DEMOCRITUS UNIVERSITY OF THRACE
SCHOOL OF HEALTH SCIENCES
DEPT. OF MOLECULAR BIOLOGY AND GENETICS

BSc Thesis

Propensities of Asn-Gly containing heptapeptides to form β-turn structures: a Molecular Dynamics simulation approach

---

Author: Dimitrios A. Mitsikas
Dept. of Molecular Biology and Genetics

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

September 2018

Abstract

Over the last two decades, molecular modelling has emerged as an invaluable tool for studying the protein folding *in silico*. Molecular mechanics calculations along with computer graphics are now widely used to visualise molecular shape and structure, and quantify steric demand. On the other hand, the most recent quantum mechanics calculations continue to play an ever increasing role in Structural Biology, promising high quality description of molecular structures. In this thesis, we examine the accuracy of Molecular Dynamics simulations and their ability to approach systems derived from quantum mechanics calculations. More specifically, three Molecular Dynamics simulations of 5 μs each in explicit water solvent were carried out for three heptapeptides containing the Asn-Gly segment, in order to study their folding and dynamics. Previous data, based on quantum mechanics calculations from Kang Y. K. and Yoo I. K., has proven that these peptides adopt β-turn and β-hairpin structures. The results from the simulations' analyses point out significant divergence from the *ab initio* models, denoting severe dynamicity in our systems. The heptapeptides show a general tendency to form β-turn conformations regarding their four-residue central part, but the results do not utterly agree with the *ab initio* models, raising the question whether Molecular Dynamics simulations are suitable for such dynamic systems.

Acknowledgements

I would like to thank my advisor Dr. Nicholas M. Glykos for his advice, encouragement, patience and most importantly for teaching me the way a scientist should work, think and confront challenges occurring in the laboratory.

My laboratory partners, the “NMG Group”, for being both excellent coworkers and friends, and for generously helping me every time I reached a tense impasse.

My friends, each one individually, for their encouragement, the beautiful experiences and the unforgettable memories that we created together.

My family, for their continuous encouragement and moral support throughout my academic studies.

Table of Contents

1. Introduction

1.1 Proteins: A Prelude

Since proteins were first recognised as a distinct class of biological molecules, they have been a major concern to the scientific society. Characterised as the most multifunctional macromolecules in living species serving a vital role in almost every biological function, scientists still strive to unravel every aspect of their nature. Over the last century, the study of protein structure and function has made a remarkable advance thanks to the growth of computer power and the consequent development of several breakthrough computational methods, such as X-ray Crystallography and Nuclear Magnetic Resonance (NMR). However, throughout the years, as computers began to replace the old "tactile models" in visualising molecular shape, the scientific interest focused on a new concern: the dynamicity of proteins. The abundance of protein structures that have been resolved during the last decades shed light on the functionality and the dynamic behaviour of these molecules, explaining their different properties and a wide range of biological mechanisms, but most importantly, decoded what today is regarded as one of the most significant axioms in Structural Biology: structure and function are two interdependent concepts. Nowadays, we know that it is of crucial importance to identify the principal components of the way proteins are being formed, in order to fully determine the various functions carried out by them.

Protein structure can be analysed into four main categories: primary structure, secondary structure, tertiary and quaternary structure[1]. Each class of this hierarchy is strictly dependent on its subordinate one. The final three-dimensional structures occur as the primary (unfolded) amino acid chains are being folded in a way to produce compact, self-contained structural domains. These domains either serve as subdomains for the construction of a greater complex or they function autonomously in certain biological processes **(Figure 1.1)**[2].

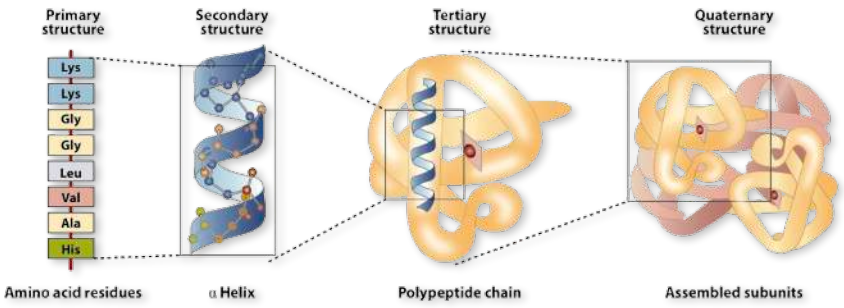


Figure 1.1 Protein structure can be classified into four main stages: primary structure (the amino acid chain), secondary structure, tertiary structure and quaternary structure. The final three-dimensional structure (native state) can possibly include different polypeptide chains, thus forming a quaternary structure. [adapted without permission from *Pearson Education Inc., 2010*]

In order to fully understand the function of proteins, one should be able to predict the three-dimensional structure from the primary amino acid sequence, that is to say, the folding process of a macromolecule. Over the last few decades, however, and despite the efforts of the scientific community, the well-known "folding problem" remains unsolved and still one of the most challenging questions in Structural and Molecular Biology.

1.2 Protein Folding

The concepts of "folding" and "denaturation" of proteins were widely known among the scientific community for almost a century, but the most notable findings came into light through the work of Cristian Anfinsen[3] and Cyrus Levinthal[4,5] during 1960. The former, studying the renaturation of a fully denatured ribonuclease, under certain conditions, proposed the "*thermodynamic hypothesis*" according to which, the native functional structure of a protein in its normal physiological environment is the one in which the Gibbs free energy of the whole system is lowest; in other words, the native conformation is determined by the entirety of interatomic interactions and thus, by the amino acid sequence[3]. The main driving force for folding water-soluble globular proteins is the packing of hydrophobic amino acids into the interior of the molecule to create a hydrophobic core. Yet, a few years earlier, Levinthal had already studied the kinetic parameters of folding process. In his attempt to interpret the factors that define how fast a protein can fold, he described this phenomenon as a random search problem. Taken literally, this means that all possible conformations of a polypeptide chain (except the native state) are equally probable, hence the native state can be found only by an unbiased random search[6]. This implies that the time a protein will consume to find its native state, depends on both the total number of possible configurations of the polypeptide chain and the time required to find each conformation (a 100-residue protein, for example, can adopt 10^{70} configurations, which leads to an enormously long folding time of about 10^{50} years)[6]. Since proteins generally fold in a timescale of milliseconds to seconds, Levinthal's statement fairly characterised as a paradox.

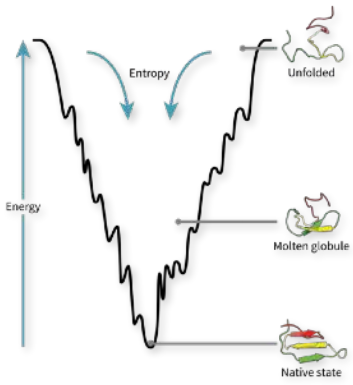


Figure 1.2 A simplified representation of what nowadays is perceived as a folding funnel. The vertical axis represents the Gibbs free energy, whereas the horizontal axis represents the entropy of the system. As it can be seen, energy landscapes are usually "rough", with several non-native local minima in which partially folded proteins may be trapped before reaching the native state. [adapted without permission from *Wikipedia*]

Levinthal's solution on the folding problem was that the folding process is speeded as there are defined pathways towards the native state; a suggestion became known as the "*kinetic theory*" [5]. Back then, most of the proposals based on known structures, to reduce the size of the conformational pool that is searched and the folding time to the experimental scale, did not contribute

much to the resolution of Levinthal's paradox. Examples include the "*nucleation growth model*"[7], the "*diffusion-collision model*"[8], the "*framework model*"[9] and the "*jigsaw-puzzle model*"[10]. During the years that followed, though, studies suggested that folding pathways are not the absolute solution to Levinthal's paradox, bringing to the foreground the landscape perspective. This theory successfully describes both the procedure of reaching a global free energy minimum (Anfinsen's "*thermodynamic hypothesis*") and the speed of folding process (Levinthal's "*kinetic hypothesis*")[11]. The landscape is depicted as a funnel, exhibiting the free energy of each configuration in relation to the degrees of freedom of the system **(Figure 1.2)**[11]. Assuming that we want to represent Levinthal's paradox, in which every protein chain mandatorily undergoes the same sequence of events until reaching the native state, based on the landscape perspective, that would be depicted as a flat funnel upon which a protein may follow almost infinite trajectories to end up eventually in the bottom of the funnel, the native state **(Figure 1.3a)**[11]. The folding funnel theory became widely popular among the scientific community, as scientists attempted to examine different perspectives of the theory **(Figure 1.3)**. Nowadays, we have reached a consensus on which protein folding is a much more heterogenous process. Each individual protein chain may follow a unique trajectory, but just like skiers descending a mountainside, they all may eventually reach the same point at the funnel's bottom, the native state **(Figure 1.3d)**[12].

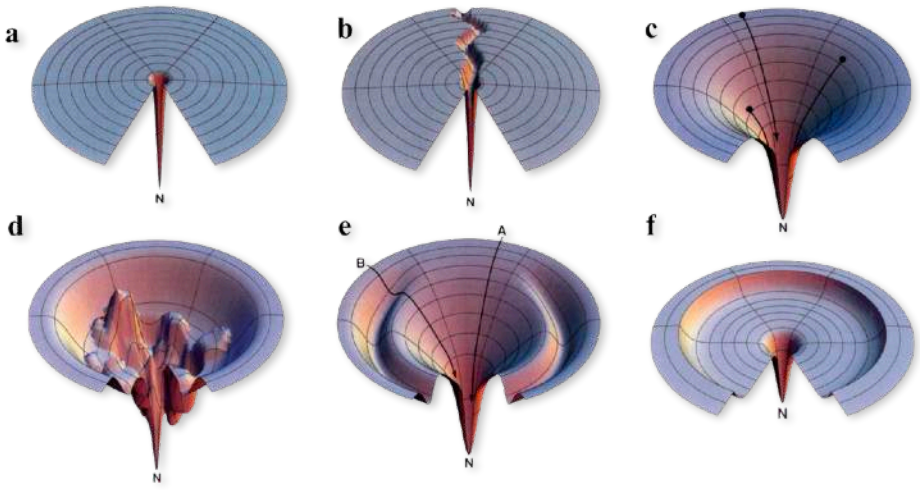


Figure 1.3 Different perspectives on the folding funnel theory (N stands for native conformation). (a) Levinthal's "golf-course" energy landscape. (b) The "grooved golf-course" landscape depicting Levinthal's pathway solution to the random search problem. (c) The "HP+" model, one of the first funnels presented as a solution to Levinthal's paradox, is an idealised landscape describing the overall decrease of the total possible conformations as the protein's free energy is reduced. (d) A coarse landscape with kinetic traps, energy barriers and numerous pathways towards native state; the "bumpy bowl" model. This representation is considered as a realistic example of a folding funnel. (e) The "moat" landscape illustrates that proteins can either fold fast (A) or slow if the protein shall first adopt a metastable conformation before reaching the native state (B). (f) The "champagne glass" funnel shows how conformational entropy can cause free energy barriers during the folding process. [adapted without permission from *Dill & Chan, Nature Stractural Biology, 1997*]

1.3 Experimental and *in silico* Methods in Protein Folding

While the folding problem remains unsolved and the *de* novo prediction of the final native conformation of a protein seems almost inevitable, experimental methods seem to be more reliable in identifying the three-dimensional structure of the protein. So far, X-ray Crystallography and Nuclear Magnetic Resonance (NMR) tend to be the two most widely used techniques in charge of discovering macromolecular structures. X-ray Crystallography, the first and most important method for identifying protein structures, is used for determining the atomic and molecular structure of a crystal, assuming that the protein can form well-defined crystals that efficiently allow X-ray diffraction[13]. NMR, on the other hand, comprises, as well, an exceptionally reliable technique providing information about the totality of a molecule's interatomic interactions[14]. Until today, various experimental methods have been developed for identifying protein structure. Such techniques include Circular Dichroism (CD)[14], Mass Spectroscopy[15], Atomic Force Microscopy (AFM)[16], Small-angle X-ray Scattering (SAXS) and Fourier Transform Infrared Spectroscopy (FT-IR)[17,18].

Nevertheless, while experimental and theoretical studies have led to the emergence of a unified protein folding mechanism and the discovery of a vast number of protein structures, recents studies indicate that proteins are even more heterogenous and complex macromolecules. Regardless the success and reliability of the above methods, the incapability of managing massive data, the errors arising from the experiment itself, and the difficulty of representing a complete, dynamic view of molecules, hinder the resolution of the complete structure, function and folding process of proteins. The most recent computational techniques, however, providing a connection link between theoretical and experimental methods, act as a complementary tool so as to fulfil the above mentioned insufficiencies. Towards this direction, the growth of computer power, the establishment of numerous accessible databases and the development of computational tools and algorithms have also contributed radically in the evolution of Molecular Biology and Bioinformatics.

In general, predicting protein structure from its amino acid sequence regardless of its natural folding process can only be accomplished using empirical methods, the most notable of which are the prediction of protein secondary structure using artificial neural networks[19], homology modelling[20], threading recognition[21] and the *de novo* prediction based on the "*thermodynamic hypothesis*"[22]. Contrary to the above empirical methods, energy-based methods attempt to predict protein structures based on mathematical models that describe the system's potential energy during the folding process. Such example are Molecular Dynamics simulations by which we can adequately examine the folding process of a protein, based on the principles of classical mechanics.

1.4 Secondary Structure

Having an overview of the long process for discovering structures, it is high time to have a closer look at the three-dimensional structure of proteins. As the amino acid chain folds, driven by the "force" of hydrophobic interactions, proteins tend to form compact structures, creating a hydrophobic core and a hydrophilic outer surface. The packing of hydrophobic residues in the interior of proteins is remarkably dense and taking into account the conformational restrictions due to the steric hindrances imposed by the side-chains, protein folding seems like a tricky puzzle[2]. In addition to this, along with side-chains, protein's backbone chain must, as well, fold into the hydrophobic interior of the molecule. Proteins' backbone, however, are highly hydrophilic because of the occurrence of imine (NH) and carbonyl (C'=O) groups in each amino acid, which act as proton donors and proton receptors respectively (**Figure 1.4**)[2]. In such hydrophobic environment, these polar groups need to be neutralized. The solution to this, is a consequent formation of local, stable conformational patterns known as secondary structure elements[2]. The most common types of secondary structures are the α-helix and the β-sheet (**Figure 1.5**). Both structures are held in shape by hydrogen bonds between the carbonyl O of one amino acid and the imino H of another.

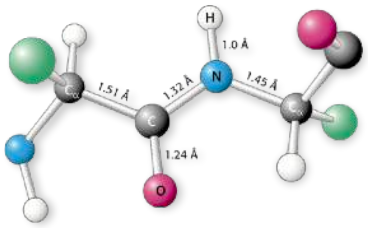


Figure 1.4 A trans peptide group showing typical distances between atoms.The imine and carbonyl groups are on either side of each Cα atom. A peptide group contains the Cα atom and C'=O group of the *n* residue, as well as the NH group and Cα atom of *n+1* residue. [adapted without permission from *Stryer, Biochemistry*]

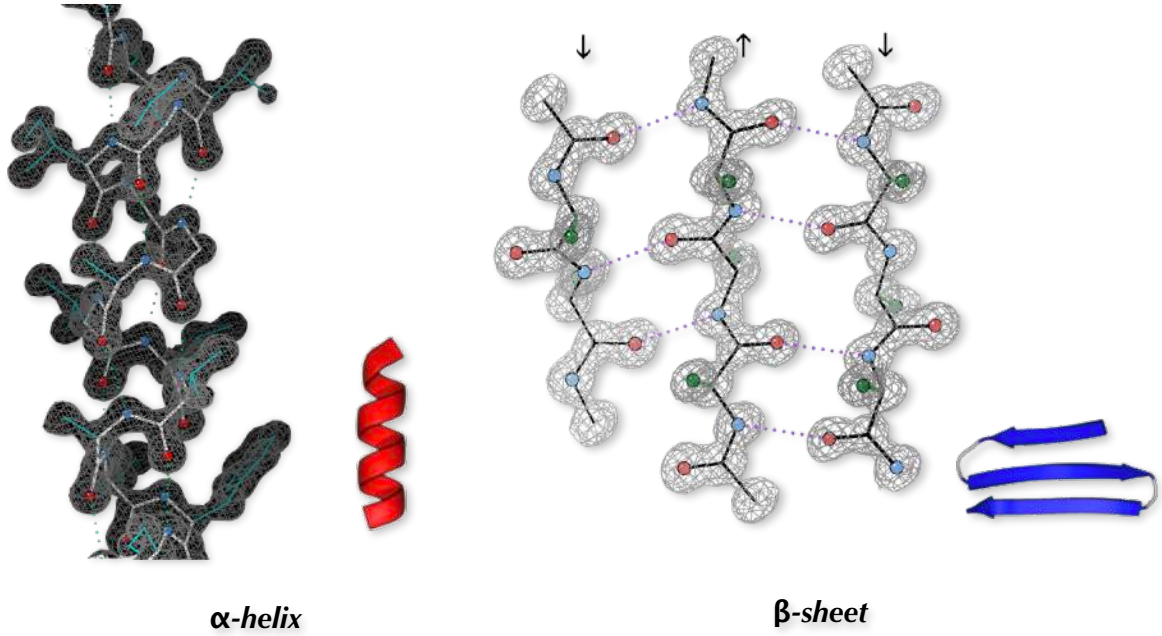


Figure 1.5 Stick figures within electron density and cartoon representations of α-helix (left) and β-sheet (right) structures. The dotted lines in each structure represent the pattern of hydrogen bonds. [adapted without permission from *Wikipedia*]

The two structural elements mentioned above are not the only ones that can be found in protein structures. Variations of them, or even completely different

conformational patterns, are parts of the secondary structure with potential significant function. Such elements include the α_L-helix, 3_{10}-helix, π-helix, β-turns and even random coils. The latter, in spite of lacking periodicity on first sight, is usually hydrophilic and can be found mainly on proteins' surface taking part in protein-protein interactions, formation of enzymes active sites and other functions.

1.5 Torsion Angles

The identification of hydrogen bonding pattern is not enough in order to strictly define a secondary structure element. Understanding the basic parameters that determine the configuration of a polypeptide chain is vital for analysing thoroughly a protein's structural motif[23]. Since peptide groups are uncharged, inflexible planes, due to the rigid nature of the amide bond, they only have two degrees of freedom corresponding to the torsions between N-Cα and Cα-C′ bonds[2,23]. These dihedral torsion angles are called *phi* (φ) and *psi* (ψ) respectively (**Figure 1.6**), and they are considered to be among the most important local structural parameters by which we can define the structure of a protein's backbone.

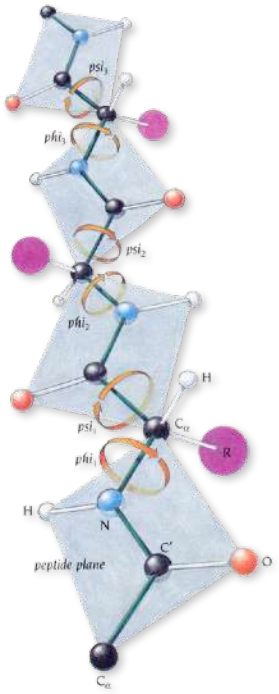


Figure 1.6 A diagram showing the planar peptide groups. While peptide bonds are rigid, the conformation of the main chain atoms is determined by the φ, ψ values of each amino acid. [adapted without permission from *Branden & Tooze, Introduction to Protein Structure*]

Both φ and ψ dihedral angles can span from -180º to 180º degrees. However, most combinations of φ and ψ values are energetically and stereochemically unfavourable and thus not permissible, due to short contacts between the side chain and main chain atoms[2]. In fact, the only residue that has clearly a broad range of allowed φ, ψ combinations is glycine; having only one hydrogen atom as a side chain, it can adopt a much wider range of configurations[2]. Glycine thus plays a quite significant structural role, as it allows many and unusual main-chain protein conformations.

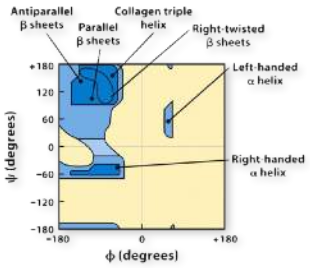


Figure 1.7 A typical Ramachandran plot showing the low-energy regions (or the so-called "allowed" regions). [adapted without permission from *Nelson & Cox, Lehninger Principles of Biochemistry*]

Torsion φ and ψ angle pairs are usually plotted against each other as dots in a two-dimensional diagram called the Ramachandran plot[2]. These diagrams proposed by the indian biophysicist G. N. Ramachandran after he calculated all the possible sterically allowed amino acid conformations[2,24]. As seen in **Figure 1.7**, the Ramachandran plot contains high-density distinct regions called "the low-energy regions", that correspond to the allowed torsion angle values of the residues that form the secondary structure elements. As a result, these plots are a convenient representation to clearly distinguish three major regions equivalent to the three major secondary structure elements: the α-helix region (lower

left quadrant), the β-sheet region (upper left quadrant) and the α_L-helix region (upper right quadrant) (in more detailed Ramachandran plots, even more secondary structure elements can bee seen, other than the above mentioned).

1.6 The β-turn Structure

β-turns are classified as a type of secondary structure elements, but unlike helices and sheets, they constitute a non-repetitive structural pattern[25,26]. Regarding their biological role, turns have a significant role in proteins providing a connection between different secondary structure elements and a direction change for the polypeptide chain, and involving in molecular recognition and protein folding[25]. All β-turns contain four residues (i to i+3) and are classified into categories based on the values of their φ and ψ angles for the second and third residue.

Venkatachalam first proposed these secondary structure elements back in 1968. While studying favourable conformations of three consecutive peptide units, he recognised three distinct conformations (I, II and III) and their main-chain mirror images (I′, II′ and III′)[27,28]. A few years later, Lewis *et. al.* (1973) broadened the number of β-turns to ten (I, I′, II, II′, III, III′, IV, V, VI and VII) defining not only φ, ψ values, but also less stringent criteria[26]. What is now widely accepted, though, is based on the work of Richardson (1981) and Thornton *et. al.* (1988, 1994) who studying the values of φ and ψ angles, defined seven categories (I, I′, II, II′, VIa, VIb and a miscellaneous category IV) and a new class of β-turn, type VIII[25,27]. The conformations of the four most common β-turn types, I, I′, II and II′, can be seen in **Figure 1.8**. The

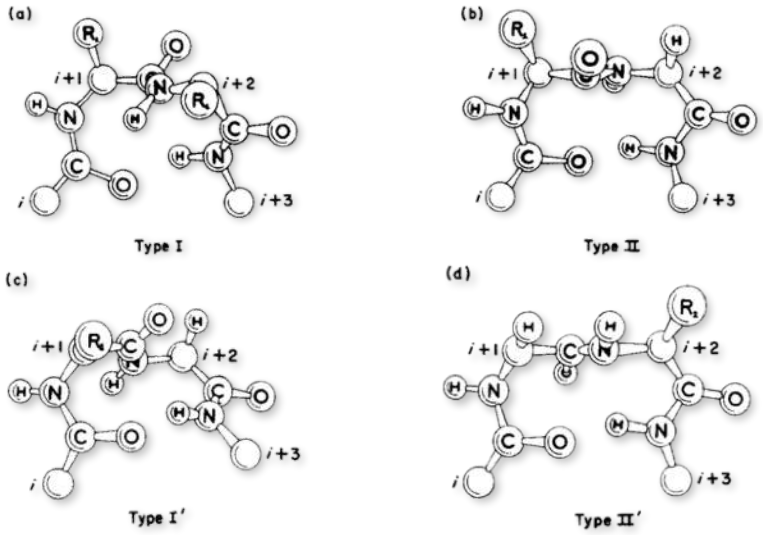


Figure 1.8 Schematic representations of type I (a), type II (b), type I′ (c) and type II′ (d) β-turns ideal conformations. R_1 and R_2 indicate the Cβ position of a side-chain. [adapted without permission from *Wilmot & Thornton, J. Mol. Biol., 1988*]

Ramachandran nomenclature in **Figure 1.9** illustrates the regions of the Ramachandran plot occupied by residues *i*+1 and *i*+2 depending on their φ and ψ angles.

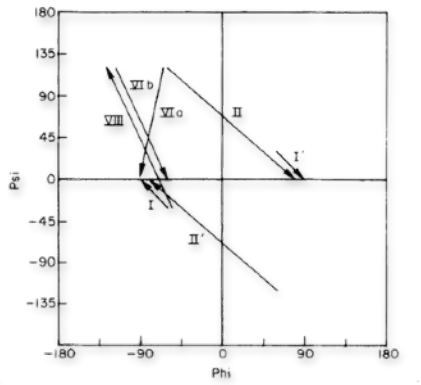


Figure 1.9 Schematic diagram showing a Ramachandran plot of classic β-turn types. [adapted without permission from *Wilmot & Thornton, J. Mol. Biol., 1988*]

Regarding their amino acid sequences, mostly hydrophilic residues are preferred, as β-turn structures are usually exposed in the outer surface of proteins. Glycine, proline, asparagine and aspartic acid exhibit a significant preference in β-turns[25]. Hydrogen bonds occurring by certain side-chains, also play an important role in stabilising β-turn structures and thus in the positional potential of each residue[25]. More specifically, proline is highly favoured at the first two positions of turns, while glycine is preferred at positions *i*+2 and *i*+3, and asparagine and aspartic acid are strongly preferred at positions *i* and *i*+2[25]. Hydrophobic residues can also be present in β-turns, especially in types I′ and II′ that frequently occur in β-hairpin structures (a structural motif where two antiparallel β-strands are connected with a turn)[25].

1.7 Purpose of Present Thesis

The main idea for this project came from a Kang & Yoo publication in which they attempted to examine whether the Asn-Gly segment promotes the formation of β-turns and β-hairpins. The propensities of certain Asn-Gly containing peptides (a tripeptide and three heptapeptides derived from X-ray structures) to form β-turns and β-hairpin structures were explored using the quantum mechanical density functional methods and the implicit solvation model in both water and CH_2Cl_2[29].

Considering the fact that quantum mechanical methods are by far more expensive than molecular mechanical calculations from a computational perspective, we attempted to study the propensities of the same Asn-Gly heptapeptides to form β-turn structures using molecular mechanical simulations and, by extension, comment on a burning question: do molecular mechanics simulations provide sufficiently good results compared to the “high quality”, but still computationally rigorous, quantum mechanical calculations?

In brief, three Molecular Dynamics simulations of 5 μs each in explicit water solvent were carried out for three heptapeptides containing the Asn-Gly segment, in order to study their structural properties. But, before we proceed to the core of this thesis, it is necessary to explain first some basic and useful concepts of molecular modelling.

2. Molecular Modelling & Molecular Dynamics Simulations

2.1 Preface to Molecular Modelling

Molecular modelling encompasses all methods, theoretical and computational, used to model or mimic the behaviour of molecules. Since all molecular modelling methods aim in an atomistic level description of the molecular system, they take advantage of physics and computer graphics instead of deducing microscopic behaviour directly from experiment. There are two common ways in specifying the atomic positions of a biological system. The most straightforward approach is to define the Cartesian (x, y, z) coordinates of all atoms, whereas the alternative is to use the internal coordinates, in which the position of each atom is described in relation to other atoms in the system[30]. The latter approach, is commonly used in quantum mechanics programs, whereas the other is quite useful in molecular mechanics calculations[30].

2.2 An Introduction to Computational Quantum Mechanics

Before we emphasise the typical methods used in this project, it is better to make a brief description of the basic elements of those quantum mechanical methods used by Kang & Yoo in their work. The starting point for every discussion concerning quantum mechanics is Schrödinger's equation, the full, time-depended form of which is

$$\left\{-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)+U\right\}\Psi(r,t)=i\hbar\frac{\partial\Psi(r,t)}{\partial t} \qquad (2.1)$$

The above equation refers to a single particle (e.g. an electron) of mass m, which is moving through space (given by a position vector $r=x\mathbf{i}+y\mathbf{j}+z\mathbf{k}$) and time ($t$) under the influence of an external field U (e.g. an electrostatic potential due to the nuclei of an atom). $\hbar$ is the Planck's constant divided by 2π and i is the square root of -1. Ψ is the wavefunction which describes the particle's motion. When the external potential U is independent of time, the wavefunction can be written as the product of a spatial part and a time part[30,31], reducing the equation finally to

$$\hat{H}\Psi = E\Psi \qquad (2.2)$$

To solve the Schrödinger equation that is operated upon by the Hamiltonian, it is necessary to find values of E and function Ψ such that, it returns the wavefunction

multiplied by the energy. In other words, E and Ψ act as eigenvalue and eigenfunction, respectively[30].

Solving the Schrödinger equation, however, for atoms with more than one electron - not to mention polyelectronic atoms or even molecules - is complicated by a number of factors[30]. Any solutions that might be found for such systems can only be approximations to the real and not the exact, true solutions of the equation[31]. One complication is that the most general form of the wavefunction that describes the properties of the system will be an infinite series of functions, indicating that there is no form more "correct" than another[31]. Another problem is that in multi-electron systems it is necessary to account for electron spin[31]. It is possible, though, to solve the equation concentrating only on the electronic motions, ignoring the motions of the nuclei[30,31]. The masses of the nuclei are considerably greater than the masses of the electrons, meaning that the electrons can be fixed almost directly to any changes in the positions of the nuclei[30]. This is called the Born-Oppenheimer approximation, and the total wavefunction of the molecule can be transformed in the following form, leading to an "electronic" Schrödinger equation:

$$\Psi_{tot}(nuclei, electrons) = \Psi(electrons)\Psi(nuclei) \tag{2.3}$$

$$\hat{H}_{electrons}\Psi_{electrons} = E_{electrons}\Psi_{electrons} \tag{2.4}$$

However, the electronic Schrödinger equation is still complex and further approximations are required. One approach is to consider that the electrons move independently of each other, assuming for each electron an average field of all other electrons[31]. The set of the molecular orbitals corresponding to the lowest energy is obtained by a process called "self-consistent-field" or SCF procedure, the archetypal form of which is the Hartree-Fock procedure[31]. SCF methods also include density functional procedures. Molecular orbitals then are being transformed into linear combinations of a finite set of basis functions (Linear Combination of Atomic Orbitals or LCAO approximation)[31]. The Hartree-Fock and LCAO approximations are finally applied to the electronic Schrödinger equation, generating the Roothaan-Hall equations, the solutions of which are termed as Hartree-Fock models or *ab initio* models[31]. The term *ab initio* ("from the beginning") models, applies generally to all models arising from "non-empirical" methods to solve the Schrödinger equation.

The most significant drawback of Hartree-Fock theory is that it fails to adequately describe electron correlation[31]. In fact, electrons' motions are correlated and they tend to normally avoid each other, something that cannot be represented using the average potential of the other electrons mentioned above. This can result to an overestimation of the electron-electron repulsion energy and thus, to an over-increased total energy[31]. One approach to treat the electron correlation is referred to as density functional theory and it is based on the electron density, contrary to the traditional wavefunction-based approaches. Density functional models are well-defined and result in successful

determination of equilibrium conformations and geometries, but they are applicable only to molecules of moderate size (50-100 atoms)[31].

It is needless to delve into more approaches of quantum mechanical calculations. Quantum mechanics is often considered to be a difficult subject and the underlying physical and mathematical background is beyond the limits of this project. It is preferable to move on to the basics of statistical mechanics and molecular dynamics simulations, upon which our study was based.

2.3 Statistical Mechanics

Molecular mechanics describe the behaviour of the system based on nuclear positions only as a function of time, contrary to quantum mechanical methods which deal with the electrons in a system[30]. This obviously means that molecular mechanics methods are computationally more efficient and less time-consuming, making them even applicable to large biological systems. Molecular mechanical calculations give rise to two major categories of simulations, Molecular Dynamics simulations (MD simulations), a useful tool for theoretical studies of the dynamic properties of a system, and Monte Carlo simulations (MC simulations), a computerised mathematical method based on statistical and probabilistic methods (the aforementioned methods are often used separately, although occasionally a combination of them is preferred in aid of computationally expensive and complex simulations - Langevin dynamics, Brownian dynamics)[32]. Usually MD simulations are convenient for studying the folding and stability of proteins, molecular recognition, and ion transportation in biological systems, as well as for drug design and structure determination.

The results obtained from MD simulations describe the behaviour of the system in an atomic level using parameters such as position and velocity[33]. Although this allows us to study many-body systems, not all properties can be directly measured in a simulation[33,34]. In fact, most of the properties that can be measured in a simulation cannot be compared with the experimental data, as no real experiment provide such detailed information[33,34]. A typical experiment measures average properties, rather than the instantaneous kinetic parameters of each atom. It is what kind of averages then we should aim to measure, if we wish to use computer simulations as the numerical equivalent of experiments[33,34]. In order to do so, we need to introduce the “language” of statistical mechanics. Statistical mechanics attempt to predict the macroscopic behaviour of a system based on the properties of each atom separately. In other words, averages corresponding to the experimental observables are defined in terms of, what we call, ensemble averages. An ensemble is a collection of all

possible systems which have different microscopic states, but an identical macroscopic or thermodynamic state. This connection between microscopic and macroscopic information can be achieved by means of a complex mathematical background[34].

2.4 Classical Mechanics and Integration Algorithms

MD simulations are based on Newton's second law. Knowing the force exerted on each atom, it is possible to determine the acceleration of each atom in the system. Integration of the equations of motion then yields a trajectory that describes the kinetic parameters of particles, such as positions, velocities and accelerations as they vary with time. Analysing this trajectory then can give us the opportunity to measure the average properties of the system. The method is deterministic; once the positions and velocities of the atoms are known, the state of the system can be predicted at every possible time, in the future or the past.

Describing the above said with mathematical terms, Newton's equation of motion is given by

$$F_i = m_i a_i \tag{2.5}$$

where F_i is the force exerted on a given particle i, m_i is the mass of the particle i and α_i is the acceleration of that particle. The force can also be expressed as the gradient of the potential energy

$$F_i = -\nabla_i V \tag{2.6}$$

The combination of equations (2.5) and (2.6) leads to the following equations

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \tag{2.7}$$

$$a_i = -\frac{1}{m_i}\frac{dV}{dr_i} \tag{2.8}$$

It is therefore possible to calculate a trajectory knowing only the initial positions of the atoms, a distribution of velocities and accelerations, determined by the gradient of the potential energy function. The initial positions can be obtained from experimental structures, resolved by techniques such as X-ray Crystallography and/or NMR Spectroscopy. The velocities, v_i, are usually obtained randomly from a Maxwell-Boltzmann or Gaussian distribution at a given temperature

$$p(v_{ix}) = \left(\frac{m_i}{2\pi k_B T}\right)^{1/2} \exp\left[-\frac{1}{2}\frac{m_i v_{ix}^2}{k_B T}\right] \tag{2.9}$$

The temperature can be calculated from the velocities using the relation

$$T = \frac{1}{(3N)} \sum_{i=1}^{N} \frac{|p_i|}{2m_i} \tag{2.10}$$

where N stands for the number of the atoms in the system. The potential energy, as we can see, is a function of the atomic positions ($3N$). Due to the complexity of this function, the equations of motion can only be solved numerically, but not analytical. That is to say, there is no exact solution to the equations of motion, either because of the complicated nature of the potential energy function, or because the computational demand is high. Therefore, an approximately numerical method is used developed upon the integration of the equations of motion. The most notable integration algorithms include the Verlet algorithm, the Leap-frog algorithm, the Velocity Verlet and the Beeman's algorithm.

In general, most of the algorithms are based on Taylor expansions, the usefulness of which lies on the reduction of an equation's terms, as they represent a function of a finite sum of terms that are calculated from the values of the function's derivatives at a single point:

$$r(t+\delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \ldots \tag{2.11}$$

$$v(t+\delta t) = v(t) + a(t)\delta t + \frac{1}{2}b(t)\delta t^2 + \ldots \tag{2.12}$$

$$a(t+\delta t) = a(t) + b(t)\delta t + \ldots \tag{2.13}$$

It is important to mention that the choice of the algorithm should be done wisely, taking into account the algorithm's computational efficiency, its capability to conserve both the energy and the momentum of the system and its integration time step, so as the results to be as accurate as possible[34].

2.5 Empirical Force Field Models

Most of the current generation potential energy functions (or force fields) provide a reasonably good agreement between accuracy and computational efficiency and thus, with the experimental data[34]. The development of a potential energy function is an extremely arduous task requiring extensive optimisation, making the construction of force fields an area of advancing research. Among the most commonly used force fields are the AMBER[35], CHARMM[36], GROMOS[37] and OPLS/AMBER[38].

So, force fields are empirical functions used for the calculation of the system's potential energy regarding particles' interactions and positions. These interactions are

expressed in terms of internal, or bonded, interactions and external, or non-bonded, interactions

$$V(R) = E_{bonded} + E_{non-bonded} \tag{2.14}$$

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

$$E_{bonded} = E_{bonds} + E_{angles} + E_{torsions} \tag{2.15}$$

The first term of the above formula is a harmonic potential representing the interaction between two atoms bonded with one covalent bond (**Figure 2.1a**)

$$E_{bonds} = \sum_{1,2\,pairs} K_b (b - b_0)^2 \tag{2.16}$$

The energy of the bond is a function of the displacement from the ideal bond length, b_0. K_b is the force constant determined by the bond's valance and represents the strength of the bond. Both of the above components are determined by the chemical type of each atom and thus, they are specific for each pair of bound atoms. The second term in equation (2.15) is also a harmonic potential referred to the variation of a bond angle from the ideal value θ_0 (**Figure 2.1b)**

$$E_{angles} = \sum_{angles} K_\theta (\theta - \theta_0)^2 \tag{2.17}$$

Values θ_0 and K_θ also depend on the chemical type of each atom, being specific for each pair of atoms constituting the angle. The last term of equation (2.15) calculates the potential energy of the system as a function of the rotations of dihedral angles (**Figure 2.1c**). This potential is considered to be periodic and is often expressed as a cosine function

$$E_{torsions} = \sum_{1,4\,pairs} K_\varphi \left(1 - \cos(n\varphi)\right) \tag{2.18}$$

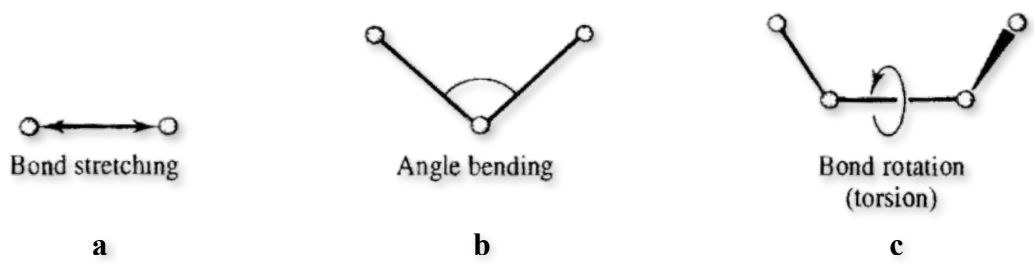


Figure 2.1 Schematic representation of the three basic internal contributions to a molecular mechanics force field: (a) bond stretching, (b) angle bending and (c) bond rotation. [adapted without permission from *Leach, Molecular Modelling*]

The energy term representing the non-bonded interactions in the potential function has two components, the van der Waals interaction energy and the electrostatic interaction energy

$$E_{non-bonded} = E_{van-der-Waals} + E_{electrostatic} \tag{2.19}$$

The van der Waals interaction between two atoms is the result of a balance between attractive and repulsive forces. As shown in **Figure 2.2**, there is a specific distance in which the potential energy reaches a minimum. This is the equilibrium distance and it depends on the chemical type of the atoms, being approximately equal to the sum of the Van der Waals radii of the atoms. If the distance between the atoms becomes shorter, the repulsive force becomes dominant due to the electron distributions interactions, whereas if the distance increases, the attractive force becomes dominant as the electron cloud of one atom gives rise to an instantaneous dipole, inducing therefore a dipole in another atom[34]. The van der Waals interactions are often represented using the Lennard-Jones 6-12 potential

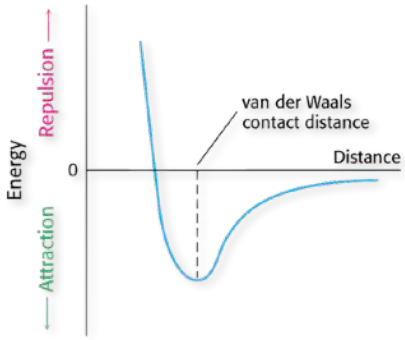


Figure 2.2 Energy of van der Waals interaction as two atom approach each other. The energy is most favourable at the van der Waals contact distance. [adapted without permission from *Stryer, Biochemistry*]

$$E_{van-der-Waals} = \sum_{\substack{nonbonded \\ pairs}} \left(\frac{A_{ik}}{r_{ik}^{12}} - \frac{C_{ik}}{r_{ik}^{6}} \right) \tag{2.20}$$

The other component of equation (2.19), the electrostatic interaction energy, is nothing more than the representation of Coulomb potential

$$E_{electrostatic} = \sum_{\substack{nonbonded \\ pairs}} \frac{q_i q_k}{D r_{ik}} \tag{2.21}$$

Despite the fact that empirical force field models - at least the ones mentioned above - share a common calculation methodology, there are differences relating to the calculations of bonded and non-bonded parameters. Consequently, their development remains an intense area of research in order to achieve even more agreement with the experimental data[34].

Last but not least, another important issue that should be considered is that the potential energy function does not include entropic contributions. This directly means that the minimum of the potential energy does not correspond to the equilibrium state. Due to the fact that experiments are generally carried out under constant temperature, pressure and space, the equilibrium state corresponds to the minimum of Gibbs free energy, G[34].

2.6 Role of Solvent in Molecular Dynamics Simulations

The use of solvent, usually water, in MD simulations has an essential influence on the structure and dynamics of biological macromolecules. One of the most important roles of solvent is the screening of electrostatic interactions[34]. Using a water solvent model, simulations and thermodynamic calculations can be applied to procedures that take place in a solution, as analogous to the cellular environment[39,40].

There are two ways to include solvent effects in an MD simulation. The first treatment is to include a dielectric constant in the electrostatic term of the potential energy function, instead of including explicitly solvent molecules[34]. This is known as the *implicit* treatment of the solvent and it provides an approximate description of the solvent behaviour. The other treatment is the *explicit* solvent model. Contrary to the implicit models, every solvent molecule is included explicitly providing a more realistic concept similar to real experiments. Apparently, this approach acquires the imposition of boundary conditions (**Figure 2.3**), in order first to prevent the diffusion of solvent molecules, and second, to calculate macroscopic properties using a limited number of solvent molecules. The imposition of periodic boundary conditions treats the system in such a way that the molecule under study is placed in a central box, the primary box, surrounded by eight neighbour boxes which are actually copies of the primary one. Every atom can interact with the neighbouring ones either in the central box, or in the surrounding boxes. Thus, if an atom leaves the primary box, then the equivalent atom from an antiparallel cell enters the primary box maintaining the periodicity of the system[34,41]. Other approaches of solvation treatment include the surrounding of the protein or just a part of a protein with a sphere of water, reducing the computational cost[34].

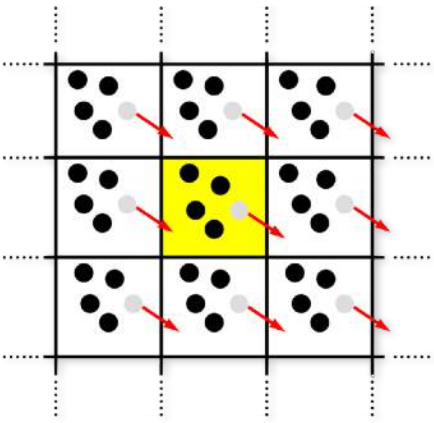


Figure 2.3 2D schematic representation of periodic boundary conditions. The box in yellow is the primary box. [adapted without permission from *Steinhauser & Hiermaier, Int. J. Mol. Sci., 2009*]

3. Preparation of the System and Methods

3.1 Technical Characteristics of our Simulations

In order to study the propensities of the Asn-Gly segment to form β-turn structures we conducted three MD simulations considering three capped heptapeptides: Ac-Ala-Ala-**Asn-Gly**-Ala-Ala-NHMe (hp_{NG}-1), Ac-Leu-Val-**Asn-Gly**-Gln-Tyr-NHMe (hp_{NG}-2, from PDB entry 1EST.ent) and Ac-Phe-Val-**Asn-Gly**-Leu-Phe-NHMe (hp_{NG}-3, derived form an octapeptide with the similar sequence Boc-Leu-Phe-Val-Aib-$_D$-Ala-Leu-Phe-Val-OMe that forms a type I' Aib-$_D$-Ala β-turn)[29]. The NAMD[42] software and the AMBER 99SB-STAR-ILDN force field were used for the peptides' simulations.

Each of the simulations 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, based on Intel's Q6600 Kentsfield 2.4 GHz quad processors and 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, with the exception of one node 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 an Nvidia GTX-460 GPU. The head node exempts from the others as it comes with four cores, 8 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 Nvidia GTX-260 GPU. Norma is presently used almost exclusively by the group of Structural and Computational Biology in Democritus University of Thrace[43].

3.2 Starting a Simulation with NAMD

Before starting the simulation, NAMD requires the following files:

- A PDB file containing the initial coordinates of the molecular system. PDB files are either accessible through the PDB database (http://www.rcsb.org/pdb/), or they can be created by the user. These files include information about atoms' number and type, residues' name and number, atomic coordinates, occupancies and temperature factors.
- The customised parameter and topology files of a compatible force field required for the calculation of the system's potential energy. In this case, the AMBER 99SB-STAR-ILDN force field was used and the parameter files were created with the LEaP program.

- An NAMD configuration file which includes the dynamic options and values that NAMD should use, in order to control exactly how the system will be simulated. The configuration file includes specific information, such as which options are enabled or disabled, the number of timesteps that must be performed, initial temperatures, etc[42]. The configuration file that was used for the three simulations of our heptapeptides can be found in the **Appendix** (**A1**).

The steps of our simulations are presented in detail in the next chapter below.

3.3 System Preparation and Simulations

The first step was the preparation of the system in each of the simulations. The PDB files for the initial extended structures of heptapeptides were generated. The peptides were solvated explicitly each one in a rectangular box of TIP3P water model. The final systems contained 1572 atoms for hp_{NG}-1, 1584 atoms for hp_{NG}-2 and 1997 atoms for hp_{NG}-3, within a box dimension of 27 Å.

Prior to the beginning of the simulations, the systems were minimized so as to remove any strong van der Waals interactions that may lead to an unstable simulation and/or a structural distortion. The entire boxes of water were then overlaid onto the proteins and the water molecules that overlapped the proteins were removed. Another energy minimisation is also necessary resulting in the fixation of the proteins' positions in their energy minima, readjusting simultaneously the water molecules to the protein molecules. Subsequent to this step is the heating phase, during which initial velocities at a low temperature are assigned to each atom and the simulation begins. Periodically, new velocities are assigned at a slightly higher temperature and the simulation continues. The procedure is repeated until the ideal temperature is achieved. In our systems the temperature was increased with a ΔT step of 20 K until the final desired temperature. This was followed by an equilibration period for the production of the NpT runs with both temperature (320 K) and pressure (1 atm) controlled using the Nosé-Hoover Langevin Dynamics and Langevin piston barostat control method, as implemented by NAMD. During this period several properties of the system are observed until they become stable with respect to time. For the production run the Verlet-I multiple time step integration algorithm was used. The inner time step was 2.5 fs. The long-range electrostatic interactions were calculated using the particle-mesh Ewald treatment. The van der Waals interactions were cut off at 9 Å and the covalent bond lengths were constrained using the SHAKE algorithm. Finally, the adaptive tempering method was implemented on the simulations and the temperature ranged between 280 K and 480 K.

Trajectories were obtained by saving the atomic coordinates of the whole systems every 0.4 ps. Each of the simulations had a total duration of 5 μs and resulted in approximately 5,000,000 frames.

3.4 Analysis of the Simulations and Programming Languages

In order to process our data and results for analysing the trajectories, we used mainly the Perl programming language. Perl was originally developed by Larry Wall in 1987 and is a high-level, multi-purpose, interpreted, dynamic programming language[44]. Although Perl is rather slow compared to other compiled programming languages such as C, yet it is among the most popular Unix scripting languages used in Bioinformatics and Computational Biology, in part because of the regular expression and string parsing abilities. The R statistical package[45], an open source programming language and environment suitable for statistical analyses and graphics, was used as well for plotting part of the data. Lastly, the fact that UNIX systems are used in the laboratory, gives us the advantage of using bash scripting and languages such AWK, for the automation of our work and quick process of our data.

The overall workflow that was followed for this project is stated below:

- Construction of the free energy landscapes of the heptapeptides via dihedral angle Principal Component Analysis and cluster isolation.
- Structural analysis of the isolated clusters taking into account 500 equally spaced structures from each cluster, using the *promotif*[46] program. Instead of analysing each structure separately we divided those 500 structures of each cluster into eight groups of approximately 65 structures each (**source code: Appendix, Script 01**). The next step was to identify the population of each turn type observed in each cluster (**source code: Appendix, Script 02**).
- Structural analysis of the representative structures of each cluster.
- Temperature based analysis of the trajectories.
- Construction of the PDB files of the *ab initio* structures from Kang & You publication using their XYZ files (**source code: Appendix, Script 03**) and conversion into DCD files.
- RMSD analysis of the trajectories using as reference structures the aforementioned *ab initio* structures.
- Construction of hierarchical dendrograms using the UPGMA algorithm (**source code: Appendix, Script 04**) and cluster isolation to find conformations similar to the experimental data.
- RMSD and population analysis of the isolated clusters in order to identify which clusters denote similarity with the *ab initio* models.

4. Results

4.1 Introduction

In this part of the thesis, we are going to analyse the trajectories derived from the MD simulations we performed on the three heptapeptides: Ac-Ala-Ala-**Asn-Gly**-Ala-Ala-NHMe (hp_{NG}-1), Ac-Leu-Val-**Asn-Gly**-Gln-Tyr-NHMe (hp_{NG}-2) and Ac-Phe-Val-**Asn-Gly**-Leu-Phe-NHMe (hp_{NG}-3)[29]. As it can be seen, the peptides are capped with an acetyl group at the N-terminal end, and an *N*-methylamide at the C-terminal end. Termini capping is a common process in computational analysis of proteins. N-terminal acetylation and C-terminal amidation reduce the overall charge of the peptide, increasing its overall solubility and thus leading to a closer mimic of the native state.

The data analysis of the simulations was carried out mainly by *carma*[47] and its GUI program *grcarma*[48]. This program requires as input a pair of a DCD/PSF files. DCD file constitutes a binary file that contains the trajectory produced by the simulation. Each set of coordinates including in the DCD file corresponds to one frame at a time. The PSF file (Protein Structure File) contains all of the atomic-specific information needed to apply a particular force filed to a molecular system. Among the atomic information included (atoms, bonds, angles, etc.), there are also details about atomic charge and mass[49]. Other programs used include *plot*[50], *PyMOL*[51], *VMD*[52] and *promotif*.

4.2 Principal Component Analysis & Clustering

Biomolecular processes such as protein folding and protein function can be described in terms of the molecule's free energy

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

where P is the probability distribution of the molecular system along some (generally multidimensional) coordinate r and P_{max} denotes its maximum, which is subtracted to ensure that ΔG=0 for the lowest Gibbs free energy minimum. The resulting free energy landscape is essential for understanding protein folding[53].

Dihedral angle Principal Component Analysis (dPCA) is a systematic approach, useful for the construction of a low-dimensional free energy landscape from a classical MD simulation[53]. In contrast to Cartesian Principal Component Analysis (cPCA), it is known that the dPCA method provides a more detailed representation of the free energy surface when it comes to studying the conformational dynamics of small peptides. That is because the substantial internal motion can successfully be separated

by the trivial overall motion (translation and overall rotation)[54]. Internal motion, despite being subtle, corresponds to the most important conformational degrees of freedom, thus illustrating molecular structure and interactions, and reducing the complex protein dynamics to its essential degrees of freedom[55].

The dPCA method is based on the dihedral angles (φ_i, ψ_i) of the peptide backbone. The reason for considering only the dihedral angles of a flexible molecule is because they undergo changes of large amplitudes (bearing in mind the scale of the forces that typically influence protein folding), contrary to other internal coordinates such as bond angles and bond lengths[54]. However, dihedral angles are circular and periodic

$$\varphi,\psi \in [0^\circ, 360^\circ] \tag{4.2}$$

unlike regular data, such as Cartesian coordinates, where

$$x \in (-\infty, +\infty) \tag{4.3}$$

which makes the definition of a metric not straightforward and as a result difficult to calculate distances or means. Thus, to recover a metric coordinate space (i.e., a linear vector space for which distances between all members of the set are well-defined), it is necessary to perform a sin- and cos-transformation, by which every angle is represented by its equivalent vector (x,y) on the unit circle[54,56]

$$\varphi \mapsto \begin{cases} x = \cos\varphi \\ y = \sin\varphi \end{cases} \tag{4.4}$$

Following the transformation, the next step is to calculate the mean and the covariance matrix. Through the diagonalization of the covariance matrix we obtain the eigenvectors v_n and the eigenvalues λ_n, which are organised in descending order, i.e., λ_1 represents the largest eigenvalue. The eigenvectors with the largest eigenvalue are the principal components and they tend to contain most of the atomic fluctuations derived from the simulation. Hence, a large part of the system's fluctuations can be described in terms of only a few PCA eigenvalues[54].

The free energy landscapes of our trajectories were constructed using *grcarma* and the dPCA method, and are shown below in **Figure 4.1-4.4**. As an initial step we decided to take into account the dihedral angles of the entire peptides (**Figure 4.1**) and afterwards the dihedral angles of the four central residues only (**Figure 4.2-4.4**). All figures show the free energy (in kcal/mol) as a function of the first three principal components. Following these, a colour-coding cluster representation based on the dPCA of the four central residues made by the *plot* program (**Figure 4.2-4.4)** and a table containing the populations of the most prominent clusters (**Table I**) are presented.

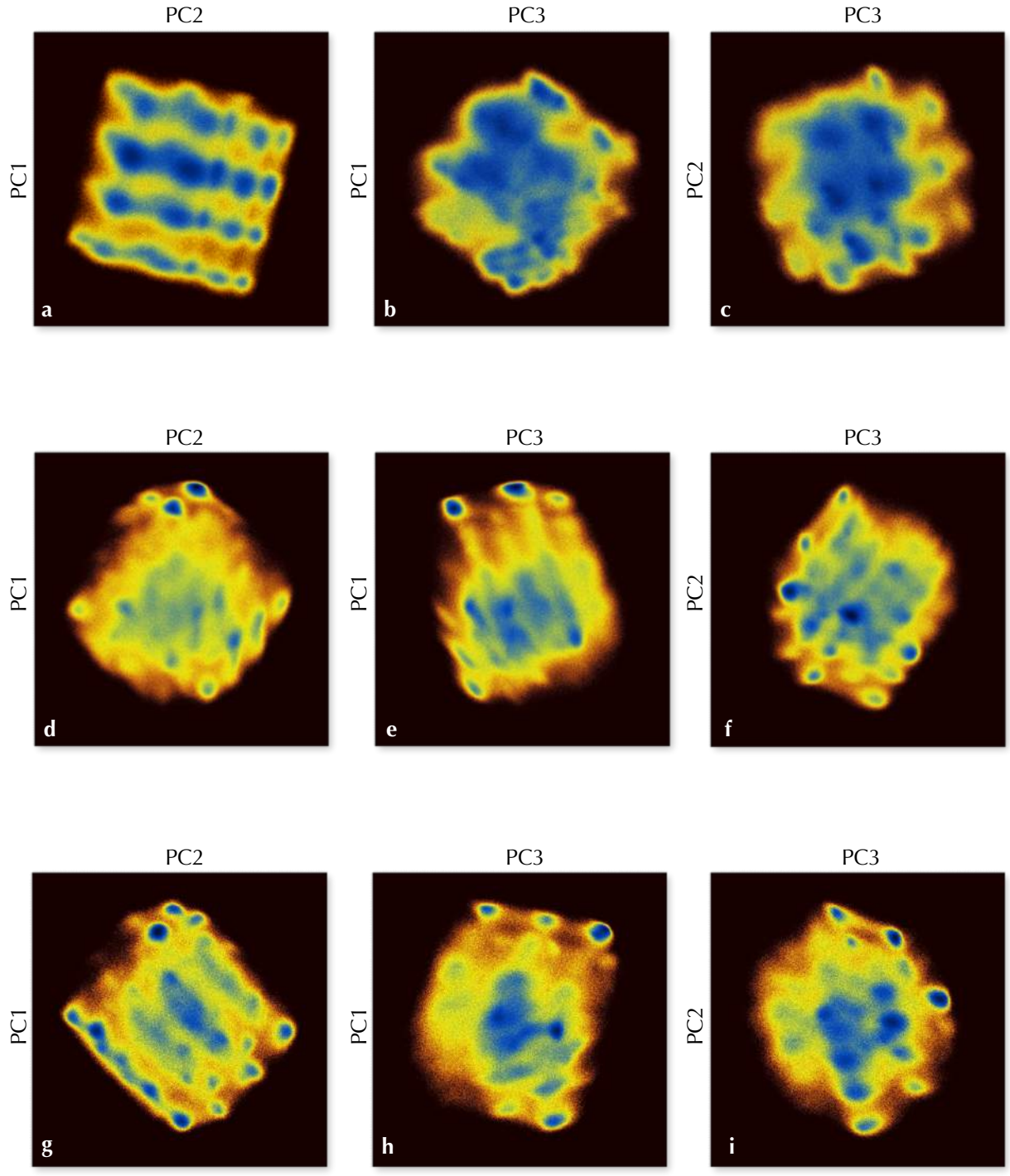


Figure 4.1 Two-dimensional representations of the free energy landscapes as obtained by the dPCA method taking into account the torsion angles of the entire heptapeptides: (a-c) ΔG plots along the first three principal components of hp_{NG}-1, (d-f) ΔG plots along the first three principal components of hp_{NG}-2, (g-i) ΔG plots along the first three principal components of hp_{NG}-3. The blue regions in the diagrams correspond to the various energy minima.

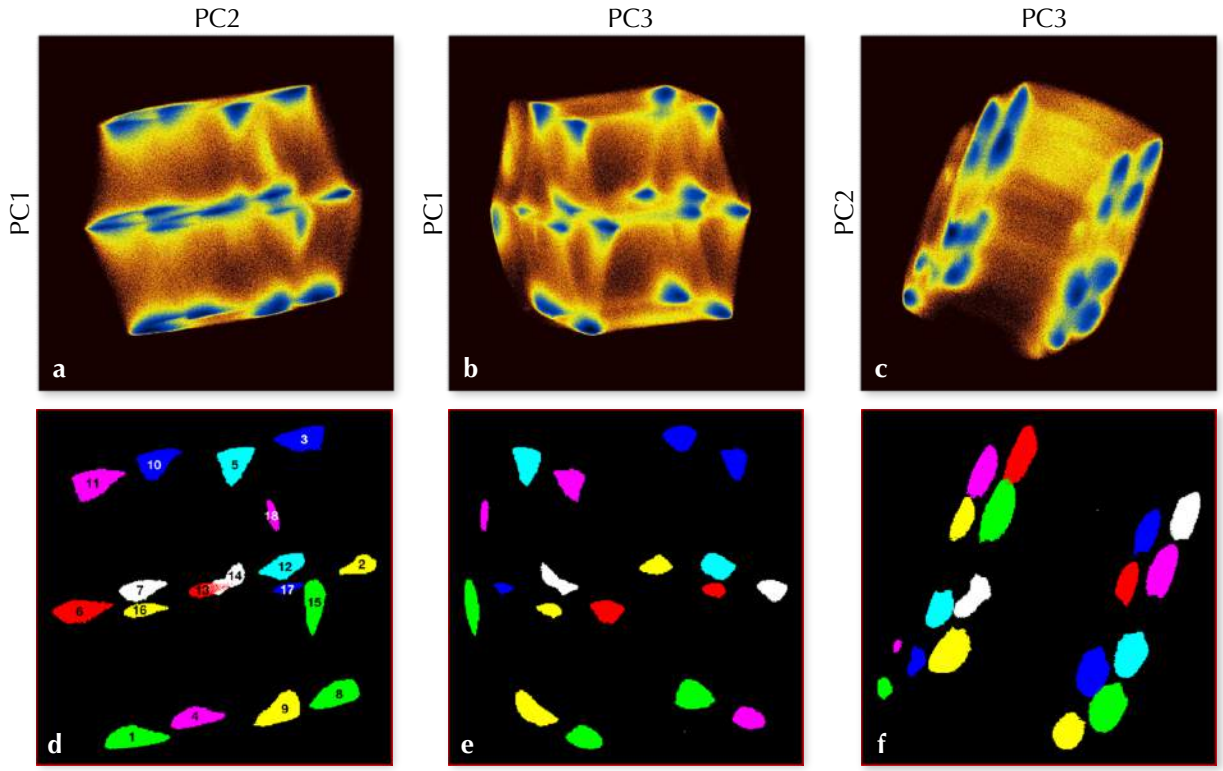


Figure 4.2 Two-dimensional representations of the free energy landscapes as obtained by the dPCA method taking into account the torsion angles of the four central residues of hp_{NG}-1: (a-c) ΔG plots along the first three principal components of the trajectory. The blue regions in the diagrams correspond to the various energy minima. (d-f) Colour-coding panels illustrating the conformational clusters obtained by the dPCA analysis of the four central residues

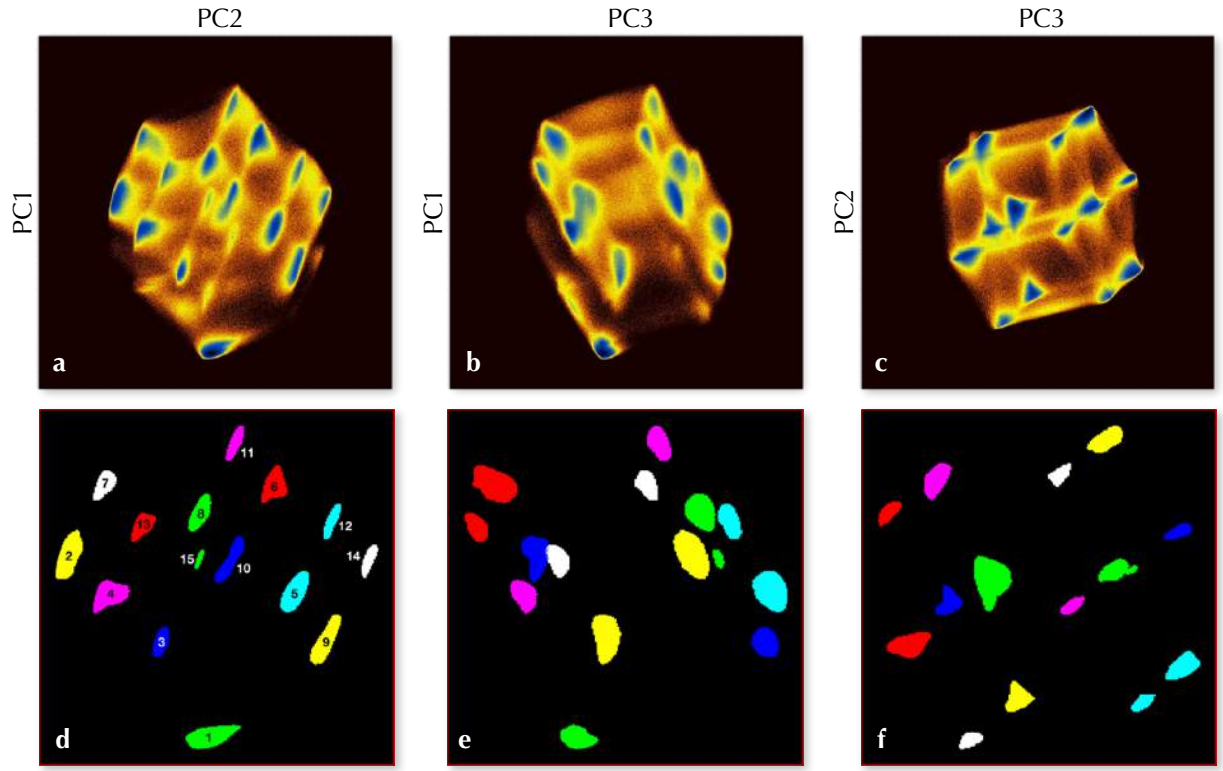


Figure 4.3 Two-dimensional representations of the free energy landscapes as obtained by the dPCA method taking into account the torsion angles of the four central residues of hp_{NG}-2: (a-c) ΔG plots along the first three principal components of the trajectory. The blue regions in the diagrams correspond to the various energy minima. (d-f) Colour-coding panels illustrating the conformational clusters obtained by the dPCA analysis of the four central residues

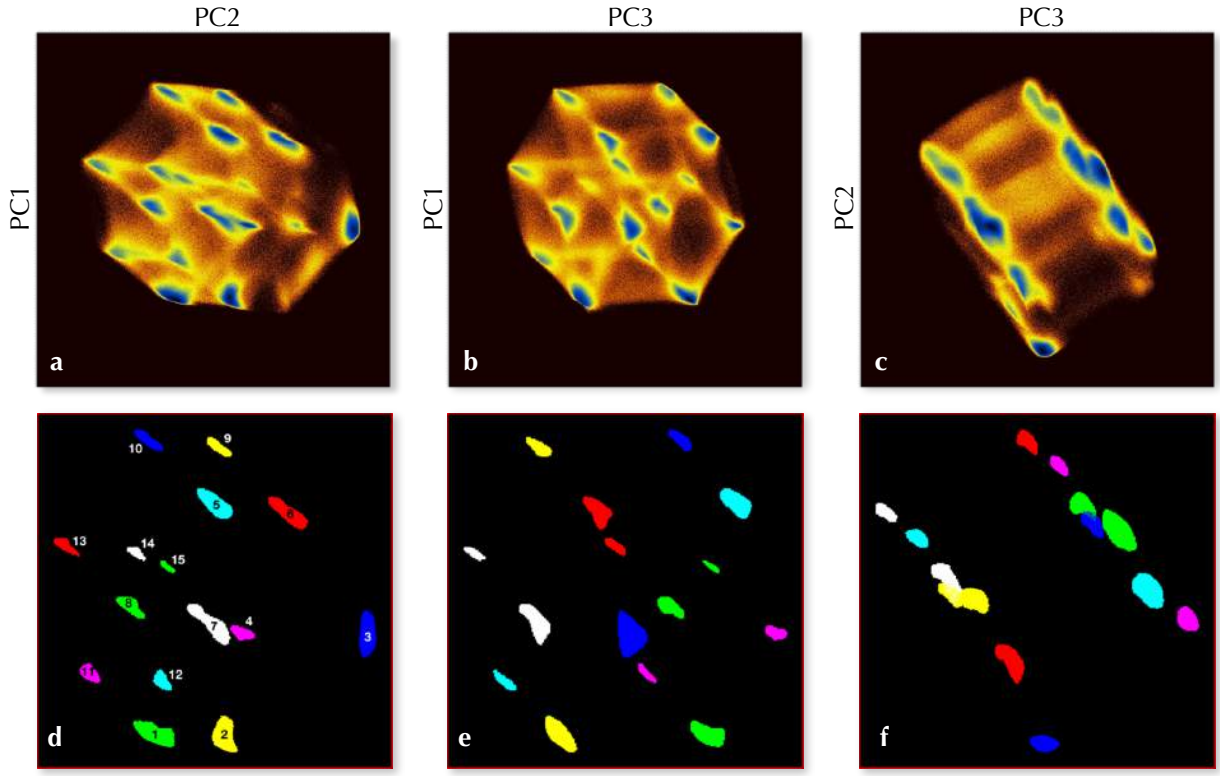


Figure 4.4 Two-dimensional representations of the free energy landscapes as obtained by the dPCA method taking into account the torsion angles of the four central residues of hp_{NG}-3: (a-c) ΔG plots along the first three principal components of the trajectory. The blue regions in the diagrams correspond to the various energy minima. (d-f) Colour-coding panels illustrating the conformational clusters obtained by the dPCA analysis of the four central residues

As it can be seen, the free energy landscapes are quite rugged with several free energy minima that correspond to specific conformational structures. As it was expected, the number of clusters that observed is being reduced as we progressively reduce the amount of dihedral angles included in the dPCA analysis. The decrease in the number of prominent structures comes in agreement with the fact that the peptides' edges present a more dynamic behaviour, whereas the central parts containing the Asn-Gly segment adopt relatively more stable conformations resulting possibly in certain secondary structure motifs with certain patterns of torsion angles and hydrogen bonds.

It is also obvious that the free energy landscapes of the entire peptides (**Figure 4.1**) have a lower signal-to-noise ratio than those of the central residues (**Figure 4.2-4.4**). This explains the fact that many unstable intermediate states being formed during the simulation and promoted by the dynamic peptides' ends, may have quite similar structural patterns with the stable ones, making the energy minima broader without well defined limits, rather than sharp and distinct to each other.

The bottom panels in **Figure 4.2-4.4** illustrate the most prominent clusters visualized upon the free energy landscapes and **Table I** below contains the populations of the most prominent conformational states of the heptapeptides in water as obtained by the dPCA analysis

Table I: Populations of the most prominent clusters of heptapeptides in water.

No. of Cluster	hp_{NG}-1 (full length)	hp_{NG}-2 (full length)	hp_{NG}-3 (full length)	hp_{NG}-1 (central part)	hp_{NG}-2 (central part)	hp_{NG}-3 (central part)
1	7.06	34.44	19.56	11.88	20.10	20.66
2	11.45	25.30	13.02	4.26	14.09	17.38
3	8.34	10.62	11.49	8.65	4.55	16.50
4	19.84	6.50	6.38	7.23	9.54	4.22
5	9.83	4.19	6.25	7.67	10.02	11.33
6	4.87	3.80	7.09	6.32	10.13	6.92
7	0.80	4.36	3.58	5.16	3.24	7.48
8	1.52	3.11	7.96	10.74	5.98	3.87
9	9.60	2.14	1.80	9.40	7.55	2.65
10	0.69	1.20	2.16	5.10	4.39	2.71
11	1.01	1.72	1.54	6.46	2.42	1.40
12	0.82	0.79	2.62	5.97	2.41	1.74
13	1.30	1.52	1.88	2.30	2.85	1.64
14	1.30	0.15	3.53	2.75	2.44	1.08
15	4.59	0.16	4.02	3.01	0.28	0.41
16	1.01		0.90	2.06		
17	0.98		1.18	0.67		
18	0.71		2.79	0.37		
19	4.67		1.05			
20	1.86		0.43			
21	1.41		0.53			
22	0.87		0.12			
23	0.19					
24	1.01					
25	0.33					
26	0.34					
27	0.65					
28	0.15					
29	0.94					
30	0.22					
31	0.16					
32	0.12					
34	0.27					

Table I: Shown are the population probability (among clustered frames in %) for the conformational clusters derived from each dPCA analysis. The table contains only resulting clusters with a population probability greater than or equal to 0.1%.

4.3 Structural Analysis

In order to identify the presence of β-turns in each cluster derived from the Principal Component Analysis we used the *promotif* program. The program provides details of the locations and types of stractural motifs that are present in proteins. Concerning β-turns, these are defined by four consecutive residues (i to i+3) with the distance between the Cα atoms of residues i and i+3 being less than 7 Å and where the central two residues are not helical[25,46]. *Promotif* can potentially assign a structure to one of the nine classes of β-turns based on the φ, ψ angles of the i+1 and i+2 residues. The ideal angles for each β-turn category are shown below in **Table II**. The φ, ψ angles were allowed to vary by $\pm 30^{o}$ form these ideal values, with one angle being allowed to deviate by as much as 40^{o}.

Table II: Ideal angles for β-turn types.

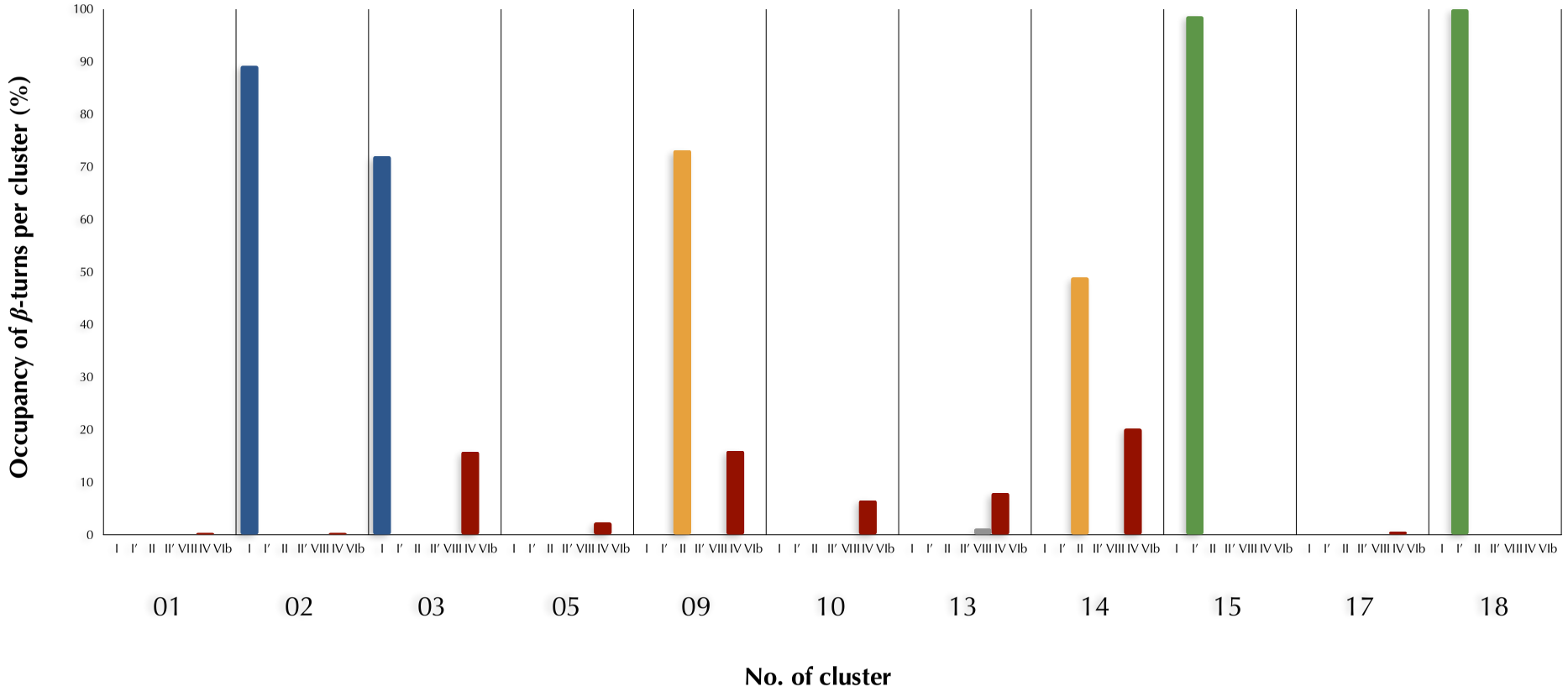
Turn Type	Dihedral Angles (^{o})			
	φ_{i+1}	ψ_{i+1}	φ_{i+2}	ψ_{i+2}
I	-60	-30	-90	0
I′	60	30	90	0
II	-60	120	80	0
II′	60	-120	80	0
VIa1*	-60	120	-90	0
VIa2*	-120	120	-60	0
VIb	-135	135	-75	160
VIII	-60	-30	-120	120
IV	*Turns excluded from the above categories*			

Table II: β-turns are divided into nine categories based on the dihedral angles of the central residues i+1 and i+2.
* VIa1 and VIa2 require *cis*-proline at position i+2 and thus this type is not being studied in this project.

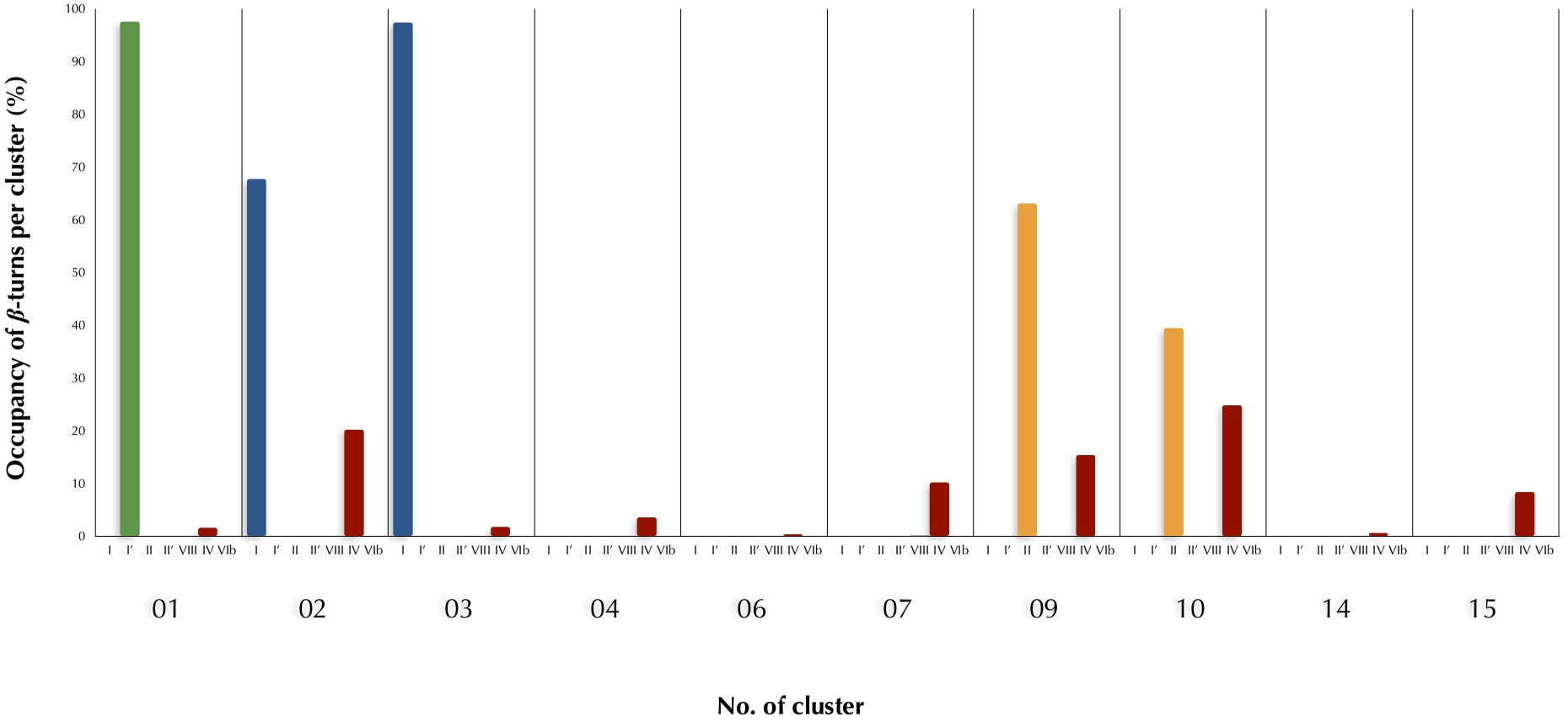
To have an overall view of the β-turns occupancy in each cluster, we studied the presence and the corresponding populations of each β-turn class in every cluster, using 500 equally spaced structures from each cluster, as obtained from the dPCA analysis of the peptides' four-residue central part. The following charts show the occupancy of β-turns among these 500 structures in every cluster for the three heptapeptides. Clusters in which β-turns were not identified are omitted. A more detailed list of the populations of β-turns in each cluster's sample dataset can be found in the **Appendix** (**A2**).

hp_{NG}-1

(Ala-**Asn**-**Gly**-Ala)

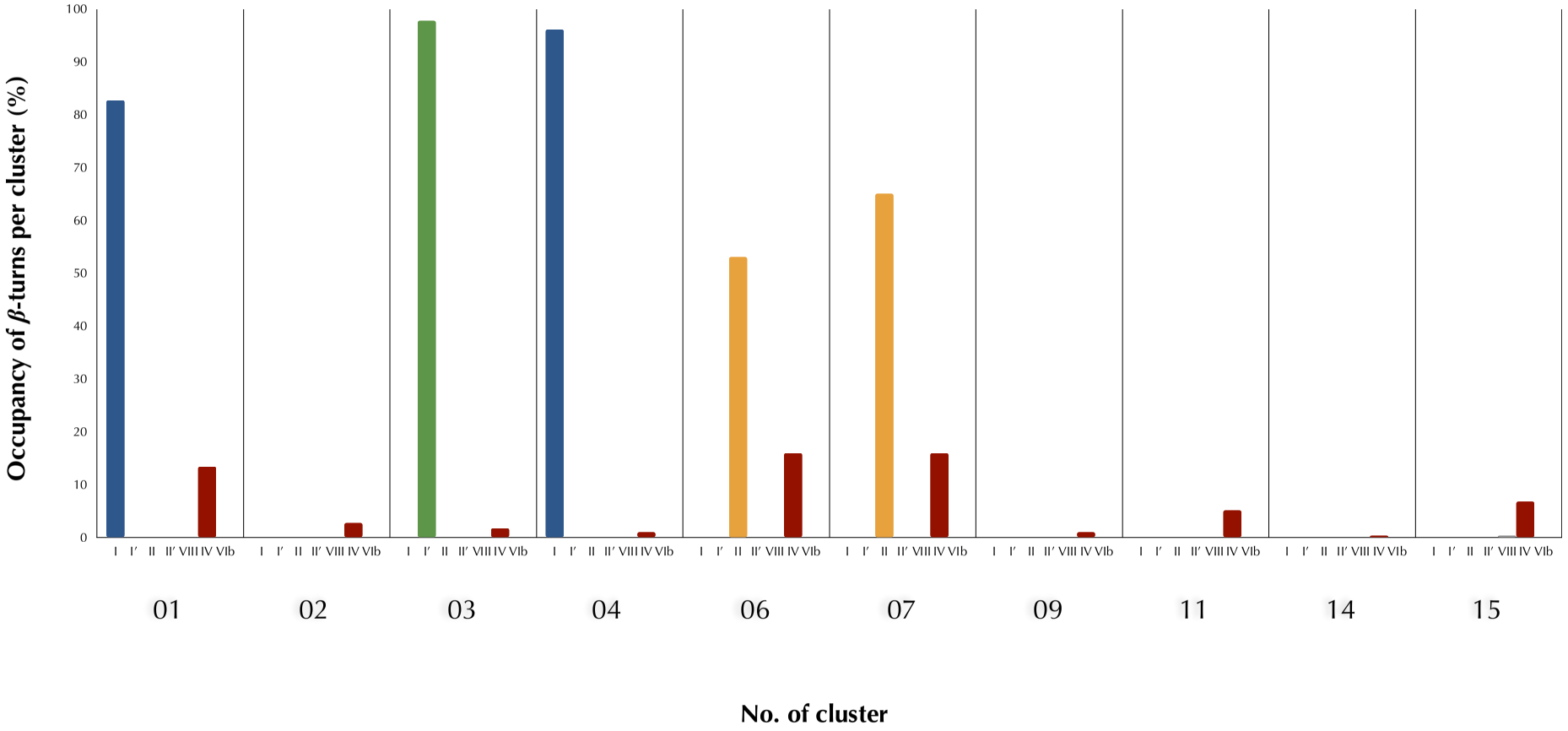


hp_{NG}-2
(Val-**Asn-Gly**-Gln)



hp_{NG}-3

(Val-**Asn**-**Gly**-Leu)



As it can bee seen, for hp_{NG}-1, two of the most populated clusters are occupied mainly by type I turns. Type I′ and II turns are also present in some less populated clusters with type II being more preferred. For hp_{NG}-2, the most populated cluster is occupied mainly by type I′ turns. The second most preferred *β*-turn type for this heptapeptide is type I, followed by type II which is present in less populated clusters. hp_{NG}-3 shows a strong preference for type I and type I′ turns. Type II turns are also present in lesser populations. Regarding type IV turns, almost every cluster is occupied by minor populations of this turn category. The population of type VIII turns in each heptapeptide is barely observable, whereas type II′ and type VIb turns are not present at all. It is important to stress again that not all frames of each cluster were analysed. This structural analysis refers to a dataset of about 500 representative structures obtained by each cluster.

The next step was to perform a structural analysis on the representative structures of each cluster that produced by *grcarma*. The results produced by *promotif* can be seen in the following tables.

hp_{NG}-1
(Ala-Asn-Gly-Ala)

No. of Representative Structure	Turn Type	Central Residues	*i*, *i*+3 hydrogen bond	Ramachandran Region
1	IV	Ala-Asn	Yes	AB
2	IV	Asn-Gly	No	Aa
8	IV	Ala-Asn	No	Aa
10	I	Gly-Ala	Yes	AA
12	I	Ala-Asn	No	Aa
14	I	Ala-Asn	Yes	AA
16	I IV	Asn-Gly Gly-Ala	Yes No	AA Aa
17	II′	Gly-Ala	Yes	A
19	I IV	Ala-Asn Asn-Gly	Yes No	AA Aa

Table III: Shown are only the representative structures of hp_{NG}-1 that form *β*-turn motifs. From left to right are listed the number of cluster of which is the representative, the turn type, the *i*+1 and *i*+2 central residues, the presence of hydrogen bond between the CO group of residue *i* and the NH group of residue *i*+3, and the regions of the Ramachandran plot occupied by residues *i*+1 and *i*+2.

hp_{NG}-2 (Val-Asn-Gly-Gln)

No. of Representative Structure	Turn Type	Central Residues	*i*, *i*+3 hydrogen bond	Ramachandran Region
1	I	Gly-Gln	Yes	AA
5	I	Val-Asn	No	AA
9	IV	Gly-Gln	No	A
10	I	Gly-Gln	Yes	AA
11	IV IV	Val-Asn Asn-Gly	No No	Aa aL
12	IV	Gly-Gln	No	Aa
13	I	Val- Asn	Yes	AA
15	IV	Asn-Gly	No	AP

Table IV: Shown are only the representative structures of hp_{NG}-2 that form β-turn motifs. From left to right are listed the number of cluster of which is the representative, the turn type, the *i*+1 and *i*+2 central residues, the presence of hydrogen bond between the CO group of residue *i* and the NH group of residue *i*+3, and the regions of the Ramachandran plot occupied by residues *i*+1 and *i*+2.

hp_{NG}-3 (Val-Asn-Gly-Leu)

No. of Representative Structure	Turn Type	Central Residues	*i*, *i*+3 hydrogen bond	Ramachandran Region
1	I	Val-Asn	No	AA
3	IV IV	Val-Asn Gly-Leu	No Yes	Aa A
4	IV	Asn-Gly	Yes	PL
5	IV	Gly-Leu	No	AA
7	IV	Val-Asn	No	AB
9	I	Gly-Leu	No	AA
13	I	Gly-Leu	Yes	AA
14	VIII	Val-Asn	No	AB
15	I IV	Val-Asn Gly-Leu	Yes No	AA A

Table V: Shown are only the representative structures of hp_{NG}-3 that form β-turn motifs. From left to right are listed the number of cluster of which is the representative, the turn type, the *i*+1 and *i*+2 central residues, the presence of hydrogen bond between the CO group of residue *i* and the NH group of residue *i*+3, and the regions of the Ramachandran plot occupied by residues *i*+1 and *i*+2.

As shown above, the results of the structural analysis between the representative cluster structures and the 500-sample groups from each cluster differ a lot. As shown in **Tables III-V** only one representative in each heptapeptide adopts a type IV β-turn structure with Asn-Gly being the central segment. The remaining representatives adopt mainly β-turn motifs with Asn or Gly being only one of the central two residues. A minority of representatives also have been found to either adopt double turns or even random coils. But why is there such difference between these two structural analyses? The representative structure refers to the structural configuration that corresponds to the centre of the cluster. Due to the increased kinetic frustration of our peptides a representative structure may differ a lot compared to a group of 500 equally spaced but random configurations of the cluster. Following, we present a superimposed sample dataset of the most populated clusters of the three heptapeptides along with the representative structure for each cluster. All 3D models were created in PyMOL.

For hp_{NG}-1 the three major clusters have a population of 11.8% (Cluster No. 1), 10.74% (Cluster No. 8) and 9.40% (Cluster No. 9). The representative structures of clusters No. 1 and No. 8 correspond to a type IV β-turn with Ala and Asn being the central residues i+1 and i+2 respectively. The representative of cluster No. 9 corresponds to a random coil instead of a turn (**Figure 4.5**).

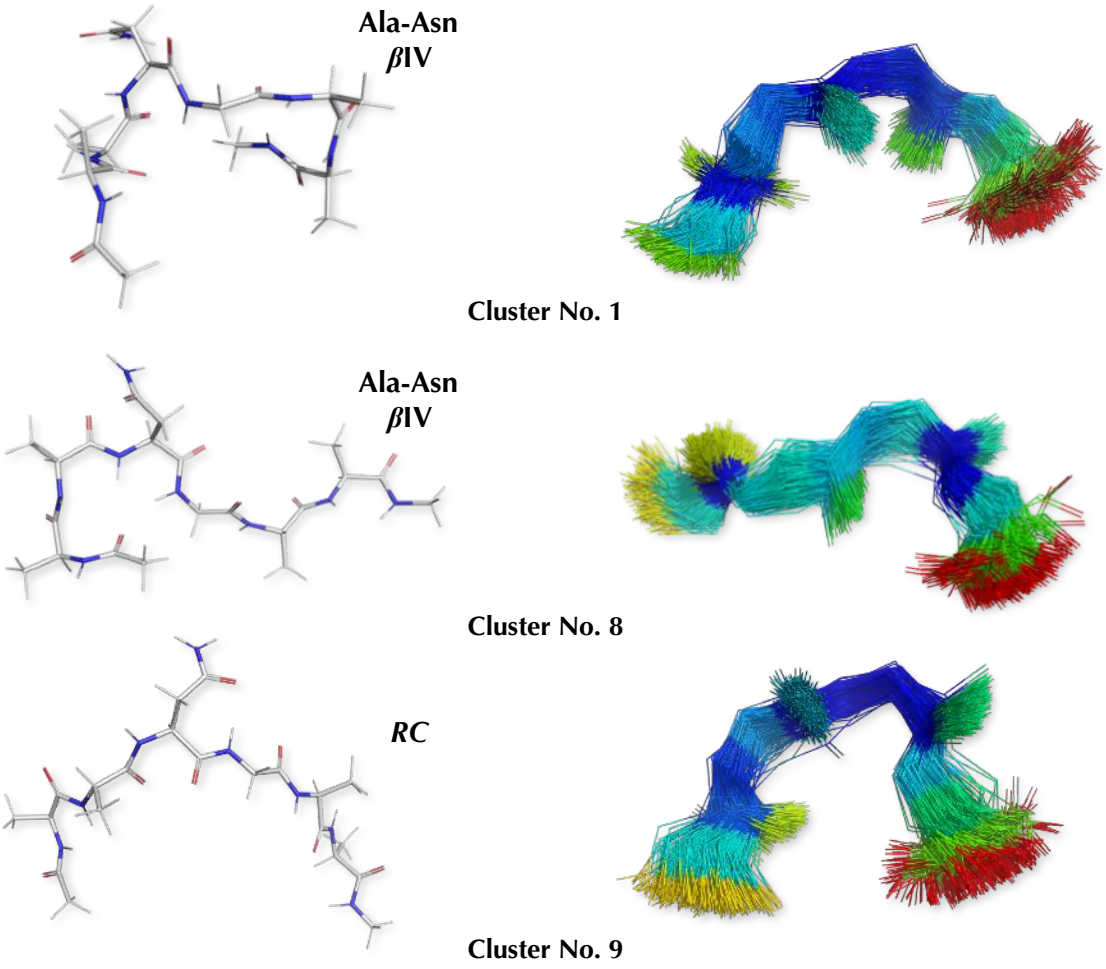


Figure 4.5 Representative structures (left) along with the 500 equally spaced superimposed structures for the four central residues (right) of the three major clusters of hp_{NG}-1. The colours of the superimposed structures denote the RMS fluctuations, varying from blue (small values of RMSF) to red (large values of RMSF).

For hp_{NG}-2 the three major clusters have a population of 20.10% (Cluster No. 1), 14.09% (Cluster No. 2) and 10.13% (Cluster No. 6). Cluster No. 5 has a population of 10.02% almost equal to the third cluster. The representative structures of clusters No. 1 and No. 5 both adopt a type I β-turn. In the first case the central segment consists of Gly-Gln, whereas in the second case of Val-Asn. Representatives from clusters No. 2 and No. 6 do not adopt any specific conformation (**Figure 4.6**).

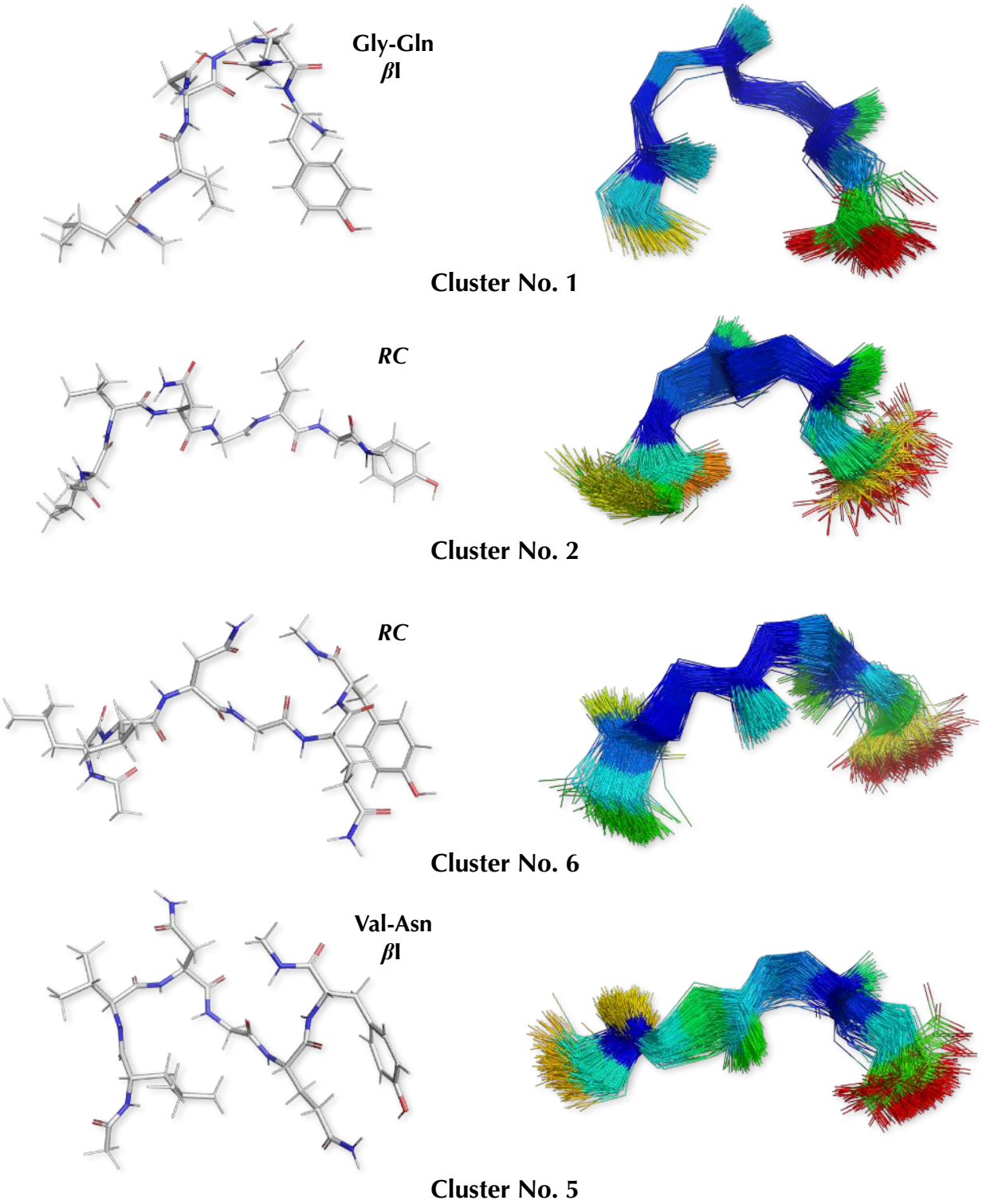


Figure 4.6 Representative structures (left) along with the 500 equally spaced superimposed structures for the four central residues (right) of the four major clusters of hp_{NG}-2. The colours of the superimposed structures denote the RMS fluctuations, varying from blue (small values of RMSF) to red (large values of RMSF).

For hp_{NG}-3 the three major clusters have a population of 20.66% (Cluster No. 1), 17.38% (Cluster No. 2) and 16.50% (Cluster No. 3). Cluster No. 5 has also a significant population of 11.33%. The representative structure of Cluster No. 1 forms a type I Val-Asn β-turn and that of Cluster No. 5 a type IV Gly-Leu β-turn. The representative of Cluster No. 3 forms a double turn, while that of Cluster No. 2 does not adopt any specific conformation (**Figure 4.7**).

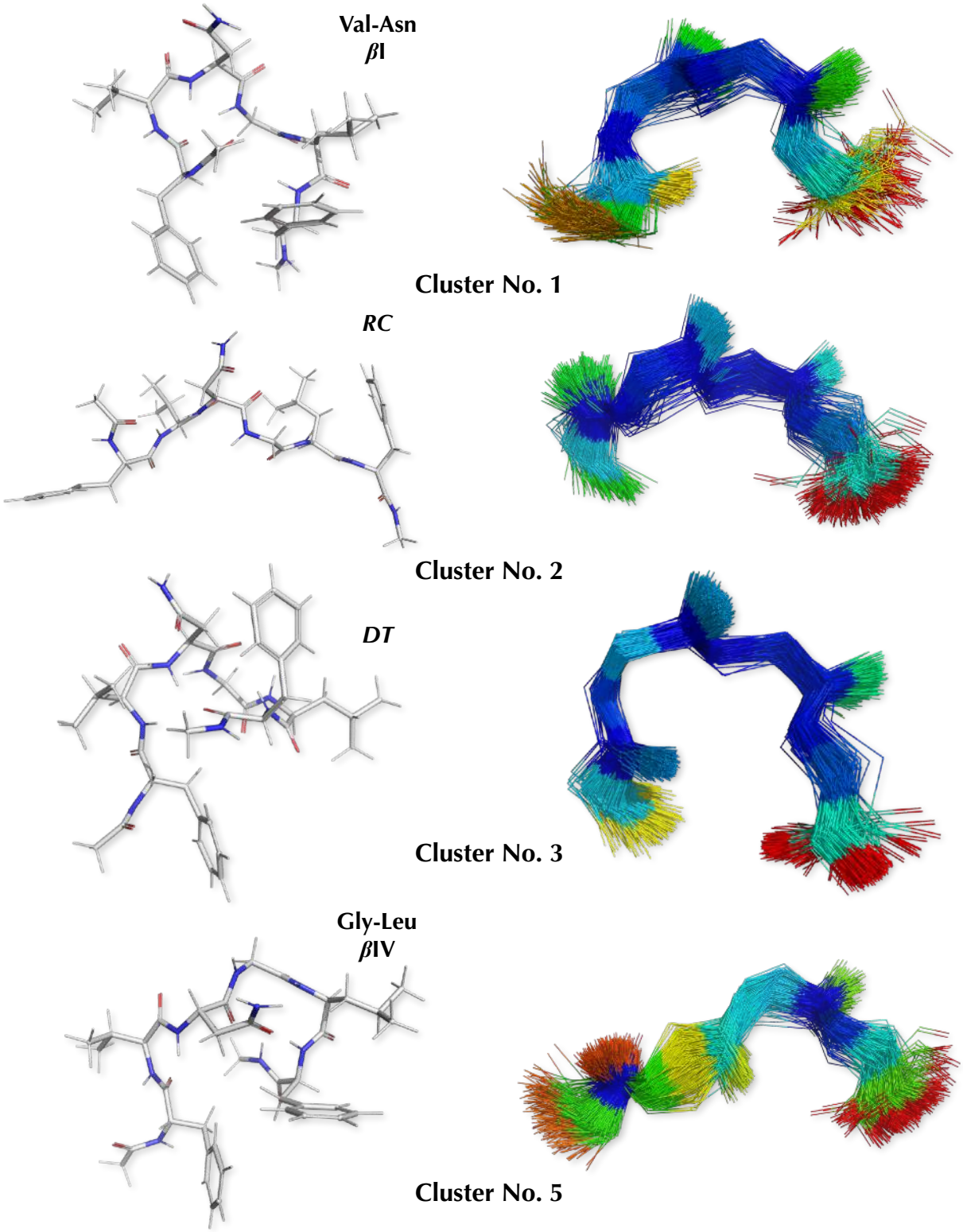
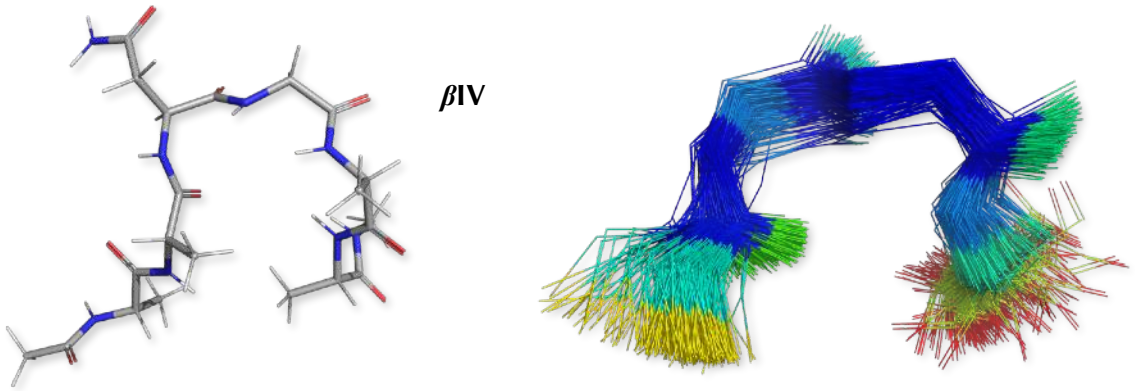
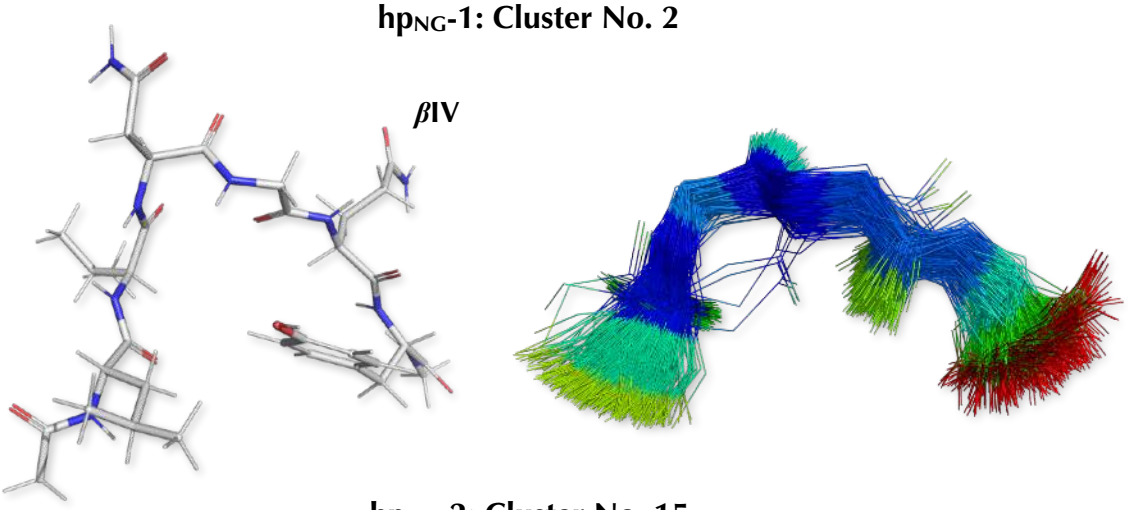


Figure 4.7 Representative structures (left) along with the 500 equally spaced superimposed structures for the four central residues (right) of the four major clusters of hp_{NG}-3. The colours of the superimposed structures denote the RMS fluctuations, varying from blue (small values of RMSF) to red (large values of RMSF).

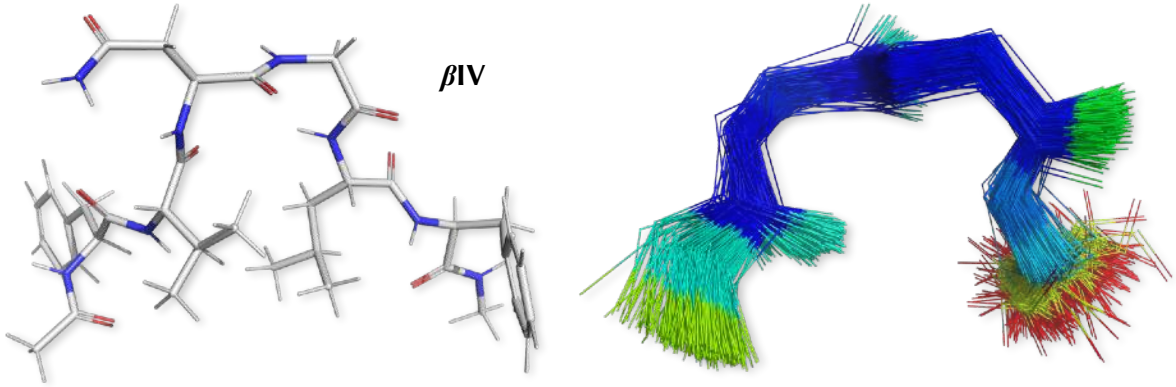
As for the representative structures that form an Asn-Gly β-turn, the analysis assigned only one representative structure in each of the heptapeptides. For hp_{NG}-1, the representative of Cluster No. 2 forms a type IV β-turn. Cluster No. 2 has a population of 4.26%. For hp_{NG}-2, the representative structure of Cluster No. 15, that has a population of 0.28%, forms a type IV β-turn. Finally, for hp_{NG}-3, the representative structure of Cluster No. 4 adopts also a type IV β-turn. The population probability for that cluster is 4.22% (**Figure 4.8**).



hp_{NG}-1: Cluster No. 2



hp_{NG}-2: Cluster No. 15



hp_{NG}-3: Cluster No. 4

Figure 4.8 Representative structures (left) along with the 500 equally spaced superimposed structures for the four central residues (right) of the clusters that correspond to Asn-Gly β-turns. The colours of the superimposed structures denote the RMS fluctuations, varying from blue (small values of RMSF) to red (large values of RMSF).

4.4 Temperature Based Analysis

For our simulations with the NAMD program the adaptive tempering method was implemented. This method dynamically updates the simulation temperature. The temperature T varies between a range of [T_{min}, T_{max}]. The general idea behind this method is that when the potential energy of a given structure is below the so far calculated average energy, the temperature is lowered. Conversely, when the current energy is higher than the average, the temperature is increased. The resulting effect is a faster conformational sampling in order to find minimum energy structures[42,57]. **Figure 4.9** below presents a temperature based analysis of our trajectories.

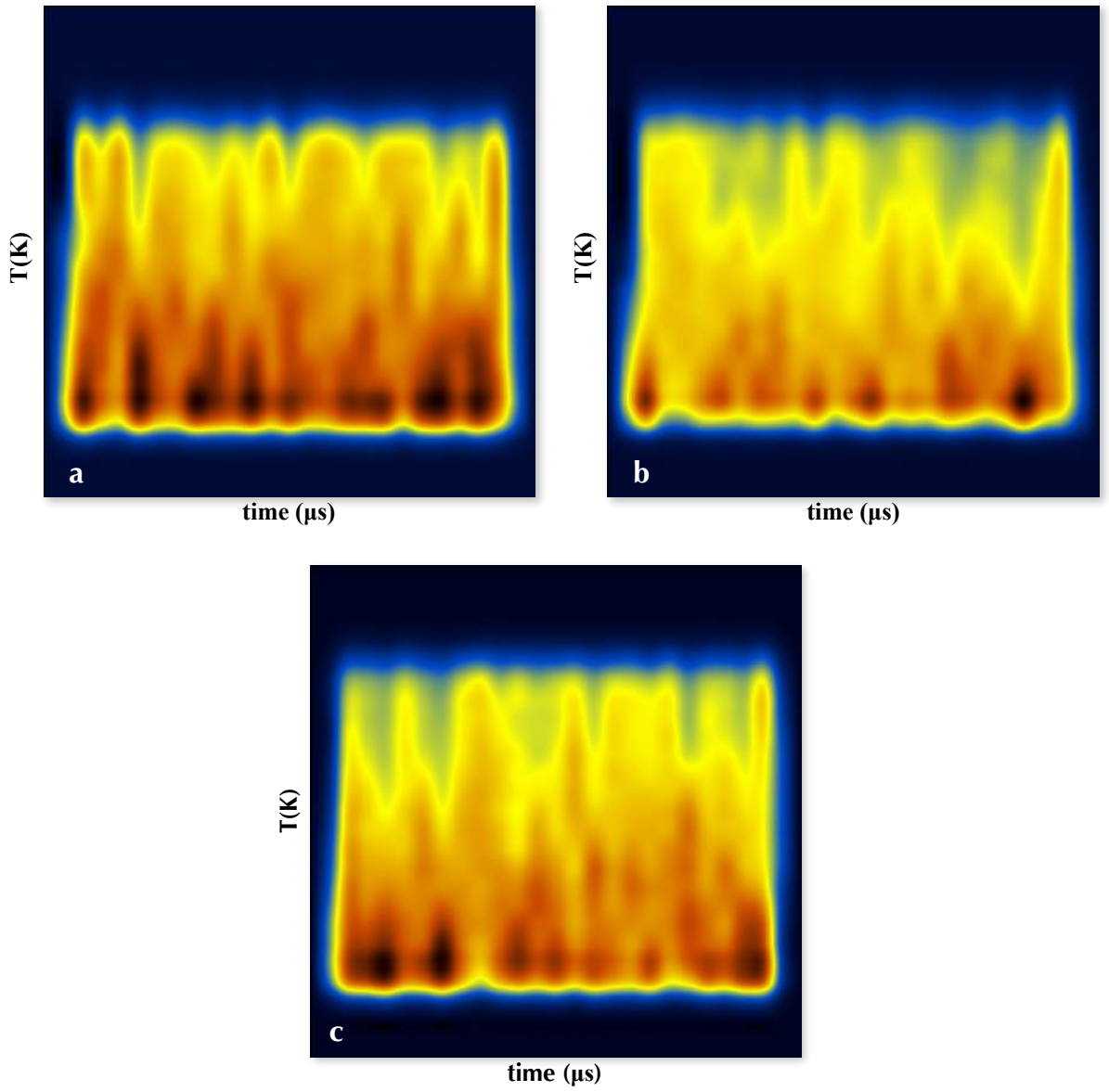


Figure 4.9 Diagrams of temperature distribution as a function of the simulation time for (a) hp_{NG}-1, (b) hp_{NG}-2 and (c) hp_{NG}-3. Blue colours correspond to a low number of conformations, while red and black to a large number of conformations. The diagrams were produced using the *plot* program.

As shown above, most of the conformations, and possibly the more stable ones, are present in low temperatures (less than 360 K). However, there are no distinct, stable folding events for any of the three heptapeptides, meaning that they present quite a dynamic behaviour during the whole time of their simulation.

4.5 Comparison with the *ab initio* Models

In this part of the thesis we are going to compare the results from our MD analyses with the *ab initio* models of the same heptapeptides presented by Kang & Yoo. In order to do so, we first performed an RMSD analysis of our trajectories using as reference structures the four β-turn models (I, I′, II, II′) and an extended one, produced by their work using quantum mechanical calculations and the DFT method.

The Root Mean Square Deviation (RMSD) is a common analysis in Structural Biology, notably when carrying out an MD simulation. The RMSD calculates the average distance between atoms of different superimposed conformations. It is thus a comparative method for analysing protein structures[58].

The RMSD can be calculated by the following equation

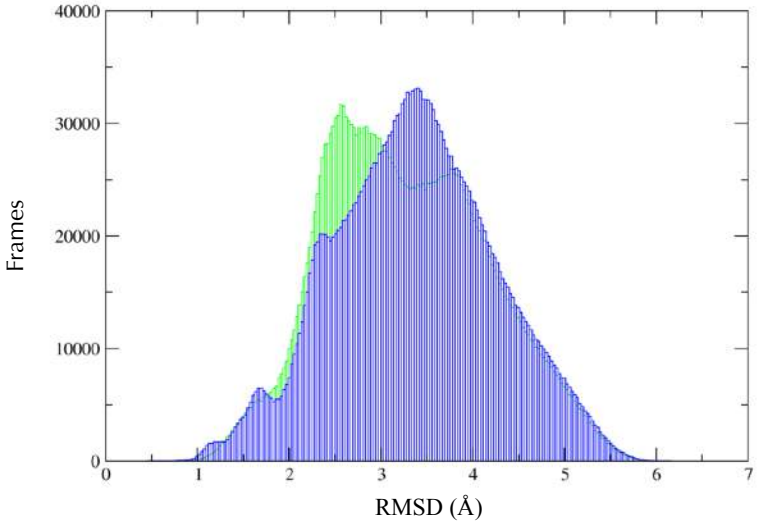
$$RMS = \left\langle \left(r_i^a - r_i^\beta \right)^2 \right\rangle^{\frac{1}{2}} = \sqrt{\frac{1}{N_i} \sum_i \left(r_i^a - r_i^\beta \right)^2} \qquad (4.5)$$

where r_i^a states the atomic coordinates of a specific structure at a specific time, r_i^β the atomic coordinates of another structure, in our case the reference structure, and N is the number of the atoms. The lesser the magnitude of the RMSD value is, the greater the similarity between the two superimposed structures. Conversely, the greater the RMSD value is, the larger the difference between the superimposed structures. In case of identical structures, the RMSD value is 0 Å. Usually, values of RMSD less than or equal to 2.0 Å denote sufficient and obvious similarity between structures.

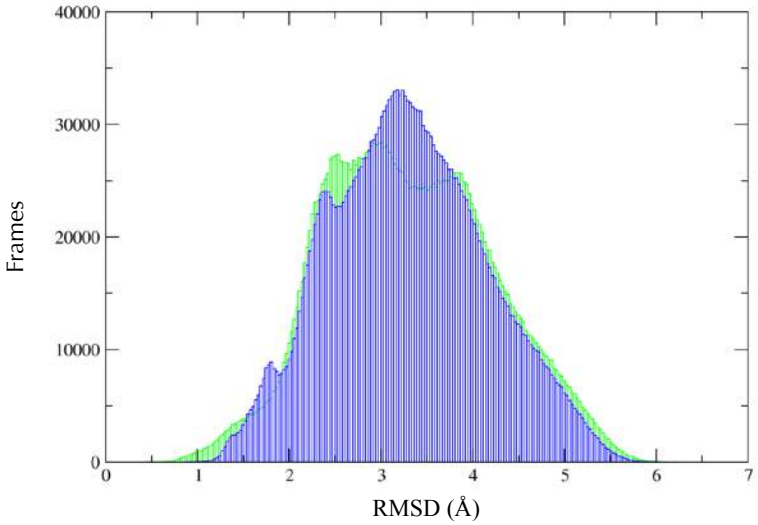
For the RMSD analysis we used only frames that correspond to temperatures of less than 360 K, in order to focus only on the more stable conformations and avoid excess noise in our results. The following figures were created with the *Grace* plotting tool[59] (**Figure 4.10-4.12**) and are histogram charts showing the distribution of RMSD fluctuations during the simulation time in reference to the *ab initio* models.

hp_{NG}-1

βI (green) and βI′ (blue) turns as reference structures



βII (green) and βII′ (blue) turns as reference structures



Extended conformation as reference structure

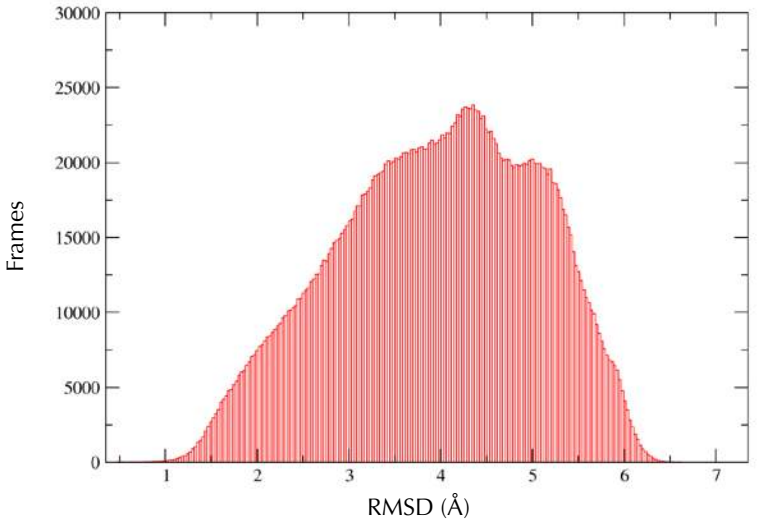


Figure 4.10 Histogram charts presenting the distribution of the RMS deviations of each hp_{NG}-1 conformation of our trajectory (< 360 K) in reference to each of the *ab initio* models.

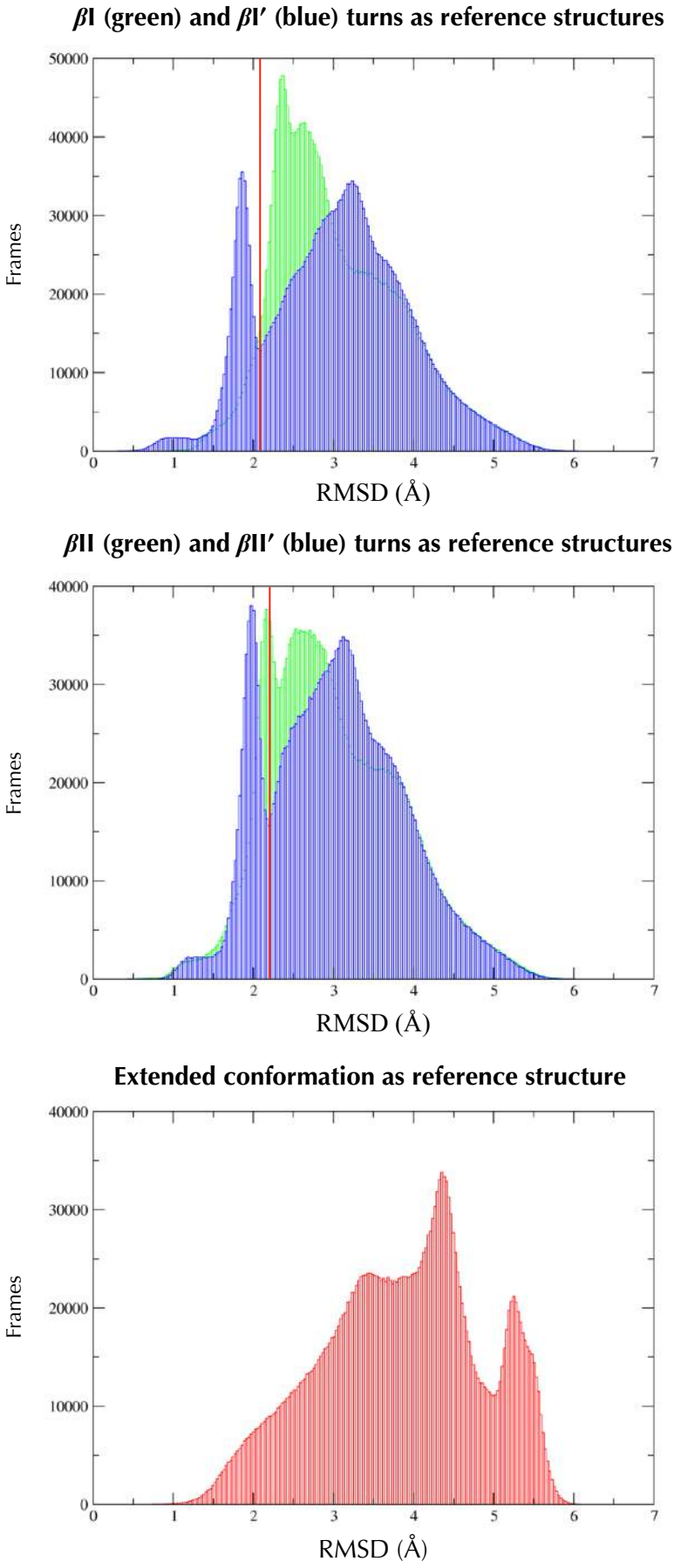
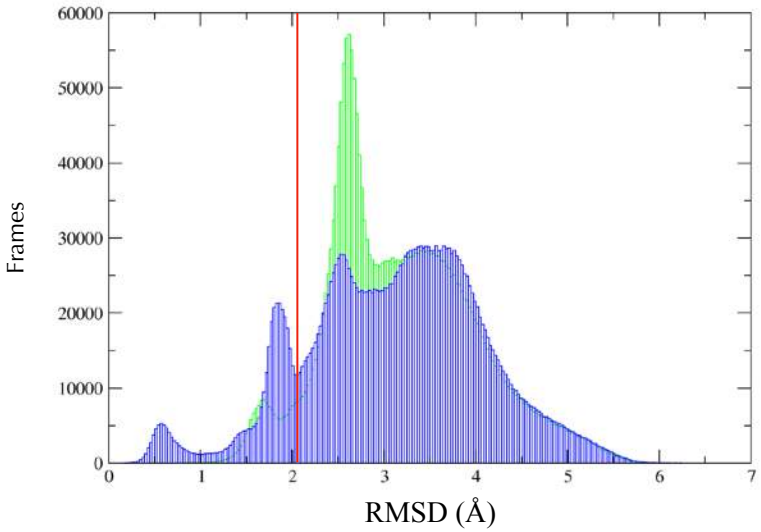


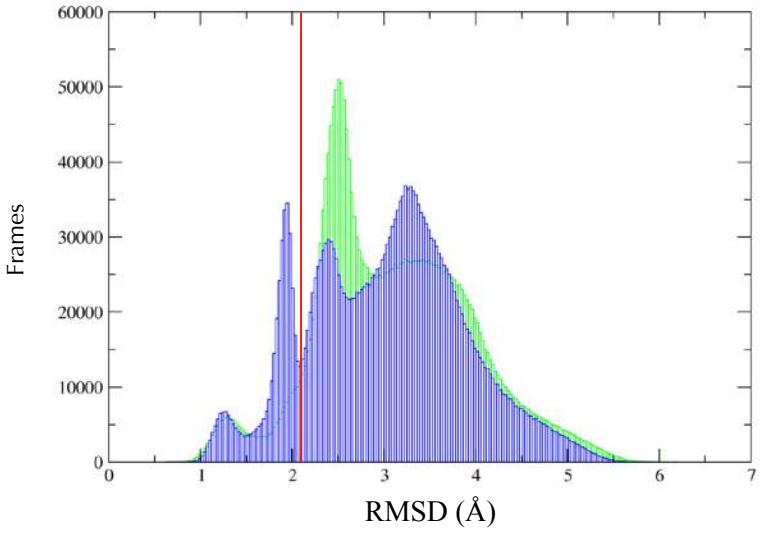
Figure 4.11 Histogram charts presenting the distribution of the RMS deviations of each hpNG-2 conformation of our trajectory (< 360 K) in reference to each of the *ab initio* models.

hp_{NG}-3

βI (green) and βI′ (blue) turns as reference structures



βII (green) and βII′ (blue) turns as reference structures



Extended conformation as reference structure

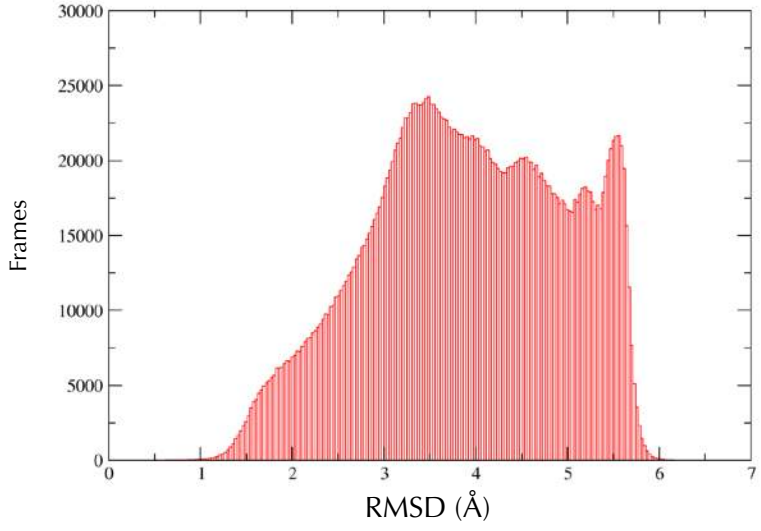


Figure 4.12 Histogram charts presenting the distribution of the RMS deviations of each hp_{NG}-3 conformation of our trajectory (< 360 K) in reference to each of the *ab initio* models.

Most of the conformations of the hp_{NG}-1 trajectory show a reduced convergence with the *ab initio* models. However, some of the RMSD histograms of hp_{NG}-2 and hp_{NG}-3 contain distinct peaks at low values of RMSD denoting that there are groups of conformations that present sufficient similarity with the *ab initio* models. In order to compare our trajectories with the *ab initio* models, we shall first choose an RMSD cutoff so as to isolate only clusters which are considerably similar to those models that were identified through quantum mechanical calculations. The RMSD cutoff that was chosen equals to 2.2 Å and was selected after observing the width of each distinct peak at low values of RMSD in each of the histogram charts.

The next step was to perform an RMSD based cluster analysis in all our three trajectories independently of the *ab initio* models. The concept of this, lies in the fact that, since there are frames in our trajectories that present such high similarity with the *ab initio* models (the above RMSD analysis presented values even less than 1 Å in some cases), there will be clusters of conformations sufficiently similar to the *ab initio* models. Therefore, we constructed a hierarchical dendrogram for each of the three trajectories using the UPGMA algorithm and choosing the above mentioned threshold of 2.2 Å (**Figure 4.13-4.15**). **Table VI** also summarises the populations of all clusters found. It is important to mention that the RMSD values for each heptapeptide were calculated every 1000th frame, thus resulting in the examination of 5033 frames from the first trajectory, 5010 frames from the second trajectory and 5027 frames from the third trajectory.

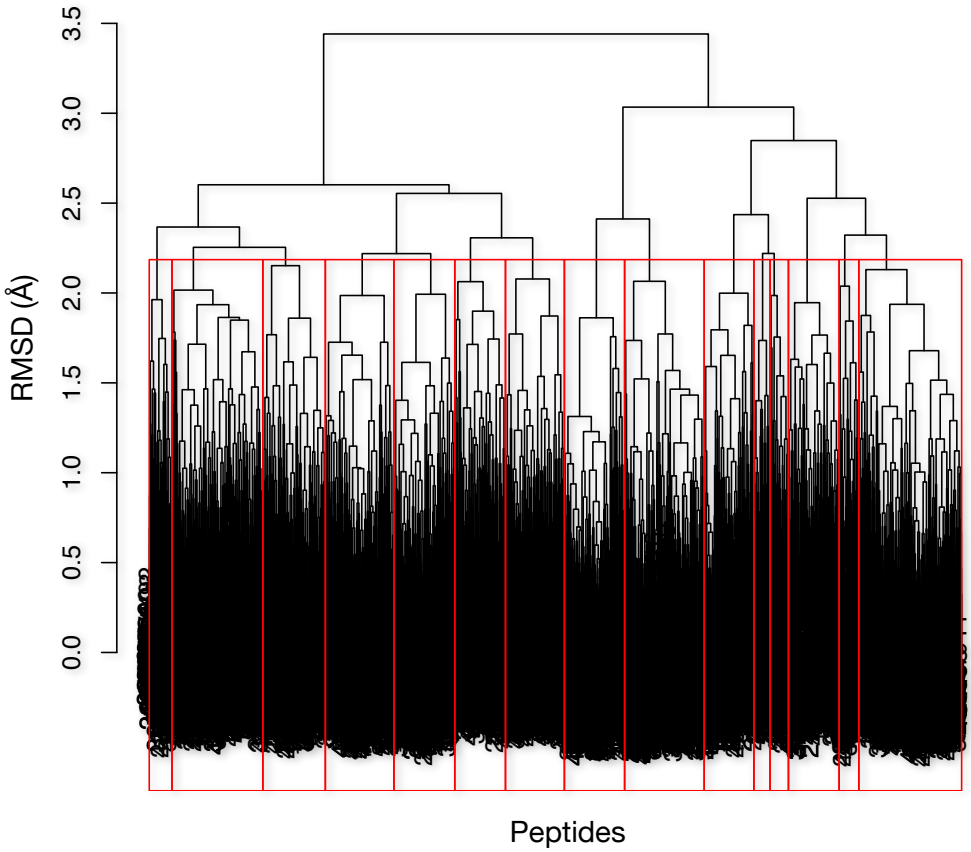


Figure 4.13 The hierarchical dendrogram produced by the RMSD analysis of hp_{NG}-1, as created using the UPGMA algorithm. The red line indicates the RMSD cutoff set for the cluster analysis and isolation, and the amount of clusters isolated.

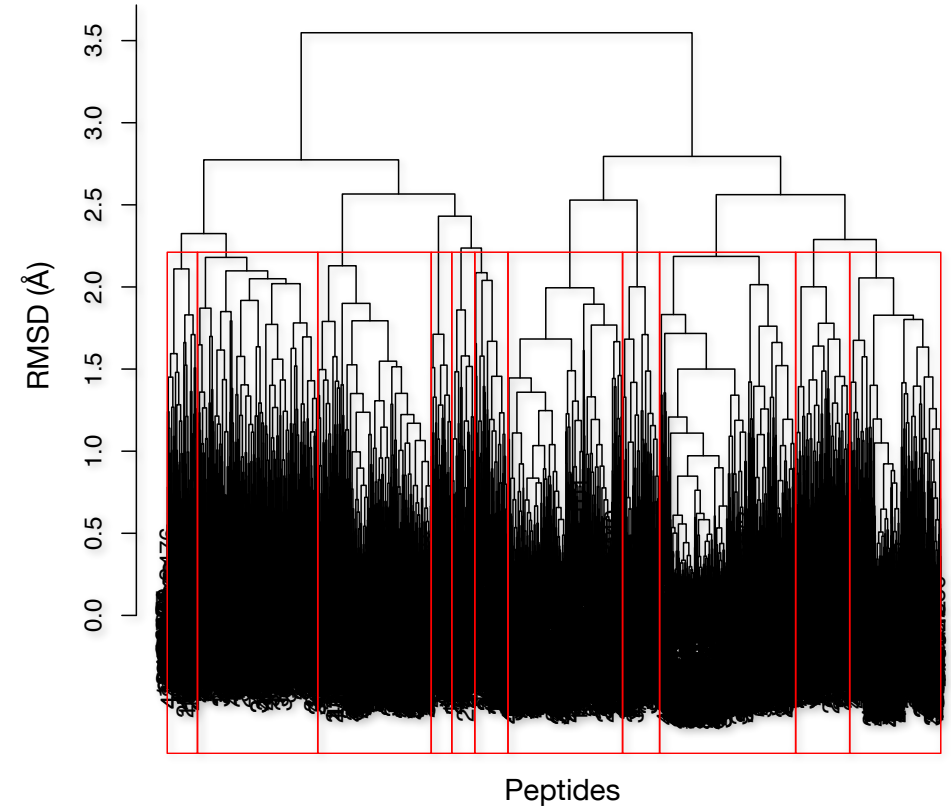


Figure 4.14 The hierarchical dendrogram produced by the RMSD analysis of hp_{NG}-2, as created using the UPGMA algorithm. The red line indicates the RMSD cutoff set for the cluster analysis and isolation, and the amount of clusters isolated.

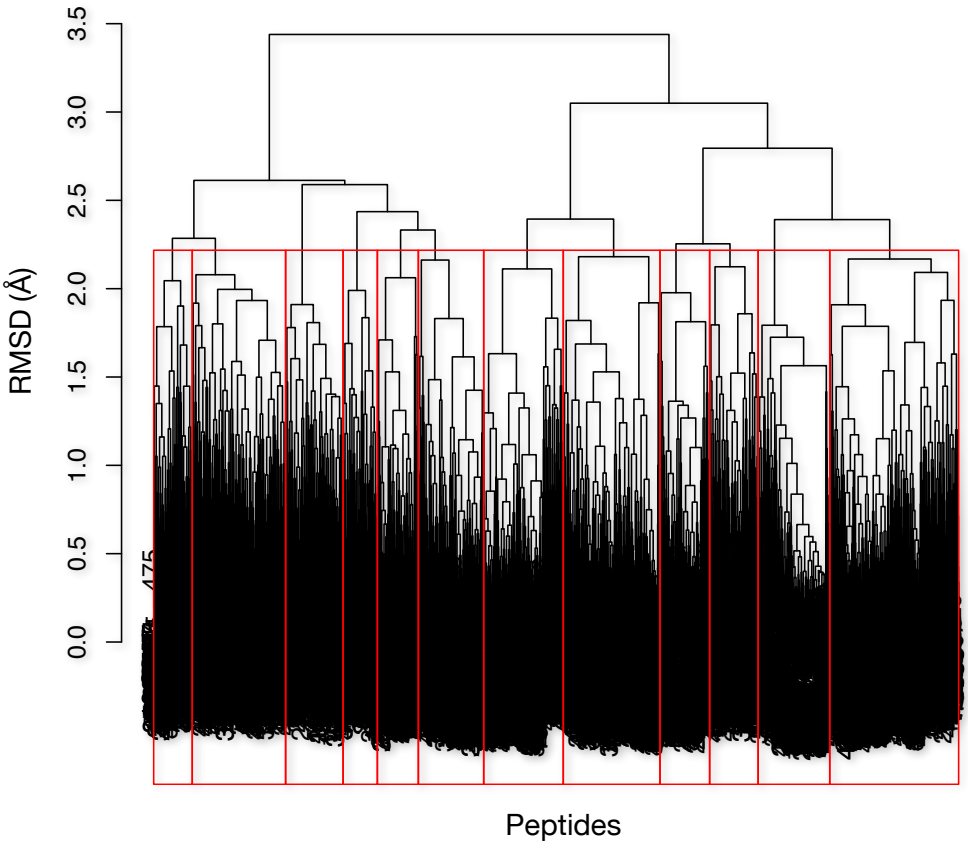


Figure 4.15 The hierarchical dendrogram produced by the RMSD analysis of hp_{NG}-3, as created using the UPGMA algorithm. The red line indicates the RMSD cutoff set for the cluster analysis and isolation, and the amount of clusters isolated.

Table VI: Populations of the clusters produced from the RMSD analysis.

No. of Cluster	hp_{NG}-1	hp_{NG}-2	hp_{NG}-3
1	2.26	14.81	16.01
2	7.69	6.99	8.89
3	1.99	11.78	4.26
4	2.42	17.58	9.89
5	11.19	14.67	12.02
6	7.25	4.79	8.12
7	7.41	15.55	11.60
8	6.24	3.91	6.03
9	12.66	2.97	6.17
10	8.44	2.65	7.14
11	6.12	4.29	5.09
12	7.49		4.79
13	6.24		
14	9.80		
15	2.80		

Table VI: Shown are the populations (among examined frames in %) of clusters from each heptapeptide as produced by the RMSD based clustering analysis.

At this point, in order to examine which clusters indicate significant structural similarity with the *ab initio* models for each of the heptapeptides, we have to proceed on an RMSD analysis of each cluster using as reference conformations the experimental models. For this purpose, we created pseudo-trajectories for each heptapeptide by concatenating sequentially each cluster's frames. The RMSD analysis was made via *grcarma* and the following graphs (**Figure 4.16-4.18**) were produced by the *Grace* plotting tool.

hp_{NG}-1

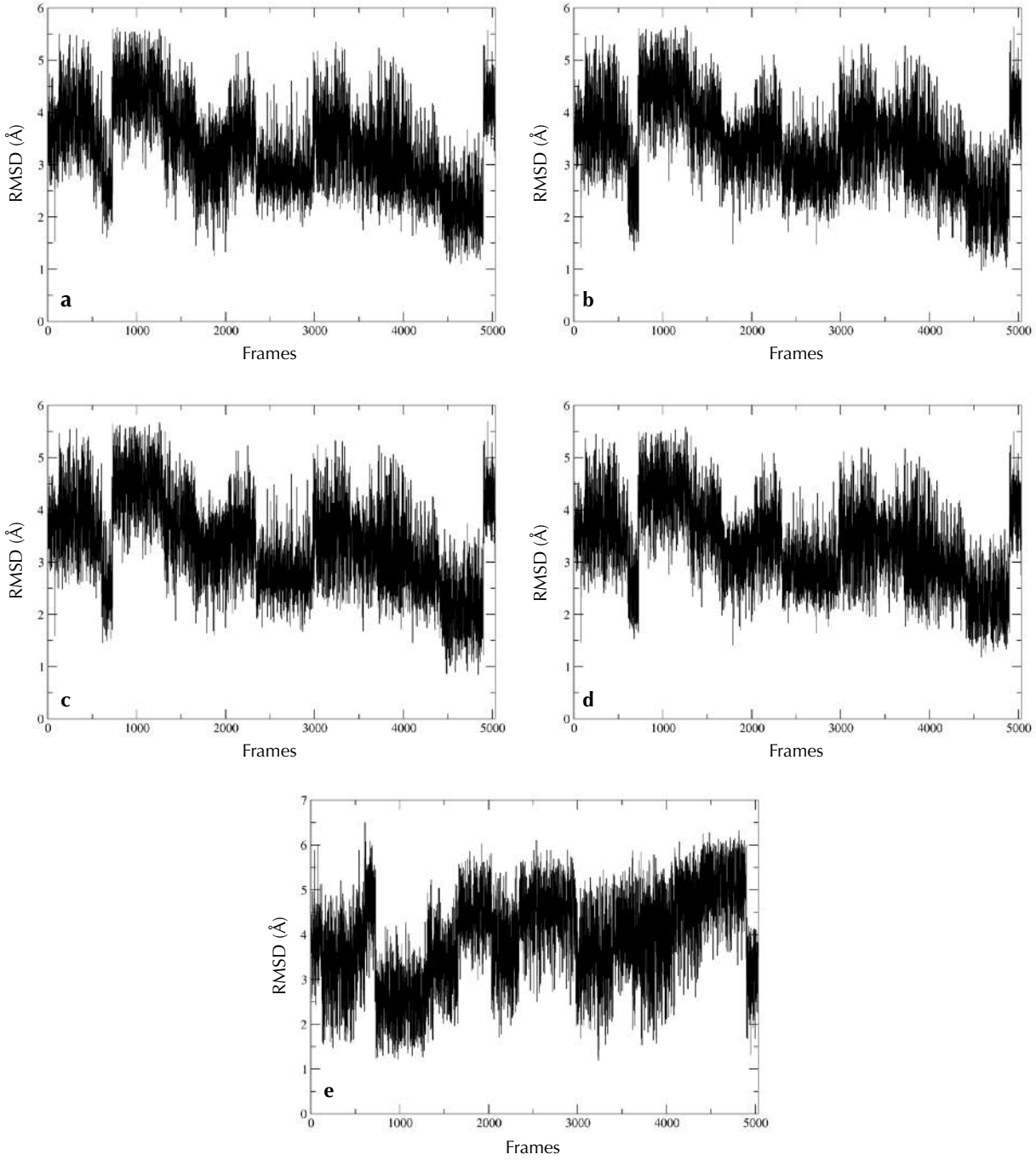


Figure 4.16 RMSD diagrams of the hp_{NG}-1 pseudo-trajectory representing all clusters and using as reference conformation the *ab initio* model for type I β-turn (a), I′ β-turn (b), II β-turn (c) II′ β-turn (d) and the extended conformation (e).

Cluster No. 14 (Frames: 4399-4892), which has a population of 9.80%, presents the greatest similarity with type I, I′, II and II′ β-turns of the *ab initio* models for hp_{NG}-1. The RMSD in each case fluctuates around 2.0 Å, 2.2 Å, 2.1 Å and 2.1 Å respectively. Cluster No. 05 (Frames: 723-1286), that has a population of 11.19%, presents the greatest similarity with the extended conformation of the *ab initio* models for hp_{NG}-1, with the RMSD values in this cluster fluctuating around 2.5 Å.

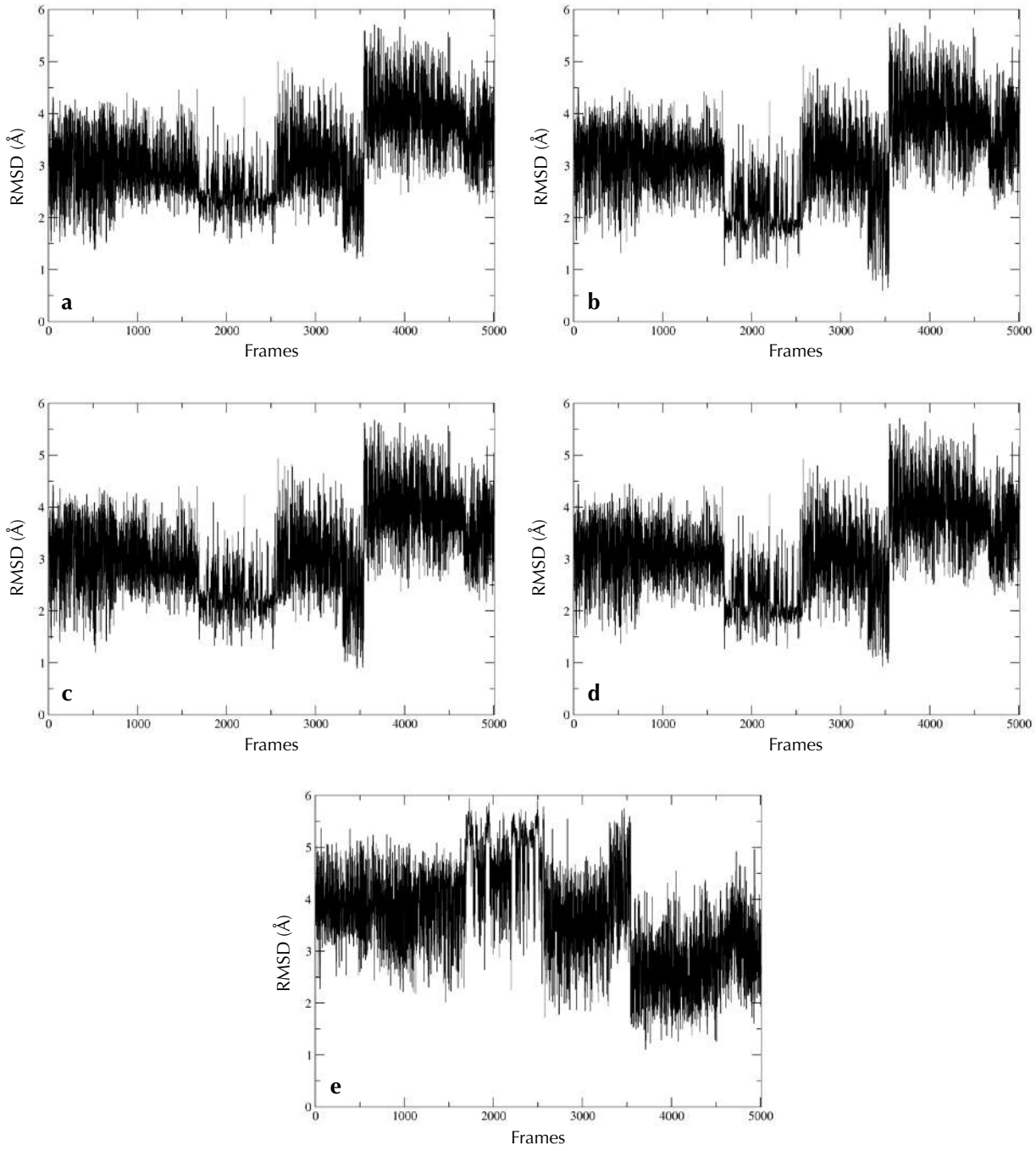


Figure 4.17 RMSD diagrams of the hp$_{NG}$-2 pseudo-trajectory representing all clusters and using as reference conformation the *ab initio* model for type I β-turn (a), I′ β-turn (b), II β-turn (c) II′ β-turn (d) and the extended conformation (e).

Cluster No. 04 (Frames: 1682-2563), that has a population of 17.58%, presents the greatest similarity with type I, I′, II and II′ β-turns of the *ab initio* models for hp$_{NG}$-2. The RMSD in each case fluctuates around 2.3 Å, 1.8 Å, 2.2 Å and 2.0 Å respectively. Cluster No. 07 (Frames: 3538-4317), that has a population of 15.55%, presents the greatest similarity with the extended conformation of the *ab initio* models for hp$_{NG}$-2, with the RMSD values in this cluster fluctuating around 2.2 Å.

hp_{NG}-3

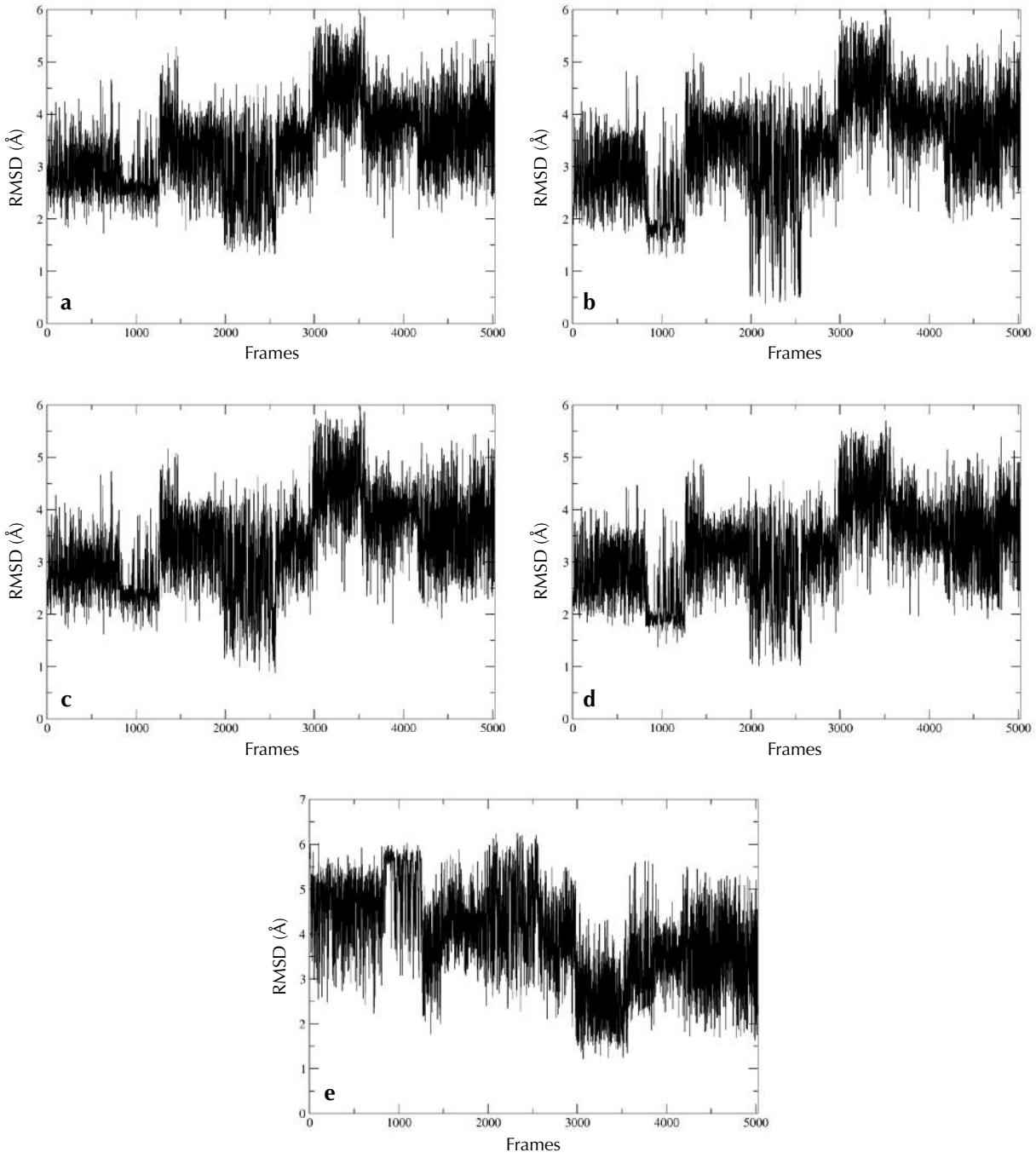


Figure 4.18 RMSD diagrams of the hp_{NG}-3 pseudo-trajectory representing all clusters and using as reference conformation the *ab initio* model for type I *β*-turn (a), I′ *β*-turn (b), II *β*-turn (c) II′ *β*-turn (d) and the extended conformation (e).

As for hp_{NG}-3, Cluster No. 05 (Frames: 1963-2567), having a population of 12.02%, denotes significant similarity with the *ab initio* type I *β*-turn, while Cluster No. 02 (Frames: 805-1252), with a population probability of 8.89%, indicates the greatest similarity with the remaining *β*-turn *ab initio* models. The RMSD in each of the above cases fluctuates around 2.1 Å, 1.8 Å, 2.4 Å and 2.0 Å regarding I, I′, II, II′ turn types respectively. Cluster No. 07 (Frames: 2975-3558), which has a population of 11.60%, presents the greatest similarity with the extended conformation of the *ab initio* models, with the RMSD values in this cluster fluctuating around 2.5 Å.

How can only one cluster denote similarity with more than one β-turn types of the *ab initio* models? Maybe the structural differences among these turn types are relatively small and along with the RMSD cutoff that has been chosen for the cluster analysis, the cluster isolation might not represent such accurate discretion among the clusters' structures and the experimental models. Nevertheless, clusters that indicate significant similarity with the *ab initio* models correspond to relatively large populations, but still they do not describe more than 20-30% of the trajectory in each case. Clusters that are similar with the extended *ab initio* configurations have considerable populations, however, they fluctuate around relatively large RMSD values. Clusters similar to the experimental β-turns denote better similarity, but still we cannot have a highly-accurate perspective of the data. The hypothesis that the Asn-Gly peptides can adopt β-turn conformations can in general be true, but further research is necessary in order to prove whether this type of structural motif is strongly favoured for these peptides.

5. Conclusions and Discussion

The primary aim of this project was to evaluate the ability of MD simulations to sufficiently predict the structure and dynamics of small peptides compared to the "high quality" quantum mechanical calculations. For this purpose, we performed three 5 μs MD simulations on three Asn-Gly heptapeptides, analysed the results and compared them with their corresponding *ab initio* models.

While quantum mechanics calculations have shown that type I′ β-turn is the most preferred motif in aqueous solution for the three heptapeptides containing the Asn-Gly segment (with some variations depending on the sequence of the peptides and the solvent polarity), the results obtained from our MD simulations seem to diverge significantly. It appears that the peptides suffer from severe kinetic frustration with many non-native structures being transiently stabilized during our simulations. The resulting free energy landscapes and the following cluster analysis have shown that there is no funnel-like gradient leading to a native state and thus, no noticeable highly populated native-like intermediate. Furthermore, peptides' central four-residue part appears to be more stable without, however, affecting extensively the general dynamic behaviour of our systems.

The clusters' structural analysis, on the other hand, implies the presence of β-turn motifs in a significant number of clusters in each heptapeptide. However, the preference in positions i+1 and i+2 in β-turns varies between the four central residues, with the Asn-Gly segment preferring the aforementioned positions in a negligible proportion in relation to the whole trajectory. For hp_{NG}-1 the most preferred structure is an Ala-Asn βIV turn motif, whereas for hp_{NG}-2 and hp_{NG}-3 the most preferred structures are a Gly-Gln and a Val-Asn βI turns respectively. The remaining structures show a strong preference in βI and βIV turns, as well as there are representative structures that form a random coil rather than a canonical secondary structure pattern. Such difference between our results and the *ab initio* models can be attributed not only in the short peptides' length and their corresponding kinetic dynamicity, but it might also emanate from the positional potential of each residue among the different turn types. For instance, the Asn residue is in general considerably favoured at position i+2 (Thornton *et. al.*, 1994). The formation of hydrogen bonds between side chains that stabilize the turns and thus the amino acid sequence of each heptapeptide might also play a particular role.

Also, the RMSD based analysis of our trajectories in comparison with the *ab initio* models denoted some similarity. Examining the frames of the whole trajectories that correspond to lower temperatures and thus to more stable configurations, and performing RMSD analyses using as reference structures the experimental models, we found notable conformation populations around low values of RMSD (2.2 Å). Consequent cluster isolation using the UPGMA algorithm and RMSD based clustering analyses have indeed shown similarity to the *ab initio* models in considerable

proportions of the trajectories. The tendency of the Asn-Gly heptapeptides to form generally β-turn structures is relatively apparent, but the tendency among the different β-turn types is not clearly distinguishable. A further analysis with a lower RMSD threshold may be able to group conformations of closer structural similarity and reduce excessive noise.

To have an overall view, we can ascertain that the Asn-Gly heptapeptides present a dynamic behaviour in aqueous solution with a general tendency to form β-turn conformations, as the four-residue central part presents a more stable behaviour. The preference of amino acids varies in positions i+1 and i+2 and the peptides are inclined to adopt mainly type I and IV β-turn conformations. Regarding the question we initially posed, whether MD simulations can provide sufficiently comparable results to quantum mechanical calculations, the answer is still not absolutely clear. MD simulations seem to apply better to larger, more stable molecules, rather than small and dynamic systems. But in order to be assured of this hypothesis, further systematic research is necessary so as to confirm our results and make a step closer to the answer of this question.

References

1. Berg, J., Tymoczko, J. and Stryer, L. (2003). *Biochemistry*. 7th ed. New York: W.H. Freeman.
2. Brändén, C. and Tooze, J. (2009). *Introduction to protein structure*. 2nd ed. New York, NY: Garland Pub.
3. Anfinsen, C. (1973). Principles that Govern the Folding of Protein Chains. *Science*, 181(4096), pp.223-230.
4. Levinthal, C. (1968). Are there pathways for protein folding?. *Journal de Chimie Physique*, 65, pp.44-45.
5. Levinthal, C. (1969). How to fold graciously. Mössbaun Spectroscopy in Biological Systems. Proceedings, Univ. of Illinois Bulletin, 67(41), pp.22-24.
6. Karplus, M. (1997). The Levinthal paradox: yesterday and today. *Folding and Design*, 2, pp.S69-S75.
7. Wetlaufer, D. (1973). Nucleation, Rapid Folding, and Globular Intrachain Regions in Proteins. *Proceedings of the National Academy of Sciences*, 70(3), pp.697-701.
8. Karplus, M. and Weaver, D. (1976). Protein-folding dynamics. *Nature*, 260(5550), pp. 404-406.
9. Fersht, A. (1995). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proceedings of the National Academy of Sciences*, 92(24), pp.10869-10873.
10. Harrison, S. and Durbin, R. (1985). Is there a single pathway for the folding of a polypeptide chain?. *Proceedings of the National Academy of Sciences*, 82(12), pp. 4028-4030.
11. Chan, H. and 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.
12. Dill, K. and Chan, H. (1997). From Levinthal to pathways to funnels. *Nature Structural & Molecular Biology*, 4(1), pp.10-19.
13. Drenth, J. (2011). *Principles of protein X-ray crystallography*. New York: Springer.
14. Dinner, A., Šali, A., Smith, L., Dobson, C. and Karplus, M. (2000). Understanding protein folding via free-energy surfaces from theory and experiment. *Trends in Biochemical Sciences*, 25(7), pp.331-339.
15. EYLES, S. (2004). Methods to study protein dynamics and folding by mass spectrometry. *Methods*, 34(1), pp.88-99.
16. Engel, A. (1999). Atomic force microscopy: a powerful tool to observe biomolecules at work. *Trends in Cell Biology*, 9(2), pp.77-80.
17. Mertens, H. and Svergun, D. (2010). Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *Journal of Structural Biology*, 172(1), pp. 128-141.
18. Glassford, S., Byrne, B. and Kazarian, S. (2013). Recent applications of ATR FTIR spectroscopy and imaging to proteins. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1834(12), pp.2849-2858.
19. Holley, L. and Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences*, 86(1), pp.152-156.

20. Brindha, S., Sailo, S., Chhakchhuak, L., Kalita, P., Gurusubramanian, G. and Kumar, N. (2011). Protein 3D structure determination using homology modelling and structure analysis. *Science Vision*, 11(3), pp.125-133.
21. Rost, B., Schneider, R. and Sander, C. (1997). Protein fold recognition by prediction-based threading. *Journal of Molecular Biology*, 270(3), pp.471-480.
22. Hardin, C., Pogorelov, T. and Luthey-Schulten, Z. (2002). Ab initio protein structure prediction. *Current Opinion in Structural Biology*, 12(2), pp.176-181.
23. Pauling, L., Corey, R. and Branson, H. (1951). The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4), pp.205-211.
24. Ramachandran, G., Ramakrishnan, C. and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1), pp.95-99.
25. Hutchinson, E. and Thornton, J. (1994). A revised set of potentials for β-turn formation in proteins. *Protein Science*, 3(12), pp.2207-2216.
26. Lewis, P., Momany, F. and Scheraga, H. (1973). Chain reversals in proteins. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 303(2), pp.211-229.
27. Wilmot, C. and 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. Venkatachalam, C. (1968). Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers*, 6(10), pp.1425-1436.
29. Kang, Y. and Yoo, I. (2016). Propensities of peptides containing the Asn-Gly segment to form β-turn and β-hairpin structures. *Biopolymers*, 105(9), pp.653-664.
30. Leach, A. (2009). *Molecular modelling*. Harlow: Pearson/Prentice Hall.
31. Hehre, W. (2003). A guide to molecular mechanics and quantum chemical calculations. Irvine, CA: Wavefunction.
32. IIzaguirre, J., Catarello, D., Wozniak, J. and Skeel, R. (2001). Langevin stabilization of molecular dynamics. *The Journal of Chemical Physics*, 114(5), pp.2090-2098.
33. Frenkel, D. (2002). Understanding molecular simulation. San Diego: Academic Press.
34. Ch.embnet.org. (2018). *Home MD*. [online] Available at: https://www.ch.embnet.org/MD_tutorial/ [Accessed 5 Jul. 2018].
35. Ambermd.org. (2018). *The Amber Molecular Dynamics Package*. [online] Available at: http://ambermd.org [Accessed 5 Jul. 2018].
36. Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. and Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2), pp.187-217.
37. Gromos.net. (2018). *Biomolecular Simulation - The GROMOS Software*. [online] Available at: http://www.gromos.net [Accessed 5 Jul. 2018].
38. Jorgensen, W. and 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.
39. Bizzarri, A. and Cannistraro, S. (2002). Molecular Dynamics of Water at the Protein−Solvent Interface. *The Journal of Physical Chemistry B*, 106(26), pp.6617-6633.
40. Yuet, P. and Blankschtein, D. (2010). Molecular Dynamics Simulation Study of Water Surfaces: Comparison of Flexible Water Models. *The Journal of Physical Chemistry B*, 114(43), pp.13786-13795.

41. Attig, Norbert & Binder, Kurt & Grubm, Helmut & Kremer, Kurt. (2004). Computational Soft Matter: From Synthetic Polymers to Proteins.
42. Phillips, J., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R., Kalé, L. and Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16), pp.1781-1802.
43. Glykos, N. M. (2016). *The Norma computing cluster*. [online] Available at: https://norma.mbg.duth.gr/index.php?id=about:intro [Accessed 9 Jul. 2018].
44. Perl.org. (2018). *The Perl Programming Language - www.perl.org*. [online] Available at: https://www.perl.org [Accessed 23 Jul. 2018].
45. R-project.org. (2018). *R: The R Project for Statistical Computing*. [online] Available at: https://www.r-project.org [Accessed 23 Jul. 2018].
46. Hutchinson, E. and Thornton, J. (2008). PROMOTIF-A program to identify and analyze structural motifs in proteins. *Protein Science*, 5(2), pp.212-220.
47. Glykos, N. (2006). Software news and updates carma: A molecular dynamics analysis program. *Journal of Computational Chemistry*, 27(14), pp.1765-1768.
48. Koukos, P. and 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.
49. Ks.uiuc.edu. (2018). *PSF Files*. [online] Available at: http://www.ks.uiuc.edu/Training/Tutorials/namd/namd-tutorial-unix-html/node23.html [Accessed 8 Jul. 2018].
50. Utopia.duth.gr. (2018). *Home - plot*. [online] Available at: https://utopia.duth.gr/glykos/plot/ [Accessed 30 Jul. 2018].
51. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
52. Humphrey, W., Dalke, A. and Schulten, K., "VMD - Visual Molecular Dynamics", *J. Molec. Graphics*, 1996, vol. 14, pp. 33-38.
53. Altis, A., Otten, M., Nguyen, P., Hegger, R. and Stock, G. (2008). Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *The Journal of Chemical Physics*, 128(24), p.245102.
54. Mu, Y., Nguyen, P. and Stock, G. (2004). Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins: Structure, Function, and Bioinformatics*, 58(1), pp.45-52.
55. Amadei, A., Linssen, A. and Berendsen, H. (1993). Essential dynamics of proteins. *Proteins: Structure, Function, and Genetics*, 17(4), pp.412-425.
56. Altis, A., Nguyen, P., Hegger, R. and Stock, G. (2007). Dihedral angle principal component analysis of molecular dynamics simulations. *The Journal of Chemical Physics*, 126(24), p. 244111.
57. Zhang, C. and Ma, J. (2010). Enhanced sampling and applications in protein folding in explicit solvent. *The Journal of Chemical Physics*, 132(24), p.244101.
58. Knapp, B., Frantal, S., Cibena, M., Schreiner, W. and Bauer, P. (2011). Is an Intuitive Convergence Definition of Molecular Dynamics Simulations Solely Based on the Root Mean Square Deviation Possible?. *Journal of Computational Biology*, 18(8), pp.997-1005.
59. Plasma-gate.weizmann.ac.il. (2018). *Grace Home*. [online] Available at: http://plasma-gate.weizmann.ac.il/Grace/ [Accessed 1 Sep. 2018].

Appendix

A1: Configuration File used for hp_{NG}-1

```
#
# Input files
#
amber                   on
readexclusions          yes
parmfile                asn-gly-1.prmtop
coordinates             asn-gly-1.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                2.5
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        27.00    0.00    0.00   # <===== CHANGE ME
cellBasisVector2         0.00   27.00    0.00   # <===== CHANGE ME
cellBasisVector3         0.00    0.00   27.00   # <===== 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                     12000                   ;# <===== CHANGE ME
output                  output/equil_ca
```

A2: Populations of β-turns per cluster

hp_{NG}-1

```
Total number of conformations in Cluster 01: 500
Total number of beta-turns: 2
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):     2   0.4 %

Total number of conformations in Cluster 02: 499
Total number of beta-turns: 465
   Type I(%):   445  89.2 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):    20   4.0 %

Total number of conformations in Cluster 03: 500
Total number of beta-turns: 439
   Type I(%):   360  72.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):    79  15.8 %

Total number of conformations in Cluster 05: 500
Total number of beta-turns: 12
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):    12   2.4 %

Total number of conformations in Cluster 09: 500
Total number of beta-turns: 446
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):   366  73.2 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):    80  16.0 %

Total number of conformations in Cluster 10: 500
Total number of beta-turns: 35
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     2   0.4 %
  Type IV(%):    33   6.6 %

Total number of conformations in Cluster 13: 502
Total number of beta-turns: 46
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     6   1.2 %
  Type IV(%):    40   8.0 %

Total number of conformations in Cluster 14: 500
Total number of beta-turns: 346
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):   245  49.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):   101  20.2 %

Total number of conformations in Cluster 15: 500
Total number of beta-turns: 493
   Type I(%):     0   0.0 %
  Type I'(%):   493  98.6 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):     0   0.0 %

Total number of conformations in Cluster 17: 489
Total number of beta-turns: 3
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):     3   0.6 %

Total number of conformations in Cluster 18: 491
Total number of beta-turns: 491
   Type I(%):     0    0.0 %
  Type I'(%):   491  100.0 %
  Type II(%):     0    0.0 %
 Type II'(%):     0    0.0 %
Type VIII(%):     0    0.0 %
  Type IV(%):     0    0.0 %
```

hp_{NG}-2

```
Total number of conformations in Cluster 01: 500
Total number of beta-turns: 496
   Type I(%):     0   0.0 %
  Type I'(%):   488  97.6 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):     8   1.6 %

Total number of conformations in Cluster 02: 499
Total number of beta-turns: 440
   Type I(%):   338  67.7 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     1   0.2 %
  Type IV(%):   101  20.2 %

Total number of conformations in Cluster 03: 501
Total number of beta-turns: 497
   Type I(%):   488  97.4 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):     9   1.8 %

Total number of conformations in Cluster 04: 500
Total number of beta-turns: 18
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):    18   3.6 %

Total number of conformations in Cluster 06: 500
Total number of beta-turns: 2
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):     2   0.4 %

Total number of conformations in Cluster 07: 500
Total number of beta-turns: 52
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     1   0.2 %
  Type IV(%):    51  10.2 %

Total number of conformations in Cluster 09: 499
Total number of beta-turns: 392
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):   315  63.1 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):    77  15.4 %

Total number of conformations in Cluster 10: 499
Total number of beta-turns: 321
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):   197  39.5 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):   124  24.8 %

Total number of conformations in Cluster 14: 499
Total number of beta-turns: 3
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):     3   0.6 %

Total number of conformations in Cluster 15: 512
Total number of beta-turns: 43
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):    43   8.4 %
```

hp_{NG}-3

```
Total number of conformations in Cluster 01: 499
Total number of beta-turns: 480
   Type I(%):   413  82.8 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):    67  13.4 %

Total number of conformations in Cluster 02: 499
Total number of beta-turns: 14
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):    14   2.8 %

Total number of conformations in Cluster 03: 499
Total number of beta-turns: 497
   Type I(%):     0   0.0 %
  Type I'(%):   488  97.8 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):     9   1.8 %

Total number of conformations in Cluster 04: 498
Total number of beta-turns: 484
   Type I(%):   479  96.2 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):     5   1.0 %

Total number of conformations in Cluster 06: 499
Total number of beta-turns: 349
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):   265  53.1 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):    84  16.8 %

Total number of conformations in Cluster 07: 499
Total number of beta-turns: 405
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):   325  65.1 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):    80  16.0 %

Total number of conformations in Cluster 09: 498
Total number of beta-turns: 5
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):     5   1.0 %

Total number of conformations in Cluster 11: 497
Total number of beta-turns: 26
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):    26   5.2 %

Total number of conformations in Cluster 14: 506
Total number of beta-turns: 2
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     0   0.0 %
  Type IV(%):     2   0.4 %

Total number of conformations in Cluster 15: 502
Total number of beta-turns: 36
   Type I(%):     0   0.0 %
  Type I'(%):     0   0.0 %
  Type II(%):     0   0.0 %
 Type II'(%):     0   0.0 %
Type VIII(%):     2   0.4 %
  Type IV(%):    34   6.8 %
```

Script 01

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

(@ARGV == 9) or die "Usage: script06.pl file1 outfile1 outfile2 outfile3 outfile4 outfile5 outfile6 outfile7 outfile8\n";

open (IN, $ARGV[0]) || die "Can not open $ARGV[0] for reading\n";
open ( OUT_1, ">$ARGV[1]") || die "Can not open $ARGV[1] for writing\n";
open ( OUT_2, ">$ARGV[2]") || die "Can not open $ARGV[2] for writing\n";
open ( OUT_3, ">$ARGV[3]") || die "Can not open $ARGV[3] for writing\n";
open ( OUT_4, ">$ARGV[4]") || die "Can not open $ARGV[4] for writing\n";
open ( OUT_5, ">$ARGV[5]") || die "Can not open $ARGV[5] for writing\n";
open ( OUT_6, ">$ARGV[6]") || die "Can not open $ARGV[6] for writing\n";
open ( OUT_7, ">$ARGV[7]") || die "Can not open $ARGV[7] for writing\n";
open ( OUT_8, ">$ARGV[8]") || die "Can not open $ARGV[8] for writing\n";

$/=undef;
$input = <IN>;
@super_pdb = split (/ENDMDL/, $input);
for ($i=0; $i<=64; $i++)
{
        printf OUT_1 "$super_pdb[$i]";
}
for ($i=65; $i<=129; $i++)
{
        printf OUT_2 "$super_pdb[$i]";
}
for ($i=130; $i<=194; $i++)
{
        printf OUT_3 "$super_pdb[$i]";
}
for ($i=195; $i<=259; $i++)
{
        printf OUT_4 "$super_pdb[$i]";
}
for ($i=260; $i<=324; $i++)
{
        printf OUT_5 "$super_pdb[$i]";
}
for ($i=325; $i<=389; $i++)
{
        printf OUT_6 "$super_pdb[$i]";
}
for ($i=390; $i<=454; $i++)
{
        printf OUT_7 "$super_pdb[$i]";
}
for ($i=455; $i<=519; $i++)
{
        printf OUT_8 "$super_pdb[$i]";
}

close( OUT_1);
close( OUT_2);
close( OUT_3);
close( OUT_4);
close( OUT_5);
close( OUT_6);
close( OUT_7);
close( OUT_8);
close (IN);

exit(0);
```

Script 02

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

(@ARGV == 2) or die "Usage: script07.pl file1 outfile\n";
open (IN, $ARGV[0]) || die "Can not open $ARGV[0] for reading\n";
open ( OUT, ">$ARGV[1]") || die "Can not open $ARGV[1] for writing\n";

$a=0;
$b=0;
$c=0;
$d=0;
$e=0;
$f=0;
$/=undef;
$input = <IN>;
@turns = split (/\n/, $input);
```

```
foreach $turns (@turns)
{
        if ($turns =~ /\sI\s/)
        {
                $a++;
                $turns=$';
        }
        elsif ($turns =~ /I\'/)
        {
                $b++;
                $turns=$';
        }
        elsif ($turns =~ /\sII\s/)
        {
                $c++;
                $turns=$';
        }
        elsif ($turns =~ /II\'/)
        {
                $d++;
                $turns=$';
        }
        elsif ($turns =~ /VIII/)
        {
                $e++;
                $turns=$';
        }
        elsif ($turns =~ /IV/)
        {
                $f++;
                $turns=$';
        }
        elsif ($turns =~ /Cluster\s+(\d+)\S\s+(\d+)/)
        {
                print OUT "Total number of conformations in Cluster $1: ", $2, "\n";
                $tot=$2;
        }
}

$sum=$a+$b+$c+$d+$e+$f;
$x1=100*$a;
$x2=100*$b;
$x3=100*$c;
$x4=100*$d;
$x5=100*$e;
$x6=100*$f;

print OUT "Total number of beta-turns: ", $sum, "\n";
printf OUT "%13s %5d %- 2.1f %1s\n", "Type I(%):", $a, ($x1/$tot), "%";
printf OUT "%13s %5d %- 2.1f %1s\n", "Type I'(%):", $b, ($x2/$tot), "%";
printf OUT "%13s %5d %- 2.1f %1s\n", "Type II(%):", $c, ($x3/$tot), "%" ;
printf OUT "%13s %5d %- 2.1f %1s\n", "Type II'(%):", $d, ($x4/$tot), "%";
printf OUT "%13s %5d %- 2.1f %1s\n", "Type VIII(%):", $e, ($x5/$tot), "%";
printf OUT "%13s %5d %- 2.1f %1s\n\n", "Type IV(%):", $f, ($x6/$tot), "%";

close ( OUT);
close (IN);

exit(0);
```

Script 03

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

open (INFILE, $ARGV[0]) || die "Can not open $ARGV[0]\n";

$i=1;
$/=undef;
$input = <INFILE>;
@pdb = split (/\n/, $input);
foreach $pdb (@pdb)
{
       if ($pdb =~ /^C\s*(\S*)\s*(\S*)\s*(\S*)\s*(\S*)\s*(\S*)\s*(\S*)/)
       {
              printf "%-6s %4d  %-3s %3s %1s %3d %11.3f %7.3f %7.3f %5.2f %5.2f %11s\n", ATOM, $i, $6, $4, A, $5, $1, $2, $3, 1, 0, C;
              $i++;
              $pdb=$';
       }
       elsif ($pdb =~ /^N\s*(\S*)\s*(\S*)\s*(\S*)\s*(\S*)\s*(\S*)\s*(\S*)/)
       {
              printf "%-6s %4d  %-3s %3s %1s %3d %11.3f %7.3f %7.3f %5.2f %5.2f %11s\n", ATOM, $i, $6, $4, A, $5, $1, $2, $3, 1, 0, N;
              $i++;
              $pdb=$';
       }
       elsif ($pdb =~ /^O\s*(\S*)\s*(\S*)\s*(\S*)\s*(\S*)\s*(\S*)\s*(\S*)/)
       {
              printf "%-6s %4d  %-3s %3s %1s %3d %11.3f %7.3f %7.3f %5.2f %5.2f %11s\n", ATOM, $i, $6, $4, A, $5, $1, $2, $3, 1, 0, O;
              $i++;
              $pdb=$';
       }
}
print "END\n";

close (INFILE);
exit(0);
```

Script 04

```
#Example of the script used for the first heptapeptide.

A <- matrix(scan("asn-gly-1.RMSD.matrix", n=5033*5033), 5033, 5033, byrow = TRUE)
hc <- hclust( as.dist(A), method="complete")
plot(hc)
postscript(hc)
plot(hc)
dev.off()
cutree( hc, h = 2.2)
rect.hclust( hc, h= 2.2, border="red")
clusters <- cutree( hc, h = 2.2)
as.data.frame(clusters)
names(a) <- NULL
write.table(a, file = "all_clusters.list", sep = " ", quote = FALSE)
q()
```