that interact with the peptide similarly to a natural membrane. This natural interaction enables the study of the peptide's interaction with the lipid environment, which is crucial for its activity.

Overall, micelles offer a more precise and natural environment for studying the Tetra-F2W-RK peptide compared to using DMSO as a solvent. However, it's worth noting that micelles are more expensive compared to the simple solvent DMSO. Simulations with micelles require more computational resources and often involve more complexity in preparing the structure. Also, the limited availability of experimental data for the structure and behavior of DMSO makes it more accessible in terms of cost. Nevertheless, while micelles are costlier, they often provide more accurate and reliable information. Therefore, the choice between the two depends on the specific research application, budget availability and study objectives.

5.2 TFE vs. Micelles

Comparing the results obtained by Zarena D. et al. (using micelles) with those from the TFE simulations, several differences emerge concerning the stability and structural preferences of the peptide. When TFE is used as a solvent, the peptide appears to exhibit significant stability over longer time intervals, specifically at 9-10 μs and 2.5-.8.3 μs among others. The STRIDE analysis indicates that the peptide predominantly adopts turns and this observation is further confirmed by the WebLogo graph. Overall, the peptide exhibits distinct structural behavior when placed in micelles compared to TFE, with TFE promoting greater stability and a preference for turns over extended time periods.

The choice between these two environments for studying the Tetra-F2W-RK peptide depends on the particular aims of the research and the data required. TFE offers an environment where the peptide appears to maintain stable structures, making it suitable for investigating the peptide’s structural properties under conditions that simulate moments of stability. On the other hand, micelles seem to provide a more natural environment closer to a biological membrane. In this environment, the peptide appears to retain stable structures for long time intervals, which may be preferable if you are interested in studying the dynamic behavior of the peptide and its interactions within a membrane-like environment. Ultimately, the preferred environment depends on the research objectives, available resources and budget constraints.

5.3 Water vs. Micelles

Comparing the results from the environments of micelles and water for the study of the Tetra-F2W-RK peptide, certain differences in stability and structure of the peptide are observed. In water, the peptide appears to adopt a stable structure for extended periods, approximately from 2 to 3.95 μs, with additional stable structures at other time intervals. The STRIDE analysis indicates that the peptide exhibits various conformations, with turns and β-sheets being predominant, while 3-10 helices and α-helices are also present but less frequently.

On the other hand, micelles as I have already mentioned, provide an environment that faithfully mimics the surrounding membrane and offers more stable structures for the Tetra-F2W-RK peptide compared to water. For the study of Tetra-F2W-RK, water is more cost-effective than micelles. Simulations in water require fewer computational resources and are less complex in terms of structure preparation compared to micelles. However, despite being more budget-friendly and providing greater stability for the peptide, water simulations lack the naturalness of the micelles' environment, which closely resembles biological conditions. The choice between the two depends on the research objectives, budget considerations and available resources for the study.

Conclusion and Discussion:

The aim of this project is to study the stability, structure and behavior of the Tetra-F2W-RK peptide in Molecular Dynamics Simulations, compared to experimental approaches, which were done in the presence of micelles. Peptide preferences were investigated under three different conditions. The solvents in each condition were DMSO, TFE and water separately. The GRCARMA program was used to produce RMSD matrices and distinct patterns were observed in the environments under consideration. In the case where DMSO was the solvent, distinct stability points of the peptide structure were observes in the time intervals: 0.3 μs, 1μs, 2.5 μs, 11.2-11.6 μs, 13.5 μs, 13.8 μs, etc. However these stable areas were relatively short. In the case where the solvent was TFE, the peptide showed more stable and longer-lasting structure. The RMSD matrix yielded stable structure in the time intervals 9-10 μs, and 2.5-8.3 μs. In addition, when water was used as the solvent, a distinct stable structure was observed for a long time, specifically in the interval 2-3.95 μs, while shorter-lived stable structures also appeared at the time points of 1 μs, 1.3 μs and 1.8 μs.

Also particularly important was the contribution of the STRIDE and WebLogo analyses. In the results of the analyses, a clear preference is observed in all three solvent conditions for turns and coils. Helical structures are mainly present in DMSO. The presence of β-sheets was also observed in water, while β-bridges and α-helices are less prominent. Furthermore, the different conformational motifs represented by the clusters indicate the existence of energy minima in each solvent.

Overall, the present study demonstrates the effect of solvent on peptide stability and conformation. The results suggest that DMSO is not a favorable solvent for the long-term stability of the peptide. On the contrary, in TFE and water solvents the peptide shows continuous stability. However the preference of peptide in these solvents is not helical structures.

Comparing these results the experimental approach, [104] Tetra-F2W-RK, shows similar stable structures to those of the experiments that Zarena D. et al. had conducted. More specifically, the Tetra-F2W-RK peptide does not show helical structures in all four conditions (DMSO, TFE, water, micelles). In DMSO, the helical structures are more distinct and the stable regions are relatively short-lived. In TFE cluster, helical structures are absent. In water the main structure is a-ribbon (strand-turn-strand), in complete disagreement with the experimental structure in micelles. Therefore the solvent in which the peptide most closely resembles the helical structure it displayed in the micelle solvent experiment is shown to be DMSO.

It is particularly important to also mention that the polarity, hydrophobicity and hydrogen bonding potential of the solvent have a crucial role in the conformational landscape of the peptide. TFE as the more polar solvent and water as a highly polar and hydrogen bonding-capable solvent might interact differently with the peptide in contrast to the polar aprotic solvent, DMSO.

In comparison the Tetra-F2W-RK peptide in micelles is more favorable. The micelles have an environment very similar to the physiological conditions of the peptide membranes, with the result that the structures by extension of the peptide are similar to the normal ones, e.g. two-turn helix. DMSO, TFE and water are simple solvents but their use cannot completely replace micelles. Nevertheless their choice depends on the purpose of studying the Tetra-F2W-RK.

In conclusion, the choice of solvent is particularly important for understanding the structural behavior of peptides. Although the present study provides important insights into the effect of solvent, an interesting potential extension of the research could be to study other factors, such as the effect of temperature on peptide stability and structure. In addition the present findings can be helpful in developing peptide-based materials and drugs whose functionalities depend on stable secondary structures.

References:

1. Carl-Ivar Brändén and Tooze, J. (2009). Introduction to protein structure. New York, Ny: Garland Pub, pp.1–3.

2. Stryer, L., Berg, J., Tymoczko, J. and Gatto, G. (2019). Biochemistry. 8th ed. New York Macmillan Learning Wh Freeman, pp.28–30.

3. Twyman, R.M. (2014). Principles of proteomics. New York: Garland Science, p.11

4. Alston, T.A. (2008). Lipmann and Anfinsen: Nobel biochemists of Beecher's anesthesia laboratory. Journal of Clinical Anesthesia, 20(1), pp.61–63.

5. Anfinsen, C.B. (1972). The formation and stabilization of protein structure. Biochemical Journal, 128(4), pp.737–749.

6. Kresge, N., Simoni, R.D. and Hill, R.L. (2006). The Thermodynamic Hypothesis of Protein Folding: the Work of Christian Anfinsen. Journal of Biological Chemistry, 281(14), pp.e11–e13.

7. Ivankov, D.N. and Finkelstein, A.V. (2020). Solution of Levinthal's Paradox and a Physical Theory of Protein Folding Times. Biomolecules, 10(2), p.250.

8. Melkikh, A.V. and Meijer, D.K.F. (2018). On a generalized Levinthal's paradox: The role of long- and short range interactions in complex bio-molecular reactions, including protein and DNA folding. Progress in Biophysics and Molecular Biology, 132, pp.57–79.

9. Zwanzig, R., Szabo, A. and Bagchi, B. (1992). Levinthal's paradox. Proceedings of the National Academy of Sciences, 89(1), pp.20–22.

10. Islam, S.A., Karplus, M. and Weaver, D.L. (2002). Application of the diffusion-collision model to the folding of three-helix bundle proteins. Journal of Molecular Biology, 318(1), pp.199–215.

11. Ahluwalia, U., Katyal, N. and Deep, S. (2012). RESEARCH EDUCATION MODELS OF PROTEIN FOLDING. J P P JOURNAL OF PROTEINS AND PROTEOMICS, [online] 3(2), pp.85–93. Available at: https://web.archive.org/web/20180411000346id_/http://www.jpp.org.in/index.php/jpp/article/viewFile/17/60 [Accessed 15 Aug. 2023].

12. Bengt Nölting and Agard, D.A. (2008). How general is the nucleation-condensation mechanism? 73(3), pp.754–764.

13. Fersht, A.R. (1997). Nucleation mechanisms in protein folding. Current Opinion in Structural Biology, 7(1), pp.3–9.

14. Banerjee, R., Sen, M., Bhattacharya, D. and Saha, P. (2003). The Jigsaw Puzzle Model: Search for Conformational Specificity in Protein Interiors. Journal of Molecular Biology, 333(1), pp.211–226.

15. Harrison, S. and Durbin, R. (1985). Is there a single pathway for the folding of a polypeptide chain? (native-like structure/microdomains/jigsaw-puzzle analogy/protein conformation/protein renaturation). Proc. Nail. Acad. Sci. USA, 82, pp.4028–4030.

16. Gassner, N.C., Baase, W.A. and Matthews, B.W. (1996). A test of the 'jigsaw puzzle' model for protein folding by multiple methionine substitutions within the core of T4 lysozyme.. 93(22), pp.12155–12158.

17. Onuchic, J.N., Socci, N.D., Luthey-Schulten, Z. and Wolynes, P.G. (1996). Protein folding funnels: the nature of the transition state ensemble. Folding and Design, 1(6), pp.441–450.

18. Dill, K.A. and Chan, H.S. (1997). From Levinthal to pathways to funnels. Nature Structural & Molecular Biology, 4(1), pp.10–19.

19. Chan, H.S. and Dill, K.A. (1998). Protein folding in the landscape perspective: Chevron plots and non-arrhenius kinetics. Proteins: Structure, Function, and Genetics, 30(1), pp.2–33.

20. Karplus, M. (1997). The Levinthal paradox: yesterday and today. Folding and Design, 2, pp.S69–S75.

21. Smyth, M.S. and Martin, J.H.J. (2000). X Ray Crystallography. Molecular Pathology, 53(1), pp.8–14.

22. Maveyraud, L. and Mourey, L. (2020). Protein X-ray Crystallography and Drug Discovery. Molecules, 25(5), p.1030.

23. Srivastava, A., Nagai, T., Srivastava, A., Miyashita, O. and Tama, F. (2018). Role of Computational Methods in Going beyond X-ray Crystallography to Explore Protein Structure and Dynamics. International Journal of Molecular Sciences, 19(11).

24. BERNAL, J.D. and CROWFOOT, D. (1934). X-Ray Photographs of Crystalline Pepsin. Nature, 133(3369), pp.794–795.

25. Aaron R, D. (2000). Understanding protein folding via free-energy surfaces from theory and experiment. Trends in Biochemical Sciences, 25(7), pp.331–339.

26. Wüthrich, K. (1986). NMR with Proteins and Nucleic Acids. Europhysics News, 17(1), pp.11–13.

27. Chen, Y., Ding, F., Nie, H., Serohijos, A.W., Sharma, S., Wilcox, K.C., Yin, S. and Dokholyan, N.V. (2008). Protein folding: Then and now. Archives of Biochemistry and Biophysics, 469(1), pp.4–19.

28. Munoz, V. and Cerminara, M. (2016). When fast is better: protein folding fundamentals and mechanisms from ultrafast approaches. Biochemical Journal, 473(17), pp.2545–2559.

29. Li, B., Fooksa, M., Heinze, S. and Meiler, J. (2017). Finding the needle in the haystack: towards solving the protein-folding problem computationally. Critical Reviews in Biochemistry and Molecular Biology, 53(1), pp.1–28.

30. Zhuravleva, A. and Korzhnev, D.M. (2017). Protein folding by NMR. Progress in Nuclear Magnetic Resonance Spectroscopy, 100, pp.52–77.

31. Smith, A.E., Zhang, Z., Pielak, G.J. and Li, C. (2015). NMR studies of protein folding and binding in cells and cell-like environments. Current Opinion in Structural Biology, 30, pp.7–16.

32. Hu, K.-N. and Tycko, R. (2010). What can solid state NMR contribute to our understanding of protein folding? Biophysical Chemistry, 151(1-2), pp.10–21.

33. Zeeb, M. and Balbach, J. (2004). Protein folding studied by real-time NMR spectroscopy. Methods (San Diego, Calif.), 34(1), pp.65–74.

34.Greenfield, N.J. (2006a). Analysis of the kinetics of folding of proteins and peptides using circular dichroism. Nature Protocols, 1(6), pp.2891–2899.

35.Greenfield, N.J. (2006b). Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. Nature Protocols, 1(6), pp.2527–2535.

36.Greenfield, N.J. (2006c). Using circular dichroism spectra to estimate protein secondary structure. Nature Protocols, 1(6), pp.2876–2890.

37. Hussain, R. and Siligardi, G. (2016). Characterisation of Conformational and Ligand Binding Properties of Membrane Proteins Using Synchrotron Radiation Circular Dichroism (SRCD). Advances in Experimental Medicine and Biology, pp.43–59.

38. Gekko, K. (2019). Synchrotron-radiation vacuum-ultraviolet circular dichroism spectroscopy in structural biology: an overview. Biophysics and Physicobiology, 16(0), pp.41–58.

39. Floudas, C.A., Fung, H.K., McAllister, S.R., Mönnigmann, M. and Rajgaria, R. (2006). Advances in protein structure prediction and de novo protein design: A review. Chemical Engineering Science, 61(3), pp.966–988.

40. Kaushik, R., Singh, A. and Jayaram, B. (2019). Ab initio Protein Structure Prediction. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/B978012809633820321X [Accessed 15 Aug. 2023].

41. Yousef, M., Abdelkader, T. and El-Bahnasy, K. (2019). Performance comparison of ab initio protein structure prediction methods. Ain Shams Engineering Journal, 10(4), pp.713–719.

42. Li, Y., Zhang, C., Yu, D.-J. and Zhang, Y. (2022). Deep learning geometrical potential for high-accuracy ab initio protein structure prediction. iScience, 25(6), p.104425.

43. Osguthorpe, D. (2000). Ab initio protein folding. Current Opinion in Structural Biology, 10(2), pp.146–152.

44. Bonneau, R. and Baker, D. (2001). Ab Initio Protein Structure Prediction: Progress and Prospects. Annual Review of Biophysics and Biomolecular Structure, 30(1), pp.173–189.

45. Han, K., Liu, Y., Xu, J., Song, J. and Yu, D.-J. (2022). Performing protein fold recognition by exploiting a stack convolutional neural network with the attention mechanism. Analytical Biochemistry, 651, p.114695.

46. Jones, D.T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. Journal of Molecular Biology, 287(4), pp.797–815.

47. Webb, B. and Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. Current Protocols in Bioinformatics, 54(1).

48. Sippl, M.J. (1999). Who solved the protein folding problem? Structure, 7(4), pp.R81–R83.

49. Allen, M. (2004). Introduction to Molecular Dynamics Simulation Introduction to Molecular Dynamics Simulation. Lecture Notes, 23, pp.1–28.

50. Hollingsworth, S.A. and Dror, R.O. (2018). Molecular Dynamics Simulation for All. Neuron, 99(6), pp.1129–1143.

51. Pechlaner, M., Wilfred, Hansen, N. and Smith, L.J. (2022). Molecular dynamics simulation or structure refinement of proteins: are solvent molecules required? A case study using hen lysozyme. European Biophysics Journal, [online] 51(3), pp.265–282.

52. Wilfred and Mark, A.E. (1998). Validation of molecular dynamics simulation. 108(15), pp.6109–6116.

53. Alder, B.J. and Wainwright, T.E. (1957). Phase Transition for a Hard Sphere System. The Journal of Chemical Physics, 27(5), pp.1208–1209.

54. Hoover, W.G., Anthony and Hoover, V.N. (1983). Historical Development and Recent Applications of Molecular Dynamics Simulation. Advances in chemistry series, pp.29–46.

55. Rapaport, D.C. (1999). Molecular dynamics simulation. Computing in Science & Engineering, 1(1), pp.70–71.

56. Zarringhalam, M., Ahmadi-Danesh-Ashtiani, H., Toghraie, D. and Fazaeli, R. (2019b). The effects of suspending Copper nanoparticles into Argon base fluid inside a microchannel under boiling flow condition by using of molecular dynamic simulation. Journal of Molecular Liquids, 293, p.111474.

57. Peng, Y., Zarringhalam, M., Barzinjy, A.A., Toghraie, D. and Afrand, M. (2020). Effects of surface roughness with the spherical shape on the fluid flow of argon atoms flowing into the microchannel, under boiling condition using molecular dynamic simulation. Journal of Molecular Liquids, 297, p.111650.

58. Zarringhalam, M., Ahmadi-Danesh-Ashtiani, H., Toghraie, D. and Fazaeli, R. (2019a). Molecular dynamic simulation to study the effects of roughness elements with cone geometry on the boiling flow inside a microchannel. International Journal of Heat and Mass Transfer, 141, pp.1–8.

59. Maroo, S.C. and Chung, J.N. (2008). Molecular dynamic simulation of platinum heater and associated nano-scale liquid argon film evaporation and colloidal adsorption characteristics. Journal of Colloid and Interface Science, 328(1), pp.134–146.

60. Toghraie, D., Mokhtari, M. and Afrand, M. (2016). Molecular dynamic simulation of Copper and Platinum nanoparticles Poiseuille flow in a nanochannels. Physica E: Low-dimensional Systems and Nanostructures, 84, pp.152–161.

61. Durrant, J.D. and McCammon, J.A. (2011). Molecular dynamics simulations and drug discovery. BMC Biology, 9(1).

62.Hünenberger, P.H. (2005). Thermostat Algorithms for Molecular Dynamics Simulations. pp.105–149.

63. Gelpi, J., Hospital, A., Goñi, R. and Orozco, M. (2015). Molecular dynamics simulations: advances and applications. Advances and Applications in Bioinformatics and Chemistry, 8, pp.37–47.

64. Freddolino, P.L., Harrison, C.B., Liu, Y. and Schulten, K. (2010). Challenges in protein folding simulations: Timescale, representation, and analysis. Nature physics, 6(10), pp.751–758.

65. Ding, Y., Yu, K. and Huang, J. (2023). Data science techniques in biomolecular force field development. 78, pp.102502–102502

66. Berendsen, H.J.C., van der Spoel, D. and van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. Computer Physics Communications, 91(1-3), pp.43–56.

67. Ponder, J.W. and Case, D.A. (2003) Force fields for protein simulations. Advances in Protein Chemistry, 66, 27- 85.

68. Larsson, P., Hess, B. and Lindahl, E. (2011). Algorithm improvements for molecular dynamics simulations. Wiley Interdisciplinary Reviews: Computational Molecular Science, 1(1), pp.93–108.

69. The Amber Molecular Dynamics Package. (2020) Available at: http://ambermd.org (Accessed 8 November 2020).

70. Onufriev, A. (2008b). Implicit Solvent Models in Molecular Dynamics Simulations: A Brief Overview. p. 125-134.

71. Anandakrishnan, R., Drozdetski, A., Walker, Ross C. and Onufriev, Alexey V. (2015). Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations. Biophysical Journal, 108(5), pp.1153–1164.

72. Rita, A. and Cannistraro, S. (2002). Molecular Dynamics of Water at the Protein−Solvent Interface. Journal of Physical Chemistry B, 106(26), pp.6617–6633.

73. Samsonov, S., Teyra, J. and Pisabarro, M.T. (2008). A molecular dynamics approach to study the importance of solvent in protein interactions. Proteins: Structure, Function, and Bioinformatics, 73(2), pp.515–525.

74. Bürgi, R., Daura, X., Mark, A., van Gunsteren, W., Bellanda, M., Mammi, S. and Peggion, E. (2001). Folding study of an Aib-rich peptide in DMSO by molecular dynamics simulations. Journal of Peptide Research, 57(2), p.107.

75. Smondyrev, A.M. and Berkowitz, M.L. (1999). Molecular Dynamics Simulation of DPPC Bilayer in DMSO. Biophysical Journal, 76(5), pp.2472–2478.

76. Duarte, A.M.S., van Mierlo, C.P.M. and Hemminga, M.A. (2008). Molecular Dynamics Study of the Solvation of an α-Helical Transmembrane Peptide by DMSO. The Journal of Physical Chemistry B, 112(29), pp.8664–8671.

77. Civera, C., Arias, C., Elorza, M.A., Elorza, B., García-Blanco, F. and Galera-Gómez, P.A. (2014). Hydrophobicity enhancement in micelles of Triton X-165 by the presence of the cosolvent 2,2,2 trifluoroethanol (TFE). Journal of Molecular Liquids, 199, pp.29–34.

78. Fraga, A.S., Esteves, A.C., Micaelo, N., Cruz, P.F., Brito, R.M.M., Nutley, M., Cooper, A., Barros, M.M.T. and Pires, E.M.V. (2012). Functional and conformational changes in the aspartic protease cardosin A induced by TFE. International Journal of Biological Macromolecules, 50(2), pp.323–330.

79. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L. and Schulten, K. (2005). Scalable molecular dynamics with NAMD. Journal of Computational Chemistry, 26(16), pp.1781–1802.

80. Glykos, N. (2020) The Norma computing cluster. Available at: https://norma.mbg.duth.gr/index.php?id=about:intro (Accessed: 16 October 2020).

81. Bank, R.P.D. (n.d.). (2010) RCSB PDB: About RCSB PDB: Enabling Breakthroughs in Scientific and Biomedical Research and Education. [online] Available at: https://www.rcsb.org/.

82. Phillips, J. and Hardy, D. (n.d.). NAMD Configuration Files. [online] Available at: https://www.ks.uiuc.edu/Training/Tutorials/namd/namd-tutorial-unix-html/node26.html#ap-configfiles [Accessed 15 Aug. 2023].

83. Phillips, J. and Hardy, D. (n.d.). What is needed (NAMD 2.14 User's Guide). [online] Available at: https://www.ks.uiuc.edu/Research/namd/2.14/ug/node8.html [Accessed 15 Aug. 2023].

84. Schlick, T. (2002). Molecular Modeling and Simulation. p.399. doi:https://doi.org/10.1007/978-0-387-22464-0.

85. Ho, B. and Dill, K.A. (2006). Folding Very Short Peptides Using Molecular Dynamics. PLOS Computational Biology, 2(4), pp.e27–e27.

86. Adamidou, T., Arvaniti, K.-O. and Glykos, N.M. (2018). Folding Simulations of a Nuclear Receptor Box-Containing Peptide Demonstrate the Structural Persistence of the LxxLL Motif Even in the Absence of Its Cognate Receptor. The Journal of Physical Chemistry B, 122(1), pp.106–116.

87. Frenkel, D. and Smit, B. (2001). Understanding Molecular Simulation: From Algorithms to Applications. San Diego: Academic Press

88. Pechlaner, M., Wilfred, Hansen, N. and Smith, L.J. (2022). Molecular dynamics simulation or structure refinement of proteins: are solvent molecules required? A case study using hen lysozyme. European Biophysics Journal, [online] 51(3), pp.265–282.

89. Price, D.J. and Brooks, C.L. (2004). A modified TIP3P water potential for simulation with Ewald summation. The Journal of Chemical Physics, 121(20), pp.10096–10103.

90. . Glykos, N. (2006). CARMA: a molecular dynamics analysis program', Journal of Computational Chemistry, 27(14), pp. 1765−1768.

91. Koukos, P. & Glykos, N. (2013) 'Grcarma: A fully automated task-oriented interface for the analysis of molecular dynamics trajectories', Journal of Computational Chemistry, 34(26), pp.2310-2312.

92. Phillips, J. and Hardy, D. (n.d.). (n.d.). PSF Files. Available at: https://www.ks.uiuc.edu/Training/Tutorials/namd/namd-tutorial-unix-html/node23.html [Accessed 15 Aug. 2023].

93. Aier, I., Varadwaj, P.K. and Raj, U. (2016). Structural insights into conformational stability of both wild-type and mutant EZH2 receptor. Scientific Reports, 6(1).

94. Heinig, M. and Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Research, 32(Web Server), pp.W500–W502.

95. Mu, Y., Nguyen, P.H. and Stock, G. (2004). Energy landscape of a small peptide revealed by dihedral angle principal component analysis. Proteins: Structure, Function, and Bioinformatics, 58(1), pp.45–52.

96. Sittel, F., Jain, A. and Stock, G. (2014). Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. The Journal of Chemical Physics, 141(1), p.014111.

97. Riccardi, L., Nguyen, P.H. and Stock, G. (2009). Free-Energy Landscape of RNA Hairpins Constructed via Dihedral Angle Principal Component Analysis. Journal of Physical Chemistry B, 113(52), pp.16660–16668.

98. Altis, A., Nguyen, P.H., Hegger, R. and Stock, G. (2007). Dihedral angle principal component analysis of molecular dynamics simulations. The Journal of Chemical Physics, 126(24), p.244111.

99. David, C.C. and Jacobs, D.J. (2013). Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. Protein Dynamics, pp.193–226.

100. Bratchell, N. (1989). Cluster analysis. Chemometrics and Intelligent Laboratory Systems, 6(2), pp.105–125.

101. Shao, J., Tanner, S.W., Thompson, N. and Cheatham, T.E. (2007). Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. Journal of Chemical Theory and Computation, 3(6), pp.2312–2334.

102. Humphrey, W., Dalke, A. and Schulten, K. (1996). VMD: Visual molecular dynamics. Journal of Molecular Graphics, 14(1), pp.33–38.

103. Pymol.org. (2019). PyMOL Available at: https://pymol.org/2/.

104. Zarena, D., Mishra, B., Lushnikova, T., Wang, F. and Wang, G. (2017). The π Configuration of the WWW Motif of a Short Trp-Rich Peptide Is Critical for Targeting Bacterial Membranes, Disrupting Preformed Biofilms, and Killing Methicillin-Resistant Staphylococcus aureus. Biochemistry, 56(31), pp.4039–4043.

105. Gkogka, I. & Glykos*, N.M. (2022), "Folding molecular dynamics simulation of T-peptide, a HIV viral entry inhibitor : Structure, dynamics, and comparison with the experimental data", J. Comput. Chem., 43, 942-952.

106. Georgoulia*, P.S. & Glykos, N.M. (2018), "Folding Molecular Dynamics Simulation of a gp41-Derived Peptide Reconcile Divergent Structure Determinations", ACS Omega, 3, 14746-14754.

107. Koukos, P.I. and Glykos, N.M. (2014). Folding Molecular Dynamics Simulations Accurately Predict the Effect of Mutations on the Stability and Structure of a Vammin-Derived Peptide. The Journal of Physical Chemistry B, 118(34), pp.10076–10084.

108. Research in Systems Neuroscience COMPUTATIONAL Journal of CHEMISTRY Organic • Inorganic • Physical Biological • Materials. (n.d.). Available at: https://utopia.duth.gr/glykos/pdf/grcarma_reprint.pdf [Accessed 16 Aug. 2023].

109. Patmanidis, I. and Glykos, N.M. (2013). As good as it gets? Folding molecular dynamics simulations of the LytA choline-binding peptide result to an exceptionally accurate model of the peptide structure. Journal of Molecular Graphics and Modelling, 41, pp.68–71.

110. Kolocouris*, A., Arkin, I. & Glykos*, N.M. (2022), "A proof-of-concept study of the secondary structure of influenza A, B M2 and MERS- and SARS-CoV E transmembrane peptides using folding molecular dynamics simulations in a membrane mimetic solvent", Phys. Chem. Chem. Phys., 24, 25391-25402.

111. Mitsikas, D.A. & Glykos*, N.M. (2020), "A molecular dynamics simulation study on the propensity of Asn-Gly-containing heptapeptides towards β-turn structures: Comparison with ab initio quantum mechanical calculations", PLoS ONE, 15(12): e0243429.

112. Stylianakis, I., Shalev, A., Scheiner, S., Sigalas, M.P., Arkin, I.T., Glykos*, N.M. & Kolocouris*, A. (2020), "The balance between side- chain and backbone- driven association in folding of the α- helical influenza A transmembrane peptide", J. Comput. Chem., 41, 2177-2188.

113. Georgoulia, P.S. & Glykos*, N.M. (2019), "Molecular simulation of peptides coming of age: Accurate prediction of folding, dynamics and structures", Arch. Biochem. Biophys., 664, 76-88.

114. Serafeim, A.-P., Salamanos, G., Patapati, K.K. & Glykos*, N.M. (2016), "Sensitivity of Folding Molecular Dynamics Simulations to Even Minor Force Field Changes", J. Chem. Inf. Model., 56, 2035-2041.

115. Baltzis, A.S., Koukos, P.I. & Glykos*, N.M. (2015), "Clustering of molecular dynamics trajectories via peak-picking in multidimensional PCA-derived distributions", arXiv:1512.04024 [q-bio.BM]

116. Georgoulia, P.S. & Glykos*, N.M. (2013), "On the Foldability of Tryptophan-Containing Tetra- and Pentapeptides: An Exhaustive Molecular Dynamics Study", J. Phys. Chem. B, 117, 5522–5532.

117. Georgoulia, P.S. & Glykos*, N.M. (2011), "Using J-coupling constants for force field validation: Application to hepta-alanine", J. Phys. Chem. B, 115, 15221–15227.

118. Patapati, K.K. & Glykos*, N.M. (2011), "Three Force Fields' Views of the 310 Helix", Biophys. J., 101, 1766-1771.

119. Patapati, K.K. & Glykos*, N.M. (2010), "Order through Disorder: Hyper-Mobile C-Terminal Residues Stabilize the Folded State of a Helical Peptide. A Molecular Dynamics Study", PLoS ONE, 5, e15290.

120. Bank, R.P.D. (2020). RCSB PDB - 6NM2: NMR Structure of WW291.

121. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J. and Back, T. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596, pp.583–589.