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

BSc Thesis

“Folding dynamics of the T-peptide in DMSO”

***Author*: Ioanna Gkogka (1943)**

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

Alexandroupolis, Greece

June 2021

Abstract

A thorough explanation of the protein folding mechanism was not available for almost 50 years. In recent years, several experimental and theoretical approaches have been applied to give new insights into the protein folding process. In the present thesis, a Molecular Dynamics simulation of peptide T in DMSO solution was performed. Peptide T is a synthetic octapeptide fragment, which corresponds to the region 185-192 of the gp120 HIV coat protein and functions as a viral entry inhibitor. In order to validate the accuracy of MD simulations we compared our results with the NMR experimental conclusions derived from the study that Picone and her colleagues had conducted: “A 500 MHz study of peptide T in a DMSO solution”. Their results suggested that a type I β-turn including the four C-terminal residues, T^5, N^6, Y^7, and T^8, in which T^8 is bonded to T^5 CO was the most prominent structure in solution. Our results have shown that peptide T is highly flexible, comprising a dynamic system. The main structural characteristics observed were turns and helices, with a greater preference for β-turns. According to our calculations, the most preferred conformation for residues 5-8 was a β-turn type IV. Moreover, the C-terminal sequence proved to be more stable than the rest of the peptide. Comparison between the experimental and simulation-derived chemical shifts verified a reasonable agreement between the two sets of data. Overall, the MD simulation managed to predict with sufficient accuracy the folding behavior, dynamic properties, and structural characteristics of peptide T, as these have been identified in the experiment.

Περίληψη

Μία αναλυτική περιγραφή του μηχανισμού της πρωτεϊνικής αναδίπλωσης δεν ήταν διαθέσιμη για περίπου 50 χρόνια. Τα τελευταία χρόνια, έχουν εφαρμοστεί αρκετές πειραματικές και θεωρητικές προσεγγίσεις με σκοπό την καλύτερη κατανόηση της διαδικασίας αναδίπλωσης των πρωτεϊνών. Στη παρούσα πτυχιακή εργασία πραγματοποιήθηκε μία προσομοίωση Μοριακής Δυναμικής του πεπτιδίου T σε διάλυμα DMSO. Το πεπτίδιο T είναι ένα συνθετικό οκταπεπτίδιο, το οποίο αντιστοιχεί στην περιοχή 185-192 της πρωτεϊνης gp120 του καλύμματος του ιού HIV και λειτουργεί ως αναστολέας της εισόδου του ιού στα κύτταρα. Με σκοπό να επιβεβαιώσουμε την ακρίβεια της μεθόδου των προσομοιώσεων Μοριακής Δυναμικής συγκρίναμε τα αποτελέσματά μας με τα συμπεράσματα που προέκυψαν από το πείραμα NMR, που πραγματοποίησαν η Picone και οι συνεργάτες της: “A 500 MHz study of peptide T in a DMSO solution”. Τα αποτελέσματά τους υπέδειξαν ότι μία β-στροφή τύπου I, που περιλαμβάνει τα τέσσερα C-τελικά αμινοξικά κατάλοιπα, T^5, N^6, Y^7, και T^8, στα οποία το T^8 σχηματίζει δεσμούς με το T^5 CO, ήταν η πιο πιθανή δομή σε διάλυμα. Τα αποτελέσματά μας έδειξαν ότι το πεπτίδιο T είναι ιδιαίτερα ευέλικτο, συνιστώντας ένα δυναμικό σύστημα. Τα κύρια δομικά χαρακτηριστικά του που παρατηρήθηκαν ήταν οι στροφές και οι έλικες, με μεγαλύτερη προτίμηση στις β-στροφές. Σύμφωνα με τους υπολογισμούς μας, η προτιμώμενη διαμόρφωση για τα κατάλοιπα 5-8 ήταν β-στροφή τύπου IV. Επιπρόσθετα, το C-τελικό τμήμα της αλληλουχίας αποδείχτηκε ότι είναι πιο σταθερό σε σχέση με το υπόλοιπο πεπτίδιο. Η σύγκριση μεταξύ των πειραματικών χημικών μετατοπίσεων και αυτών που προέκυψαν από την προσομοίωση επαλήθευσε ότι υπάρχει ικανοποιητική συμφωνία μεταξύ τους. Συνολικά, η προσομοίωση Μοριακής Δυναμικής προέβλεψε με ικανοποιητική ακρίβεια το πρότυπο αναδίπλωσης, τις δυναμικές ιδιότητες και τα δομικά χαρακτηριστικά του πεπτιδίου T, όπως αυτά αποδίδονται στο πείραμα.

Acknowledgments

I would like to express my deep gratitude and respect to my advisor Dr. Nicholas M. Glykos for the opportunity he gave me to work in an exciting field of science, for his support, guidance, and for providing invaluable feedback throughout this project.

I would also like to thank my family for their support, understanding, and encouragement throughout my studies.

Table of contents

"If we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms."

-Richard P. Feyman

NOBEL LECTURE, 1965

1. Introduction

1.1 Proteins

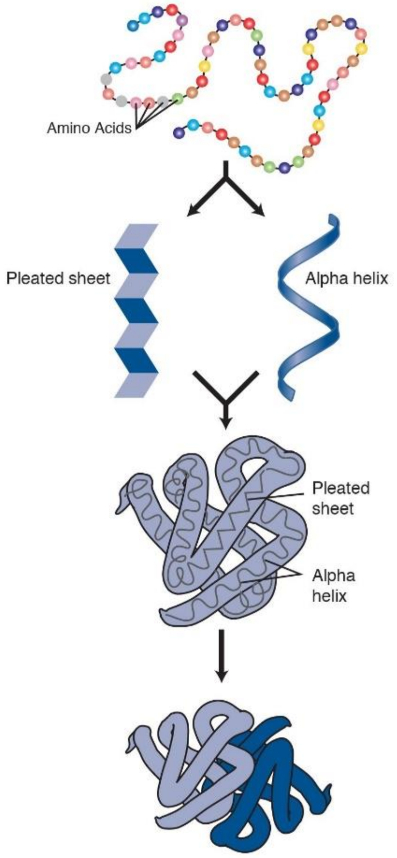


Figure 1.1: The four levels of protein structure: primary structure, secondary structure, tertiary structure and quaternary structure. (adapted without permission from *National Human Genome Research Institute*)

The word protein was first mentioned in a letter sent by the Swedish chemist Jöns Jakob Berzelius to Gerhardus Johannes Mulder on July 10, 1838. He wrote: "*The name protein that I propose for the organic oxide of fibrin and albumin, I wanted to derive from the Greek word πρώτειος because it appears to be the principal substance of animal nutrition"*[1].

Proteins are the most abundant biological macromolecules, characterized by high heterogeneity and demonstrate a wide range of biological functions, including catalyzing metabolic reactions, providing mechanical support and immune protection, transporting and storing molecules, such as oxygen, generating movement, transmitting signals and, regulating cell growth and differentiation[2][3].

The functional properties of proteins are linked with their three-dimensional structure. Protein structure can be examined into four major categories (primary structure, secondary structure, tertiary and quaternary structure) **(Figure 1.1)**. The building blocks of proteins are amino acids. Primary structure refers to the amino acid sequence of a polypeptide chain. Specific parts of the polypeptide chain can form secondary structures. The most common secondary structure elements are α-helices and β-sheets. Secondary structures are stabilized by hydrogen bonds between the main-chain peptide groups. Tertiary structure refers to the three-dimensional shape of a polypeptide chain. Protein molecules that consist of more than one polypeptide chain form a quaternary structure. The formation of tertiary or quaternary structures, causes distant amino acids to approach each other to form a functional region, an active site. The three-dimensional structure occurs because polypeptide chains fold and form compact and self-contained structural regions, known as domains.

In order to be able to perceive the biological function of proteins, we should manage to predict the three-dimensional structure from the amino acid sequence. This is commonly known as the "protein folding problem", one of the major challenges in the field of Molecular Biology[4].

1.2 The Protein Folding Problem

The protein folding problem refers to how a protein's amino acid sequence determines its 3D structure. The mechanism by which a polypeptide chain folds from the denatured random coil state to the native protein structure is a fundamental aspect of Structural Biology. A thorough explanation of the mechanism by which proteins fold was not available for almost 50 years, despite the substantial effort that had been dedicated to this problem[5]. Notable research activity in this field was conducted by Christian Anfinsen and Cyrus Levinthal.

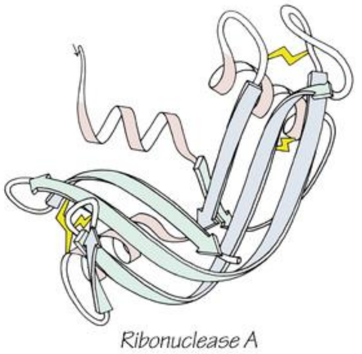


Figure 1.2: Three-dimensional structure of Ribonuclease-A. (reproduced without permission from *Wikipedia*)

Christian Anfinsen performed a series of denaturation-renaturation experiments using the enzyme Ribonuclease-A **(Figure 1.2)** and by 1962 he had developed the "*thermodynamic hypothesis of protein folding*", according to which the native or natural conformation of a protein is thermodynamically the most stable (i.e., the Gibbs free energy of the whole system is lowest) in the intracellular environment. These experiments proved that the native conformation of a protein is defined by the total of the interatomic interactions and therefore, by the amino acid sequence, in a given environment[6].

In 1968 Cyrus Levinthal, in an effort to define the kinetic parameters that determine protein folding, noted that a protein chain of ordinary size would require an enormously long folding time to find the native state by a random search among all possible configurations. Indeed, for a polypeptide chain of 150 residues with three possible conformations for every residue, the time needed to search all possible conformations of the chain is 10^{48} years. Since protein folding time ranges between 0.1 to 1000 seconds, Levinthal's statement is known as "*Levinthal's paradox*"[4][5][7]. Levinthal proposed that there is a pathway of folding, which means that there is "*a well-defined sequence of events which follow one another so as to carry the protein from the unfolded random coil to a uniquely folded metastable state*"[8].

1.3 Protein Folding Models

Several models of protein folding have been suggested. The proposals were based mainly on known structures, to restrict the conformational space that is examined and the folding duration to the experimental scale. Examples include the "*nucleation-growth*" model, the *"diffusion-collision"* model, the *"nucleation-condensation"* model, and the *"jigsaw-puzzle"* model. The majority of these models are descriptive and do not offer a means of calculating the approximate folding time, which is a crucial factor for the resolution of Levinthal's paradox[5].

The *"nucleation-growth"* model proposes that the initial stages of 3D structure formation (nucleation) take place autonomously in distinct regions of globular protein molecules. The growth of the nucleus occurs by adding peptide chain segments that are close to the nucleus in the amino acid sequence. This process would create native protein structures with separate regions of continuous polypeptide chain[9].

According to the *"diffusion-collision"* model, proteins consist of distinct parts (elementary microdomains), each sufficiently short for all conformational alternatives to be examined rapidly. The microdomains diffuse and microdomain-microdomain collisions occur, leading to the formation of higher aggregates[10].

The *"nucleation-condensation"* model suggests that secondary and tertiary structure interactions happen simultaneously. The nucleation-condensation mechanism uses diffuse, extended regions rather than classic nuclei, which are well-formed elements of structure present in ground states, and has the minimum accumulation of intermediates since the nucleus does not form until the transition state[11][12].

The *"jigsaw-puzzle"* model suggests that the evolution of amino acid sequences would benefit multiple paths to the folded-native state. The analogy of a jigsaw puzzle, with multiple routes to a single solution, seems to be suitable[13].

Recently, a so-called "New View" has emerged, replacing the concept of *"folding pathways"* with *"energy landscapes"*. According to the *"energy landscape theory"*, *"folding pathways"* are not the correct explanation for the kinetic problem Levinthal raised. The *"energy landscape theory"* describes the process of reaching a global free energy minimum (satisfying Anfinsen's experiments) and fulfills Levinthal's kinetic problem, by multiple folding routes on funnel-like energy landscapes[14]. An energy landscape is the free energy of each conformation as a function of the degrees of freedom. The vertical axis of the funnel represents the 'internal free energy' of a particular chain conformation. The many lateral axes represent the conformational coordinates. Each conformation is depicted by a point on the multidimensional energy surface. High energy conformations are displayed as hills, whereas low energy conformations as valleys. The kinetic process of folding and unfolding a protein can be compared to the movement of a ball on this energy surface, so that each protein molecule corresponds to a ball rolling on the energy landscape, moving through the hills and valleys, finally ending in the bottom of the funnel, the native state[15]. Different types of energy landscapes are presented in **Figure 1.3**.

Levinthal's "golf course" energy landscape **(Figure 1.3a)** represents Levinthal's random search problem. When a ball rolls randomly on a flat course, it requires a long time to find, and fall in, the hole. The "grooved golf course" landscape **(Figure 1.3b)** presents the 'pathway' solution to Levinthal's random search problem. Starting from a denatured conformation A, the folding molecule goes through a tunnel on the landscape, to the native structure N. The "HP+" landscape **(Figure 1.3c)** is an idealization showing that the decrease of a protein's internal free energy, leads to a reduction of its conformational freedom. The "bumpy bowl" model **(Figure 1.3d)** is a rugged energy landscape with kinetic traps, energy barriers, and narrow paths, which lead to the native structure. The "moat" energy landscape **(Figure 1.3e)** depicts that a protein can follow a slow or a fast folding process. The "champagne glass" model **(Figure 1.3f)** illustrates how conformational entropy can cause free energy barriers to the folding process. In this model, the rate-limiting factor for folding is the wandering on the flat plateau as the polypeptide chain attempts to locate its way downhill[15].

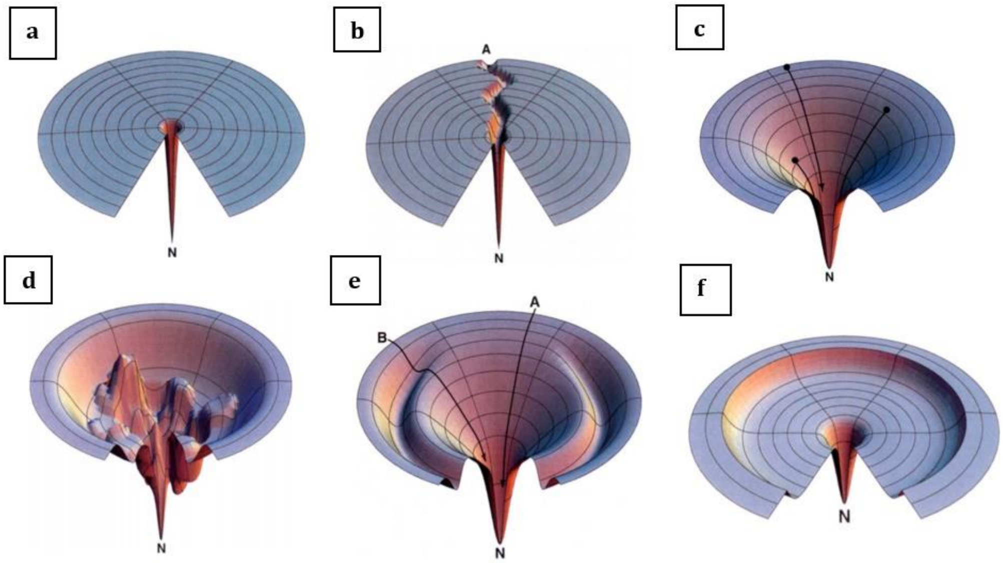


Figure 1.3: Different energy landscapes. (a) The Levinthal 's "golf course" landscape. N is the native conformation. The chain searches for N randomly. (b) The "grooved golf course" landscape presents the 'pathway' solution to the random search problem. A pathway leads from a denatured conformation A to the native conformation N. (c) The "HP+" landscape is an idealized funnel landscape. The reduction of the protein's free energy leads to the native structure. (d) The "bumpy bowl" energy landscape is a rugged energy landscape with kinetic traps, energy barriers and some narrow paths, which lead to the native structure. (e) The "moat" landscape illustrates that a protein could have a fast-folding process (A), or, when a kinetic trap is present, a slow-folding process (B). (f) The "champagne glass" landscape illustrates how conformational entropy can cause free energy barriers to the folding process. (adapted without permission from *Dill & Chan, Nature Structural Biology, 1997)*

1.4 Experimental methods to study protein folding

Many biophysical methods have been applied over the past years to describe the folding of a variety of proteins[16]. Common experimental methods include X-ray crystallography (XRC), Nuclear Magnetic Resonance (NMR) spectroscopy, Circular Dichroism (CD), and Cryogenic Electron Microscopy (Cryo-EM).

X-ray crystallography is currently the most preferred method for structure prediction of proteins and biological macromolecules. This method aims to determine the 3D molecular structure from a crystal. A crystalline sample is exposed to an *x*-ray beam and then the resulting diffraction patterns are analyzed. The pattern of the diffraction spots provides information regarding the crystal packing symmetry and the size of the repeating unit that forms the crystal. The intensities of the spots can be utilized to produce an electron density map. This map is then used to obtain the molecular structure of the protein[17].

NMR spectroscopy has been widely used to analyze the structure of small molecules, small proteins, or protein domains. The only prerequisite of this method is a small volume of concentrated protein solution that is placed in a strong magnetic field. Certain atomic nuclei, and particularly those of hydrogen atoms, have a magnetic moment or spin. The spin aligns along the strong magnetic field. In response to applied radiofrequency (RF) pulses of electromagnetic radiation the spin changes to a misaligned, excited state. With the return of the excited hydrogen nuclei to their aligned state, RF radiation is emitted. This radiation can be calculated and displayed as a spectrum. The environment of each hydrogen nucleus influences the nature of the emitted radiation. When a nucleus is excited, it affects the absorption and emission of radiation by nearby nuclei. Two-dimensional NMR (2D NMR) is a set of NMR techniques giving data plotted in a space defined by two frequency axes rather than one. 2D NMR spectra supply more information about a molecule than one-dimensional NMR spectra and are particularly valuable for structure determination. It is possible by 2D NMR to differentiate the signals from hydrogen nuclei in different amino acid residues and to recognize and calculate the small shifts in these signals that appear when these hydrogen nuclei are in close proximity, for interaction to take place. The size of these shifts represents the distance between the interacting pair of hydrogen atoms and consequently, gives information about the distances between the parts of the protein molecule. Combining this information with the amino acid sequence of the protein molecule, it is feasible to obtain the 3D structure of the protein[18]. The usefulness of the NMR method lies in its specificity at the level of distinct atoms. This method can define the distribution of structures in a conformational ensemble from parameters extracted from the spectra. The calculation of Nuclear Overhauser Effects (NOEs) can detect the proximity of pairs of atoms. Less specific information is available from the measurement of chemical shifts[19].

Circular dichroism (CD) is used for the definition of secondary structure elements and folding characteristics of proteins that have been acquired using recombinant techniques or purified from tissues. The most common applications are to find out whether an expressed, purified protein is folded, or if a mutation disturbs its configuration or stability and to study protein-protein interactions[20].

Cryogenic Electron Microscopy (Cryo-EM) is a method for imaging frozen-hydrated samples at cryogenic temperatures by electron microscopy. Cryo-EM examines the specimen without the use of artificial treatments such as fixing, dehydration, and staining, allowing the representation of the native state of biological structures[21].

1.5 Computational methods to study protein folding

Although experimental techniques for determining protein structure, provide high-resolution structural information about a subset of proteins, they are difficult, expensive, and time-consuming. Protein structures that could not be found using experimental techniques, can be predicted by various computational approaches[22]. There are three main categories of computational protein structure prediction methods: comparative modeling, ab initio methods, and fold recognition methods.

Comparative modeling, also known as homology modeling, is considered to be the most reliable method for protein structure prediction. In comparative modeling, the structure of a protein is predicted by comparing its amino acid sequence with the sequences of proteins of known structure. The assumption is that proteins with similar sequences have similar structures. If a strong similarity is found, it can be assumed that the proteins have similar structures. If no strong similarities can be found, then comparative modeling cannot be used[23]. The accuracy of predictions made by comparative modeling greatly relies on the degree of sequence similarity. If the target and the template have more than 50% sequence identity, predictions are usually of high quality, whereas when they share less than 30% of their sequences, the prediction will possibly contain significant errors[24].

Ab initio methods use only the information in the target sequence. There are two branches of ab initio prediction: knowledge-based methods and simulation methods. Knowledge-based prediction methods predict structures by applying rules, which are derived from observations made on known protein structures. Simulation methods attempt to predict protein structures by simulating the protein folding process using basic physics. The central principle of simulation-based protein structure prediction is that the native fold of a protein can be identified by discovering the configuration of the protein molecule with the lowest energy as determined by an appropriate potential energy function[23].

Fold recognition methods rely on the concept that structure is evolutionary more conserved than sequence. Therefore, the variety of different folds is more restricted than suggested by sequence diversity. These methods attempt to detect a model fold for a given target sequence among the known folds even if no sequence similarity can be identified[24].

A worldwide experiment for protein structure prediction, the Critical Assessment of protein Structure Prediction (CASP) is taking place every two years since 1994 has helped in the advancement of methods of identifying protein structure from sequence[25].

1.6 Secondary structure elements

During protein folding the hydrophobic side chains are being packed towards the center of the protein molecule, forming a hydrophobic core and a hydrophilic outer surface. The packing of hydrophobic side chains towards the interior of proteins is extremely dense. Given the conformational restrictions caused by the steric hindrances generated by the side-chains, protein folding resembles a three-dimensional puzzle. Concurrently, the protein's backbone chain must also fold into the hydrophobic core of the molecule. However, the presence of imine groups (NH) and carbonyl groups (C'=O) in each peptide group, which act as proton donors and proton receptors, respectively, results in high hydrophilicity of the main chain. In a hydrophobic environment, these polar groups must be neutralized, via the formation of hydrogen bonds. As a result, stable conformational patterns form, known as secondary structure elements. The most common types of secondary structures are α-helices and β-sheets **(Figure 1.4)**.

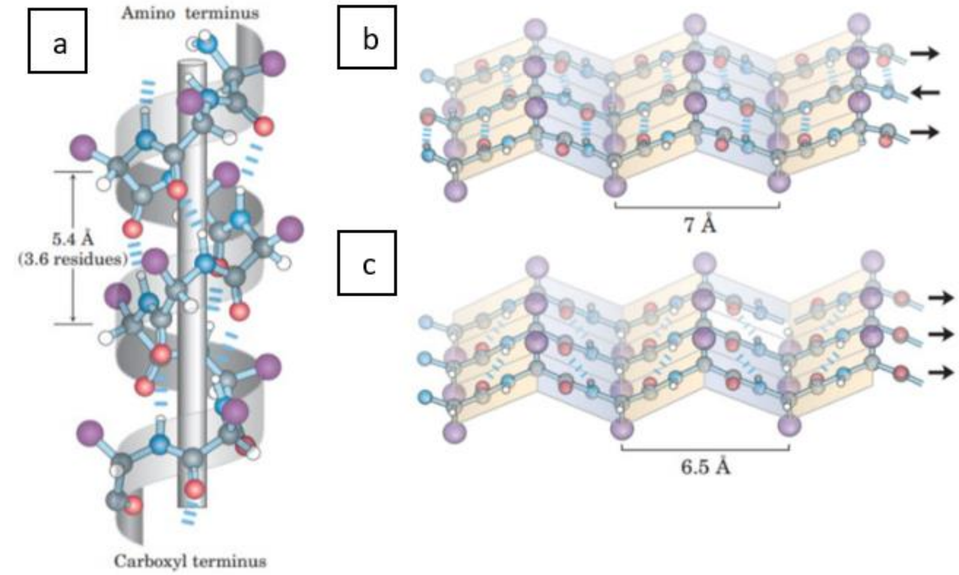


Figure 1.4: (a) Ball-and-strick model of an α-helix showing the interchain hydrogen bonds. (b) Parallel β-sheet structure. (c) Anti-parallel β-sheet structure. The color coding is the following: grey: carbon atoms, white: hydrogen atoms, red: oxygen atoms, blue: nitrogen atoms and purple: R groups. (adapted without permission from *Nelson & Cox, Lehninger Principles of Biochemistry*).

Both structures are characterized by hydrogen bonds between NH and C' =O groups of the main chain[4]. A helix is the simplest arrangement the polypeptide chain can adopt. The repeating unit is a single turn of the helix, which extends about 5.4 Å along the long axis. Each helical turn includes 3.6 amino acid residues. In all proteins, the helical twist of the helix is right-handed. The β-sheet is formed of β-strands, extended conformations of 5 to 10 residues. The adjacent polypeptide chains in a β-sheet can be either parallel or antiparallel, meaning that they have the same or opposite amino-to-carboxyl orientations, respectively. The repeating unit is shorter for the parallel conformation (6.5 Å, vs. 7 Å for antiparallel) and the hydrogen-bonding patterns differ as well[2].

The two structural elements analyzed above are not the only ones present in protein structures. Examples of other secondary structure elements include α_L-helices or left-handed α-helices, 3_{10}-helices (3 residues and 10 atoms per turn), π-helices (4.1 residues per turn), β-turns, and random coils[4].

1.7 φ, ψ dihedral angles

In order to describe secondary structure elements, it is important to understand the main parameters that define the conformation of a polypeptide chain. Assuming a dipeptide of residues n and n+1, the peptide group includes the Cα carbon atom and the C'=O group of residue n, as well as the NH group and Cα carbon atom of residue n+1. Since peptide groups lack flexibility, due to the inflexible nature of the C'-N peptide bond, they only have two degrees of freedom that correspond to the torsion angles of N-Cα and Cα-C' bonds. These dihedral torsion angles are called φ and ψ, respectively[4] (**Figure 1.5**). In principle, φ and ψ angles can have any value between -180° and +180°, but many values are prohibited by steric interference between amino acid side chains and main-chain atoms[2].

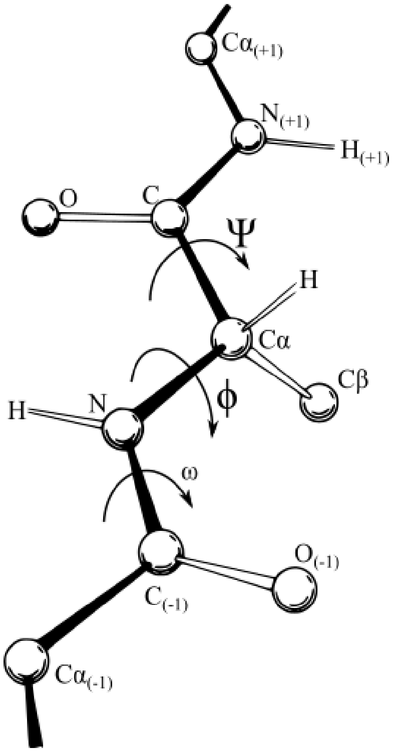


Figure 1.5: The φ and ψ torsion angles. (reproduced without permission from *Wikipedia*)

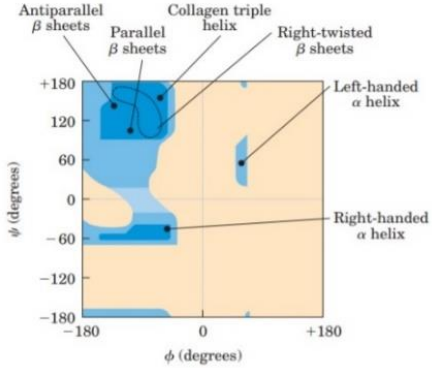


Figure 1.6: A Ramachandran plot showing the allowed secondary structures. (reproduced without permission from Nelson & Cox, Lehningher, principles of Biochemistry).

Allowed values for φ and ψ become evident when they are plotted against each other in a Ramachandran plot (**Figure 1.6**), introduced by the Indian biophysicist G. N. Ramachandran. Each dot of the plot corresponds to the φ, ψ values of one amino acid residue, which is part of a well-defined structure. In the Ramachandran plot the allowed φ, ψ angle values of the residues that form the various secondary structure patterns, are represented as distinct regions. More specifically, the α-helix region is depicted in the lower-left quadrant, the β-sheet region in the upper left quadrant, and the α_L-helix region in the upper right quadrant of the plot[4].

1.8 β-turns

β-turns are the most common type of nonrepetitive structural pattern present in proteins, comprising on average 25% of the residues[26]. Turns have a significant role in proteins; they are the connecting elements that link different secondary structure elements (for example the ends of two adjacent segments of an antiparallel β-sheet),[2] they provide a direction change for the polypeptide chain and have been associated with molecular recognition and protein folding. Therefore, since β-turns were first recognized substantial effort has been dedicated to their analysis and the prediction of turns from the amino acid sequence[26]. The structure is a 180^o turn involving four amino acid residues (i, i+1, i+2, i+3)[2].

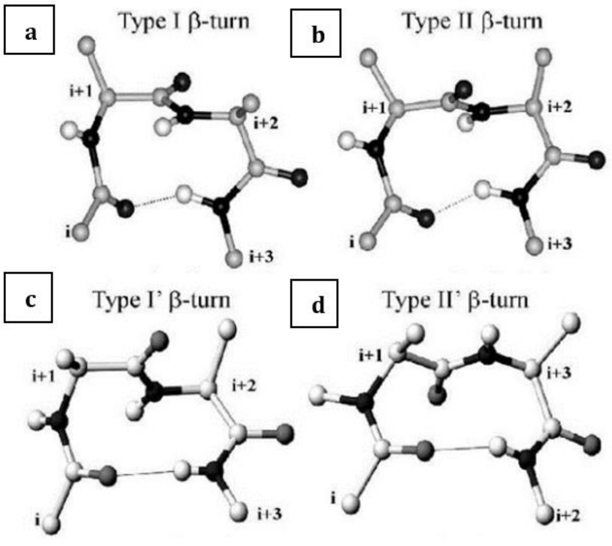


Figure 1.7: The four most common β-turn types (a) type I, (b) type II, (c) type I', (d) type II'. (adapted without permission from *Appavu et al., Transcriptomics: Open Access, 2016*)

β-turns (**Figure 1.7**) were first recognized by Venkatachalam in 1968, who was studying favorable conformations of three consecutive peptide units using model-building techniques. He recognized three conformations (I, II, and III) that formed a hydrogen bond between the main chain carboxyl oxygen of the first residue (i) and the amino-group hydrogen of the fourth (i+3) and their main-chain mirror images (I', II', III') that would be disfavored due to steric interactions. In 1973 Lewis *et al.* found that 25% of β-turns do not possess the hydrogen bond proposed by Venkatachalam and they extended the definition of β-turns to incorporate these examples. The definition states that the distance between C_α atoms of residues i and i+3 is less than 7 Å and that the chain is not in a helical conformation. Using this they broadened the number of turn types to 10 (I, I', II, II', III, III', IV, V, VI, VII) some categories were based on *φ, ψ* angles, whereas others were based on less-stringent criteria. The classification of *β*-turns, using *φ, ψ* angles, into seven conventional turn types (I, I', II, II', IV, Via, VIb) and a new class of *β*-turn, type VIII is now widely accepted[27].

In terms of the positional potentials, generally, hydrophilic residues are more likely to be present in turns than hydrophobic residues. This is because turns are usually on the solvent-exposed, outer surface of proteins. Residues such as glycine, proline, asparagine, and aspartic acid are strongly preferred in turn conformations. There are significant differences between the potentials at each position; for instance, proline is favored at the first 2 positions of turns (i, i+1), whereas glycine is favored at the last 2 positions (i+2, i+3), and asparagine and aspartic acid are strongly favored at positions i and i+2[28].

1.9 Peptide folding simulations

Peptides are short chains of between two and fifty amino acids, linked by peptide bonds. Studies based on experimental techniques and molecular dynamics simulations have proven that peptide fragments tend to form native-like secondary structures. Experimental techniques, such as NMR, demonstrate that long peptide fragments adopt native-like conformations. This also applies to some short peptides in solution. Consequently, peptide conformational propensities that are extracted from the protein databank (PDB) are extensively used in protein-structure prediction methods[29].

Peptides are small systems that mimic many of the characteristics and complexities of larger molecular systems, such as proteins. Peptide folding simulations and experiments describe the dynamics and molecular mechanisms of primary events of protein folding. In terms of computational power, peptides are a more tractable system than proteins, and experimentally, they fold at extremely fast rates. Therefore, peptide systems provide a connection link between theoretical and experimental understanding of protein folding[30][31].

It is also important to mention that peptide simulations aim for a better understanding of the folding mechanisms and play a vital role in the improvement of physics-based three-dimensional structure prediction methods. More specifically, peptides are widely used for the development, advancement, validation, and optimization of force fields, to improve the ability of simulations to represent physical reality[31].

1.10 Human Immunodeficiency Virus

Human immunodeficiency viruses HIV-1 and HIV-2 are two species of Lentiviruses that infect humans[32]. HIV, along with related retroviruses belong to a class of enveloped fusogenic viruses that include coronaviruses, paramyxoviruses, and orthomyxoviruses. The activation of all of the above viruses depends on post-translational cleavage. HIV causes the depletion of CD4$^+$ T helper lymphocytes in their hosts and as a result, the acquired immunodeficiency syndrome (AIDS) develops[33]. AIDS was first reported as a new disease in the United States in 1981 when increasing numbers of young homosexual men yielded to rare infections and malignancies[34]. The isolation of HIV-1 followed, 2 years later. The disease was found to be established in heterosexual populations of central and east Africa[35]. The viral infection can have disastrous effects on host defense mechanisms, causing a wide range of opportunistic infections and neoplasms. HIV is one of the most extensively studied viruses and innovations in structural biology and molecular immunology have led to substantial advances and a better understanding of its mechanisms of function and effects on the human immune system[36]. HIV entry into the host cell is a complex mechanism that consists of various steps (**Figure 1.8**).

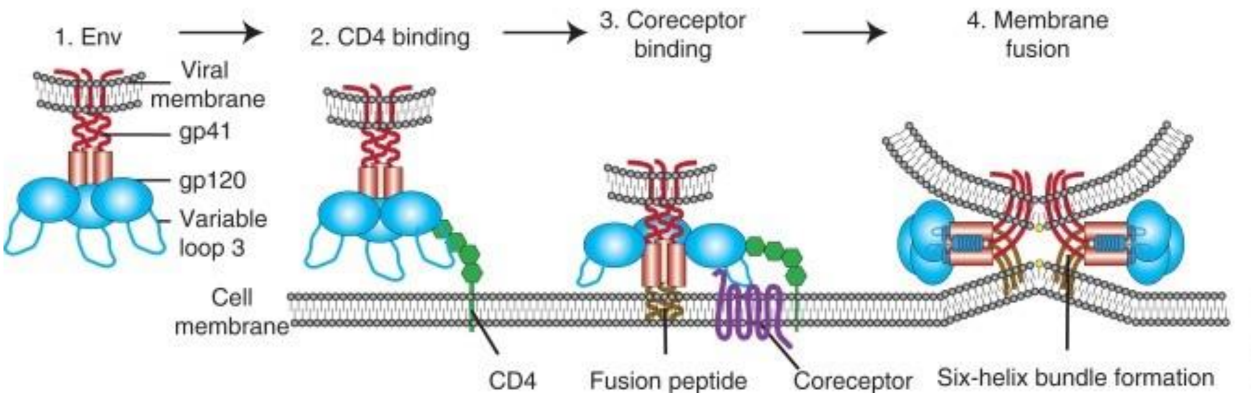


Figure 1.8: HIV entry into the host cell. (1) Attachment to the host cell. (2) Binding to the host protein CDC4. (3) Conformational changes in Env, allowing coreceptor binding, membrane fusion initiates. (4) Membrane fusion. (adapted without permission from *Wilen et al., Cold Spring Harbor Perspectives in Medicine, 2012*)

The first step of the viral replication cycle includes the adhesion of the virus to the host cell and the fusion of the cell and the viral membranes with the subsequent entry of the viral core into the cytoplasm of the host cell. HIV delivers its genome into the host cell cytoplasm following a complex series of steps, while at the same time evading the host immune response. Firstly, virions must attach to the target cell either by the viral envelope protein (Env) or host cell membrane proteins integrated into the virion. The attachment of HIV to the host cell brings the protein Env close to the viral receptor and coreceptor, therefore increasing the efficiency of infection. The second step of viral entry involves the attachment of Env to the host protein CD4, its primary receptor. Env is a highly glycosylated trimer of the exterior envelope glycoprotein, gp120, and the transmembrane envelope glycoprotein, gp41 heterodimers. The role of the gp120 subunit is to assist with receptor binding. The third step entails coreceptor binding (either CCR-5 or CXCR-4) and is considered to be the trigger that stimulates the membrane fusion potential of Env. The fourth step involves the movement of the virus particle to the productive binding site. In the final step membrane fusion induced by Env occurs[37]. All these steps of viral entry are shown in Figure 1.8.

1.11 gp120

The exterior envelope glycoprotein gp120 plays a significant role in receptor binding and interactions with neutralizing antibodies. Structural information regarding gp120 is essential for the determination of the mechanism of HIV infection and the design of new therapeutic approaches.

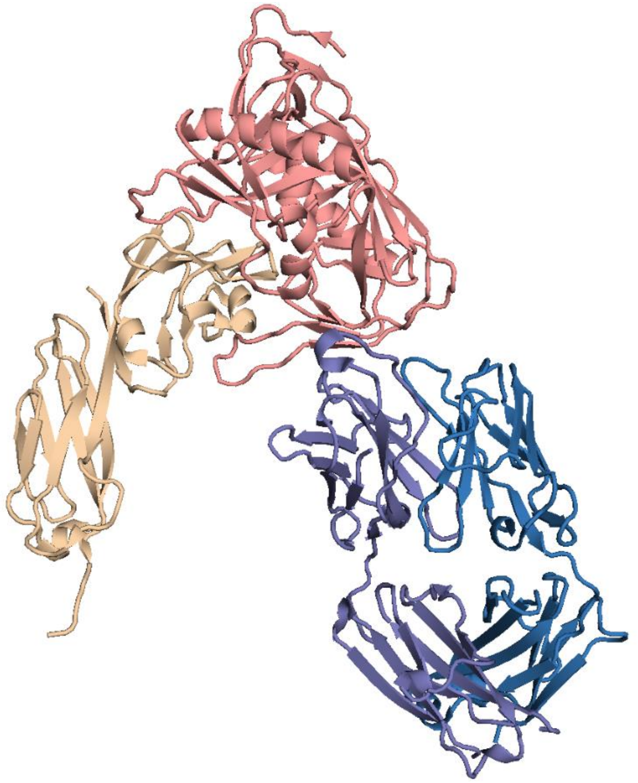


Figure 1.9: 3D structure of the gp120-CD4-Fab ternary complex. The cartoon diagram illustrates gp120 in pink, the N-terminal two domains of CD4 in yellow, and the Fab in blue and purple. PDB ID: 1GC1

The structure of gp120 protein has been determined as a complex with the N-terminal two domains of CD4, and a Fab from the human neutralizing monoclonal antibody 17b, which partially mimics the HIV-1 co-receptor. The 3D structure of this ternary complex is presented in **Figure 1.9**.

The core of gp120 is comprised of 25 β-strands, 5 α-helices, and 10 defined loop segments. The protein's polypeptide chain is folded into two major domains, the inner and outer domains. The inner domain includes a two-helix, two-strand bundle with a small five-stranded β-sandwich at its C-terminal end. The outer domain includes a stacked double barrel. The proximal barrel of the outer domain consists of a six-stranded, mixed-directional β-sheet that is twisted to embrace helix α2, while the distal barrel of the outer domain is a seven-stranded antiparallel β-barrel. The outer domain is situated in such a way so that the outer barrel and inner bundle axes are almost parallel.

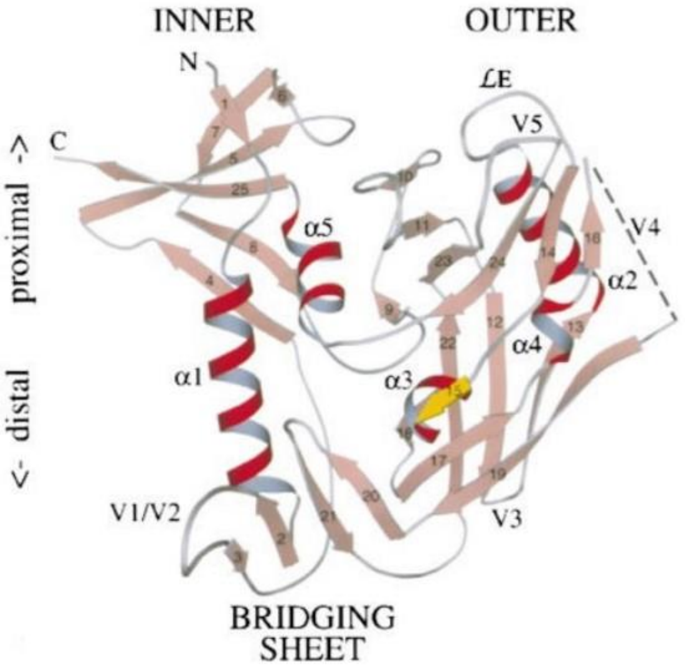


Figure 1.10: Ribbon diagram representing the structure of core gp120. The color-coding is red for α-helices, salmon for β-strands, and grey for loops. (adapted without permission from *Kwong et al., Nature, 1998*)

The proximal end of the outer domain includes variable loops V4 and V5 and the loops LD and LE which also seem to have a sequence variability, as well as loop LC, which is close in the three-dimensional space with loop LA of the inner domain. The distal end of the outer domain includes loop V3, as well as loop LF, which helps in the formation of a β20-β21 β-hairpin. The β-hairpin forms hydrogen bonds with the V1/V2 stem of the inner domain. This leads to the formation of an antiparallel, four-stranded "bridging sheet", which acts as a minidomain[33]. **Figure 1.10** is a 3D representation of core gp120.

1.12 Peptide T

It has been suggested that the interaction between HIV-1 and its host cell receptors could entail the region 185-192 of the gp120 coat protein[38], which corresponds to the gp120 V2 region[39]. The synthetic octapeptide fragment with the sequence: ASTTTNYT, is known as peptide T due to its high threonine content and it was proven to function as a viral entry inhibitor by blocking the binding of both isolated gp120 and HIV-1 with the CD4 receptor[38][39][40]. Later studies have suggested that both the CD4 receptor and a co-receptor are needed for the invasion of healthy cells by HIV-1. The core fragment of gp120 presented in Figure 1.10 is depleted of some variable regions, including the variable V2 loop, which includes peptide T. Nevertheless, the likely structure of the V1/V2 stem is modeled and located above the gp120 core, in close proximity with the antibody fragment. Furthermore, the authors of the study that determined the ternary structure of the complex presented in Figure 1.9, proposed that CD4 binding induces a conformational change in gp120, which translocates the V1/V2 loop even closer to the co-receptor, implying that the V2 region may be involved in the interaction with the chemokine co-receptor, and/or the CD4 receptor[38]. A later study reported the structure of V1/V2 in complex with a human antibody, PG9 and they concluded that V1/V2 forms a four-stranded β-sheet domain, involving four anti-parallel β-strands[41].

Peptide T is endowed with a strong chemotactic activity on human monocytes, which is associated with the blockage of CD4 binding[38]. The chemotactic activity could be inhibited by anti-CD4 monoclonal antibodies (Mabs). The core peptide required for chemotactic activity is the pentapeptide: Thr-Thr-Asn-Tyr-Thr[42]. Other biological activities include its pharmacological ability to prevent neuronal cell death produced by the envelope protein, which leads to the evaluation of peptide T as a potential therapeutic agent for the neuropsychiatric and neurological deficits of AIDS[43]. Peptide T has also a potential capability to resolve psoriatic lesions[44]. Psoriasis affects 2–4% of the population, and an increased occurrence of psoriasis has been identified in patients infected with HIV. Although peptide T seems to have positive outcomes in the treatment of psoriasis, little is known regarding the mechanism of peptide T action in treating this disease. Examples of the possible mechanisms that peptide T may use include its ability to interact with somatostatin, with the vasoactive intestinal polypeptide (VIP), and with the epidermal growth factor to regulate cell growth, its capacity to affect the synthesis of somatostatin, etc[45]. It is also important to note that peptide T has the ability to weaken neuroinflammation associated with Alzheimer's disease (AD)[44]. AD is correlated with aging and is defined by brain inflammation leading to neocortical atrophy. Neocortical atrophy and the depletion of large neocortical neurons are also frequent characteristics of HIV infection of the brain, implying convergence of pathogenic pathways. The capacity of chronic Dala1-peptide T-amide (DAPTA), the modified analog of peptide T, treatment to prohibit reductions in cortical thickness and loss of supplementing cortical neurons suggests novel research areas related to the pathogenesis of dementia that involves neuronal loss, even when triggered by different etiologic agents (AIDS vs. AD)[46].

Several clinical trials have been conducted to evaluate peptide T as a possible treatment. Between 1987 and 1989 the National Institutes of Mental Health (NIMH) conducted phase I clinical trials to test immunological parameters, along with viral load changes. The results revealed that the synthetic peptide was entirely non-toxic and also exhibited several symptomatic improvements. However, a randomized double-blind placebo clinical trial of internasal peptide T for the treatment of painful distal neuropathy associated with AIDS concluded that there was no significant difference in pain scores. Many placebo-controlled trials have proven that peptide T has clinical benefits in reversing memory loss and cognition related to HIV infection[46]. Since then, several controversial data have been reported on the possible therapeutic applications of the peptide, which resulted in a loss of interest in peptide T for the treatment of AIDS. The discovery of the role of chemokine receptors introduced a re-examination of the peptide and its analogs[38][46].

1.13 Purpose of the present thesis

During a disease outbreak, alternative therapies such as drug repurposing, vaccination, and immunotherapy are more effective than traditional drug discovery, which is inherently slow to cope with the need for timely therapeutic solutions. Both vaccination and immunotherapy are based on peptide targets. Peptide-based therapies can be more rapid and cost-effective in clinical settings. Other advantages of peptides include their high biological activity, specificity and affinity to desired targets, and low toxicity, due to the limited possibility for accumulation in the body[47][48]. Thus, a better understanding of the structure and function of peptides has proven to be a crucial factor that can contribute to the improvement of human health and disease management[48].

The purpose of the present thesis is to study the folding mechanism of peptide T through Molecular Dynamics simulations and to compare our results with the results from the NMR experiment that Picone *et al.* had conducted: “A 500 MHz study of peptide T in a DMSO solution”. In this publication, they studied peptide T as a zwitterion in DMSO solution by means of proton NMR spectroscopy at 500MHz. More specifically, NMR spectra were obtained at 500 MHz, double-quantum-filtered (DFQ) COSY, and NOESY spectra were run and chemical shifts of all backbone protons and temperature coefficients of the labile protons were reported. The chemical shift data indicated a non-random conformational state. They found that the residues S^2 and T^8, whose resonances of the NH groups are broader than the other five, might adopt a singular conformational position and that the side chains of two of the four threonines, whose methyl groups both resonate at 1.03 ppm, are in a similar environment. To their surprise, the NOESY spectrum showed effects between NHs of T^4 and T^5 and Y^7 and T^8, findings that indicate the presence of well-defined conformers. They concluded that peptide T presents an unusual degree of conformational order in DMSO solution. The minimal value of the T^8 chemical shift in the range of 298-330K, the diagnostic NOE between the NHs of Y^7 and T^8, and variable temperature data were consistent with a type I β-turn including the four C-terminal residues, T^5, N^6, Y^7, and T^8, in which T^8 is bonded to T^5 CO. Although this conformation seemed to be the most prominent, they believed that it was not the only one present in solution. It seemed to be the only one detectable due to the non-linear dependence of NOE on interatomic distances[49].

In the literature reporting computer simulations, it is quite common to find a statement such as “the simulation results are in agreement with the experimental results”[30]. The statement mentioned previously has sometimes been used vaguely or without much consideration. The validity of the methodology or calculations is often supported just by a qualitative or incidental agreement[30]. One of the major concerns when evaluating the effectiveness of MD simulations of proteins or peptides is the degree to which the simulations sufficiently sample the conformational space of the protein or peptide. If a property is inadequately sampled over the MD simulations, the results obtained will be of limited value[50].

The aim of this project was to analyze the molecular dynamics trajectory, study the structural properties of peptide T and examine which are the most likely conformations that it can adopt using physics-based methods, under the scope of comparing the simulation results with the physical reality. We also calculated the simulation-derived chemical shifts and tried to measure the divergence from the experimentally determined chemical shift values. A more detailed description of the workflow that we followed can be found in the 4.1 section of this thesis.

2. Molecular Dynamics Simulations

2.1 Introduction

Molecular dynamics simulations were first introduced by Alder and Wainwright in the late 1950s to analyze the interactions of hard spheres. The first molecular dynamics simulation of a realistic system was the simulation of liquid water, which was carried out by Rahman and Stillinger in 1974. The first protein simulation was the simulation of the bovine pancreatic trypsin inhibitor (BPTI) in 1977. Nowadays, this method has become a routine and a variety of molecular dynamics experiments have been conducted, such as molecular dynamics simulations of solvated proteins, protein-DNA complexes as well as lipid systems, addressing a wide range of issues, including the thermodynamics of ligand binding and the folding of small proteins. Also, molecular dynamics simulation techniques are extensively used in experimental techniques such as X-ray crystallography and NMR structure prediction[51].

There are two main categories of simulation techniques: Molecular Dynamics Simulations (MD) and Monte Carlo Simulations (MC). Additionally, there is a whole series of hybrid techniques that combine characteristics from both of the above-mentioned methods. The main advantage of MD over MC is that it allows us to investigate the dynamical properties of the system, such as transport coefficients, time-dependent responses to perturbations, rheological properties, and spectra[52].

By following the dynamics of a molecular system in space and time, we can gain new insights concerning the structural and dynamic properties of our system, such as molecular geometries and energies, mean atomic fluctuations, local fluctuations, enzyme-substrate binding, rates of configurational changes, free energies, and the nature of different types of concerted motions[53]. Moreover, computer simulations act as a bridge between microscopic length and time scales and the macroscopic world, as well as between theory and experiment. MD simulations allow us to access information that would be difficult to obtain through classical experiments, make comparisons with experimental results, or test a new theory[52].

2.2 Statistical Mechanics

Molecular dynamics simulations describe a system at the microscopic level, using variables such as atomic positions and velocities. The conversion of this microscopic information to macroscopic observables requires statistical mechanics. Statistical mechanics is the field of physical sciences that focuses on understanding macroscopic systems from the properties of individual molecules comprising the system. Time-independent statistical averages are often introduced, to link the macroscopic to the microscopic system[51].

Nevertheless, not all properties of a system can be directly measured in a simulation. Most of the aspects that can be measured in a simulation cannot be compared to experimental data, since none of the experimental approaches provides such detailed information. Instead of using such explicit information, a typical experiment calculates average properties, averaged over a large number of particles and over the time of the measurement. If we intend to use computer simulations as the numerical equivalent of experiments, we must be aware of what kind of averages we should aim to measure, and to do so, we need to introduce the "language" of statistical mechanics[54].

2.3 Classical Mechanics and Integration Algorithms

MD simulations are based on Newton's second law of motion. Knowing the force applied to each atom, it is feasible to define the acceleration of each atom in the system. Integration of the equations of motion then produces a trajectory that describes the positions, velocities, and accelerations of the particles as they change with time. Then, from the trajectory analysis, we can determine the average values of properties of the system. The method is deterministic; once the positions and velocities of every atom are known, the state of the system can be easily predicted at any time.

Using mathematical terms to describe the above said, Newton's equation of motion is given by

$$F_i = m_i a_i \tag{2.1}$$

where F_i is the force applied on particle *i*, m_i is the mass and a_i the acceleration of particle *i*. The force can also be expressed as a gradient of potential energy

$$F_i = -\nabla_i V \tag{2.2}$$

The combination of equations (2.1) and (2.2) leads to the following two equations

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \tag{2.3}$$

$$a_i = -\frac{1}{m_i}\frac{dV}{dr_i} \tag{2.4}$$

Therefore, it is possible to calculate a trajectory knowing only the initial positions of atoms, an initial distribution of velocities, and the acceleration, which is defined by the gradient of the potential energy function. As mentioned previously, the equations of motion are deterministic. This means that knowing the positions and velocities at time zero, then it is possible to calculate the positions and velocities of the system at any other given time, t. The initial positions can be derived from experimental structures, which were solved by experimental techniques such as X-ray Crystallography and NMR spectroscopy. The velocities, *vi,* are often selected randomly from a Maxwell-Boltzmann or Gaussian distribution at a given temperature

$$p(v_{ix}) = \left(\frac{m_i}{2\pi k_B T}\right)^{1/2} exp\left[-\frac{1}{2}\frac{m_i v_{ix}^2}{k_B T}\right] \tag{2.5}$$

where k_B is Boltzmann's constant and T is the temperature of the system. The temperature can be calculated using the relation

$$T = \frac{1}{(3N)}\sum_{i=1}^{N}\frac{|p_i|}{2m_i} \tag{2.6}$$

where *N* is the number of atoms in the system. The potential energy is a function of the atomic positions (3N) of all the atoms in the system. Due to the complexity of this function, the equations of motion can only be solved numerically, but not analytically. Alternatively stated, there is no analytical solution to the equations of motion, either because of the complicated nature of the potential energy function or because of the extended computational time needed. As a consequence, numerical algorithms have been developed for the integration of the equations of motion. The most noteworthy integration algorithms are the Verlet algorithm, the Leap-frog algorithm, the Velocity Verlet, and Beeman's algorithm.

The majority of the above algorithms are based on Taylor's series of expansion, the main advantage of which is the reduction of an equation's terms

$$r(t + \delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \cdots \tag{2.7}$$

$$v(t + \delta t) = v(t) + \alpha(t)\delta t + \frac{1}{2}b(t)\delta t^2 + \cdots \tag{2.8}$$

$$a(t + \delta t) = a(t) + b(t)\delta t + \cdots \tag{2.9}$$

Regarding the proper selection of the algorithm, it is essential to consider the algorithm's ability to conserve energy and momentum, its computational efficiency, and its integration time step, thereby the results will be as close to reality as possible, though a level of inaccuracy will inevitably be present[51].

2.4 Force Fields

Force fields are empirical functions that calculate the energy of a system as a function of the nuclear positions[55]. The majority of the current generation force fields (or potential energy functions) provide a fairly sufficient compromise between computational efficiency and accuracy[51]. However, the development of parameter sets is a very challenging task, demanding rigorous optimization and parameterization and substantial efforts will be required to further improve their accuracy[51][52]. Among the most commonly known and widely used force fields are the Assisted Model Building for Energy Refinement (AMBER)[56], Chemistry at Harvard Macromolecular Mechanics (CHARMM)[57], Groningen Molecular Simulation (GROMOS)[58], and Optimized Potentials for Liquid Simulation (OPLS)[59]. The value of the energy is calculated as a sum of internal, or bonded, interactions and external, or non-bonded, interactions

$$V(R) = E_{bonded} + E_{non\text{-}bonded} \tag{2.10}$$

E_{bonded} is a sum of three terms corresponding to three types of atom movement, bond stretching, angle bending, and bond rotation

$$E_{bonded} = E_{bond\text{-}stretch} + E_{angle\text{-}bend} + E_{rotate\text{-}along\text{-}bond} \tag{2.11}$$

The first term of the above equation is a harmonic potential representing the interaction between two atoms bonded with a covalent bond (**Figure 2.1a**). The energy of the bond is a function of the displacement from the ideal bond length, b_0. K_b is the force constant that determines the strength of the bond. Both the ideal bond length and the force constant are specific for each pair of bound atoms

$$E_{bond\text{-}stretch} = \sum_{1,2\ pairs} K_b(b - b_0)^2 \tag{2.12}$$

The second term in equation 2.11 is also a harmonic potential, which refers to the angle deviation of a bond angle θ from the ideal value θ_0 (**Figure 2.1b**). Values θ_0 and K_θ vary depending on the chemical type of each atom comprising the angle

$$E_{bond\text{-}bend} = \sum_{angles} K_\theta (\theta - \theta_0)^2 \tag{2.13}$$

The last term in equation 2.11 calculates the potential energy of the system as a function of the rotations of dihedral angles (**Figure 2.1c**). This potential is periodic, it describes the steric barriers between atoms separated by 3 covalent bonds and is frequently expressed as a cosine function

$$E_{rotate\text{-}along\text{-}bond} = \sum_{1\text{-}4\ pairs} K_\varphi (1 - \cos(n\varphi)) \tag{2.14}$$

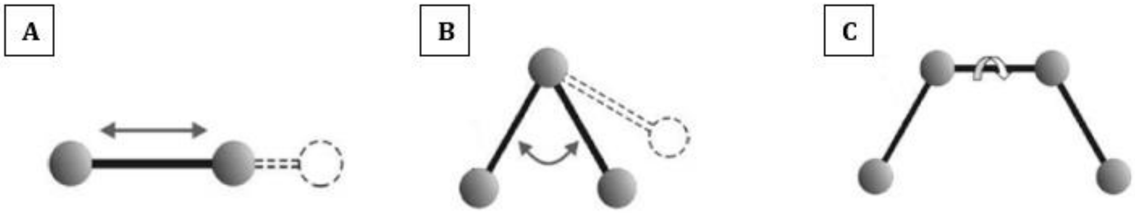


Figure 2.1: Schematic representations of three types of molecular vibrations. Bonded interactions include: (A) covalent bond stretching, (B) angle bending and (C) rotation around bonds. (adapted without permission from *Doh & Lee, Computers & Structures, 2016*)

The second term in equation 2.10 is the energy term that represents the nonbonded interactions, and it has two components, the van der Waals interaction energy, and the electrostatic interaction energy

$$E_{non\text{-}bonded} = E_{van\text{-}der\text{-}Waals} + E_{electrostatic} \tag{2.15}$$

The van der Waals interaction between two atoms is the result of an equilibrium between repulsive and attractive forces. The van der Waals interaction is modeled using the Lennard-Jones potential

$$E_{van\text{-}der\text{-}Waals} = \sum_{\substack{nonbonded \\ pairs}} \left(\frac{A_{ik}}{r_{ik}^{12}} - \frac{C_{ik}}{r_{ik}^{6}} \right) \tag{2.16}$$

where A and C are atom-dependent constants. The possibility of interaction between two atoms increases as the distance between them decreases. There is a specific distance, called the equilibrium distance in which the potential energy reaches a minimum value. If the distance between the two atoms becomes shorter than the equilibrium distance, repulsive forces become dominant, whereas if the distance increases, then attractive forces become dominant.

The second component of equation 2.15 is the electrostatic interaction energy and is represented by Coulomb's potential

$$E_{electrostatic} = \sum_{\substack{nonbonded \\ pairs}} \frac{q_i q_k}{D r_{ik}} \tag{2.17}$$

where D is the effective dielectric constant and r is the distance between two atoms having charges q_i and q_k[51].

2.5 The Role of Solvent in Molecular Dynamics Simulations

The use of solvent in MD simulations has a great impact on the structure, dynamics, and thermodynamics of biological molecules. One of the most crucial properties of the solvent is the screening of electrostatic interactions[51]. In the present project, we studied the folding process of peptide T in a Dimethyl Sulfoxide (DMSO) solution. DMSO is thought to solvate backbone NHs very strongly, favor extended conformers[38], and usually, reduce, or even prevent the formation of stable secondary structure conformations[60]. However, in some cases, e.g. peptide T, DMSO favors folded conformations. Also, the viscosity of the DMSO medium is higher than water and as a result, it can mimic at least one of the physicochemical features of membranes and cytoplasm[38].

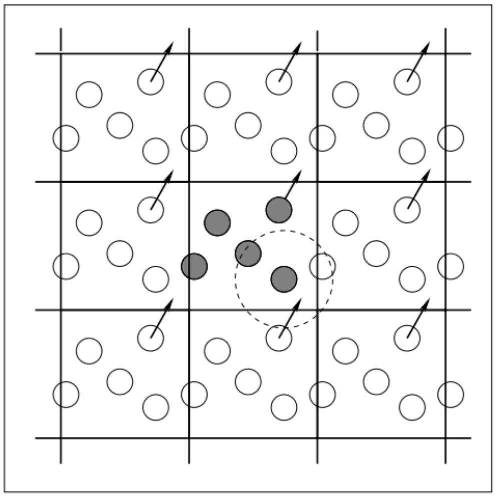


Figure 2.2: Periodic boundary conditions. The box in the center is the primary box. (reproduced without permission from *Attig et. al., NIC series, 2004*)

There are two main ways to incorporate solvent effects in an MD simulation, the implicit solvent models, and the explicit solvent models. In the implicit treatment of the solvent, an effective dielectric constant is included in the electrostatic term of the potential energy function. In the explicit treatment of the solvent, solvent molecules are included in the simulation. In this method, solvent boundary conditions must be imposed to prevent the diffusion of solvent molecules away from the protein and also, to allow calculation of macroscopic properties of the system. The most common treatment of the boundary is the periodic boundary conditions method (**Figure 2.2**). Periodic boundary conditions allow a simulation to be performed using a comparatively small number of particles so that the particles experience forces as if they were in a bulk solution. According to this method, the molecule under study is placed in the central box, defined as the primary box. The primary box is surrounded by eight neighboring boxes[51]. Every atom can interact with its neighbors and thus, if an atom moves out of the primary simulation box, an image particle enters the primary box, replacing it[52].

2.6 AMBER

Amber is a collection of programs applied to perform and analyze MD simulations. It also refers to a series of classical molecular mechanics force fields, used for biomolecular simulations. The most commonly used versions of the Amber force field are the ff94, ff99SB, ff03, and GAFF[61]. The ff94 version has been the most widely used with the AMBER suite of programs since its publication. This version introduced the set of parameters needed for all-atom protein simulations. Special features of the ff94 version include explicit use of all hydrogen atoms, fixed partial charges on atom centers, no specific functional form for hydrogen bonding, and dihedral parameters suitable for relative quantum-mechanical (QM) energies of alternative rotamers of small molecules. Although ff994 has been extensively used, certain limitations were reported, such as the over-stabilization of α-helices. The ff96 and ff99 versions aimed to improve the parameters used for the calculation of φ, ψ dihedral angles. A new parameter set, denoted as ff99SB replaced the existing parameters for backbone dihedrals angles. This parameter set accomplishes a better balance of secondary structure elements[62]. The improvement of the amino acid side-chain torsion potentials of the Amber ff99SB force field, lead to new force fields, ff99SB-ILDN[63], ff99SB-STAR[64], and ff99SB-STAR-ILDN[62][63][64].

3. Methods

3.1 Introduction

The study of the folding mechanism of peptide T was conducted through a Molecular Dynamics simulation using the NAMD program, a software developed for high-performance simulations of large biomolecular systems[65]. For the peptide's simulation, we used the AMBER force field, 99SB-STAR-ILDN[62][63][64].

Reliable simulations are computationally extremely intensive. In order to increase the computing power and reduce the computational cost, the use of parallel computers, which create a cluster, is usually preferred[66]. The simulation was carried out by Norma, a stateless Beowulf-class computing cluster based on the Caos NSA GNU/Linux distribution. Norma consists of 40 CPU cores, 46 GB of physical memory, and 6 GPGPUs distributed over 10 nodes which are based on Intel's Q6600 Kentsfield 2.4 GHz quad processors and are connected via a dedicated HP ProCurve 1800-24G gigabit ethernet switch. Each of the nine nodes offers four cores, 4 GB of physical memory, and two (gigabit) network interfaces. Only one node is based on Intel's i7 965 extreme which offers 6GB of physical memory plus a CUDA-capable GTX-295 card. Of the eight Q6600-based nodes, four are equipped with Nvidia GTX-460 GPU. The head node comes with four cores, eight GB of physical memory, 1.5 TB of storage in the form of a RAID-5 array of four disks, three (gigabit) network interfaces, and a Nnvidia GTX-260 GPU. Norma, which is located at the Department of Molecular Biology and Genetics of Democritus University of Thrace, is presently used almost exclusively for computational biology and crystallography projects of the Structural and Computational Biology group[67].

3.2 Starting molecular dynamics simulations with NAMD

In order to start an MD simulation, NAMD requires at least three files:

- A PDB file which contains the atomic coordinates and/or velocities for the system. PDB files are available through the PDB database (https://www.rcsb.org/), or they can be generated by the user. These files include information about the record type, atom ID, atom name, residue name, residue ID, x, y, z coordinates, occupancy, and temperature factor[68].
- A force field parameter file, which contains all the numerical constants required for the evaluation of forces and energies, given a structure file and atomic coordinates. The parameter file defines bond strengths, equilibrium lengths, etc. NAMD is able to

use different types of force fields (CHARMM, X-PLOR, AMBER, GROMACS)[68]. We used the parameter file of the AMBER 99SB-STAR-ILDN force field.

- A configuration file that specifies to NAMD how the simulation should be run. This file includes the dynamic options and values that NAMD should use, such as the number of timesteps to perform, initial coordinates, etc. It also determines what features are active or inactive and how long the simulation should continue[68][69]. A detailed description of the steps of our simulation is presented in the following section.

3.3 System preparation and simulation protocol

The first step in conducting the simulation was the preparation of the system. The PDB file for the initial fully extended state of peptide T was generated. This step was followed by the solvation and ionization of the system, which was performed using the program LEAP from the AMBER tools distribution. The simulation was conducted using periodic boundary conditions with a cubic unit cell sufficiently large to guarantee a minimum separation between the neighboring cells of at least 16 Å. We studied the dynamics of the folding simulation of peptide T using a mixed DMSO/water system. The adaptive tempering method was used, and the temperature ranged between 280 and 380K.

Prior to the start of the simulation, an energy minimization step was conducted in order to remove any strong Van der Waals interactions, which could cause structural distortion and as a result, lead to an unstable simulation. An entire box of solvent was then added onto the peptide and those solvent molecules that overlapped with the peptide were removed. Then the heating phase followed during which initial velocities at a low temperature were assigned to each atom of the system and the simulation begins. Periodically, new velocities were assigned at a slightly higher temperature and the simulation continues. This procedure is repeated until the preferred temperature is reached. When the desired temperature is reached, the simulation continues and during this phase, several properties are examined, such as the pressure, the structure, the temperature, and the energy. During the equilibration phase, the simulation is run until these properties become stable with respect to time. If there is a significant increase or decrease in temperature, the velocities can be scaled accordingly, so that the temperature returns close to its desired value[51]. The final step of the simulation is the production phase, during which the simulation is run for the time length required, which can range from several hundred ps to ns or more. During this step, system coordinates at different times are stored in the form of trajectories and are then used for different calculations (calculation of the mean energy, root mean square fluctuations between structures, etc.)[70].

Our system was first energy minimized (**Appendix, A1**) for 2000 conjugate gradient steps and then the temperature was increased (Appendix, A1) with a ΔT step of 20K until the final desired temperature of 320K. Subsequently, the system was equilibrated under constant temperature and pressure (NpT conditions) until the volume was equilibrated. The temperature and pressure were controlled using the Nosé-Hoover Langevin dynamics and Langevin piston barostat control methods, as implemented by the NAMD program. For the production phase (**Appendix, A2**), the Verlet-I multiple-step integration algorithm was used. The inner timestep was 2.5 fs, with nonbonded interactions being calculated every one step. The long-range electrostatic interactions were calculated every two timesteps, using the Particle Mesh Ewald method (PME). The cutoff for the Van der Waals interactions was set at 8 Å and the SHAKE algorithm was used to restrict all bonds involving hydrogen atoms.

The trajectory was obtained by saving the atomic coordinates every 0.8 ps. The simulation had a total duration of 4.04 μs and resulted in 5056000 frames.

3.4 Trajectory analysis and Programming Languages

The programs CARMA[71] and its GUI program GRCARMA[72] along with custom scripts have been used for the majority of our analyses, including the removal of overall rotations/translations, calculation of *φ, ψ* dihedral angles, calculation of RMSD's from a chosen reference structure, dihedral space principal component analysis (dPCA) and corresponding cluster analysis, calculation of average structures and production of PDB files from the trajectories. CARMA requires as input two files, a DCD and a Protein Structure File (PSF) file. The DCD file is the trajectory file, a binary file that contains sets of coordinates for the system. Each set of coordinates corresponds to one frame in simulation time[68]. The PSF file includes all of the molecule-specific information needed to apply a particular force field to a molecular system (atoms, bonds, angles, dihedrals, impropers, etc.)[73]. Secondary structure assignments were calculated using the program STRIDE[74]. Other structural analyses were performed using the promotif[75] program. The molecular graphics representations and figure preparations were performed with VMD[76], a molecular visualization program that uses 3-D graphics and built-in scripting[77], CARMA, and WebLogo[78]. For the production of colour-coding scatter plots, which illustrate the conformational clusters obtained by the dPCA analysis we used the plotting program, plot[79]. Chemical shifts were calculated using the program SPARTA+[80].

In order to process our data, analyze our results, and compare the experimental and simulation-derived chemical shifts, we used the Perl (Practical Extraction and Report Language) programming language. Perl is a high-level, general-purpose, interpreted, dynamic programming language. It was originally developed by Larry Wall in 1987[81]. The R statistical package, an open-source programming language environment suitable for statistical computing and graphics[82], was also used for plotting part of our data.

4. Results

4.1 Introduction

In this section of the thesis, we are going to focus on the analysis of the trajectory derived from the MD simulation we performed on peptide T. The overall workflow that we followed for this project is presented below:

- Secondary structure analysis using the algorithm STRIDE[74] and the logo-generating program WebLogo[78].
- Further structural analysis of turns and helices using the promotif[75] program.
- Construction of free energy landscapes of peptide T via dihedral angle Principal Component Analysis and cluster isolation.
- Association of high-density peaks with distinct peptide conformers, visualization of this structural information with schematic diagrams of representative and superposition structures.
- Application of Good-Turing[83] statistics to ascertain the extent of sampling and statistical significance.
- Calculation of chemical shifts and comparison with the experimental data.

4.2 Secondary structure analysis

The assignment of secondary structural elements is a crucial step in the determination of three-dimensional protein structures. There are several secondary structure assignment methods, which use different approaches and produce different assignments. Examples of these approaches include detection of patterns in inter-C^{α} distances, analysis of bond angles and lengths between consecutive C^{α} atoms, analysis of hydrogen bonding patterns, comparison of interatomic distance matrices of structural fragments with reference distances.

Intending to identify the peptide’s basic structural characteristics, we performed a secondary structure analysis using the algorithm STRIDE (secondary STRuctural IDEntification). STRIDE is an automated algorithm for protein secondary structure assignment from atomic coordinates, which is based on both hydrogen bond energy (E_{hb}) and statistically derived backbone torsional angle information[74]. The color coding used for the STRIDE-derived per residue secondary structure assignments **(Figure 4.1A)** is the following: pink for α-helices, purple for 3_{10}-helices, cyan for turns, and white for random coil. For a better understanding of the peptide’s secondary structure preferences, we produced two WebLogo graphs **(Figure 4.1B, Figure 4.1C)**. WebLogo is a program that generates sequence logos. Each logo consists of stacks of letters, one for each position in the sequence, and the height of the symbols within the stack represents the relative frequency of the corresponding amino acid at that specific position[78]. The first WebLogo graph **(Figure 4.1B)** is a representation of the per residue SRIDE-derived secondary structure assignments and corresponds to all the frames of the simulation, whereas the second WebLogo graph **(Figure 4.1C)** to the frames with an adaptive tempering temperature of less than 320 K and as a result represents more stable conformations. The isolation of stable conformers was possible because the simulation was performed using the adaptive tempering method which automatically adjusts the thermostat according to the energy of the system. The symbols used correspond to α-helices (H), 3_{10}-helices (G), turns (T), and random coils (C).

According to the results obtained from STRIDE and WebLogo*,* we can observe that peptide T is highly flexible and that the majority of residues are being assigned to turn or coil states. Assignments to helical structures (α-helices and 3_{10}-helices) are very rare, with only some minor occurrences. When we observe more closely the WebLogo diagrams, we can see that there are no considerable differences between them and that the first and last residues are quite flexible. This behavior is expected for such a short peptide. Residues 3-5 tend to form mostly turns, while we can identify some minor occurrences of coil, 3_{10}-helical, and even α-helical structures.

The experimental data of Picone et.al. suggested that peptide T could adopt fairly stable conformations and proposed that a helical segment (either α-helical or 3_{10}-helical) could be present, but the most tenable hypothesis was the one of a 5-8 β-turn[49]. These are the main structural characteristics observed in our analysis as well, but the difference between those two analyses is that our results suggest a significant degree of flexibility in the system.

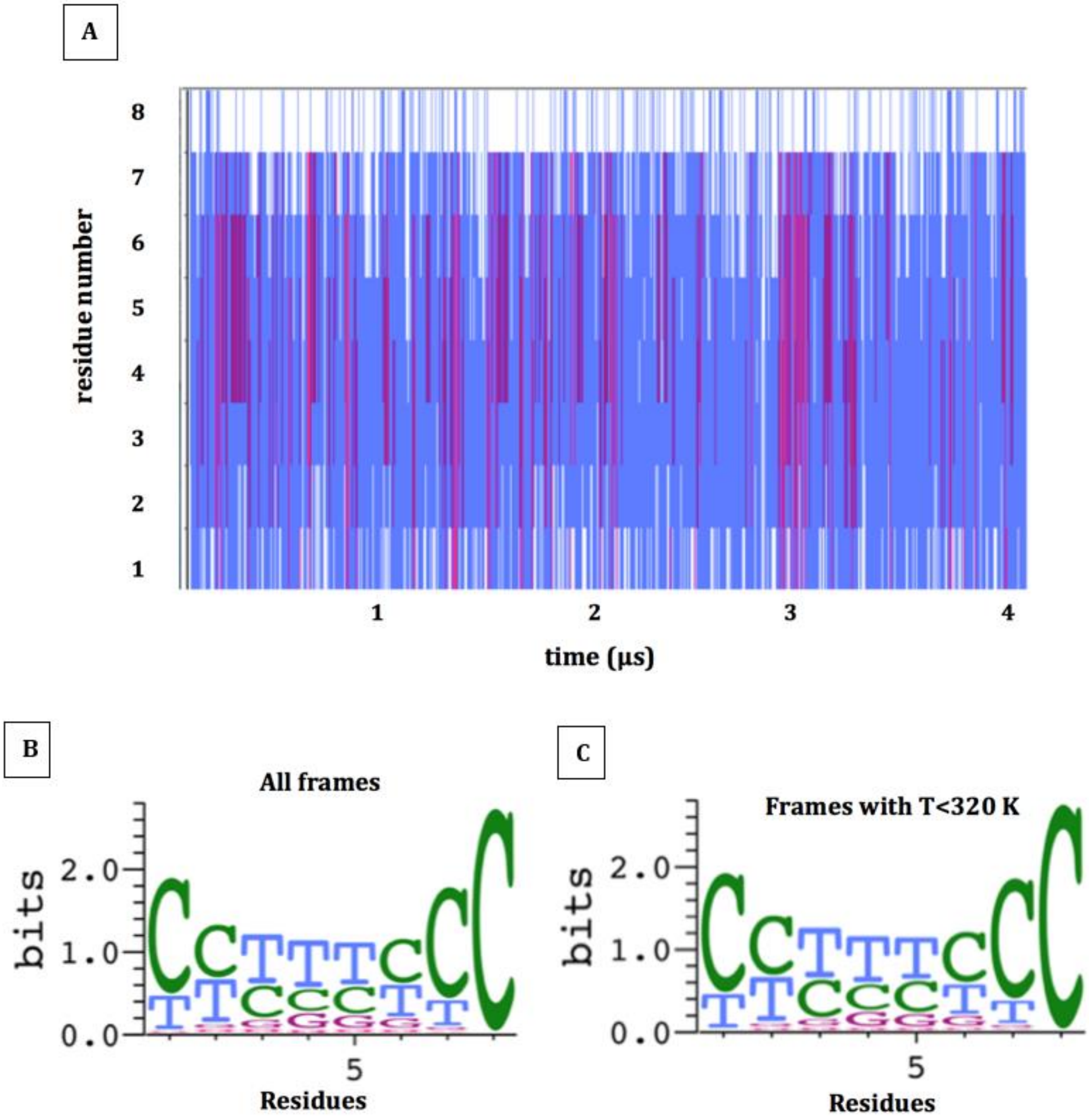


Figure 4.1: Secondary structure analysis overview. (A) Secondary structure diagram produced by the program STRIDE. The color coding is pink for α-helices, purple for 3_{10}-helices, cyan for turns and white for random coil. (B) The WebLogo-derived representations of the per residue STRIDE-derived secondary structure assignments corresponding to all the frames of the simulation. (C) The WebLogo-derived representations of the per residue STRIDE-derived secondary structure assignments corresponding to the frames of the simulation with an adaptive tempering temperature of less than 320 K. For (B) and (C) each letter corresponds to a different secondary structure element: H to α-helices, G to 3_{10}-helices, T to turns and C to random coil.

4.3 Structural analysis of turns and helices

In order to further analyze the main structural characteristics (turns and helices) of peptide T, we used the promotif program. The program analyzes a protein coordinate file and provides information about the structural motifs present in the protein. This program provides details about the following structural features in proteins: secondary structure, β, and γ turns, helical geometry and interactions, β-strands and β-sheet topology, β-bulges, β-hairpins, β-α-β units, ψ-loops, disulfide bridges, and main-chain hydrogen bonding patterns. Promotif creates postscript files for each type of motif in the protein, and a summary page, which gives a short description of each motif.

First, we analyzed the β-turns, which are defined as four consecutive residues (i, i+1, i+2, i+3) where the distance between the C_α atoms of residues i and i+3 is less than 7 Å and the central two residues are not helical. β-turns are organized in different categories according to the φ, ψ angles of residues i+1, and i+2. The ideal angles for each β-turn category are shown in **Table I**. The φ, ψ angles were allowed to vary by ±30^o from these ideal values, with one angle being allowed to deviate by 40^o[75].

Turn type ***β*-turns**	φ_{i+1}	ψ_{i+1}	φ_{i+2}	ψ_{i+2}
I	-60^o	-30^o	-90^o	0^o
II	-60^o	120^o	80^o	0^o
VIII	-60^o	-30^o	-120^o	120^o
I’	60^o	30^o	90^o	0^o
II’	60^o	-120^o	-80^o	0^o
VIa1^a	-60^o	120^o	-90^o	0^o
VIa2^a	-120^o	120^o	-60^o	0^o
VIba	-135^o	135^o	-75^o	160^o
IV	*Turns excluded from above categories*			

Table I: Ideal φ, ψ dihedral angles for the nine categories of β-turns. VIa1^a and VIa2^a require *cis*-proline at position *i*+2, therefore this turn type is not being studied in this project.

We studied the presence and the corresponding frequency of each β-turn type, using the whole trajectory. The table below (**Table II**) shows the frequency of each β-turn class for every possible sequential amino acid tetrad. The five different possible combinations were the following: a. 1-Ala-2-Ser-3-Thr-4-Thr, b. 2-Ser-3-Thr-4-Thr-5-Thr, c. 3-Thr-4-Thr-5-Thr-6-Asn, d. 4-Thr-5-Thr-6-Asn-7-Tyr, e. 5-Thr-6-Asn-7-Tyr-8-Thr.

Residues	I	II	VIII	I'	II'	IV
1 A 2 S 3 T 4 T	11.86%	0.19%	1%	0.17%	0.02%	7.17%
2 S 3 T 4 T 5 T	14.83%	0.2%	0.9%	0.37%	0.02%	12.01%
3 T 4 T 5 T 6 N	11.27%	-	0.51%	-	0.03%	11.81%
4 T 5 T 6 N 7 Y	5.93%	0.0001%	0.97%	-	0.005%	6.7%
5 T 6 N 7 Y 8 T	4.41%	0.12%	0.68%	0.05%	0.03%	5.72%

Table II: The frequency of each β-turn type for every possible sequential amino acid tetrad.

As it can be seen, the most preferred β-turn types are type I and IV, while β-turns type II, VIII, I' and II' are not so frequent. According to the above results, the most prominent β-turn type for the amino acid sequences: 1-Ala-2-Ser-3-Thr-4-Thr and 2-Ser-3-Thr-4-Thr-5-Thr is a type I, while the second most preferred β-turn type is type IV. For the sequences 4-Thr-5-Thr-6-Asn-7-Tyr and 5-Thr-6-Asn-7-Tyr-8-Thr, the most preferred β-turn type is type IV, followed by type I.

The next step was to analyze the second most prominent, according to our calculations, secondary structure element, helices. Promotif generates a table, which gives basic information about each type of helix recognized by the secondary structure assignment program, such as the helix number, the start, and end residue, helix type, the number of residues and the amino acid sequence, helix length and unit rise (both in Å), the number of residues per turn, the helix pitch in Å, and a measure of the deviation of the helix geometry from an ideal helix (in degrees)[75]. The table below (**Table III**) shows the frequency of each helix type for every possible combination of sequential amino acid residues.

	3_{10}-helix	α-helix	π helix
1 A 3	-	-	-
2 A 4	3.44%	-	-
3 A 5	3.34%	-	-
4 A 6	5.84%	-	-
5 A 7	1.19%	-	-
6 A 8	-	-	-
1 A 4	-	-	-
2 A 5	1.19%	3.11%	-
3 A 6	1.5%	1.64%	-
4 A 7	1.04%	3.52%	-
5 A 8	-	-	-
1 A 5	-	-	-
2 A 6	1.09%	0.59%	0.01%
3 A 7	0.37%	0.64%	0.01%
4 A 8	-	-	-
1 A 6	-	-	-
2 A 7	0.32%	0.51%	0.0002%
3 A 8	-	-	-
1 A 7	-	-	-
2 A 8	-	-	-
1 A 8	-	-	-

Table III: The percentages for each type of helix for every possible combination of sequential amino acid residues.

According to the results presented in **Table III**, the most preferred type of helix is 3_{10}-helix, followed by α-helix, while π-helix is extremely rare with low percentages. A comparison between the results obtained from tables II and III clearly shows a preference for β-turns rather than helices. This result is in agreement with our previous results obtained from the secondary structure analysis, which are presented in Figure 4.1. Also, it is important to mention that according to our calculations, the most preferred conformation for the amino acid sequence 5-Thr-6-Asn-7-Tyr-8-Thr is a β-turn type IV (5.72%), followed by a β-turn type I (4.41%), while no helical conformations were observed for this combination and the 4-8 one. This observation agrees with the experimental conclusions, where it is stated that the most prominent conformation is a 5-8 β-turn rather than a 4-8 helical segment[49]. But unlike our calculations, the experimental results state that the most likely cyclic structure is a type I β-turn[49] rather than a type IV.

4.4 Principal Component Analysis and Clustering

Secondary structure analyses enabled us to recognize the basic characteristics of the simulation. In this part of the project, we intend to place our observations in a more structurally oriented framework, by identifying the most prominent peptide conformations.

Biomolecular processes such as molecular recognition, protein folding, and aggregation can be explained in terms of the molecule's free energy

$$\Delta G(r) = -\, k_B T[\ln P(r) - \ln P_{\max}] \tag{4.1}$$

where P is the probability distribution of the molecular system along some, usually multidimensional, coordinate r and $P_{\max}$ denotes its maximum, which is subtracted to ensure that ΔG=0 for the lowest free energy minimum. The resulting free energy landscape is vital for gaining an understanding of protein folding[84].

Principal Component Analysis (PCA), also known as quasiharmonic analysis or essential dynamics method, is a way to systematically reduce the dimensionality of a complex system. The approach is based on the covariance matrix, which provides information related to the two-point correlations of the system. PCA represents a linear transformation that diagonalizes the covariance matrix, therefore removes the instantaneous linear correlations among the variables[84][85].

Dihedral angle Principal Component Analysis (dPCA) is based on the dihedral angles (φ_n, ψ_n) of the peptide backbone. In order to eliminate problems arising from the circularity of these variables, a transformation from the space of dihedral angles to a linear metric coordinate space (i.e., a vector space with a well-defined distance between any two points) using the trigonometric functions sin φ_n and cos φ_n is essential. The dPCA method is attractive because other internal coordinates such as bond lengths and bond angles usually do not experience changes of large amplitudes. As a result, the analysis starts with the relevant part of the dynamics, preventing unnecessary noise. Moreover, since dPCA is based on the backbone dihedral angles, it can easily distinguish between the kinetically well separated main conformational states of the peptide, such as the α_R helical and the β extended conformations[84][86].

Molecular dynamics simulation methods produce trajectories of atomic positions (and optionally velocities and energies) as a function of time and demonstrate the sampling of a molecule's energetically accessible conformational ensemble[87]. Due to the advances in computer speed and algorithm efficiency, MD simulations are sampling larger amounts of molecular and biomolecular conformations. Being able to sift these conformations qualitatively and quantitatively into groups is a demanding and important task[88]. Data-mining methods, like clustering, provide a way to group and understand the information in the trajectory[87].

Cluster analysis is a term applied to several techniques that aim to divide a set of objects into different groups, also called clusters so that objects within the same group have more similarities with each other than objects that are part of different groups. Cluster analysis is utilized when the dimensionality of the data prevents detailed visual examination. Multidimensional scaling techniques such as principal component analysis, principal coordinates analysis, or non-linear mapping can be used for visual examination of the data. Both principal component analysis and principal coordinates analysis can be utilized to reduce the dimensionality of the data, for instance when the number of variables surpasses the number of objects or when there are linear relationships among the variables. It is also important to note that cluster analysis can be applied to a subset of the largest principal components[89].

The free energy landscapes of the trajectory were constructed using GRCARMA and the dPCA method **(Figures 4.2, 4.3)**. Initially, we considered the dihedral angles of the entire peptide **(Figure 4.2)**, and afterward, we decided to limit the dPCA analysis to residues 5-8. Figure 4.2 and the top row of Figure 4.3 show the free energy (in kcal/mol) as a function of the first three principal components. High-density peaks, which are illustrated dark blue correspond to clusters of structures with similar principal component values, and as a result similar dihedral angles and backbone structures. The second row of representations in Figure 4.3 is a color-coding cluster representation made by the plot program and **Table IV** contains the populations of each cluster produced by this analysis.

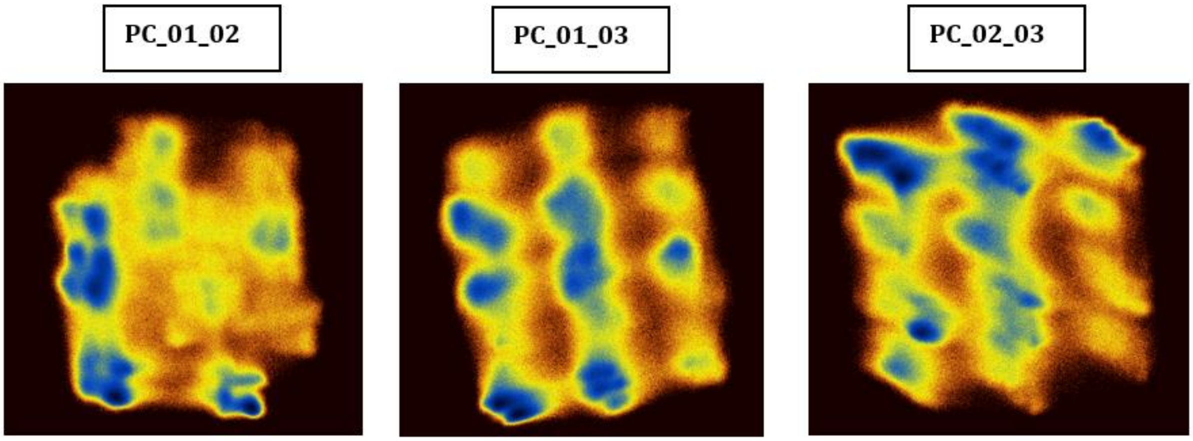


Figure 4.2: Two-dimensional representations of the free energy landscapes obtained by the dPCA method taking into consideration the torsion angles of the entire peptide.

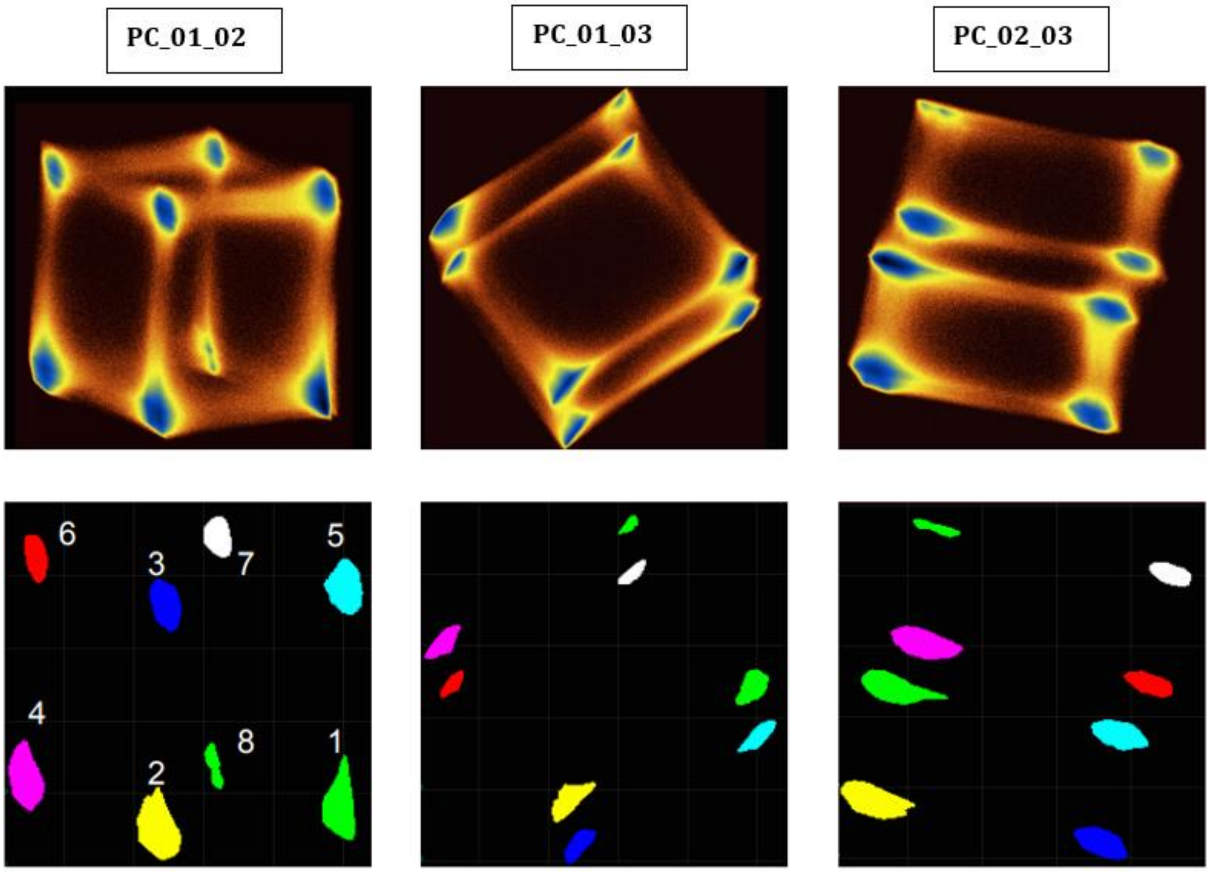


Figure 4.3: Two-dimensional representations of the free energy landscapes obtained by the dPCA method taking into consideration the torsion angles of residues 5-8 of peptide T. Top row: ΔG plots along the first three principal components of the trajectory (blue corresponds to high density). Second row: Color-coding panels depicting the conformational clusters obtained by the dPCA analysis of residues 5-8.

As demonstrated above, the energy landscapes in Figure 4.2 are rugged, while in Figure 4.3 we can observe that there is a limited number of peaks, this is indicative of a well-defined system with few free energy minima, that correspond to distinct conformational structures. The low number of prominent structures suggests that this segment of peptide T adopts more stable conformations, which correspond to distinct secondary structures with specific torsion angles and hydrogen bond patterns. These results are in agreement with the experimental conclusions, that residues 5-8 of peptide T adopt a stable conformation.

The table below (Table IV) displays the populations of each cluster obtained by the second dPCA analysis.

Cluster	Frames (out of 5056000)	Percentages
1	829017	16.4%
2	627294	12.41%
3	360755	7.14%
4	456049	9.02%
5	310782	6.15%
6	144079	2.85%
7	128426	2.54%
8	37878	0.75%

Table IV: The populations of the eight clusters produced by dPCA along with the percentage of clustered structures.

In total 2894280 out of 5056000 (57.24%) were included in clusters. The three main clusters are cluster 1 with 829017 frames out of 5056000 (16.4%), cluster 2 with 627294 configurations out of 5056000 (12.41%), and cluster 4 with 456049 frames out of 5056000 (9.02%).

Having acquired the distribution of the principal components from the previous analysis, it is feasible to relate high-density peaks with distinct peptide conformations. Average structures for each cluster were calculated and in order to identify the representative structures we selected the frame from the trajectory which had the lowest Root Mean Square Deviation (RMSD) from the corresponding average structure. RMSD is the most widely used quantitative measure of the similarity between two superimposed atomic coordinates[90]. The RMSD values are presented in Å and can be calculated using the following equation

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(r_i^x - r_i^y\right)^2} \tag{4.2}$$

where N is the number of atoms, i the current atom, r^x the target structure, and $r^y$the reference structure[91]. The lower the RMSD value is, the more similar the two structures. When there are no structural differences, the RMSD value is 0.0 Å. Generally, two structures share sufficient similarities when the RMSD value is less than 2.0 Å.

Intending to visualize more precisely the structural flexibility that is present in our system, we followed this process not only for the single most populated motif but for the five most populated motifs that were closer to the average structure. **Figure 4.4** shows schematic (cartoon) diagrams of these representative structures corresponding to the eight clusters of the trajectory. The color coding is blue for 3_{10}-helices, cyan for turns, and white for random coil. In the center of this figure is the log density projection of the trajectory on the first two principal components derived from dPCA. The marked peaks (1 to 8) are in a one-to-one correspondence with the structural diagrams at the periphery of the diagram. The numbers below each schematic structure representation are the relative percentages of each cluster.

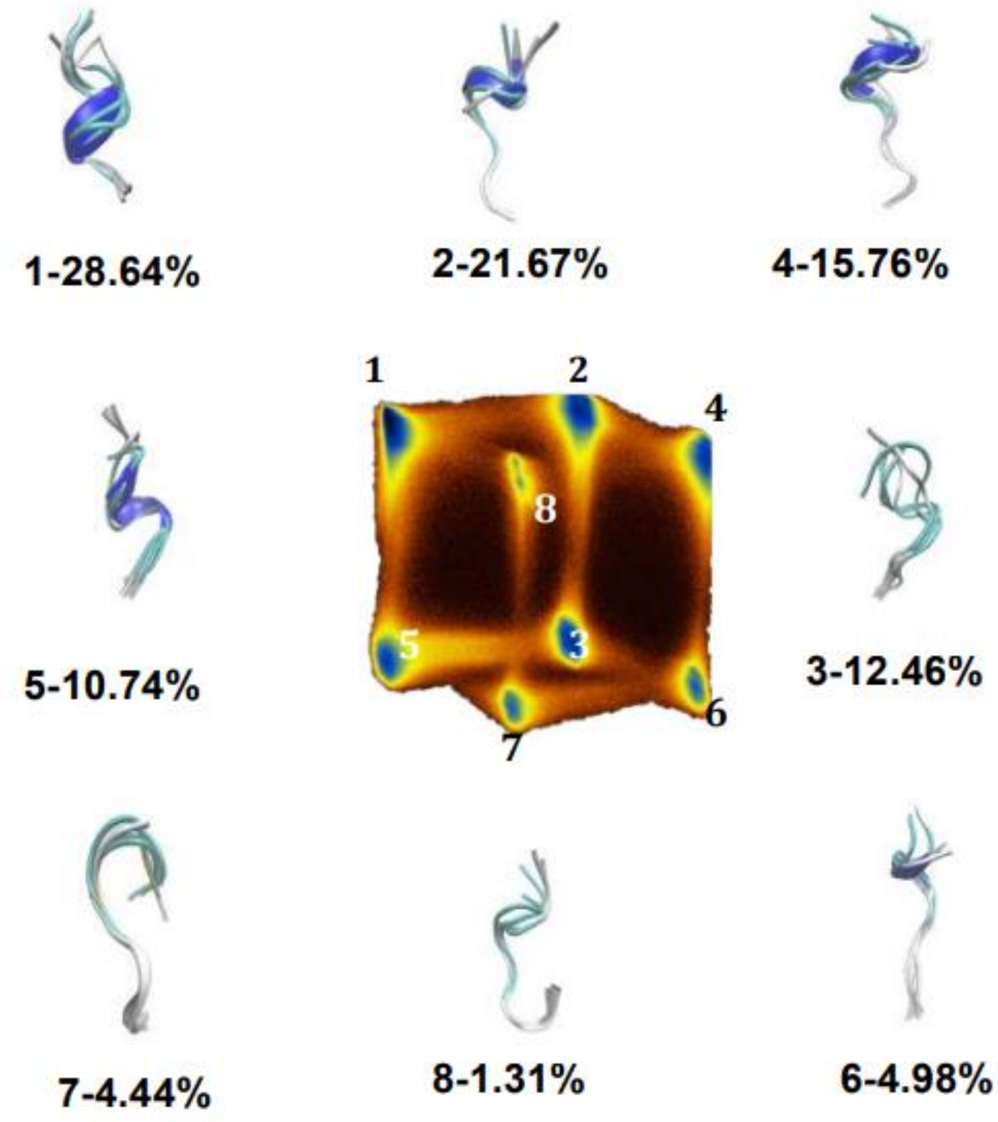


Figure 4.4: Clusters and their relative frequencies. The diagram in the center is the log density projection of the trajectory on the first two principal components. The clusters are marked with numbers 1-8. The structure schematics in the periphery are superpositions of the five representative structures that are closer to an average structure. The colour coding is blue for 3_{10}-helices, cyan for turns and white for random coil. In all diagrams, the N-terminus is toward the upper part, while the C-terminus toward the lower part of the figure. The percentages below the structures are the relative populations of the corresponding clusters. The structures have been placed so that the most populated clusters are presented at the top of the figure.

Overall, the main structural characteristics are turns and helices, as mentioned already in our previous analyses. Due to the increased kinetic frustration of the system, the representative structures differ between the clusters, while we can observe many coil conformations as well. The table below (**Table V**) presents the per residue SRIDE-derived secondary structure assignments of the five most populated motifs for each cluster. The symbols used correspond to 3_{10}-helices (G), turns (T), random coils (C), and β-bridge (B).

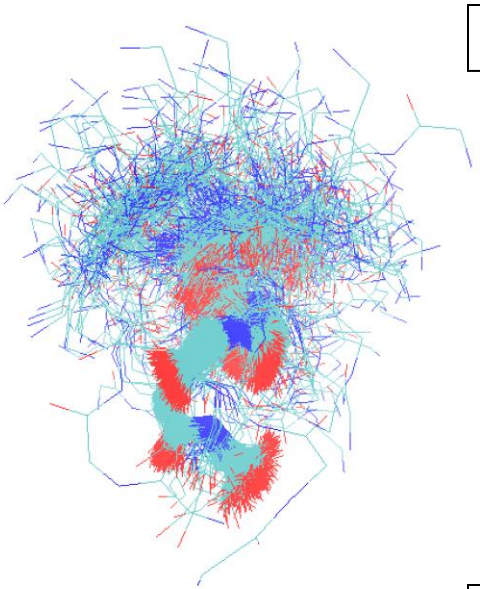
Most populated motifs	**Clusters**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1st	CCCGGGCC	CCCCCCCC	CTTTTTTCC	CCCCCCCC	CCTTTTTC	CCCCCCCC	CCCCCCCC	CCCCCCCC
2nd	CCTTTTTC	CTTTTTCC	CCCCCCCC	TTTTTCCC	TTTTTTTC	TTTTTCCC	CTTTTCCC	CTTTTCCC
3rd	TTTTTTTC	CCTTTTCC	CCTTTTCC	CTTTTCCC	CCCGGGCC	CTTTTCCC	TTTTTCCC	TTTTCCCC
4th	CCCTTTTC	TTTTTTCC	TTTTCCCC	TTTTCCCC	CTTTTTTC	TTTTCCCC	TTTTCCCC	TTTTTCCC
5th	TTTGGGCC	CGGGCCCC	TTTTTTCC	CGGGCCCC	CCGGGCCC	CGGGCCCC	CBTTBCCC	CCTTTTCC

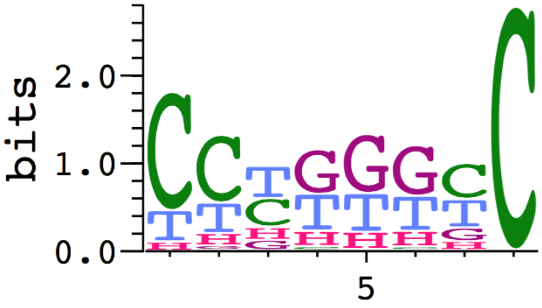
Table V: Per residue SRIDE-derived secondary structure assignments of the five most populated motifs for each cluster.

It is important to mention that the structure diagrams in Figure 4.4 are simplified representations and thus, do not sufficiently depict the actual amount of structural variability that is present in the clusters, nor the precise folding path that the peptide backbone follows. In order to represent more realistically the structural variability present in our system, we created the superposition figures presented in **Figure 4.5**. Figure 4.5 is an image from the superposition of 500 structures that belong to each cluster derived from the dPCA analysis. The backbone structures are shown as sticks and the color coding is cyan for

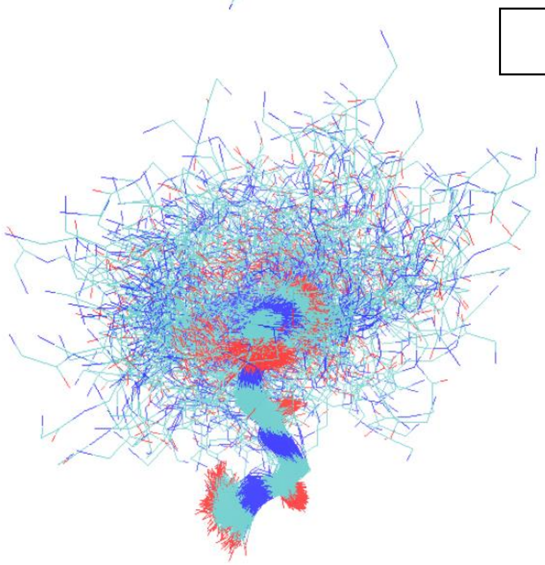
C atoms, blue for N atoms, and red for O atoms. Next to each superposition figure, there is a WebLogo diagram specific for each cluster. The 3D models were created in VMD and the sequence logos were created using WebLogo. The superpositions are complex and noisy and as a result, it is difficult to understand the structural content that is present in the clusters. Nevertheless, it is clear that the C-terminus forms more compact structures than the N-terminus.

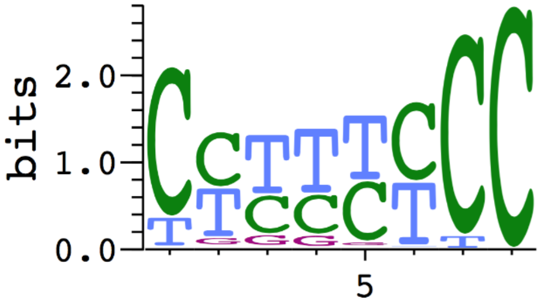
Cluster 1



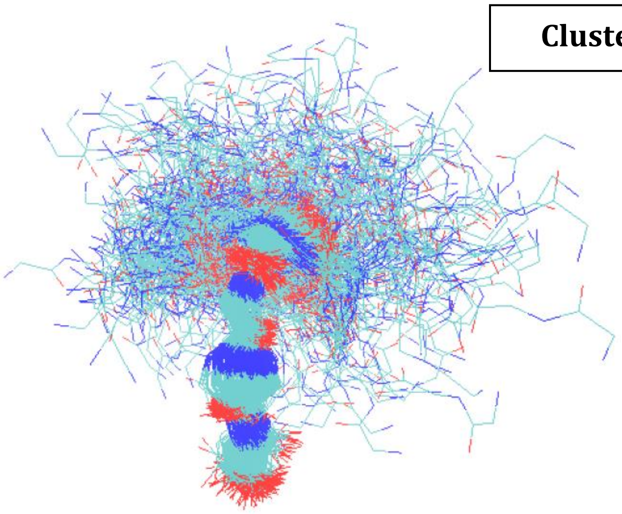


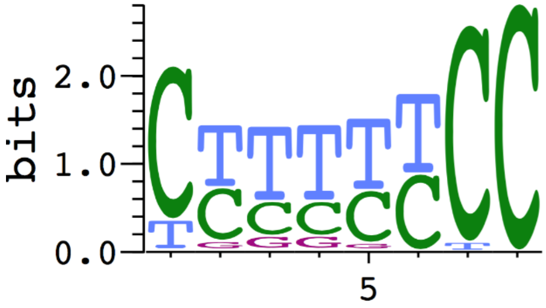
Cluster 2



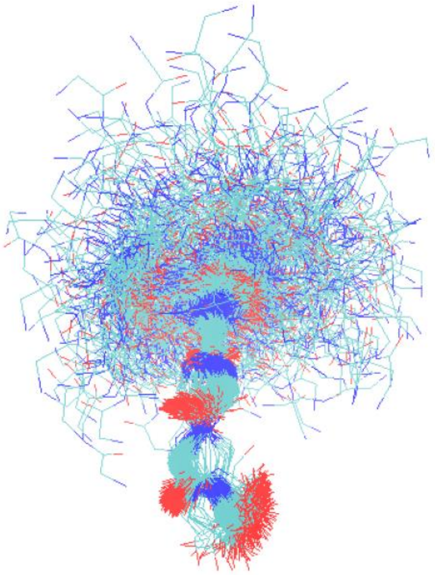


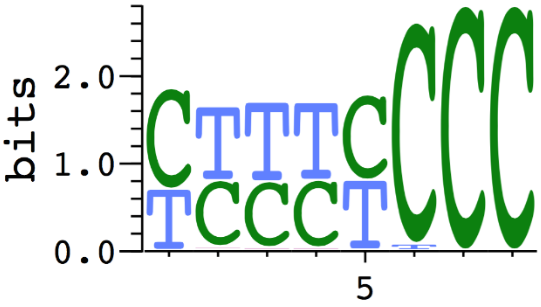
Cluster 3



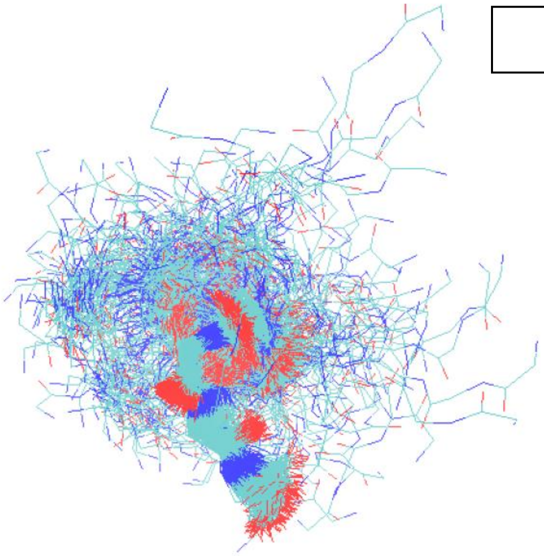


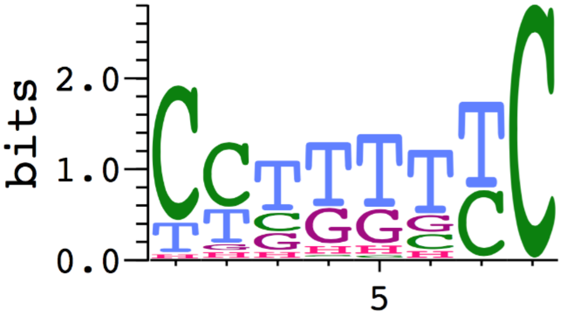
Cluster 4



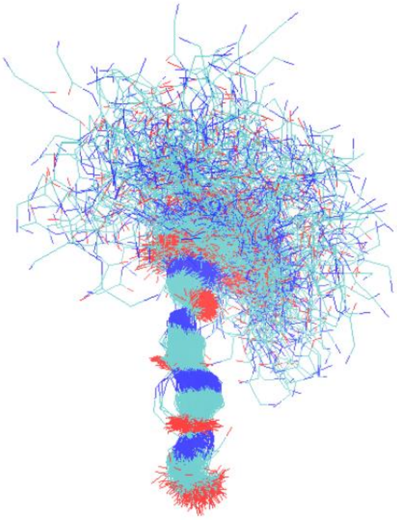


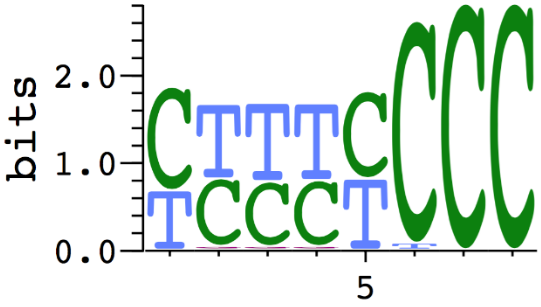
Cluster 5

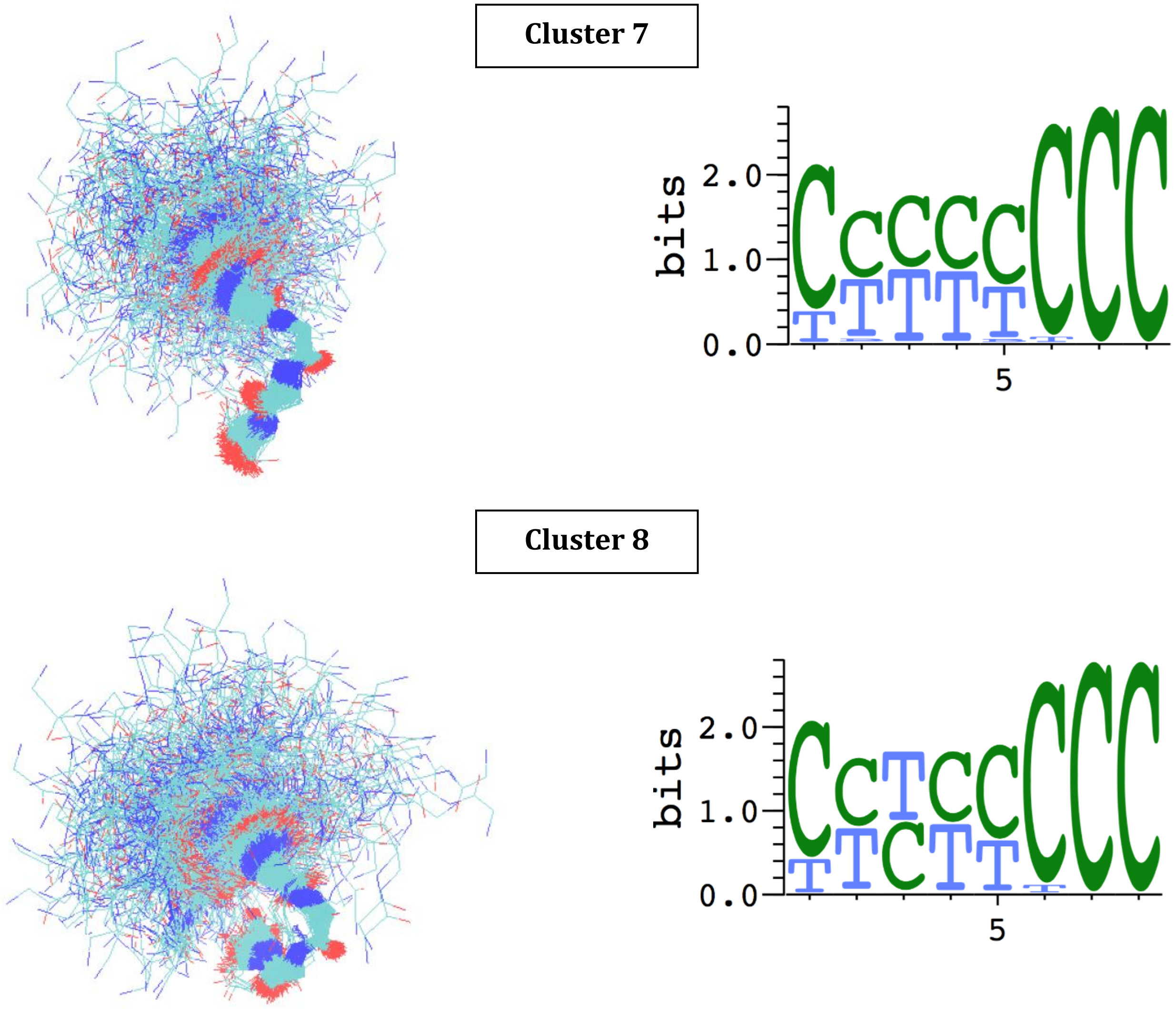




Cluster 6






Cluster 7

Cluster 8

Figure 4.5: Superposition of 500 structures (left) that belong to each cluster derived from the dPCA analysis. The color coding is cyan for C atoms, blue for N atoms and red for O atoms. WebLogo diagrams (right) of the per residue STRIDE-derived secondary structure assignments.

4.5 Extent of Sampling and Statistical Significance

According to our findings, peptide T is highly flexible, suggesting that a folding molecular dynamics simulation must visit a vast conformational space associated with the disordered state. Although we applied the adaptive tempering method, intending to increase the sampling competence of the simulation, it is highly unlikely to obtain a statistically significant sample. To ascertain whether our simulation was efficiently sampled we applied Good-Turing statistics. This method estimates the probability of unidentified conformations as a function of the RMSD from the conformations that have already been identified in the simulation[83]. We applied this method to our trajectory and our results are presented in the figure below (**Figure 4.6**). The black upper curve is the estimate obtained using the Cα atoms of all residues of the peptide and the blue curve is the estimate obtained using only residues 5-8.

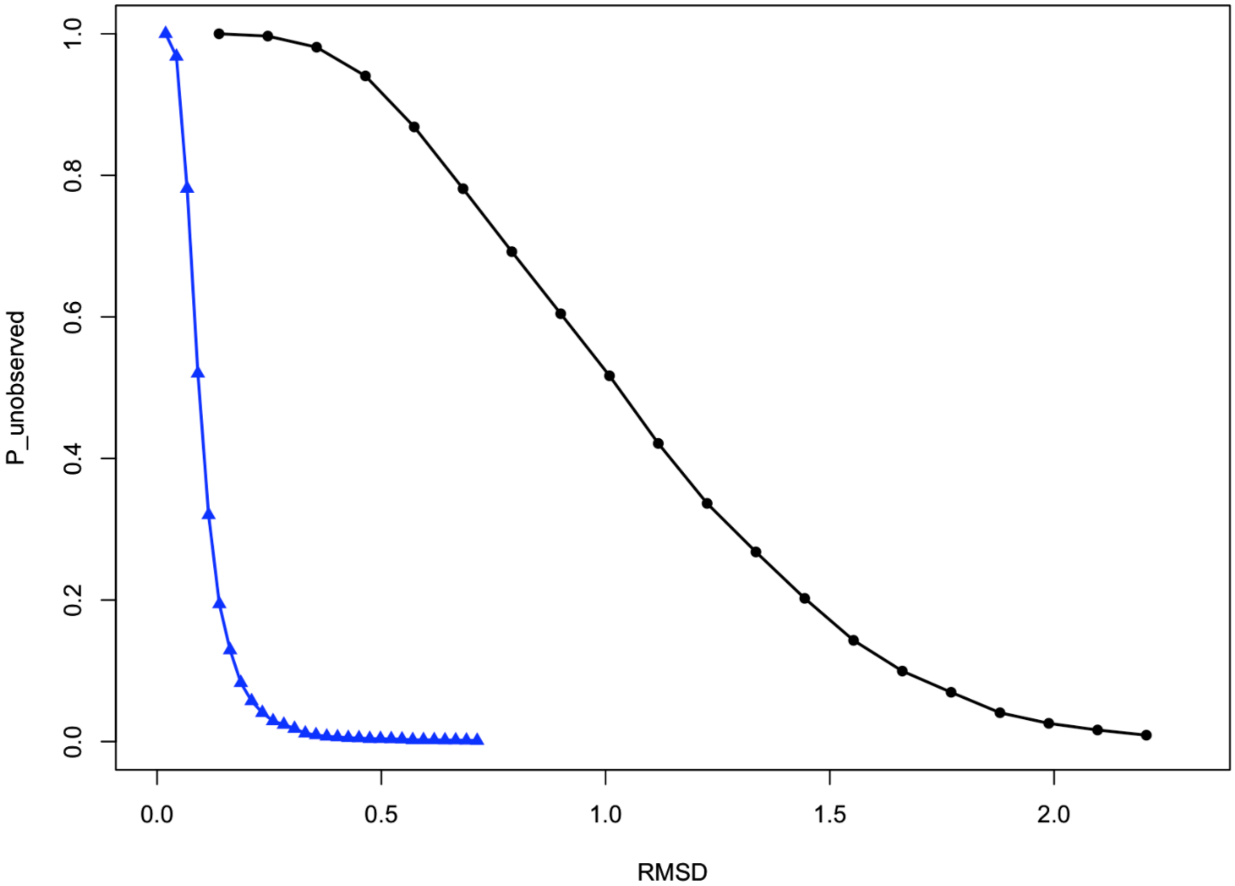


Figure 4.6: Extent of sampling and statistical significance. The black upper curve is the estimate obtained using the Cα atoms of all residues of the peptide and the blue curve is the estimate obtained using only residues 5-8.

We can observe that at low RMSDs, both these curves have high probability values and for gradually increasing RMSDs, the curves asymptotically approach low probability values. The results from the analysis performed using all residues, suggest that the most different structure we should expect to observe if we doubled the simulation time would differ by no more than approximately 2.0 +- 0.1 Angstrom (RMSD) from those already observed. For example (black curve in Figure 4.6), we would expect that on average one out of 25 previously unobserved structures ($P_{unobserved}$ = 0.04) would differ by an RMSD value of at least 1.8 Angstrom from the already observed structures.

The results from the analysis performed using only residues 5-8 (blue curve in Figure 4.6) suggest that the most different structure we should expect to observe if we doubled the simulation time would differ by no more than approximately 0.5 +- 0.07 Angstrom (RMSD) from those already observed. It can be seen that by limiting the residue selection to the amino acids comprising the C-terminal part of the peptide, the curve approaches low probability values faster, demonstrating better sampling. These results are in agreement with our previous results discussed in 4.4, that residues 5-8 correspond to more stable peptide conformers.

4.6 Chemical shifts

In this part of the project, we are going to make quantitative comparisons between the experimental results and the simulation-derived results. More specifically, we compared the experimentally determined chemical and secondary chemical shifts with those derived from the simulation via the application of the SPARTA+[80] program.

In organic chemistry, NMR chemical shifts have been used to recognize and define the covalent structure of small organic compounds for over 60 years[92][93]. Nuclear chemical shifts are powerful indicators of the structural types that biopolymers can adopt. The main goal of NMR experiments is the amplification of the resolution and sensitivity with which the chemical shift of a nucleus can be calculated[94]. Nowadays, it is feasible to use protein chemical shift data to distinguish protein secondary and super-secondary structure, to calculate backbone and side-chain torsion angles, to define residue-specific surface areas, to measure protein flexibility, to produce protein structure models, and to accurately predict protein structures[92]. Moreover, chemical shifts offer comprehensive information about the nature of hydrogen exchange dynamics, ionization and oxidation states, the ring current influence of aromatic residues, and hydrogen bonding interactions[94].

Each of the three types of secondary structure elements in proteins (α-helices, β-sheets, and random coils) corresponds to representative chemical shifts, for every atom in these 20 amino acids. The specific chemical shifts for amino acid residues in random coils or short unstructured polypeptides are formally known as “random coil” chemical shifts.

Random coil chemical shifts are possibly the most important "reference" shifts in protein NMR as they are crucial for the determination of secondary chemical shifts ($\Delta\delta$ shifts)[93]. Secondary shifts, also known as conformation-dependent shifts[95], are defined as the difference between the observed amino acid chemical shift (δ_{obs}) and the corresponding random coil (δ_{rc}) value:

$$\Delta\delta = \delta_{obs} - \delta_{rc.} \tag{4.3}$$

Secondary shifts are predominantly influenced by non-covalent interactions and contain the most dynamic information about proteins[95][93].

Protein chemical shifts have been found to follow certain nucleus-specific and residue-specific standards. More precisely, $^1H_\alpha$ chemical shifts tend to range from 3.5 to 5.5 ppm, while ^1HN chemical shifts range from 6.5 to 10.0 ppm. Except for glycine, when any of the remaining 19 amino acid residues are in helices, their $^1H_\alpha$ shifts are shifted upfield concerning their random coil values by an average of 0.30 ppm., while in β-sheets the $^1H_\alpha$ chemical shift values are shifted downfield by an average of 0.46 ppm. In terms of the ^1HN shifts, these seem to be much more sensitive to their environment than $^1H_\alpha$ shifts[93].

When performing an NMR experiment, it is important to consider the effects of the solvent of resonant nuclei on their resonance spectra[96]. Additionally, the understanding of medium effects may provide new tools for the determination of molecular structure. Solvent effects are described as the result of intermolecular forces. The significance of solvent effects to NMR derives from the relatively high concentrations needed, in comparison with other spectroscopic methods[97]. Regarding proton NMR, the majority of H chemical shifts are of protons bound to carbon atoms. Protons bound to other atoms tend to exhibit chemical shift alterations with solvent resulting from hydrogen bonding interactions[98]. The dependence of H chemical shifts on solvent has been thoroughly studied since the beginning of high-resolution proton NMR[99]. Buckingham *et al.*[100] described four interactions responsible for solvent effects, namely: hydrogen bonding, the anisotropy of the solvent molecules, polar and van der Waals effects

$$\Delta\delta = \delta_{HB} + \delta_{A} + \delta_{E} + \delta_{W} \tag{4.4}$$

The impact of each one of these contributions can vary significantly. A major interaction with solvent effects of up to 5 ppm occurs when hydrogen bonds form with protic solutes[99]. Some molecules exhibit anisotropy in their molecular susceptibilities because they behave as magnetic dipoles when subject to an external magnetic field and produce secondary fields at other molecules[96]. Solvent effects caused by the electric field of the polar solute molecule, have been analyzed using variations of the Onsager reaction field model. Finally, van der Waals forces are significant in gas-to-solvent shifts even for non-polar molecules in non-polar solvents[99].

Intending to obtain the experimental secondary chemical shifts, we used the experimental chemical shifts from the Picone *et al.* publication[49] and the ^1H random coil chemical shifts for peptides of sequence GGXAGG (where X is any of the 20 naturally occurring amino acids or the modified amino acid 4-hydroxyproline) measured in DMSO from the Tremblay and Banks publication[101]. For the calculation of the simulation-derived secondary chemical shifts, we first calculated the simulation-derived chemical shifts using the programs CARMA[71] and SPARTA+[80] through a perl script **(Appendix, A3)**, taking into account all 5056000 frames of our trajectory and the random coil chemical shifts from the Tremblay and Banks publication[101]. This perl script produces pdb files for each frame of the trajectory, the pdb files are read by SPARTA+ and then the chemical shifts, as well as the mean and the standard deviation values are calculated. Our results are presented in the tables below (**Table VI** and **Table VII**). SPARTA+ is a chemical shift prediction program that is based on artificial neural networking. The neural network is well-trained to create quantitative relations between chemical shifts and protein structures, including backbone and side-chain conformation, H-bonding, electric fields, and ring-current effects[80]. To our knowledge, there is no tested protein structure prediction package that accurately estimates H chemical shifts in organic solvents. Ideally, we would like to use a program that could be extended to provide a satisfactory prediction of the chemical shifts in DMSO.

Table VI shows the experimental and the simulation derived $^1H_\alpha$ and HN chemical shifts, along with the "reference", or random coil chemical shifts, and the standard deviation value (σ), which indicates the amount of dispersion of the values from the mean of the set of values. Table VII shows the experimental and simulation-derived $^1H_\alpha$ and HN secondary chemical shifts, which are denoted as $\Delta\delta$ experimental and $\Delta\delta$ simulation, respectively. To allow better visualization of our data, we used the R statistical package and created the bar charts shown in **Figures 4.7a-4.7d**, where it is easier to make residue-by-residue comparisons between the experimentally determined and simulation-derived secondary chemical shifts. **Figure 4.7a** shows the experimental HA secondary shifts and **Figure 4.7b** the HA secondary chemical shifts, which were derived from the molecular dynamics simulation. **Figures 4.7c** and **4.7d** show the experimentally determined and the simulation-derived HN secondary chemical shifts, respectively. The source code of the R scripts can be found in the Appendix (**Appendix, Script 1, Script 2, Script 3, Script 4**).

Residue number	Residue	Atom	Experimental chemical shift	Simulation chemical shift	Random coil chemical shift	σ
1	A	HA	3.81	4.3184	4.47	0.0564
2	S	HA	4.42	4.4790	4.50	0.2247
2	S	HN	8.69	8.4032	8.37	0.2750
3	T	HA	4.31	4.4231	4.41	0.1811
3	T	HN	7.90	8.1580	8.00	0.3477
4	T	HA	4.30	4.3827	4.41	0.2593
4	T	HN	7.85	8.1835	8.00	0.3099
5	T	HA	4.25	4.4052	4.41	0.1881
5	T	HN	7.69	8.1068	8.00	0.3016
6	N	HA	4.51	4.7351	4.74	0.1830
6	N	HN	8.02	8.2170	8.40	0.3514
7	Y	HA	4.37	4.6196	4.60	0.1891
7	Y	HN	8.00	8.1916	8.38	0.4360
8	T	HA	3.96	4.3220	4.41	0.1010
8	T	HN	7.54	8.1998	8.00	0.3078

Table VI: Experimental and simulation-derived HA and HN chemical shifts, along with the random coil and standard deviation values.

Residue number	Residue	Atom	$\Delta\delta$ experimental	$\Delta\delta$ simulation
1	A	HA	-0.66	-0.1516
2	S	HA	-0.08	-0.0210
2	S	HN	0.32	0.0332
3	T	HA	-0.1	0.0131
3	T	HN	-0.1	0.1580
4	T	HA	-0.11	-0.0273
4	T	HN	-0.15	0.1835
5	T	HA	-0.16	-0.0048
5	T	HN	-0.31	0.1068
6	N	HA	-0.23	-0.0049
6	N	HN	-0.38	-0.1830
7	Y	HA	-0.23	0.0196
7	Y	HN	-0.38	-0.1884
8	T	HA	-0.45	-0.0880
8	T	HN	-0.46	0.1998

Table VII: Experimental and simulation-derived HA and HN secondary chemical shifts.

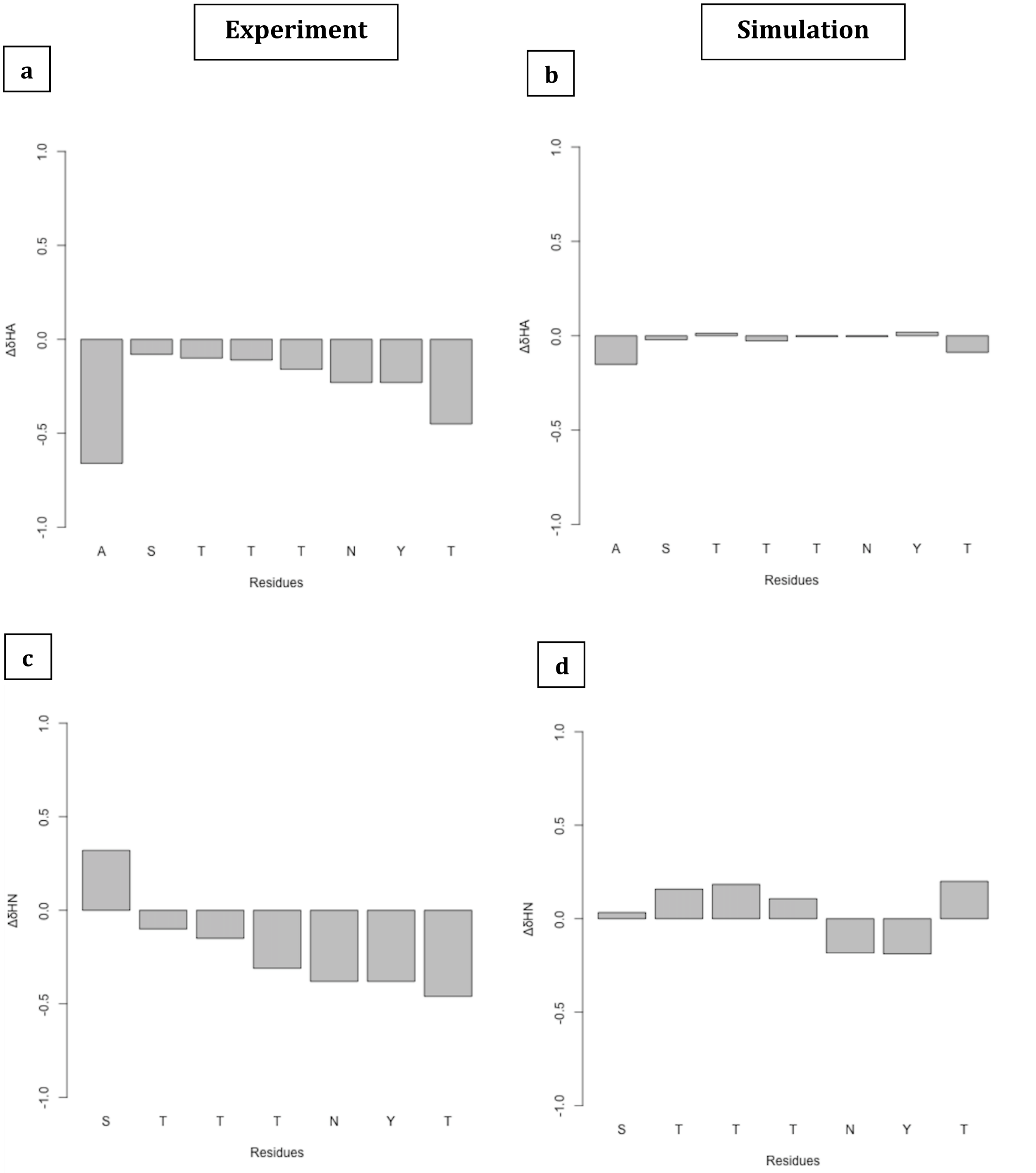


Figure 4.7: Per-residue comparisons between the experimental and simulation-derived secondary shifts. (a) Experimental HA secondary shifts. (b) Simulation-derived HA secondary shifts. (c) Experimental HN secondary shifts. (d) Simulation-derived HN secondary shifts.

At this point, in order to be able to evaluate our results and make quantitative comparisons between the experimentally determined and the simulation-derived chemical shifts, we used two statistical analyses: the reduced x^2 and the linear correlation coefficient. Both these analyses are used to define the relationship between two sets of data.

The reduced x^2 value can be calculated using the following formula:

$$x_{red}^2 = \frac{x^2}{v} = \frac{1}{v}\sum\frac{(O-E)^2}{\sigma^2} \tag{4.5}$$

where Σ is the sum, O is the observed value which corresponds to the simulation derived chemical and secondary chemical shifts, E is the expected value which corresponds to the experimentally determined chemical and secondary chemical shifts, σ^2 is the error variance of the observed values and v are the degrees of freedom. In general when:

- x_{red}^2 = 1, then the observed and expected values are in agreement with the distribution and there is a high degree of correlation.
- x_{red}^2 > 1, then either we do not have a complete correlation between the data, or the distribution values have been underrated.
- x_{red}^2 < 1, then the model is “over-fitting” the data: either the model is improperly fitting noise, or the error variance has been overestimated[102].

The correlation coefficient *r,* or Pearson product correlation coefficient values can be calculated using the following formula:

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{4.6}$$

where Σ is the sum, x_i is the value of the first set of data for position i, y_i is the value of the second set of data for position i, $\bar{x}$ is the mean of *x* values and $\bar{y}$ the mean of *y* values. The correlation coefficient is used when we want to identify whether the variations in the observed values of one quantity *y* are correlated with the variations in the measured values of another quantity *x*. The value of *r* is such that -1 ≤ *r* ≤ 1. The positive and negative values correspond to positive and negative linear correlations, respectively. In general when:

- *r* ≅ +1, then x and y have a strong positive linear correlation, an *r* value of exactly +1 indicates a perfect positive fit.
- *r* ≅ -1, then x and y have a strong negative linear correlation, an *r* value of exactly -1 indicates a perfect negative fit.
- *r* ≅ 0, then there is a random, nonlinear relationship between the two variables.

The reduced x^2 and correlation coefficient values were calculated using two perl scripts (**Appendix, Script 5, Script 6**). We performed these two statical analyses for all secondary chemical shifts, for HA secondary shifts, HN secondary shifts, HA chemical shifts, and HN chemical shifts. Our results are presented in the table below (**Table VIII**).

	$\Delta\delta$	**HA**	**HN**	**$\Delta\delta$HA**	**$\Delta\delta$HN**
x^2	7.7434	14.0855	1.6348	14.0855	1.6348
r	0.3428	0.7798	0.8765	0.8892	0.1803

Table VIII: Reduced x^2 and correlation coefficient values.

As mentioned previously, protein chemical shifts differ significantly depending on the solvent in which they are measured[93], since there was not a way to incorporate the effect of DMSO on the prediction of chemical shifts, our calculation may present certain limitations. According to the results obtained, it can be seen that the reduced x^2 values for HA and ΔδHA chemical shifts are significantly higher as opposed to HN and ΔδHN shifts, due to the remarkably lower σ values for the HA chemical shifts compared to the σ values for the HN chemical shifts. Therefore, the dispersion values may have been underrated. This seems to be particularly true for the σ value of the HA chemical shift of residue 1, which is extremely lower relative to the other σ values.

The values of the linear correlation coefficient for the HA and HN chemical shifts were found to be 0.7798 and 0.8765 respectively, verifying a reasonable agreement between the experimental and the simulation-derived data. In Figure 4.7 it is clearly indicated that the experimental and simulation-derived HA secondary chemical shifts are quite similar. This is especially true for residues 2, 3, 4, for which, as it can be seen in Table VII, the experimental values are: -0.08, -0.1 and -0.11 and the simulation-derived values are: -0.021, 0.0131, and -0.0273. While for the ΔδHA values, there are certain similarities present between the simulation and the experiment, the ΔδHN experimental and simulation-derived values differ considerably. These differentiations can also be confirmed by the linear correlation coefficient value (Table VIII) of 0.1803, whereas for the ΔδHA chemical shifts, the linear correlation coefficient has a value of 0.8892. The low correlation coefficient value for the ΔδHN chemical shifts does not indicate a random relationship between the two variables, but rather, it is the consequence of the greater differences between the values of the simulation-derived HN chemical shifts and the random coil HN chemical shifts, compared to the differences between the simulation-derived HA chemical shifts and the random coil HA chemical shifts.

5. Conclusions and Discussion

The prime objective of this project was the evaluation of the validity of Molecular Dynamics simulations to predict the structure, folding process, and dynamics of peptide T, in comparison with experimental approaches, and more specifically, NMR spectroscopy. We used physics-based methods and aimed to compare our results with the results from the NMR experiment that Picone *et. al.* had conducted[49].

The synthetic octapeptide fragment with the sequence: ASTTTNYT, is known as peptide T due to its high threonine content and it was proven to function as a viral entry inhibitor. Peptide T is the fragment corresponding to the region 185-192 of the gp120 HIV coat protein[38][39][40]. Picone *et al.* studied peptide T as a zwitterion in DMSO solution by means of NMR spectroscopy at 500 MHz. Their results suggested that a type I β-turn including the four C-terminal residues, T^5, N^6, Y^7, and T^8 was the most prominent structure. However, they also noted that this conformation was not the only one present in solution and seemed to be the only one detectable due to the non-linear dependence of NOE on interatomic distances[49].

Secondary structure analysis using the programs STRIDE[74] and Weblogo[78] implied that peptide T is highly flexible and that it comprises a dynamic system. The majority of residues were assigned to turn or coil states, while assignments to helical structures were very rare. Both WebLogo diagrams indicated that the first and last residues are quite flexible and correspond to coil states. Residues 3-5 tend to form mostly turns, while some minor occurrences of coil, 3_{10}-helical, and even α-helical structures were also identified. The above-mentioned main structural characteristics were also observed by Picone and her colleagues[49], but unlike their findings, our results suggested a significant degree of flexibility in the system.

The structural analysis of turns and helices performed using the promotif[75] program helped us gain a more detailed view of the specific types of turns and helices that peptide T could adopt. According to our results, the most preferred β-turn types were types I and IV, while β-turns type II, VIII, I' and II' were not so frequent. In more detail, the most prominent β-turn type for the amino acid sequences: 1-Ala-2-Ser-3-Thr-4-Thr and 2-Ser-3-Thr-4-Thr-5-Thr was a type I β-turn, while the second most preferred β-turn type was type IV. For the sequences 4-Thr-5-Thr-6-Asn-7-Tyr and 5-Thr-6-Asn-7-Tyr-8-Thr, the most preferred β-turn type was type IV, followed by type I. Regarding the helices, the most preferred type of helix was 3_{10}-helix, followed by α-helix, while π-helix was extremely rare. Overall, our calculations clearly showed a preference for β-turns rather than helices. Also, according to our calculations, the most preferred conformation for the amino acid sequence 5-Thr-6-Asn-7-Tyr-8-Thr was a β-turn type IV, followed by a β-turn type I, while no helical conformations were observed for this combination and the 4-8 one. This observation is in agreement with the experimental conclusions, where it is stated that the most prominent conformation is a 5-8 β-turn rather than a 4-8 helical segment[49]. But unlike our findings, the experimental results state that the most likely cyclic structure is a type I β-turn rather than a type IV.

The two dPCA analyses suggested that the 5-8 amino acid residue segment of peptide T adopts more stable conformations, which correspond to distinct secondary structures with specific torsion angles and hydrogen bond patterns. These results are in agreement with the experimental conclusions. The association of high-density peaks with distinct peptide conformers demonstrated once again that the main structural characteristics were turns and helices. Due to the increased kinetic frustration of the system, the representative structures differed between the clusters, while many coil conformations were apparent as well. In terms of the superposition diagrams, these were complex and noisy and as a result, it was difficult to understand the structural content that was present in the clusters. Nevertheless, it was clear that the C-terminus tended to adopt more compact structures compared to the N-terminus.

To ascertain whether our simulation was efficiently sampled we could have applied Good-Turing statistics[83]. We applied this method using the Cα atoms of all residues of the peptide and then we limited the residue selection to residues 5-8. Our results clearly indicate that the structural variability of this part of the peptide has been sufficiently sampled, confirming that the C-terminal part of the peptide corresponds to more stable conformations.

In the last part of the project, we made quantitative comparisons between the experimental results and the simulation-derived results. More specifically, we compared the experimentally determined chemical and secondary chemical shifts with those derived from the simulation. The values of the linear correlation coefficient for the HA and HN chemical shifts were close to 1, verifying a reasonable agreement between the experimental and the simulation-derived data.

Overall, the MD simulation managed to predict with sufficient accuracy the structural characteristics of peptide T, as have been identified in the experiment. Protein chemical shifts differ significantly depending on the solvent in which they are measured[93], since there was not a way to incorporate the effect of DMSO on the prediction of chemical shifts, our calculation may present certain limitations. Moreover, conformational studies are more accurate when NOEs between pairs of α-CH and NH protons of adjacent residues are also identified[49]. The calculation of NOEs can detect the proximity of pairs of atoms and the information that they provide is often considered more specific than the measurement of chemical shift values[19]. In our project, it was not feasible to make quantitative comparisons of the simulation-derived and the experimentally-determined NOEs, due to the lack of experimental numerical values of NOEs. Our findings could be further supported when these values become available.

References

1. Hartley, H. (1951) 'Origin of the word protein', *Nature,* 168(4267), p. 244.

2. Nelson, D. & Cox, M. (2008) *Lehninger principles of biochemistry*. 5th edn. New York: W.H. Freeman.

3. Berg, J., Tymoczko, J. & Stryer, L. (2006) *Biochemistry*. 6th edn. New York: W.H. Freeman.

4. Branden, C. & Tooze, J. (2009) *Introduction to protein structure*. 2nd edn. New York, NY: Garland Pub.

5. Karplus, M. (1997) 'The Levinthal paradox: yesterday and today', *Folding and Design,* 2, pp. S69-S75.

6. Anfinsen, C. (1973) 'Principles that govern the folding of protein chains', *Science,* 181(4096), pp.223-230.

7. Berezovsky, I. & Trifonov, E. (2002) 'Loop fold structure of proteins: resolution of Levinthal's paradox', *Journal of Biomolecular Structure & Dynamics*, 20 (1), pp. 5–6.

8. Levinthal, C. (1968) 'Are there pathways for protein folding?', *Journal de Chimie Physique*, 65(1), pp.44-45.

9. Wetlaufer, D. (1973) 'Nucleation, rapid folding, and globular intrachain regions in proteins', *Proceedings of the National Academy of Sciences of the United States of America*, 70(3), pp.697-701.

10. Karplus, M. & Weaver, D. (1994) 'Protein folding dynamics: the diffusion-collision model and experimental data', *Protein Science,* 3(4), pp. 650-668.

11. Nolting, B. & Agard, D. (2008) 'How general is the nucleation-condensation mechanism?', *Proteins*, 73(3), pp. 754-764.

12. Fersht, A. (1997) 'Nucleation mechanisms in protein folding', *Current Opinion in Structural Biology*, 7(1), pp. 3-9.

13. Harrison, S. & Durbin R. (1985) 'Is there a single pathway for the folding of a polypeptide chain?', *Proceedings of the National Academy of Sciences of the United States of America,* 82(12), pp. 4028-4030.

14. Chan, H. & Dill, K. (1998) 'Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics', *Proteins: Structure, Function, and Genetics*, 30(1), pp. 2-33.

15. Dill, K. & Chan H. (1997) 'From Levinthal to pathways to funnels', *Nature Structural Biology,* 4(1), pp. 10-1

16. Plaxco, K. & Dobson, C. (1996) 'Time-resolved biophysical methods in the study of protein folding', *Current Opinion in Structural Biology,* 6(5), pp. 630-636.

17. Smyth, M. & Martin, J. (2000). 'x Ray crystallography', *Molecular Pathology*, 53(1), pp. 8-14.

18. Alberts, B. *et al.* (2002) *Molecular Biology of the Cell*. 4th edn. New York: Garland Science.

19. Dinner, A. *et al.* (2000) 'Understanding protein folding via free-energy surfaces from theory and experiment', *Trends in Biochemical Sciences*, 25(7), pp.331-339.

20. Greenfield, N. (2006) 'Using circular dichroism spectra to estimate protein secondary structure', *Nature Protocols,* 1(6), pp.876-2890.

21. Ryu, W. (2016) *Molecular Virology of Human Pathogenic Viruses.* 1st edn. London, San Diego: Academic Press.
22. Baker, D. & Sali, A. (2001) 'Protein structure prediction and structural genomics', *Science,* 294(5540), pp. 93-96.
23. Orengo, C., Jones, D. & Thornton, J. (2003) *Bioinformatics genes, proteins & computes*. 1st edn. London: Taylor and Francis Group.
24. Floudas, C. *et al.* (2006) 'Advances in protein structure prediction and de novo protein design: a review', *Chemical Engineering Science,* 61(3), pp. 966-988.
25. Moult, J. *et al.* (1995) 'A large-scale experiment to assess protein structure prediction methods', *Proteins*, 23(3), pp. ii–iv.
26. Hutchinson, E. & Thornton, J. (1994) 'A revised set of potentials for β-turn formation in proteins', *Protein Science*, 3(12), pp.2207-2216.
27. Wilmot, C. & Thornton, J. (1988) 'Analysis and prediction of the different types of β-turn in proteins', *Journal of Molecular Biology*, 203(1), pp.221-232.
28. Hutchinson, E. & Thornton, J. (1994) 'A revised set of potentials for β-turn formation in proteins', *Protein Science*, 3(12), pp.2207-2216.
29. Ho, B. & Dill, K. (2006) 'Folding very short peptides using molecular dynamics', *PLoS Computational. Biology,* 2(4), pp. 0228-0237.
30. Gnanakaran, S. *et al.* (2003) 'Peptide folding simulations', *Current Opinion Structural Biology,* 13(2), pp. 168–174.
31. Georgoulia, P. & Glykos, N. (2019) 'Molecular simulation of peptides coming of age: accurate prediction of folding, dynamics and structures', *Archives of Biochemistry and Biophysics*, 664, pp. 76-88.
32. Weiss, R. (1993) 'How does HIV cause AIDS?', *Science,* 260(5112), pp. 1273-1279.
33. Kwong, P. *et al.* (1998) 'Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody', *Nature,* 393 (6686), pp. 648-659.
34. Sharp, P. & Hahn, B. (2011) 'Origins of HIV and the AIDS pandemic', *Cold Spring Harbor Perspectives in Medicine,* 1(1), pp. 1-22.
35. Faria, N. *et al.* (2014) 'HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations', *Science,* 346(6205), pp. 56-61.
36. Schwetz, T. & Fauci, A. (2019) 'The extended impact of human immunodeficiency virus/AIDS research', *The Journal of Infectious Diseases,* 219(1), pp. 6–9.
37. Wilen, C., Tilton, J. & Doms, R. (2012) 'HIV: cell binding and entry', *Cold Spring Harbor Perspectives in Medicine,* 2(8), pp. 1-13.
38. Picone, D. *et al.* (2001) 'Peptide T revisited: conformational mimicry of epitopes of anti-HIV proteins', *Journal of Peptide Science*, 7(4), pp. 197-207.
39. Ruff, M. *et al.* (2001) 'Peptide T inhibits HIV-1 infection mediated by the chemokine receptor-5 (CCR5)', *Antiviral Research,* 52(1), pp. 63-75.
40. Pert, C. *et al.* (1986) 'Octapeptides deduced from the neuropeptide receptor-like pattern of antigen T4 in brain potently inhibit the human immunodeficiency virus receptor binding and T-cell infectivity', *Proceedings of the National Academy of Sciences of the United States of America,* 83(23), pp. 9254–9258.

41. McLellan, J. *et al.* (2011) 'Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9', *Nature,* 480(7377), pp. 336-43.
42. Ruff, M. *et al.* (1987) 'CD4 receptor binding peptides that block HIV infectivity cause human monocyte chemotaxis', *FEBS Letters,* 211(1), pp. 17 –22.
43. Brenneman, D. *et al*. (1988) 'Peptide T sequences prevent neuronal cell death produced by the envelope protein (gp120) of the human immunodeficiency virus', *Drug Development Research,* 15(4), pp. 361-369.
44. Yang, T. *et al.* (2009) 'Peptide T exhibits a well-defined structure in fluorinated solvent at low temperature', *Journal of Peptide Science,* 15(12), pp. 818-823.
45. Liapi, C. *et al.* (1998) 'Effects of [d-ala1] peptide t-nh2 and hiv envelope glycoprotein gp120 on cyclic amp dependent protein kinases in normal and psoriatic human fibroblasts', *Journal of Investigative Dermatology,* 110(4), pp. 332–337.
46. Ruff, M. *et al.* (2003) 'Update on D-ala-peptide T-amide (DAPTA): a viral entry inhibitor that blocks CCR5 chemokine receptors', *Current HIV research,* 1(1), pp. 51-67.
47. VanPatten, S. *et al.* (2020) 'Evidence supporting the use of peptides and peptidomimetics as potential SARS-CoV-2 (COVID-19) therapeutics', *Future Medicinal Chemistry,* 12(18), pp. 1647-1656.
48. Hassan Baig, M. *et al.* (2018) 'Peptide based therapeutics and their use for the treatment of neurodegenerative and other diseases', *Biomedicine & Pharmacotherapy,* 103, pp. 574-581.
49. Picone, D. *et al.* (1988) 'A 500 MHz study of peptide T in a DMSO solution', *FEBS letters,* 231(1), pp. 159-163.
50. Zhang, Z., Shi, Y. & Liu, H. (2003) 'Molecular dynamics simulations of peptides and proteins with amplified collective motions', *Biophysical Journal,* 84(6), pp. 3583-3593.
51. Stote, R. *et al.* (1999) *Tutorial-EMBnet*. Available at: https://embnet.vital-it.ch/MD_tutorial/ (Accessed: 5 November 2020).
52. Attig,N. *et al.* (2004) 'Computational soft matter: from synthetic polymers to proteins'.
53. Schlick, T. (2002) *Molecular modelling and simulation: an interdisciplinary guide.* 2nd edn. Heidelberg: Springer-Verlag.
54. Frenkel, D. (2002) *Understanding molecular simulation*. *From algorithms to applications.* San Diego: Academic Press.
55. Leach, A. (2009) *Molecular modelling*. Harlow: Pearson/Prentice Hall.
56. *The Amber Molecular Dynamics Package.* (2020) Available at: http://ambermd.org (Accessed 8 November 2020).
57. Brooks, B. (1983) 'CHARMM: A program for macromolecular energy, minimization, and dynamics calculations', *Journal of Computational Chemistry,* 4(2), pp.187-217.
58. *Biomolecular Simulation - The GROMOS Software.* (2011) Available at: http://www.gromos.net (Accessed 8 November 2020).
59. Jorgensen, W. & Tirado-Rives, J. (1988) 'The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin', *Journal of the American Chemical Society,* 110(6), pp.1657-1666.

60. Burgi, R. *et al.* (2001) ‘Folding study of an Aib-rich peptide in DMSO by molecular dynamics simulations’, *The Journal of Peptide Research,* 57(2), pp.107-118.
61. Salomon-Ferrer, R., Case, D. & Walker, R. (2012) ‘An overview of the Amber biomolecular simulation package’, *Computational Molecular Science,* 3(2), pp. 198- 210.
62. Hornak, V. *et al.* (2006) ‘Comparison of multiple Amber force fields and development of improved protein backbone parameters’, *Proteins,* 65(3), pp. 712-725.
63. Lindorff-Larsen, K. *et al*. (2010) ‘Improved side-chain torsion potentials for the Amber ff99SB protein force field’, *PROTEINS: Structure, Function and Bioinformatics,* 78(8), pp. 1950-1958.
64. Best, R., & Hummer, G. (2009) ‘Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides’ *The Journal of Physical Chemistry B*, 113(26), pp. 9004–9015.
65. Phillips, J. *et al.* (2005) ‘Scalable molecular dynamics with NAMD’, *Journal of Computational Chemistry*, 26(16), pp.1781-1802.
66. Sterling, T. *et al.* (1995) *beowulf: a parallel workstation for scientific computation*. Available at: https://webhome.phy.duke.edu/~rgb/brahma/Resources/beowulf/papers/ICPP95/icpp95.html (Accessed: 26 November 2020).
67. Glykos, N. (2020) *The Norma computing cluster*. Available at: https://norma.mbg.duth.gr/index.php?id=about:intro (Accessed: 16 October 2020).
68. Phillips, J. & Hardy, D. (2017) ‘NAMD Tutorial’.
69. Bernardi, R. *et al.* (2012) *NAMD User’s Guide – Theoretical and Computational Biophysics Group.* Available at: http://www.ks.uiuc.edu/Research/namd/cvs/ug/ (Accessed: 26 November 2020).
70. Gkeka, P. & Cournia, Z. (2015) ‘Molecular dynamics simulations of lysozyme in water’.
71. Glykos, N. (2006) ‘Software news and updates. CARMA: a molecular dynamics analysis program’, *Journal of Computational Chemistry,* 27(14), pp. 1765−1768.
72. 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.
73. Hsin, J. *et al.* (2008) ‘Using VMD: an introductory tutorial’, *Current Protocols in Bioinformatics*, 23(4), pp. 566-579.
74. Frishman, D. & Argos, P. (1995) ‘Knowledge-based protein secondary structure assignment’, *PROTEINS: Structure, Function and Genetics,* 23(4), pp. 566- 579.
75. Hutchinson, E. & Thornton, J. (2008) ‘PROMOTIF-A program to identify and analyze structural motifs in proteins’, *Protein Science,* 5(2), pp.212-220.
76. Humphrey, W., Dalke, A. & Schulten, K. (1996) ‘VMD: visual molecular dynamics’, *Journal of Molecular Graphics,* 14(1), pp. 33−38.
77. *VMD – Visual Molecular Dynamics.* (2020) Available at: https://www.ks.uiuc.edu/Reasearch/vmd (Accessed: 27 November 2020).
78. Crooks, G. *et al.* (2004) ‘WebLogo: A Sequence Logo Generator’, *Genome Research,* 14(6), pp. 1188-1190.

79. Glykos, N. (2017) *Home – plot – Utopia.* Available at: https://utopia.duth.gr/glykos/plot/ (Accessed: 26 November 2020).

80. Shen, Y. & Bax, A. (2010) 'SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network', *Journal of Biomolecular NMR,* 48(1), pp. 13–22.

81. *The Perl Programming Language - www.perl.org.* (2006) Available at: https://www.perl.org (Accessed: 26 November 2020).

82. *R: The R Project for Statistical Computing.* (2020). Available at: https://www.r-project.org (Accessed: 26 November 2020).

83. Koukos, P. & Glykos, N. (2014) 'On the application of Good−Turing statistics to quantify convergence of biomolecular simulations', *Journal of Chemical Information and Modeling,* 54(1), pp. 209–217.

84. Altis, A. *et al.* (2008) 'Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis', *The Journal of Chemical Physics*, 128(24), p.245102.

85. Altis, A. *et al.* (2007) 'Dihedral angle principal component analysis of molecular dynamics simulations', *The Journal of Chemical Physics*, 126(24), p. 244111.

86. Mu, Y., Nguyen, P. & 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.

87. Shao, J. *et al.* (2007) 'Clustering Molecular Dynamics Trajectories: 1. Characterizing the performance of different clustering algorithms', *Journal of Chemical Theory and Computation,* 3(6), pp. 2312-2334.

88. Wolf, A. & Kirschner, K. (2013) 'Principal component and clustering analysis on molecular dynamics data of the ribosomal L11·23S subdomain', *Journal of molecular modeling*, *19*(2), pp. 539–549.

89. Bratchell, N. (1989) 'Cluster analysis', *Chemometrics and Intelligent Laboratory Systems,* 6(2), pp. 105-125.

90. Kufareva, I. & Abagyan, R. (2012) 'Methods of protein structure comparison', *Methods in molecular biology,* 857, pp. 231–257.

91. Knapp, B. *et al.* (2011) 'Is an intuitive convergence definition of molecular dynamics simulations solely based on the root mean square deviation possible?', *Journal of computational biology: a journal of computational molecular cell biology*, 18(8), pp. 997–1005.

92. Berjanskii, M. & Wishart, D. (2017) 'Unraveling the meaning of chemical shifts in protein NMR', *Biochimica et Biophysica Acta. Proteins and Proteomics,* 1865(11), pp. 1564-1576.

93. Wishart, D. (2011) 'Interpreting protein chemical shift data', *Progress in Nuclear Magnetic Resonance Spectroscopy,* 58(1-2), pp. 62-87.

94. Mielke, S. & Krishnan, V. (2003) 'Protein structural class identification directly from NMR spectra using averaged chemical shifts', *Bioinformatics,* 19(16), pp. 2054-2064.

95. Mielke, S. & Krishnan, V. (2009) ‘Characterization of protein secondary structure from NMR chemical shifts’, *Progress in Nuclear Magnetic Resonance Spectroscopy,* 54(3-4), pp. 141-165.
96. Homer, J. (1975) ‘Solvent effects on nuclear magnetic resonance chemical shifts’, *Applied Spectroscopy Reviews,* 9(1), pp. 1-132.
97. Laszlo, P. (1967) ‘Chapter 6 solvent effects and nuclear magnetic resonance’, *Progress in Nuclear Magnetic Resonance Spectroscopy*, 3, pp. 231-402.
98. Abraham, R. & Mobili, M. (2004) ‘The prediction of 1H NMR chemical shifts in organic compounds’, *Spectroscopy Europe,* pp. 16-22.
99. Abraham, R. *et al.* (2006) ‘1H chemical shifts in NMR: Part 23, the effect of dimethyl sulphoxide versus chloroform solvent on 1H chemical shifts’, *Magnetic Resonance in Chemistry,* 44(5), pp. 491-509.
100. Buckingham, A., Schaefer, T. & Schneider, W. (1960) ‘Solvent effects in nuclear magnetic resonance’, *The Journal of Chemical Physics,* 32(4), pp. 1227-1233.
101. Tremblay, M., Banks, A. & Rainey, J. (2010) ‘The predictive accuracy of secondary structure chemical shifts is more affected by protein secondary structure in solvent environment’, *Journal of Biomolecular NMR,* 46(4), pp. 257-270.
102. Bevington, P. (1969) *Data Reduction and Error Analysis for the Physical Sciences*. 3rd edn. New York: McGraw-Hill.

Appendix

A1: Energy minimization and heating up script (heat.namd)

```
#
# Input files
#
amber                   on
readexclusions          yes
parmfile                Tpept.prmtop
coordinates             Tpept.pdb

#
# Output files & writing frequency for DCD
# and restart files
#
outputname              output/heat_out
binaryoutput            off
restartname             output/restart
restartfreq             1000
binaryrestart           yes
dcdFile                 output/heat_out.dcd
dcdFreq                 400

#
# Frequencies for logs and the xst file
#
outputEnergies          400
outputTiming           1600
xstFreq                 400

#
# Timestep & friends
#
timestep                2.0
stepsPerCycle           20
nonBondedFreq           1
fullElectFrequency      2


#
# Simulation space partitioning
#
switching               on
switchDist              7
cutoff                  8
pairlistdist            9

#
# Basic dynamics
#
```

```
temperature             0
COMmotion               no
dielectric              1.0
exclude                 scaled1-4
1-4scaling              0.833333
rigidbonds              all

#
# Particle Mesh Ewald parameters.
#
Pme                     on
PmeGridsizeX            32                      # <===== CHANGE ME
PmeGridsizeY            32                      # <===== CHANGE ME
PmeGridsizeZ            32                      # <===== CHANGE ME


#
# Periodic boundary things
#
wrapWater               on
wrapNearest             on
wrapAll                 on


cellBasisVector1        32.00    0.00    0.00   # <===== CHANGE ME
cellBasisVector2         0.00   32.00    0.00   # <===== CHANGE ME
cellBasisVector3         0.00    0.00   32.00   # <===== CHANGE ME
cellOrigin               0.00    0.00    0.00   # <===== CHANGE ME

#
# Langevin dynamics parameters
#
langevin                on
langevinDamping         10
langevinTemp            320                     # <===== Check me
langevinHydrogen        off

langevinPiston          on
langevinPistonTarget    1.01325
langevinPistonPeriod    200
langevinPistonDecay     100
langevinPistonTemp      320                     # <===== Check me

useGroupPressure        yes



#
# run one step to get into scripting mode
#
minimize                0
```

```
langevinPiston          off

#
# minimize nonbackbone atoms
#
minimize                2000                     ;# <===== CHANGE ME
output                  output/min_fix

#
# heat with CAs restrained
#
set temp 20;
while { $temp < 321 } {                         ;# <===== Check me
langevinTemp            $temp
run                     1000                     ;# <===== CHANGE ME
output                  output/heat_ca
set temp [expr $temp + 20]
}

#
# equilibrate volume with CAs restrained
#
langevinPiston          on
run                     500000                     ;# <===== CHANGE ME
output                  output/equil_ca
```

A2: Equilibration script (equi.namd)

```
#
# Input files
#
amber                   on
readexclusions          yes
parmfile                Tpept.prmtop
coordinates             heat_out.coor
bincoordinates          restart.coor
binvelocities           restart.vel
extendedSystem          restart.xsc

#
# Adaptive ...
#
adaptTempMD             on
adaptTempTmin           280
adaptTempTmax           380
adaptTempBins           1000
adaptTempRestartFile    output/restart.tempering
adaptTempRestartFreq    100000
adaptTempLangevin       on
adaptTempRescaling      off
adaptTempOutFreq        400
# adaptTempDt             0.0000500
adaptTempCgamma         0


#
# Output files & writing frequency for DCD
# and restart files
#
outputname              output/equi_out
binaryoutput            off
restartname             output/restart
restartfreq             100000
binaryrestart           yes
dcdFile                 output/equi_out.dcd
dcdFreq                 400
DCDunitcell             yes


#
# Frequencies for logs and the xst file
#
outputEnergies          400
outputTiming            1600
xstFreq                 400

#
```

```
# Timestep & friends
#
timestep                2.5
stepsPerCycle           20
nonBondedFreq           1
fullElectFrequency      2

#
# Simulation space partitioning
#
switching               on
switchDist              7
cutoff                  8
pairlistdist            9
twoAwayX                yes
margin              1.0

#
# Basic dynamics
#
COMmotion               no
dielectric              1.0
exclude                 scaled1-4
1-4scaling              0.833333
rigidbonds              all

#
# Particle Mesh Ewald parameters.
#
Pme                     on
PmeGridsizeX            32                      # <===== CHANGE ME
PmeGridsizeY            32                      # <===== CHANGE ME
PmeGridsizeZ            32                      # <===== CHANGE ME

usePMECUDA            no



#
# Periodic boundary things
#
wrapWater               on
wrapNearest             on
wrapAll                 on



#
# Langevin dynamics parameters
#
langevin                on
```

```
langevinDamping         1
langevinTemp            320                     # <===== Check me
langevinHydrogen        off

langevinPiston          on
langevinPistonTarget    1.01325
langevinPistonPeriod    400
langevinPistonDecay     200
langevinPistonTemp      320                     # <===== Check me

useGroupPressure        yes

firsttimestep           522500000
run                     500000000              ;# <===== CHANGE ME
```

A3: calc_shifts.pl

```
#!/usr/bin/perl -w

(@ARGV == 2) or die "Usage: calc_shifts <dcd> <psf>\n";

#
# How many shifts we will be collecting ?
#

(`carmanox -atmid ALLID -pdb -first 1 -last 1 $ARGV[0] $ARGV[1]` eq "" ) or die "Carma
made a boo-boo. Too bad ...\n";
`sparta+ -in $ARGV[0].*.pdb >& /dev/null`;
`/bin/rm -rf $ARGV[0].*.pdb $ARGV[1].*.pdb`;

open ( IN, "pred.tab" ) or die "Can not open pred.tab. Usage: calc_shifts <dcd> <psf>\n";
while ( $line = <IN> )
  {
    if ( $line =~ /^FORMAT/ )
      {
        last;
      }
  }

$line = <IN>;
$tot = 0;
while ( $line = <IN> )
  {
    $ids[ $tot ] = substr( $line, 0, 14 );
    $tot++;
  }

close( IN );

`/bin/rm -rf *.tab`;

if ( $tot < 1 )
  {
    print "Too few atoms for calculating shifts. Something is wrong. Bye.\n";
    exit;
  }


print "Will be collecting data for $tot atoms. Starting ...\n";

#
# Will do it in sets of 800 structures ...
#

`mkdir tmp1`;
```

```
`mkdir tmp2`;
`mkdir tmp3`;
`mkdir tmp4`;
`mkdir tmp5`;
`mkdir tmp6`;
`mkdir tmp7`;
`mkdir tmp8`;

$first = 1;

print "Now processing set starting at frame          ";
while( 1 )
{

%8d", $first );

$last = $first + 99;
`cd tmp1 ; carmanox -atmid ALLID -pdb -first $first -last $last ../$ARGV[0] ../$ARGV[1]`;
`cd tmp1 ; sparta+ -in $ARGV[0].*.pdb >& /dev/null &`;

$first += 100;
$last = $first + 99;
`cd tmp2 ; carmanox -atmid ALLID -pdb -first $first -last $last ../$ARGV[0] ../$ARGV[1]`;
`cd tmp2 ; sparta+ -in $ARGV[0].*.pdb >& /dev/null &`;

$first += 100;
$last = $first + 99;
`cd tmp3 ; carmanox -atmid ALLID -pdb -first $first -last $last ../$ARGV[0] ../$ARGV[1]`;
`cd tmp3 ; sparta+ -in $ARGV[0].*.pdb >& /dev/null &`;

$first += 100;
$last = $first + 99;
`cd tmp4 ; carmanox -atmid ALLID -pdb -first $first -last $last ../$ARGV[0] ../$ARGV[1]`;
`cd tmp4 ; sparta+ -in $ARGV[0].*.pdb >& /dev/null &`;

$first += 100;
$last = $first + 99;
`cd tmp5 ; carmanox -atmid ALLID -pdb -first $first -last $last ../$ARGV[0] ../$ARGV[1]`;
`cd tmp5 ; sparta+ -in $ARGV[0].*.pdb >& /dev/null &`;

$first += 100;
$last = $first + 99;
`cd tmp6 ; carmanox -atmid ALLID -pdb -first $first -last $last ../$ARGV[0] ../$ARGV[1]`;
`cd tmp6 ; sparta+ -in $ARGV[0].*.pdb >& /dev/null &`;

$first += 100;
$last = $first + 99;
`cd tmp7 ; carmanox -atmid ALLID -pdb -first $first -last $last ../$ARGV[0] ../$ARGV[1]`;
`cd tmp7 ; sparta+ -in $ARGV[0].*.pdb >& /dev/null &`;

$first += 100;
```

```
$last = $first + 99;
`cd tmp8 ; carmanox -atmid ALLID -pdb -first $first -last $last ../$ARGV[0] ../$ARGV[1]`;
`cd tmp8 ; sparta+ -in $ARGV[0].*.pdb >& /dev/null &`;


$first += 100;

$procs = `ps -aef | grep 'sparta+' | wc -l`;
while( $procs > 2 )
  {
    sleep(1);
    $procs = `ps -aef | grep 'sparta+' | wc -l`;
  }

`cd tmp1 ; /bin/rm -rf $ARGV[0].*.pdb $ARGV[1].*.pdb *_struct.tab ; mv * ../ >&
/dev/null`;
`cd tmp2 ; /bin/rm -rf $ARGV[0].*.pdb $ARGV[1].*.pdb *_struct.tab ; mv * ../ >&
/dev/null`;
`cd tmp3 ; /bin/rm -rf $ARGV[0].*.pdb $ARGV[1].*.pdb *_struct.tab ; mv * ../ >&
/dev/null`;
`cd tmp4 ; /bin/rm -rf $ARGV[0].*.pdb $ARGV[1].*.pdb *_struct.tab ; mv * ../ >&
/dev/null`;
`cd tmp5 ; /bin/rm -rf $ARGV[0].*.pdb $ARGV[1].*.pdb *_struct.tab ; mv * ../ >&
/dev/null`;
`cd tmp6 ; /bin/rm -rf $ARGV[0].*.pdb $ARGV[1].*.pdb *_struct.tab ; mv * ../ >&
/dev/null`;
`cd tmp7 ; /bin/rm -rf $ARGV[0].*.pdb $ARGV[1].*.pdb *_struct.tab ; mv * ../ >&
/dev/null`;
`cd tmp8 ; /bin/rm -rf $ARGV[0].*.pdb $ARGV[1].*.pdb *_struct.tab ; mv * ../ >&
/dev/null`;


@files = glob("$ARGV[0]*.tab");

if ( @files == 0 )
  {
    last;
  }

foreach $file ( @files )
{

`tail -$tot $file | awk '{printf "%8.3f ", \$5}' >> SHIFTS`;
`echo >> SHIFTS`;
}

`/bin/rm -rf $ARGV[0]*.tab`;



}
```

```
`rmdir tmp1 tmp2 tmp3 tmp4 tmp5 tmp6 tmp7 tmp8`;
print "\n\n";

#
# Calculate means + sigmas
#
open ( IN, "SHIFTS" ) or die "Can not open SHIFTS ??? How did this happen ???\n";

for ( $i=0 ; $i < $tot ; $i++ )
{
       $mean= 0.0;
       $nof_lines = 0;
       $std = 0.0;
       while ( $line = <IN> )
         {
           @data = split( ' ', $line );

           $nof_lines++;
           $delta = $data[ $i ] - $mean;
           $mean += $delta / $nof_lines;
           $std += $delta * ($data[ $i ] - $mean);
         }

        printf "%s    %8.4f %8.4f\n", $ids[ $i ], $mean, sqrt( $std / ($nof_lines -1));
        seek( IN, 0, 0 );
}

close( IN );

print "\nAll done.\n\n";
```

Script 1: ΔδHA_experimental.R

```
HAexperimental <- c(-0.66, -0.08, -0.1, -0.11, -0.16, -0.23, -0.45)
residues <- c("A", "S" "T" "T" "T" "N"  "Y"  "T")
png (file= "barchart1.png")
barplot (HAexperimental, names.arg= residues, xlab= "Residues", ylab= "ΔδHA", ylim= c(-
1,+1), border= "black")
dev.off()
```

Script 2: ΔδHA_simulation.R

```
HAsimulation <- c(-0.1516, -0.0210, 0.0131, -0.0273, -0.0049, 0.0196, -0.0880)
residues <- c("A", "S" "T" "T" "T" "N"  "Y"  "T")
png (file= "barchart2.png")
barplot (HAsimulation, names.arg= residues, xlab= "Residues", ylab= "ΔδHA", ylim= c(-
1,+1), border= "black")
dev.off()
```

Script 3: ΔδHN_experimental.R

```
HNexperimental <- c(0.32, -0.1, -0.15, -0.31, -0.38, -0.38, -0.46)
residues <- c("S" "T" "T" "T" "N"  "Y"  "T")
png (file= "barchart3.png")
barplot (HNexperimental, names.arg= residues, xlab= "Residues", ylab= "ΔδHN", ylim= c(-
1,+1), border= "black")
dev.off()
```

Script 4: ΔδHN_simulation.R

```
HNsimulation <- c(0.0332, 0.1580, 0.1835, 0.1068, -0.1830, -0.1884, 0.1998)
residues <- c("S" "T" "T" "T" "N"  "Y"  "T")
png (file= "barchart4.png")
barplot (HNsimulation, names.arg= residues, xlab= "Residues", ylab= "ΔδHN", ylim= c(-
1,+1), border= "black")
dev.off()
```

Script 5: calc_chi-square.pl

```
!/usr/bin/perl -w
open ( IN , "$ARGV[0]") or die "Usage: calc_chi-square <input_file>\n";
$sum = 0;
$num_of_lines= 0;
while ( $line = <IN> )
{
 @data = split (' ', $line);
   $col_num = @data;
    if ( $col_num == 3 )
     {
       $num_of_lines++;
       $subtract = $data[0] - $data[2];
       $val = $subtract * $subtract / ($data[1] * $data[1]);
       $sum += $val;
     }
    else
     {
       die "Not 3 columns in input file\n";
     }
}

close (IN);
print"The reduced chi-square value is\t", $sum / ($num_of_lines - 1), "\n";
```

Script 6: calc_correlation-coefficient.pl

```
#!/usr/bin/perl -w

(@ARGV == 2) or die "Usage: calc_corr <file1> <file2>\n";
open (IN_1, "$ARGV[0]") or die "Cannot open <file1>\n";
open (IN_2, "$ARGV[1]") or die "Cannot open <file2>\n";

@file1 = <IN_1>;
@file2 = <IN_2>;

close (IN_1);
close (IN_2);

$N = @file1;
$sum1 = 0;
$sum2 = 0;

for ( $i = 0; $i < $N; $i++)
{

 $sum1 += $file1[$i];
 $sum2 += $file2[$i];
}
```

```
$mean1 = $sum1 / $N;
$mean2 = $sum2 / $N;

$sum_xy = 0;
$sum_x_square = 0;
$sum_y_square = 0;

for ( $i = 0; $i < $N; $i++)
{
  $x = $file1[$i] - $mean1;
  $y = $file2[$i] - $mean2;
  $xy = $x * $y;
  $sum_xy += $xy;
  $sum_x_square += $x * $x;
  $sum_y_square += $y * $y;
}

print "corr =\t", $sum_xy / sqrt($sum_x_square * $sum_y_square), "\n";
```