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

Final year thesis

“Folding of an FTZ-derived peptide by molecular dynamics”

Adamidou Triantafyllia
Advisors: Dr. Nicholas M. Glykos
Dr. Katsani Aikaterini

Alexandroupolis 2017

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΓΕΝΕΤΙΚΗΣ

Διπλωματική Εργασία

«Μελέτη αναδίπλωσης μέσω προσομοιώσεων μοριακής δυναμικής ενός πεπτιδίου προερχόμενο από την πρωτεΐνη FTZ»

Αδαμίδου Τριανταφυλλιά

Επιβλέποντες καθηγητές: Δρ. Γλυκός Νικόλαος
Δρ. Κατσάνη Αικατερίνη

Αλεξανδρούπολη 2017

Acknowledgments

I would like to thank my two supervisors, Dr. Nicholaos M. Glykos and Dr. Aikaterini Katsani for helping me to improve a critical way of thinking and teaching me to always serve the truth. They have been both a great inspiration for me. I would also like to thank NMG group for creating a friendly and team spirit environment. Moreover, I would like to thank personally Athanasios Baltzis for his assistance and useful advices. My friends and my family deserve my gratitude, because each and one of them strongly supported me in this journey.

“Never forget where you started from and never doubt how far you can go.”

Table of contents

Chapter 1: Introduction

Chapter 2: Molecular Dynamics Simulation

Chapter 3: Methods

Chapter 4: Results

Chapter 5: Discussion

Abstract

Molecular Dynamics (MD) are being used extensively for the identification of molecules' structure and folding procedure. In the present thesis we examine the accuracy of this method and its ability to approach the experimentally identified structures. More specifically, a 8.87 μs folding simulation was carried out, using AMBER99SB-STAR-ILDN force field and TIP3P water model, for FTZpep peptide. FTZpep is a synthetic peptide containing LXXLL-related motifs of FTZ. This peptide takes part in the formation of parasegments in Drosophila melanogaster embryos. FTZpep is composed of 19 residues: VEERPSTLRALLTNPVKKL. It has been proved through X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy experiments that it shows a nascent helical conformation in aqueous solution and it forms a long stretch of α-helix in the presence of trifluoroethanol (TFE) and receptor FTZ-F1. The results from the simulation confirm the experimental results, with the RMSD matrix showing dynamic behavior and the secondary structure analysis implying helical conformations. NMR distance restraints and J-couplings that were produced by the simulation, mostly agree with the experimental data.

Keywords: Molecular Dynamics, force fields, AMBER99SB-STAR-ILDN, FTZpep, fushi tarazu, folding procedure, NMR, NOEs, J-couplings

Περίληψη

Οι προσομοιώσεις Μοριακής Δυναμικής έχουν χρησιμοποιηθεί εκτενώς με σκοπό τον προσδιορισμό της δομής μορίων και την μελέτη της αναδίπλωσής τους. Στόχος μας είναι η εξακρίβωση των ικανοτήτων των προσομοιώσεων Μοριακής Δυναμικής να αναπαραστήσουν τα εργαστηριακά πειράματα. Πραγματοποιήθηκε προσομοίωση 8.87 μs σε ένα συνθετικό πεπτίδιο (FTZpep) προερχώμενο από την πρωτεΐνη FTZ, με την χρήση του δυναμικού πεδίου AMBER99SB-STAR-ILDN και του μοντέλου νερού TIP3P. Το FTZpep περιέχει δομικά μοτίβα LXXLL και λαμβάνει μέρος στη ρύθμιση του σχηματισμού των παραμεταμερών στο έμβρυο της Drosophila melanogaster. Αποτελείται από τα εξής 19 αμινοξέα: VEERPSTLRALLTNPVKKL. Έχει αποδειχθεί πειραματικά μέσω κρυσταλλογραφίας ακτίνων Χ και φασματογραφίας NMR πως το συγκεκριμένο πεπτίδιο εμφανίζει ελικοειδή διαμόρφωση όταν βρίσκεται σε υδατικό διάλυμα, ενώ διαμορφώνεται α-έλικα παρουσία διαλύματος TFE και του υποδοχέα FTZ-F1. Τα αποτελέσματα της προσομοίωσης ταυτίζονται με τα πειραματικά δεδομένα, με τον πίνακα RMSD να επιβεβαιώνει την δυναμική συμπεριφορά του πεπτιδίου στο νερό ενώ το γράφημα δευτεροταγούς δομής καταδεικνύει ελικοειδείς διαμορφώσεις και παρουσία α-έλικας για συγκεκριμένα κατάλοιπα. Τα δεδομένα για distance restraints και J-couplings που προκύπτουν από την προσομοίωση εμφανίζουν μεγάλο ποσοστό συμφωνίας με τα πειραματικά.

Λέξεις κλειδιά: Μοριακή Δυναμική, δυναμικά πεδία, διαδικασία αναδίπλωσης, AMBER99SB-STAR-ILDN, FTZpep, fushi tarazu, NMR, NOEs, J-couplings

Chapter 1: Introduction

1.1 Proteins

Proteins are large biomolecules which are composed of amino acids (referred also as residues). Their participation in many reactions such as catalysis, DNA replication/transcription, molecule transportation, etc, is essential for life. They differ between them due to their residues and consequently to their shape. Polypeptides are linear amino acid sequences and each protein is constituted of at least one polypeptide. Amino acids are bonded together between the -NH2 of one amino acid and the -COOH of another, through peptide bonds, to form the primary structure. Secondary structure is referred to the local formations such as α-helix or β-sheet. These local structures are formed through hydrogen bonds between the oxygen of the C=O group of a peptide bond and the hydrogen of the N-H group of another peptide bond. [1]

A polypeptide folds into it’s 3-dimensional structure through the procedure of protein folding. Amino acids interact with each other through bonds in order to form the tertiary structure of the peptide which is called the native state of the protein. In the native state, the protein is folded and functional. According to Anfisen's hypothesis, the tertiary structure is determined by the amino acids' sequence of the primary structure of the protein. [2] The discovery of the way a protein is folded into it’s native state is essential for the understanding of protein function. Although numerous of researchers have dedicated their research on protein folding, it has not been fully discovered yet, the way a protein folds into it’s tertiary structure. This is called the “protein folding problem” and it still remains unsolved. However, many theories have been developed in order to explain this procedure.

1.2 The chronicles of the protein folding problem

Christian P. Anfinsen conducted in 1961 denaturation and annealing experiments of ribonuclease. It came into light through those experiments, that the tertiary structure of a protein is determined by the primary structure and amino acids' sequence, under certain conditions. [3] It is prefered spontaneously by the protein, the lowest free energy state (ΔGfolding < 0). In 1968, Cyrus Levinthal noted that an unfolded polypeptide chain has an astronomical number of possible conformations, due to a huge number of degrees of freedom. That means that it should take a lot of time to fold into it’s native state. If a polypeptide had to sample all the possible folding paths in order to acquire it’s native state, that would take longer than the age of the universe! However, a polypeptide is folded into its native state in a few milliseconds or sometimes even microseconds. Levinthal came to the conclusion that during the folding procedure, small amino acid sequences are forming structures which stabilize locally the protein and guide it to a specific folding path. [4] The state in which the polypeptide is partially folded into these local formations is called transition state and it can also be explained through funnel-like energy landscape. [5,11]

Those two milestones led to the conduction of a numerous of experiments and many theories have been developed. However, the folding procedure is not described fully and precisely by none of them.

1.3 Models of protein folding

1.3.a Hydrophobic collapse hypothesis

This hypothesis is based on the fact that spherical proteins consist of a hydrophobic core. This core is placed on the inner part of the protein and it is created from side chains of non-polar amino acids. Most of the polar amino acids are placed in the outer space of the protein and they are exposed to solutions, such as water. According to “hydrophobic collapse hypothesis”, initial secondary structures are formed due to the hydrophobic interactions. Due to those interactions, a transition state is formed which has a lower free energy from the unfolded peptide. It’s free energy though is still higher than the free energy of the native state of the protein. [6]

1.3.b Diffusion – collision hypothesis

Martin Karplus and David L. Weaver formulated the diffusion – collision hypothesis and they stated that a protein consists of several smaller microdomains which fold and collide with each other, leading to the folding of the whole protein into it’s native state. This procedure enables time-saving, as the protein does not sample all the possible folding paths during it’s folding procedure and it is guided by already folded microdomains. [7]

1.3.c Nucleation – Condensation mechanism

Secondary structure formations happen at the same time with tertiary formation structures and they interact with each other. Structures are formed due to hydrophobic interactions, into a core which guides the folding of the whole protein around it. [8]

1.3.d Folding funnels – Energy landscapes

The energy difference between the unfolded and folded state of a protein is defined by Enthalpy and Entropy. Using the second law of thermodynamics, the unfolded protein is more favorable when entropy is higher. On the other hand, the folded state is favored by enthalpy. Hydrogen bonds, ionic and van der Waals interactions are present in the well defined and stable native state of the protein. Entropy decreases when the protein is folding into the native state. [9] The difference between entropy and enthalpy is called Gibbs free energy and it's magnitude determines if a protein will be in it's folded or unfolded state.

$$\mathbf{\Delta G = \Delta H - T\Delta S}$$

In respect with the folding funnel hypothesis, it is assumed a protein's native state is acquired in cell conditions when Gibbs free energy's magnitude reaches it's minimum (negative magnitude). The energy minimum is placed on the bottom of the funnel and it represents a unique tertiary structure. Energy landscapes consist of various local minima which are related to non-native structures. Thermodynamically non-stable structures (higher free energy) are placed on the hills of the funnel, while more stable structures (lower free energy) can be found in the valleys. [10]

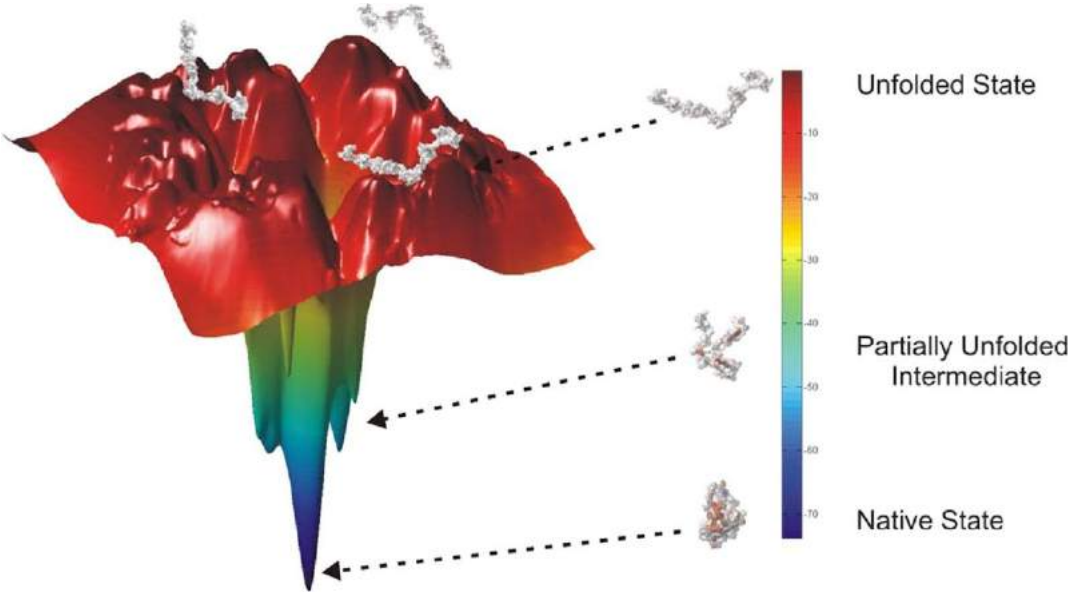


Figure 1: Energy landscape. (Reproduced without permission from A Quintas, OA Biochemistry (UK), 2013)

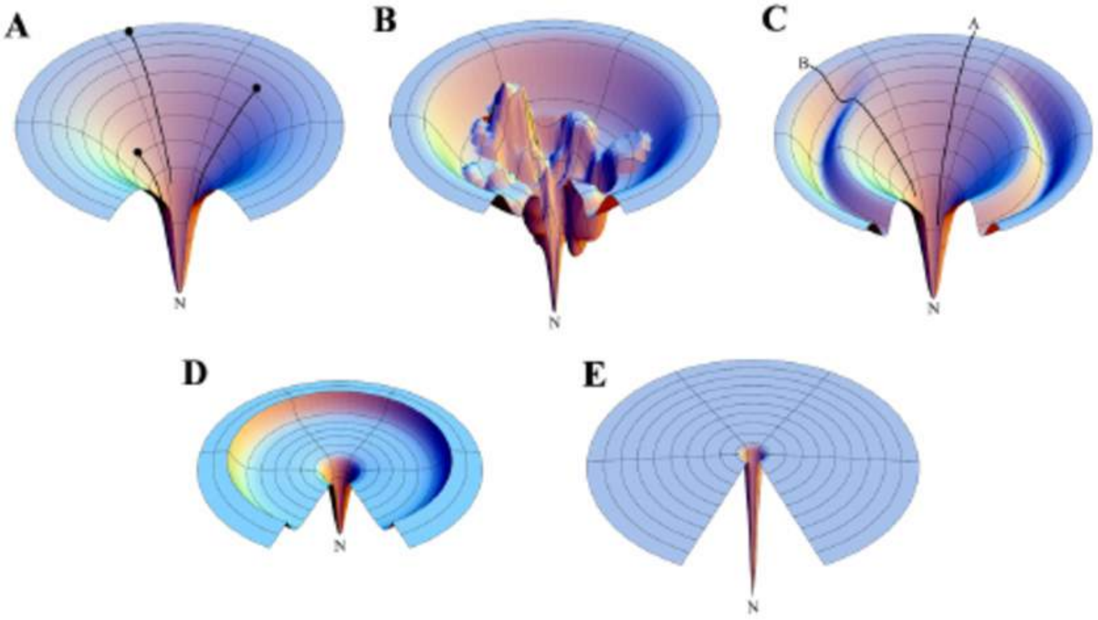


Figure 2: Energy landscapes. (Reproduced without permission from Dill, Ken A., and Hue Sun Chan. "From Levinthal To Pathways To Funnels". *Nature Structural & Molecular Biology* (1997))

1.4 Protein Folding Experiments

Many proteins' structures have been discovered through classical experiments. Using X-ray crystallography, it can be identified the secondary and tertiary structure, as long as the protein can form well-defined crystals that permit X-ray diffraction. NMR is another technique which is widely used in structure experiments. There should be mentioned also Circular Dichroism (CD), Mass Spectroscopy, AFM, SAXS, FR-IR Spectroscopy, etc. All the techniques mentioned above, are conducted in order to predict a molecule's structure and the data that are mined are used further for computational simulations. Computational studies are the connection link between experiments and theory by identifying protein structure and protein folding paths. Molecular Dynamics Simulation is a precise computational tool for both basic and applied research (molecular docking, etc).

1.5 FTZpep and FTZ – F1 receptor

Fushi tarazu (ftz) is a pair rule gene which is essential for the formation of parasegments in Drosophila melanogaster embryos. Fushi tarazu is expressed in vertical stripes very early in the development of Drosophila melanogaster embryos and it’s apparent complement gene is even-skipped. Seven stripes of ftz are interspersed with seven stripes of even-skipped forming a total of fourteen evenly spaced alternating bands that define the boundaries between future body segments in the adult fly. This is the reason why mutant embryos possess only half of the normal number of body segments. FTZ is a gene activator and it is the product of transcription and translation of fushi tarazu gene. [12, 53]

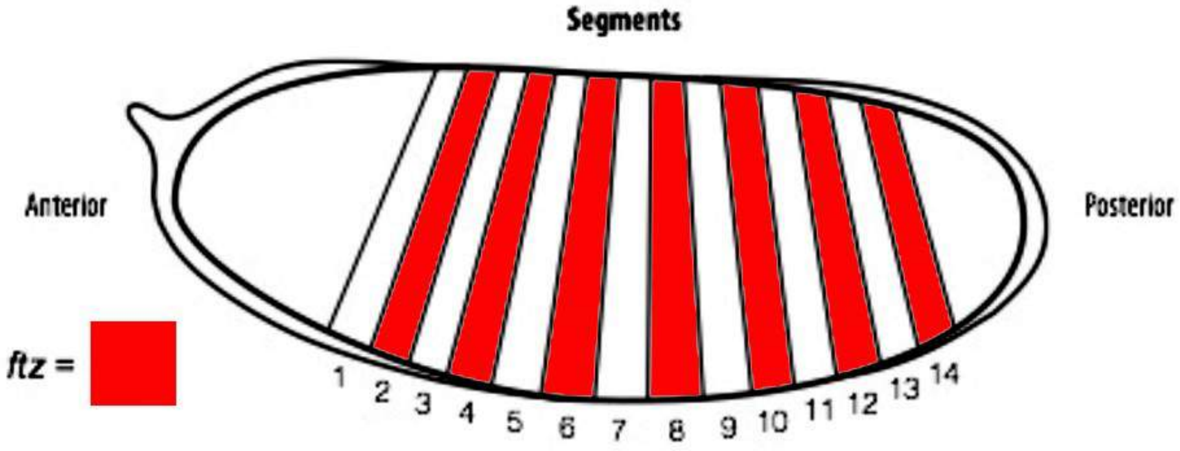


Figure 3: Fushi tarazu and even-skipped stripes in Drosophila melanogaster embryo.(Reproduced without permission from the British Society of Developmental Biology, http://thenode.biologists.com/bsdb-gurdon-summer-studentship-report-4/research/)

Fushi tarazu factor 1 (FTZ-F1) is an orphan nuclear receptor which is uniformly expressed in Drosophila melanogaster embryos and interacts with FTZ to define the segmental regions in Drosophila embryo. FTZ-F1 acts as an activator of fushi tarazu in cooperation with FTZ. It's transcription process is activated through the binding of FTZ to the ligand-binding domain (LBD) of FTZ-F1 while it has also a DNA binding domain. FTZ and FTZ-F1 cooperate to regulate target gene expression. It has been shown that FTZ, as a transcriptional co-activator, regulates transcriptional signals through binding to nuclear receptors using conserved LXXLL-related motifs (NR boxes).

Ji-Hye Yun, Chul-Jin Lee, Jin-Won Jung, and Weontae Lee determined solution structures by NMR spectroscopy of the cofactor peptide (FTZpep) with residues VEERPSTLRALLTNPVKKL. FTZpep is a synthetic peptide which is extracted from FTZ and contains LXXLL-related motifs of FTZ. The structure of FTZpep shows a nascent helical conformation in aqueous solution. A long stretch of α-helix is formed in the binding with the receptor protein and in the presence of TFE, imitating the native structure. An α-helix is exhibited with a bend near proline at +8 position in the solution structure of FTZpep when it is binded to the receptor FTZ-F1. [13]

1.6 Purpose of the present thesis

The purpose of the present thesis is to study the folding procedure of FTZpep through Molecular Dynamics simulations. Furthermore, we aim to compare our results with NMR and CD structure results of Ji-Hey Yun et al, in order to confirm whether computational simulations do assist and in which way in protein structure prediction and in protein folding path understanding.

Chapter 2: Molecular Dynamics Simulation

2.1 Introduction

Molecular Dynamics are used widely in molecular structure identification experiments and as an attempt to link classical experiments with computational methods. Moreover, MD simulations is a useful tool for the theoretical study of biomolecules concerning their behavior over time, their structure and interactions between molecules. Accessing information that could not be obtained through classical experiments only, comparing and contrasting theory with experimental data which may be derived from NMR, CD, X-ray crystallography experiments and helping researchers to create an initial idea of their molecule’s structure and behavior are only a few of the many potentials that are emerged by Molecular Dynamics simulations. [14]

There are two main categories of Molecular Dynamics: MD simulations (Molecular Dynamics Simulations) and MC (Monte Carlo Simulations). MD simulations are dealing with the dynamic properties of systems and MC simulations are based on statistical and probabilistic methods. They are used mostly separately, although they can be both combined in some algorithms, such as Langevin's Dynamics and Brownian Dynamics which are used for complex computational simulations.

2.2 Molecular Interactions

The equation of motion and Newton’s second law synthesize the basis of Molecular Dynamics simulations. Knowing the force of each particle that is part of the system, it can be determined its kinetic parameters (velocity, acceleration) and its time depended position, creating a trajectory of this system in which are described the positions, velocities and accelerations of the particles on time dependence.

According to Newton’s second law:

$$\mathbf{F = m\ a\ (\ 1\ )}$$

With F standing for the force, m for the mass and a for the acceleration of the particle.

The force can be also expressed depending on the potential energy, as follow:

$$\mathbf{F = -\ dV\ /\ dr\ (\ 2\ )}$$

With V standing for potential energy and r for particle’s position.

The combination of the those two equations leads to:

$$\mathbf{a = - 1 / m \;\; dV / dr \; (3)} \text{ and } \mathbf{– dV / dr = m \; d^2r / d^2t \; (4)}$$

Formula (3) links derivative of potential energy with acceleration on dependence with time and formula (4) links the derivative of potential energy with a position on dependence with time. Equations (3) and (4) prove that the situation prediction of a system at every possible time is attainable, insomuch initial positions, velocities' allocations and accelerations of the particles are known. The initial positions are known through NMR and/or X-ray Crystallography experiments and velocities' allocations are calculated through Maxwell - Boltzmann or Gaussian formula:

$$\mathbf{p(v) = (m / 2\pi k_B T)^{1/2} \exp (- m \, v^2 / 2k_B T) \; (5)}$$

with v standing for velocity, k_B is Boltzmann's constant and T states the temperature of the system.

Acceleration is calculated through potential energy calculation, using force fields. The calculation of acceleration is a computationally demanding process. For this reason, it is satisfactorily approached by a numerous of algorithms, some of them are the following:

- Verlet
- Leap – frog
- Velocity Verlet
- Beeman’s

Most of those algorithms are based on Taylor’s series which is based on the reduction of an equation’s terms:

$$\mathbf{r(t + dt) = r(t) + v(t)\ dt + ½\ a(t)\ dt^2 +\dots}$$

$$\mathbf{v(t + dt) = v(t) + a(t)\ dt + ½\ b(t)\ dt^2 +\dots}$$

$$\mathbf{a(t + dt) = a(t) + b(t)\ dt +...}$$

Where r is the position, v is the velocity (the first derivative with respect to time), a is the acceleration (the second derivative with respect to time), etc

For example, using Verlet algorithm, positions for time t+dt can be determined by using positions and accelerations for t and t-dt, such as:

$$\mathbf{r(t + dt) = r(t) + v(t)\ dt + ½\ a(t)\ dt^2}$$

$$\mathbf{r(t – dt) = r(t) – v(t)\ dt + ½\ a(t)\ dt^2}$$

Summing the two equations above:

$$\mathbf{r(t + dt) = 2r(t) – r\ (t – dt) + a(t)\ dt^2 \ (\ 6\ )}$$

The algorithms mentioned above are not fully accurate because they are used in order to approach the acceleration of particles. Due to that, the choice of which algorithm will be used should be done wisely. Each algorithm should be tested in order to produce data that are as close as possible to reality.

2.3 Force Fields

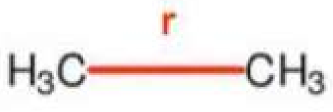
Force fields are empirical equations that are used in Molecular Dynamics simulations for the calculation of potential energy according to particles' position and interactions. Those interactions are separated into two categories: internal or bonded and external or non-bonded. Bond stretch, angle bend and torsion angle are described by bonded interactions and van der Waals interaction energy and electrostatic interaction energy are described by non-bonded interactions. The sum of both bonded and non-bonded terms equals the potential energy of the system. [15]

$$\mathbf{V = E_{bonded} + E_{non\text{-}bonded}\ (\ 7\ )}$$

The term E_{bonded} is calculated by the following formula:

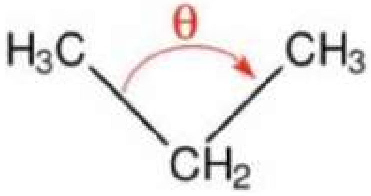
$$\mathbf{E_{bonded} = E_{bond\text{-}stretch} + E_{angle\text{-}bend} + E_{rotate\text{-}along\text{-}bond}} \quad \mathbf{(\ 8\ )}$$

The first term of formula (8) represents the interaction between two atoms which are bonded through a covalent bond and it is depending on the transposition of the atoms from the initial length r_0. K_b is a constant which is determined by bond valence.



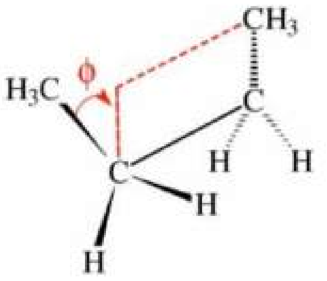
$$E_{bond\text{-}stretch} = K_b(r\text{-}r_0)^2$$

The second term is referred to the angle variation.

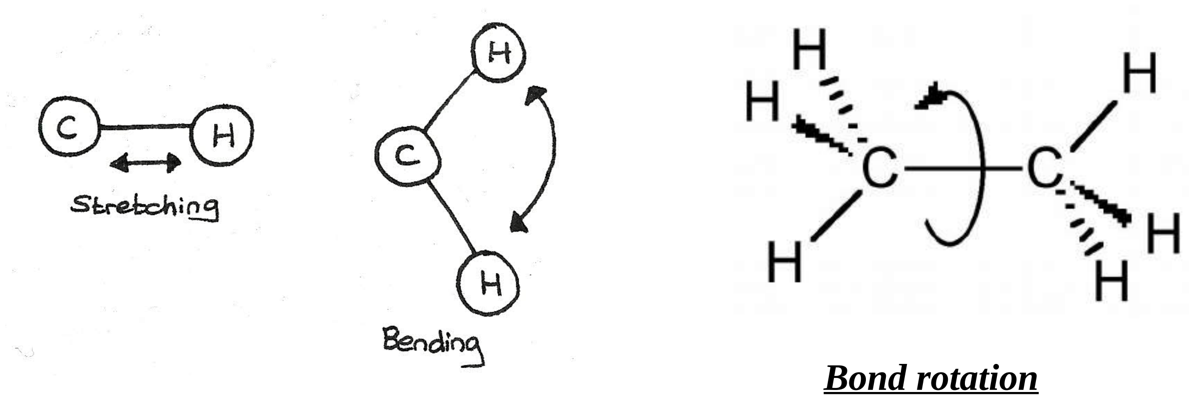


$$E_{angle\text{-}bend} = K_\theta\,(\theta\text{-}\theta_0)^2$$

The last term calculates the potential energy of the system through the torsions over a dihedral angle.



$$E_{rotate\text{-}along\text{-}bond} = K_\varphi[1+\cos(n\varphi\text{-}\delta)]$$



Bond rotation

Figure 4: Reproduced without permission from https://revise.im/chemistry/cer/analysis and http://archive.cnx.org/contents/895d8b18-b41d-49d1-ba68-3abfbd9af48d@3/stereochemistry

The term $E_{non\text{-}bonded}$ is calculated by the following formula and is referred to atoms that are separated by 3 or more bonds or to atoms that belong to different molecules.

$$\mathbf{E_{non\text{-}bonded} = E_{van\text{-}der\text{-}Waals} + E_{electrostatic}} \quad \mathbf{(\ 9\ )}$$

$E_{van\text{-}der\text{-}Waals}$ is also known as Leonnard-Jones potential:

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

The possibility of interaction between two atoms increases while it’s distance “r” decreases. There is a specific distance “r” in which the potential energy acquires zero value and it keeps decreasing as the distance decreases as well. Potential energy will get the minimum possible when the distance between the two atoms is small enough. This is the distance in which the equilibrium is achieved. If the two atoms will approach each other more than the equilibrium distance, repelling forces are created between them and the potential energy begins to increase. Figure 5 shows the diagram of the potential energy and the distance “r” between two hydrogen atoms.

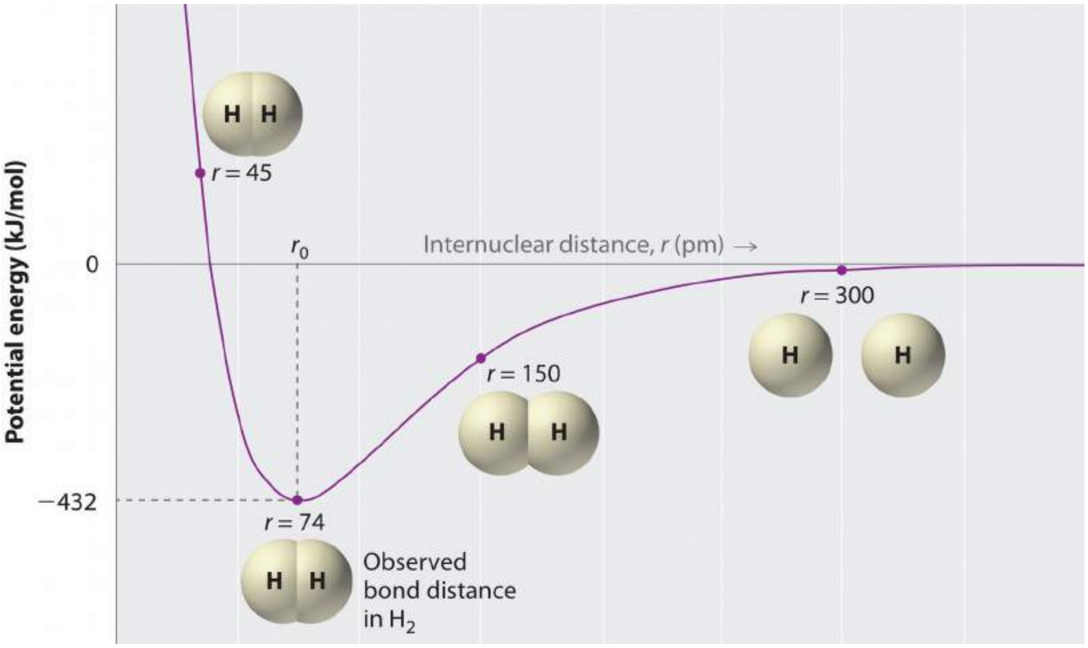


Figure 5: Reproduced without permission. Found online: http://2012books.lardbucket.org/books/principles-of-general-chemistry-v1.0/s12-ionic-versus-covalent-bonding.html

The second term, $E_{electrostatic}$ is referred to the electrostatic interactions and is described by the Coulmb’s law:

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

Among the most used force fields are the following:

- AMBER (Assisted Model Building with Energy Refinement) [16]
- CHARMM (Chemistry at Harvard Macromolecular Mechanics) [17]
- GROMOS (Groningen Molecular Simulation) [18]
- OPLS (Optimized Potentials for Liquid Simulations) [19]

Despite the fact that the force fields mentioned above share a common calculation method for potential energy, there are a few differences between them regarding the parameters and the calculation of bonded and non-bonded interactions. Force fields are continuously updated and evolved in order to achieve a greater agreement between Molecular Dynamics data and the experimental data.

2.4 Solvent in Molecular Dynamics Simulations

Solvent models consist a variety of methods within the field of computational biology and simulations in order to imitate the behavior of the solvent in the experiment. The most common solvent is water. The use of a solvent in Molecular Dynamics simulations is essential because of it's effects on the structure of the molecule, on the thermodynamical parameters of a biological system and on the electrostatic interactions between the molecules. [22] When a water model is used, simulations and thermodynamic calculations can be applied to procedures which take place in a solution. [20, 21] There are many water models that are used in simulations. Some of them are the flexible, extended simple point charge (SPCE-F) model and the flexible three-center (F3C) model. [15] In Molecular Dynamics simulation, can be used either implicit or explicit water solvent which both of them have advantages and disadvantages.

2.4.a Implicit water models

Implicit (continuum) solvents are models in which it is considered that solvent molecules can be replaced by a medium that is continuous and homogeneously able to be polarized. The medium has to be characterized to a good approximation of equivalent properties. No explicit solvent molecules are present and no explicit solvent coordinates are given. Only a small number of parameters is necessary for implicit water solvent in order to be accurate. This is caused by the thermally averaged and isotropic character of the solvent. The main parameter of implicit water models is dielectric constant (ε), and it is often accompanied by other parameters, such as surface tension. Implicit water models are considered to be computationally more efficient and provide a sensible description of the solvent behavior. While the solvent is ordering itself around a solute molecule, it's density may vary. The disadvantage of implicit water models in that they fail to predict the local fluctuations of solvent's density around a solute molecule.

2.4.b Explicit water models

In contrast with implicit water models, explicit solvent models take under consideration the molecular details of each solvent molecules, offering a more realistic picture of the experiment. There are direct and specific solvent interactions with a solute, including the coordinates and usually some of the molecular degrees of freedom. These models are frequently used in molecular mechanics (MM) and dynamics (MD) or Monte Carlo (MC) simulations. A great advantage of explicit water models is the ability of reduction of the degrees of freedom which are measured in the energy calculation, without a significant loss in the overall accuracy of the results at the same time. This characteristic enables explicit solvent water models to be considered as idealized models. However, due to that reason, some of those models can seem useful only under certain circumstances. TIPXP (where X is the number of sites used for energy evaluation) [33] and the simple point charge model (SPC) of water have been used extensively. A typical model of this kind uses a fixed number of sites (often 3 for water). We used TIP3P explicit water model. Explicit water models are considered geometrically strict because they have some geometrical parameters fixed, such as the bond length or angles. Explicit models require usually more computational power than implicit water models but they can provide more accurate and close tor reality results and the describe better the real solvent.

Chapter 3: Methods

3.1 Technical characteristics of our computational system

The study of the folding procedure of FTZpep was done through Molecular Dynamics simulation using NAMD program. [23] NAMD is a software developed for simulations of big biomolecular systems. NAMD is compatible with AMBER and CHARMM force fields. In our simulation, we used AMBER force field and more specifically we used 99SB-STAR-ILDN [50]. MD simulation is a demanding procedure which requires high computational power. Parallel connection of computers for the creation of a cluster is a way to lower the computational cost and increase the performance of the simulation, due to the apportionment of the computational tasks.

Norma is a computational cluster in which this simulation was carried out. [24] It is used for computational biology and crystallography experiments by the members of structural and computational biology group. [25] Norma is a stateless Beowulf-class computing cluster based on the Caos NSA GNU/Linux distribution and it includes 40 CPU cores, 46 Gbytes of physical memory and 6 GPGPUs dispensed over 10 nodes which are based on Intel's Q6600 Kentsfield 2.4 GHz quad processors and are connected via a dedicated HP ProCurve 1800-24G Gigabit ethernet switch. Each of the nine nodes offers four cores, four Gbytes of physical memory and two (gigabit) network interfaces. Only one node constitutes an exception because it is based on Intel's i7 965 extreme and offers six Gbytes 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 comes with four cores, eight Gbytes of physical memory, 1.5 Tbytes of storage in the form of a RAID-5 array of four disks, three (gigabit) network interfaces, and an nvidia GTX-260 GPU. The cluster of Norma is located at the Department of Molecular Biology and Genetics of Democritus University of Thrace in Alexandroupolis, Greece.

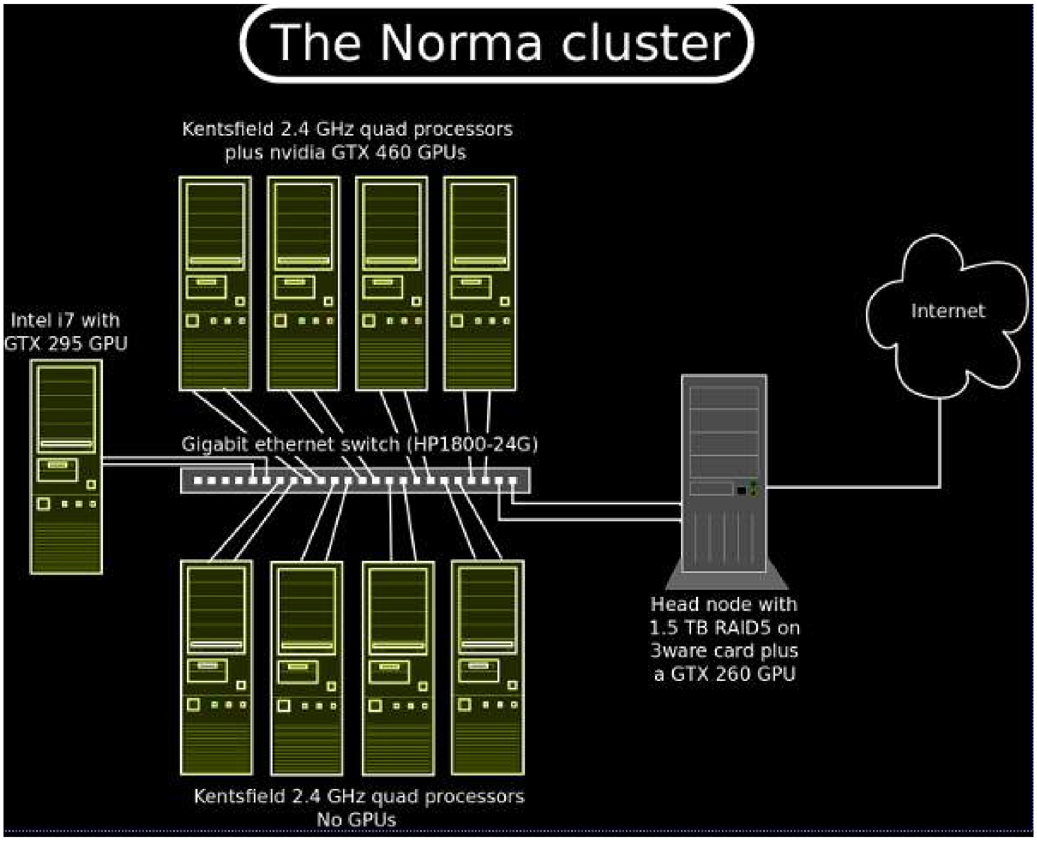
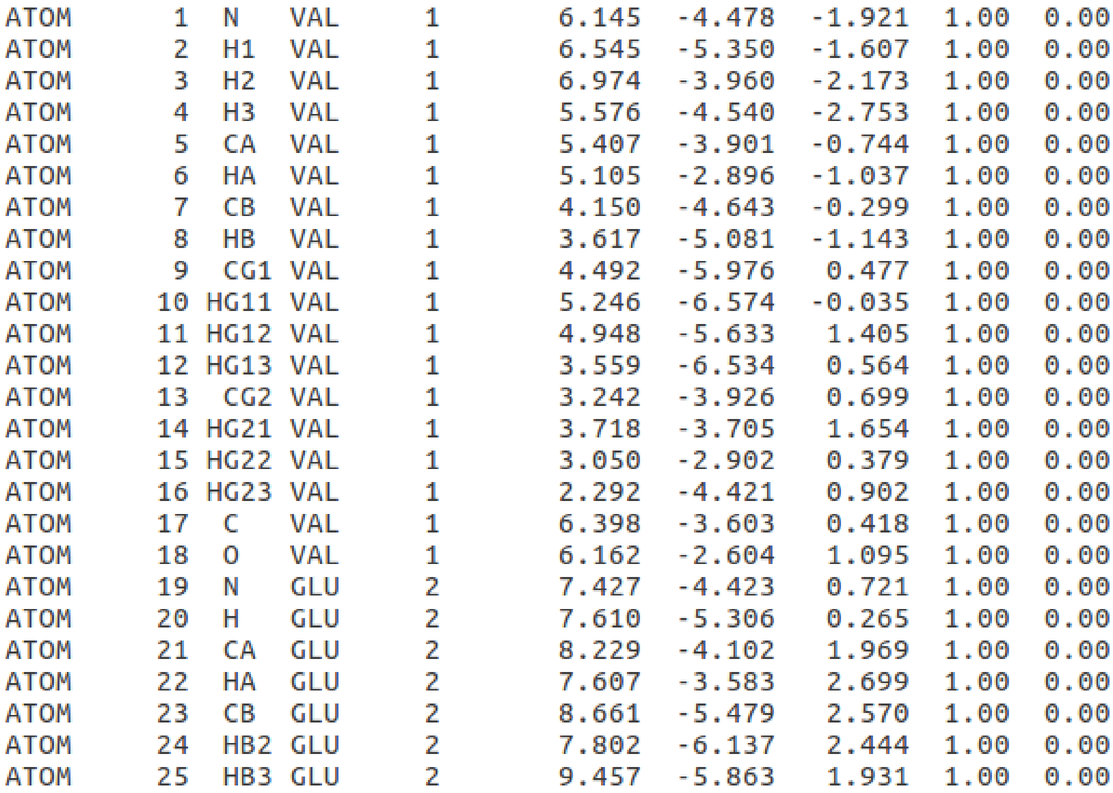


Figure 6: Schematic diagram of Norma cluster.

3.2 Simulation with NAMD

MD simulations using NAMD and AMBER force field require at least three documents:

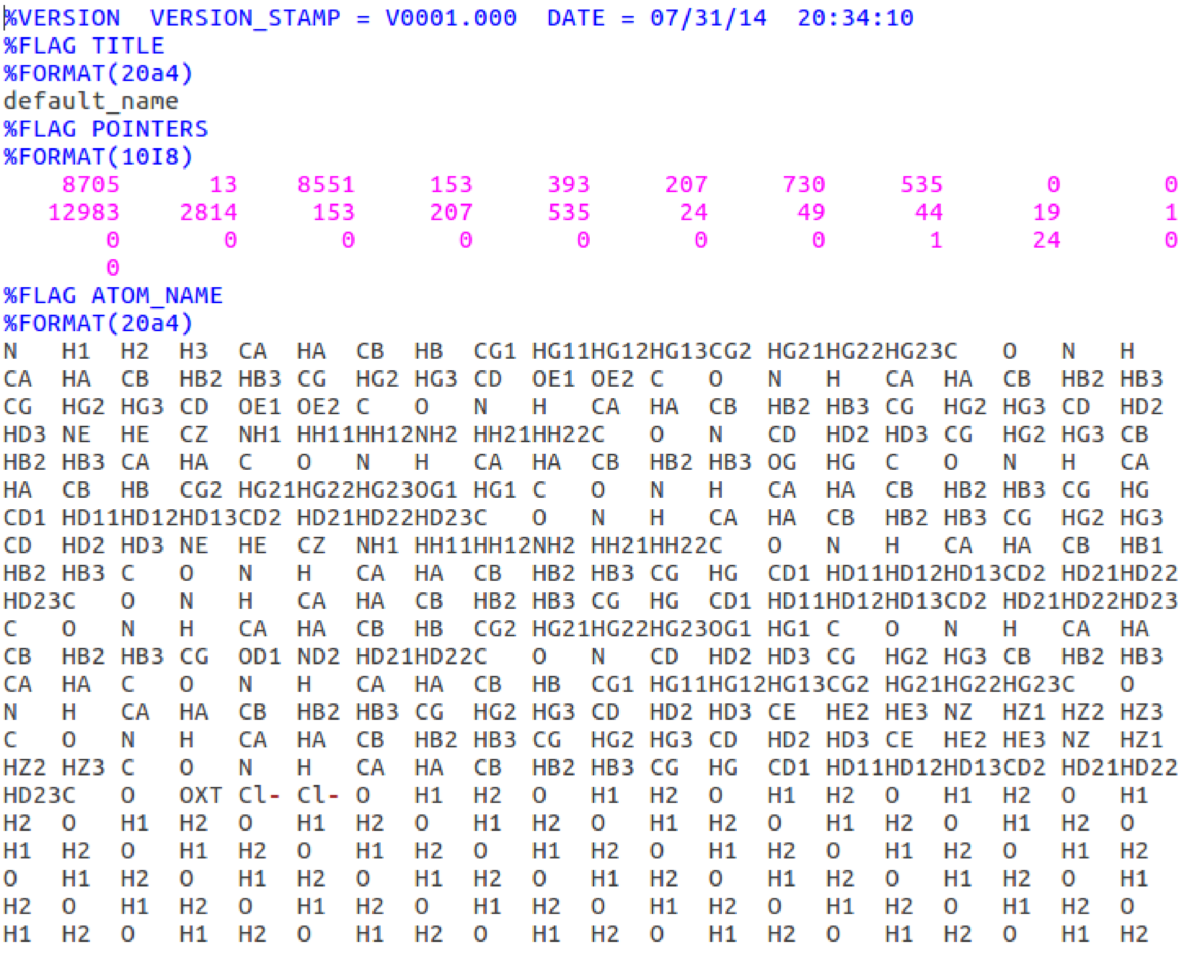
- ✔ A .pdb document (Protein Data Bank) which contains the coordinates of all atoms (and the coordinates of the heterogeneous atoms) of the system and/or their velocities. PDB files can be accessed through PDB or they can be created from the user. Part of the pdb file that was used for FTZpep simulation is displayed bellow:



```
ATOM      1  N    VAL     1       6.145  -4.478  -1.921  1.00  0.00
ATOM      2  H1   VAL     1       6.545  -5.350  -1.607  1.00  0.00
ATOM      3  H2   VAL     1       6.974  -3.960  -2.173  1.00  0.00
ATOM      4  H3   VAL     1       5.576  -4.540  -2.753  1.00  0.00
ATOM      5  CA   VAL     1       5.407  -3.901  -0.744  1.00  0.00
ATOM      6  HA   VAL     1       5.105  -2.896  -1.037  1.00  0.00
ATOM      7  CB   VAL     1       4.150  -4.643  -0.299  1.00  0.00
ATOM      8  HB   VAL     1       3.617  -5.081  -1.143  1.00  0.00
ATOM      9  CG1  VAL     1       4.492  -5.976   0.477  1.00  0.00
ATOM     10 HG11  VAL     1       5.246  -6.574  -0.035  1.00  0.00
ATOM     11 HG12  VAL     1       4.948  -5.633   1.405  1.00  0.00
ATOM     12 HG13  VAL     1       3.559  -6.534   0.564  1.00  0.00
ATOM     13  CG2  VAL     1       3.242  -3.926   0.699  1.00  0.00
ATOM     14 HG21  VAL     1       3.718  -3.705   1.654  1.00  0.00
ATOM     15 HG22  VAL     1       3.050  -2.902   0.379  1.00  0.00
ATOM     16 HG23  VAL     1       2.292  -4.421   0.902  1.00  0.00
ATOM     17  C    VAL     1       6.398  -3.603   0.418  1.00  0.00
ATOM     18  O    VAL     1       6.162  -2.604   1.095  1.00  0.00
ATOM     19  N    GLU     2       7.427  -4.423   0.721  1.00  0.00
ATOM     20  H    GLU     2       7.610  -5.306   0.265  1.00  0.00
ATOM     21  CA   GLU     2       8.229  -4.102   1.969  1.00  0.00
ATOM     22  HA   GLU     2       7.607  -3.583   2.699  1.00  0.00
ATOM     23  CB   GLU     2       8.661  -5.479   2.570  1.00  0.00
ATOM     24  HB2  GLU     2       7.802  -6.137   2.444  1.00  0.00
ATOM     25  HB3  GLU     2       9.457  -5.863   1.931  1.00  0.00
```

Figure 7: Presentation of FTZpep pdb file. Going from the left column to the right: record type, atom ID, atom name, residue name, residue ID, x, y, and z coordinates, occupancy, temperature factor.

- ✔ A customization file of AMBER force field (AMBER format PRMTOP file) which includes all the necessary parameters needed for the calculation of potential energy of the system. PRMTOP files are created from LEaP program. We used AMBER ff99SB-STAR-ILDN. Part of the PRMTOP file used for FTZpep MD simulation is presented below:



```
%VERSION  VERSION_STAMP = V0001.000  DATE = 07/31/14  20:34:10
%FLAG TITLE
%FORMAT(20a4)
default_name
%FLAG POINTERS
%FORMAT(10I8)
    8705      13    8551     153     393     207     730     535       0       0
   12983    2814     153     207     535      24      49      44      19       1
       0       0       0       0       0       0       0       1      24       0
       0
%FLAG ATOM_NAME
%FORMAT(20a4)
N   H1  H2  H3  CA  HA  CB  HB  CG1 HG11HG12HG13CG2 HG21HG22HG23C   O   N   H
CA  HA  CB  HB2 HB3 CG  HG2 HG3 CD  OE1 OE2 C   O   N   H   CA  HA  CB  HB2 HB3
CG  HG2 HG3 CD  OE1 OE2 C   O   N   H   CA  HA  CB  HB2 HB3 CG  HG2 HG3 CD  HD2
HD3 NE  HE  CZ  NH1 HH11HH12NH2 HH21HH22C   O   N   CD  HD2 HD3 CG  HG2 HG3 CB
HB2 HB3 CA  HA  C   O   N   H   CA  HA  CB  HB2 HB3 OG  HG  C   O   N   H   CA
HA  CB  HB  CG2 HG21HG22HG23OG1 HG1 C   O   N   H   CA  HA  CB  HB2 HB3 CG  HG
CD1 HD11HD12HD13CD2 HD21HD22HD23C   O   N   H   CA  HA  CB  HB2 HB3 CG  HG2 HG3
CD  HD2 HD3 NE  HE  CZ  NH1 HH11HH12NH2 HH21HH22C   O   N   H   CA  HA  CB  HB1
HB2 HB3 C   O   N   H   CA  HA  CB  HB2 HB3 CG  HG  CD1 HD11HD12HD13CD2 HD21HD22
HD23C   O   N   H   CA  HA  CB  HB2 HB3 CG  HG  CD1 HD11HD12HD13CD2 HD21HD22HD23
C   O   N   H   CA  HA  CB  HB  CG2 HG21HG22HG23OG1 HG1 C   O   N   H   CA  HA
CB  HB2 HB3 CG  OD1 ND2 HD21HD22C   O   N   CD  HD2 HD3 CG  HG2 HG3 CB  HB2 HB3
CA  HA  C   O   N   H   CA  HA  CB  HB  CG1 HG11HG12HG13CG2 HG21HG22HG23C   O
N   H   CA  HA  CB  HB2 HB3 CG  HG2 HG3 CD  HD2 HD3 CE  HE2 HE3 NZ  HZ1 HZ2 HZ3
C   O   N   H   CA  HA  CB  HB2 HB3 CG  HG2 HG3 CD  HD2 HD3 CE  HE2 HE3 NZ  HZ1
HZ2 HZ3 C   O   N   H   CA  HA  CB  HB2 HB3 CG  HG  CD1 HD11HD12HD13CD2 HD21HD22
HD23C   O   OXT Cl- Cl- O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1
H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O
H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2
O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1
H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O
H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2  O   H1  H2
```

Figure 8: PRMTOP file.

✔ A configuration file which provided to NAMD all the necessary information about the simulation. The file that was used for FTZpep MD simulation is presented below:

```
#
# Input files
#
amber                   on
readexclusions          yes
parmfile                ftzpep.prmtop
coordinates             heat_out.coor
velocities              heat_out.vel
extendedSystem          heat_out.xsc

#
# Adaptive ...
#
adaptTempMD             on
adaptTempTmin           280
adaptTempTmax           380
adaptTempBins           1000
adaptTempRestartFile    output/restart.tempering
adaptTempRestartFreq    10000
adaptTempLangevin       on
adaptTempRescaling      off
adaptTempOutFreq        400
adaptTempDt                0.000050
```

```
#
# Output files & writing frequency for DCD
# and restart files
#
outputname              output/equi_out
binaryoutput            off
restartname             output/restart
restartfreq             10000
binaryrestart           yes
dcdFile                 output/equi_out.dcd
dcdFreq                 400
DCDunitcell             yes

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

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

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

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

#
# Particle Mesh Ewald parameters.
#
Pme                     on
PmeGridsizeX            48                       # <===== CHANGE ME
PmeGridsizeY            48                       # <===== CHANGE ME
PmeGridsizeZ            48                       # <===== CHANGE ME
#
# Periodic boundary things
#
wrapWater               on
wrapNearest             on
```

```
wrapAll                 on

#
# Langevin dynamics parameters
#
langevin                on
langevinDamping         1
langevinTemp            320                     # <===== Check me
langevinHydrogen        off

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

useGroupPressure        yes

firsttimestep           10000                 # <===== CHANGE ME
run                     500000000              ;# <===== CHANGE ME
```

3.3 System Preparation

Undergoing an MD simulation requires a specific sequence of steps such as initialization of coordinates, minimization of structure, assignment of initial velocities, heating dynamics, equilibration dynamics, rescale velocities if the temperature is not suitable, production dynamics and finally the analysis of the trajectories. [26]

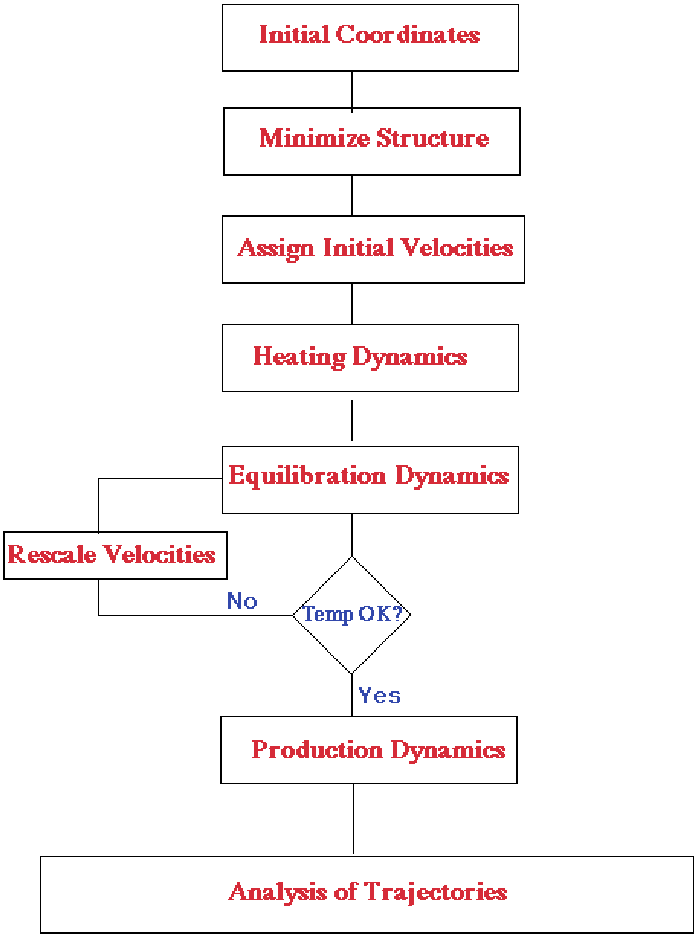


Figure 9: Molecular Dynamics Simulation steps. (Reproduced without permission from tutorial of CHARMM MD simulation tutorial, http://www.ch.embnet.org/MD_tutorial/pages/MD.Part3.html)

An initial configuration of the system must be defined before an MD simulation takes place. In most biomolecules’ MD simulation cases it’s used an already solved structure by X-ray diffraction or NMR as an initial structure. However, it can also be used a theoretical structure which is extracted by homology modeling. The selection of the initial structure must be done wisely, otherwise the quality of simulation can be effected. The FTZpep structure was determined by NMR. [13] An energy minimization should be done in order to remove any strong Van der Waals interactions which could alter the produced data. During the minimization step local energy minima are sought out by changing systematically the positions of the atoms and calculating every time the local energy. This procedure was conducted in 1000 steps. A quick heating followed and the procedure was repeated for another 1000 steps. During heating step, initial velocities are combined with low temperature and the simulation begins. New velocities are set periodically at a slightly higher temperature while the simulation continuous. This step is repeated until the ideal temperature is achieved. [27]

In our simulation, the temperature range was 280 – 380 K, according to the adaptive tempering method of NAMD. According to this method, if the potential energy of a produced structure is lower than the average value of system’s potential energy, then the temperature is decreased. On the other hand, if the potential energy of a produced structure is higher than the potential energy of the system, then the temperature is increased. This is achieved by Langevin thermostat. [27]

Before the productive phase, the phase of equilibrium must take place. Equilibrium phase is referred to Newton’s second law which is applied to every atom of the system imposing it’s trajectory. As soon as the ideal temperature is achieved, the simulation of protein and water system continuous. Characteristics such us structure, pressure, temperature, and energy are observed during this step. The simulation is taking place until the characteristics mentioned above become stable over time. [28]

The final step of the simulation is the productive phase in which the simulation will be performed for a specific time period. This period may differ according to the personalized features of each different simulation. The coordinates, energy, and velocities that were recorded during equilibrium phase will be used as input for the beginning of the productive phase. For the extraction of the trajectories, NAMD has been set to save atoms’ coordinates every 400 steps. The calculation of electrostatic interaction was conducted by PME method (Particle Mesh Ewald) and SHAKE algorithm was used for the restriction of all the bonds between hydrogen atoms and other atoms of the system. Verlet – I algorithm has been used for the calculation of atoms’ velocities and coordinates over time and water model was TIP3P [32]. The simulation produced 11.087.250 frames and simulation time was 8.87 μs.

Chapter 4: Results

4.1 Introduction

The data analysis of the MD simulation was conducted by *carma* [29] and using a graphic interface named *grcarma* [30]. This program receives as input two files, one DCD file and one PSF file. The DCD file is trajectory file. It is a binary file that contains the trajectory that was produced by the simulation, which refers to the coordinates of all atoms during the simulation. Each set of coordinates corresponds to one frame at a time. [34] The PSF file (Protein Structure File) is created by using the residues descriptions in the RTF (Residue Topology File). It contains a complete description of the topology of the system. In other words, in a PSF file is listed structural information such as atoms, bonds and angles, dihedrals, etc. There are also described charges for each nucleus as well as the nuclear mass.

NMR experiments of Yun, Ji-Hye et al [13] showed that *"...FTZpep in the absence of FTZ-F1, contains a small population of helical conformation due to intrinsic dynamics nature of small peptide."*. Circular dichroism (CD) results are similar to the mentioned from NMR experiments results and confirm that the peptide acquires a small population of helical conformation in water. Moreover, the fewer NOEs signals, result into high rmsd for backbone atoms and suggest that FTZpep has a dynamic behavior in water solution. A helical conformation is formed between residues 6-14, leaving the terminals of the peptide to be unstable. However, the peptide acquires a high population of helical conformation both in trifluoroethanol (TFE) solution and in the presence of FTZ-F1. In the presence of FTZ-F1 in 90% H2O/10% D2O, pH 6.5 and temperature 25 oC, FTZpep displays α-helix structure for the residues 5 – 18 (-3 Pro +11 Lys according to Yun et. al. [13]). The following measurements are conducted for both the shorter residues part that seems to acquire helical conformation in water solution (residues 6-14) and the whole peptide (1-19 residues).

4.2 RMSD analysis

Root Mean Square Deviation (RMSD) is an analysis that is commonly used in MD simulation data analysis. Using RMSD we can calculate the average distance between atoms of different conformations that are superimposed and is used widely in Structural Biology projects in order to compare protein structures. RMSD is calculated by the following formula:

$$\mathbf{RMSD = \surd\Sigma(x_i - x_{ref})^2 /N \ (10)}$$

Where x_i states the coordinates of the atoms for a specific time, x_{ref} states the coordinates of the reference structure and N states the atoms number. [35] The smaller the magnitude of the RMSD, the greater the similarity between the two superimposed structures. RMSD's magnitude should be 0.0Å in case the two structures are identical. RMSD magnitude is displayed by colors. Blue color implies low RMSD magnitude, red color implies high RMSD magnitude and yellow color implies medium RMSD magintude. Blue regions that are placed on the diagonal line of the matrix represent a structure that remains stable for a period of time that is analogous to the lengh of that region. The RMSD matrix that was produced by GRCARMA is displayed bellow:

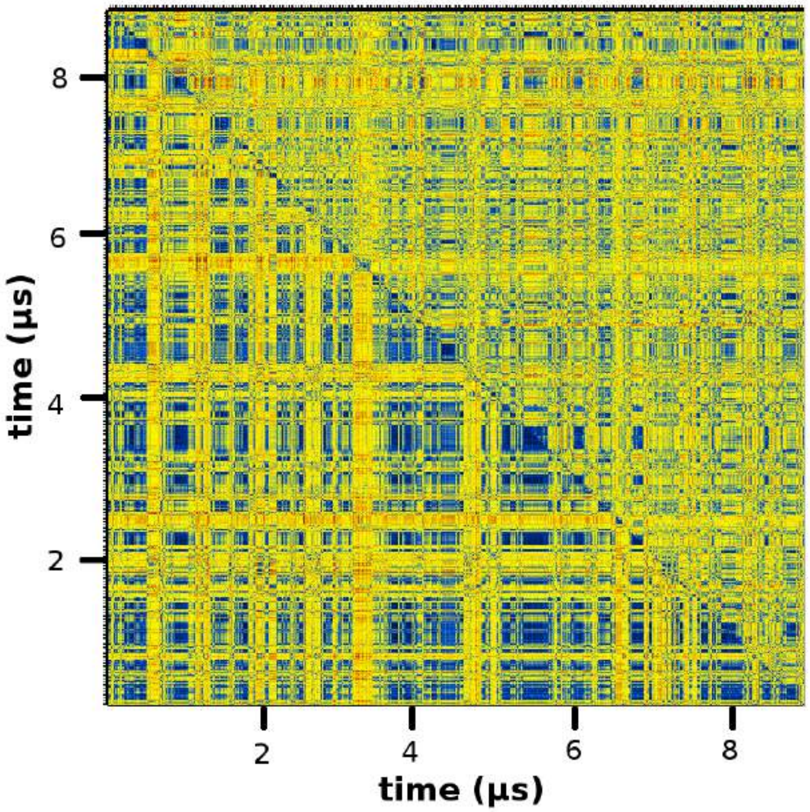


Figure 10: RMSD diagram for the Ca atoms of 11,087,250 trajectories. The upper half of the diagram displays RMSD for the whole FTZpep peptide (1-19 residues) while the lower half of the diagram displays RMSD calculations for one part of the FTZpep (6-14 residues).

RMSD matrix for 1-19 residues implies that the peptide does not prefer a specific structure during the simulation. Blue regions are not wide. However, the inner part of the peptide that is composed of residues, 6-14 shows a bigger stability over time. Specifically, we can observe a stabilization of the protein for the following time periods of the simulation: 0.8 – 1 μs, 1.8 – 2 μs, 4.3 – 4.5 μs, 5.5 – 6 μs, 8 – 8.2 μs. The whole peptide (1 – 19 residues) remains unstable. These results agree with Yun, Ji-Hye et al results, which signify that the peptide is unstable in the presence of aqueous solution. NMR experiments showed i, i+3 and i, i+4 interactions which imply the presence of α – helix. When FTZ-F1 is absent, FTZpep shows helix conformation. However, FTZpep shows fewer NOEs in aqueous solution. This means that FTZpep shows a dynamic behavior in aqueous solution. This could be the reason why there was not observed a specific stable structure for a significant period by RMSD matrix analysis.

4.3 Secondary structure prediction

Secondary structure prediction calculates the secondary structures that are adopted by the peptide during the simulation. It is a helpful tool for the prediction of the tertiary structure of the peptide and its folding process. We used STRIDE (STRuctural IDEintification) method [31]. STRIDE is based on an algorithm which combines the energy of hydrogen bonds and dihedral angles of the peptide's backbone. It produces a text file which contains the secondary structure assignments. G*rcarma* produces a diagram which depicts secondary structures of the peptide with colors, using the text file produced by STRIDE. For our calculations, the step between frames was 370.

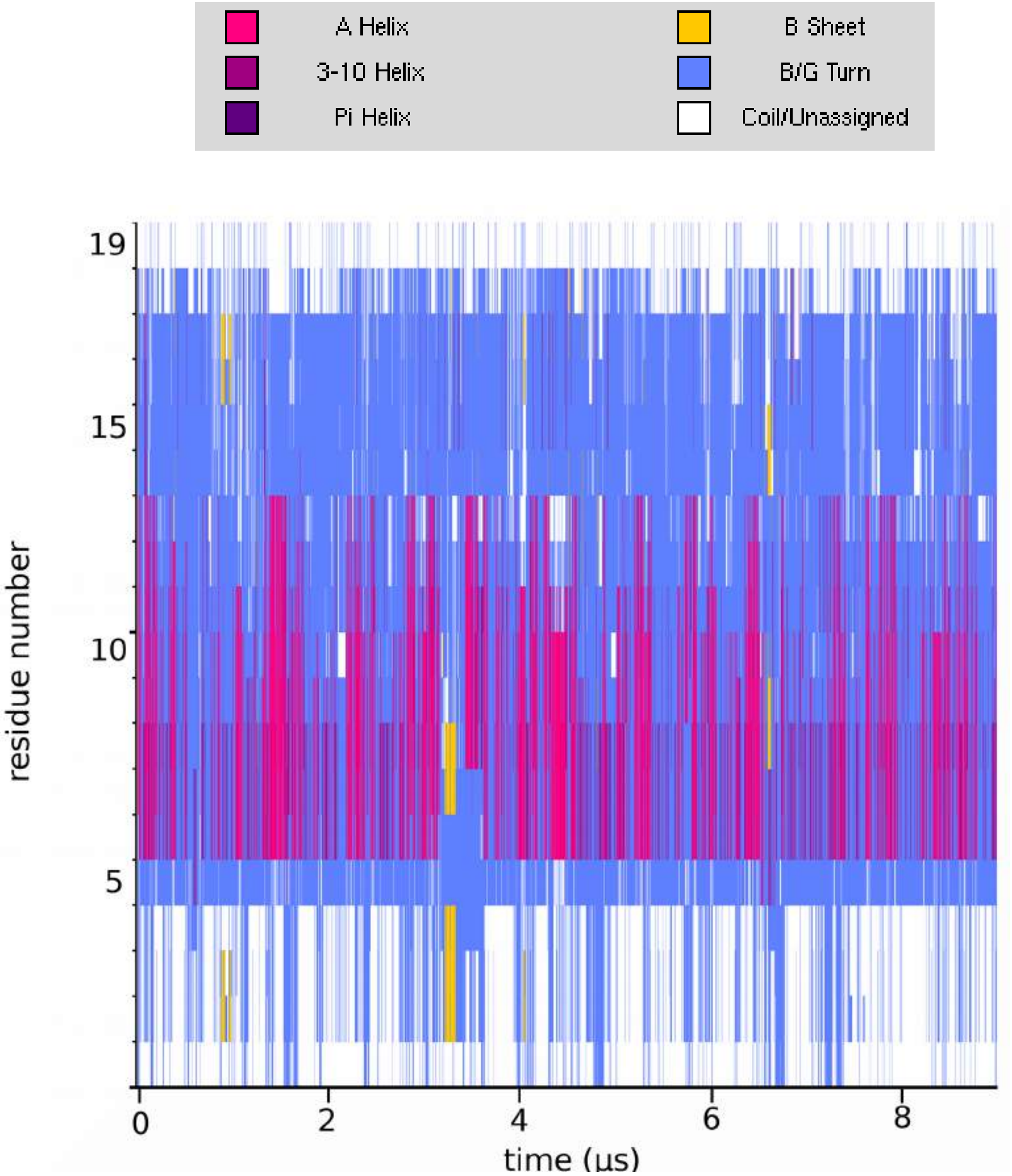


Figure 11: Secondary structure diagram produced by STRIDE. Pink color shows α – helix conformation, blue color shows turn, yellow color shows β – sheet, white color shows random coil and purple color shows 3_{10} helix. The secondary structure was produced for the whole peptide (1 – 19 residues)

WebLogo is a graphical method to represent sequences. It consists of stacks of letters. Their height is depended on their frequency. Each stack represents one position of the sequence, in our case one residue. The letter with the higher frequency observed is stacked on the top. [36, 37]

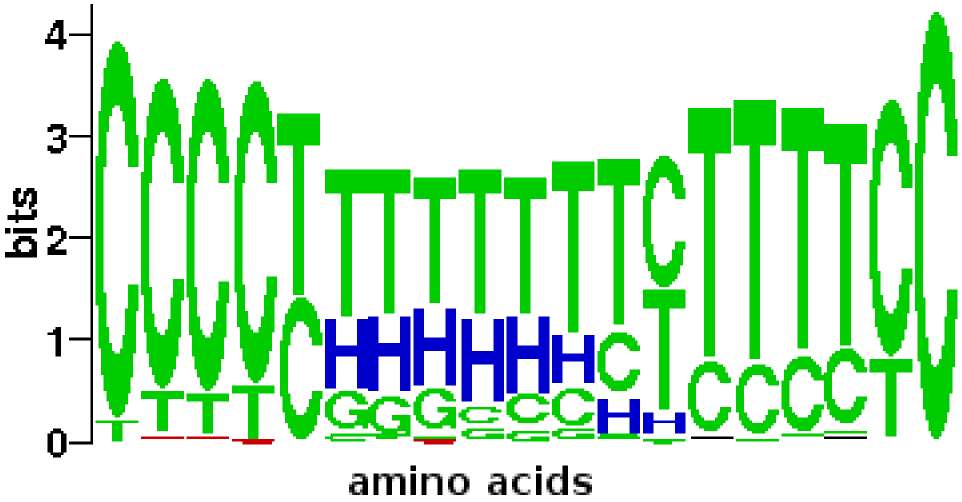


Figure 12: WebLogo graph for the peptide FTZpep (1-19 residues)

- H: α – helix
- G: 3_{10} helix
- I: π – helix
- E: β – sheet
- B: β – bridge
- T: turn
- C: coil (none of the conformations mentioned above) [36, 37]

Our results show a clear preference for α – helix conformation and turns. Observing Figure 11 we can see that pink and blue color are most present imposing α – helix and turn conformations respectively. There are a few β – sheets present, without having a great impact though on FTZpep's structure. Conformation of α – helix is conspicuous for residues 6 – 14. This result matches with the experimental data of Ji-Hye Yun et.al., which imply α – helix conformation for residues 6 – 14 in aqueous solution. The group identified many intense NOEs for α – helix conformation while being in TFE solution. FTZpep and FTZ-F1 interactions lead to α – helix conformation from residue-5 (Pro, P) till residue-18 (Lys, K) and the helix has a bend near residue-15 (Pro, P). We can observe similar data during secondary structure analysis due to the formation of a – helix for residues 6 – 14. Secondary structure diagram combined with RMSD matrixes for the whole peptide and for residues 6-14 is presented in the next page in Figure 13.

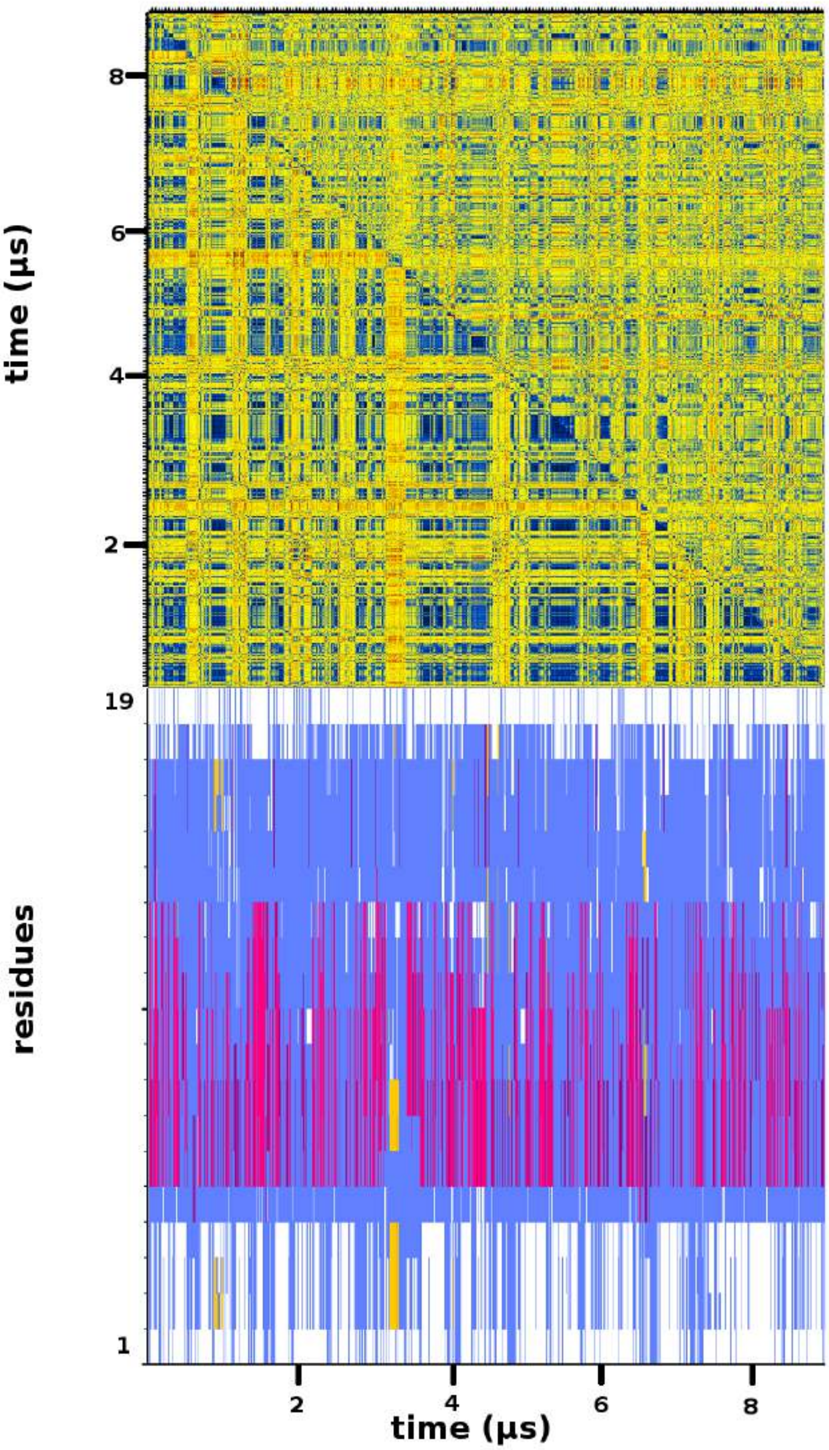


Figure 13: Parallel representation of RMSD matrix for 6-14 residues/1-19 residues and secondary structure graph.

4.4 Principal Component Analysis and Clustering

Principal component analysis (PCA) is a statistical process that can be used for identification of patterns within a variety of data and observation of their similarities and differences. In PCA a set of observed and possibly related variables is converted into a set of linearly unrelated variables called principal components. PCA is a very useful tool for multi dimensional data analysis, because dimensions can be reduced, without severe loss of the mined information. PCA is being used extensively in the field of Molecular Dynamics, due to it's characteristics.

More specifically, dihedral Principal Component Analysis (dPCA) and cartesian Principal Component Analysis (cPCA) are being used in MD simulations' data analysis. Dihedral PCA is based on the dihedral angles of the peptide backbone (φ, ψ), while cartesian PCA on the cartesian coordinates of the atoms. Cartesian PCA, although it is a useful method to investigate a protein's structure, can lead to the creation of a few artifacts. The mixing of internal and overall motions leads to artifacts, which lead to the failure of discrimination of the conformations. In the case of large amplitude motion, it is impossible to define with accuracy, a single reference structure for the elimination of overall rotations. [42] Dihedral PCA can produce more accurate data and amounts to one-to-one representation of the original angle distribution. It is essential that they should both be combined in PCA analysis of the trajectories produced by a protein simulation. [38, 39] PCA data can be further categorized into clusters, depending on data similarities. The PCA analyses that follow, have been produced through *carma* and its graphic interface *grcarma*.

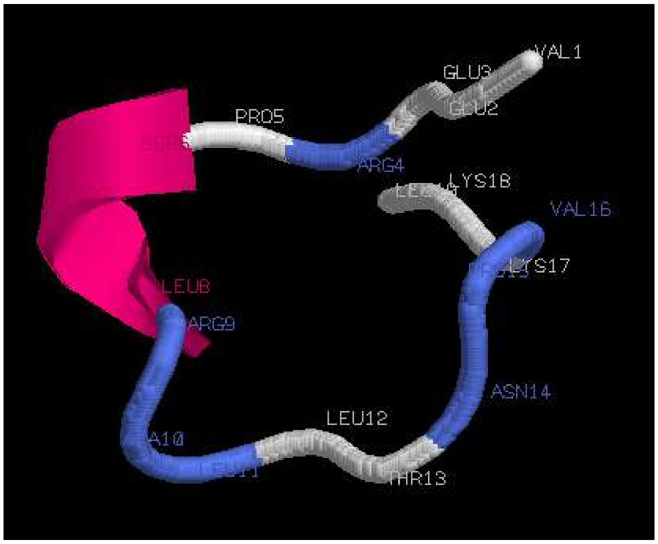
Dihedral PCA has been chosen for the first step of PCA and cluster analysis and a set of prominent clusters has been chosen for further analysis through cartesian PCA for backbone atoms only. The new clusters that occurred, have been further analyzed through cartesian PCA for non-hydrogen atoms. PDB files have been produced and representative structures are depicted below. Average structures for each cluster have been calculated and they have been compared with the frames from the trajectories. The representative structure is the frame from the trajectory with the lowest RMS deviation compared with the average structure. [40] The same sequence of PCAs has been followed for the whole peptide and residues 6-14 as well. There have been conducted three rounds of different PCAs for the whole peptide, depending on the temperature that the frames were produced during simulation. As it has already been mentioned, the simulation took place under adaptive tempering situation, so the frames have been produced on different temperatures. Representative structures will be presented from PCAs for all the frames, for frames produced below 320K and 300K respectively. RasMol has been used to depict the PDB files produced by PCAs. [41]

4.4.a PCA for all frames

Dihedral PCA produced 32 clusters. The two most populated clusters were further analyzed with the procedure described above. The most populated cluster (the 3rd of the 32) contained 653,715 frames out of 11,087,250 (4.3%) and the second most populated cluster (the 1st of the 32) contained 475,196 frames out of 11,087,250 (5.9%). The most populated clusters of cartesian PCA for backbone atoms were further analyzed through cartesian PCA for non-hydrogen atoms. The final results and figures come from the most populated clusters of cartesian PCA for non-hydrogen atoms.

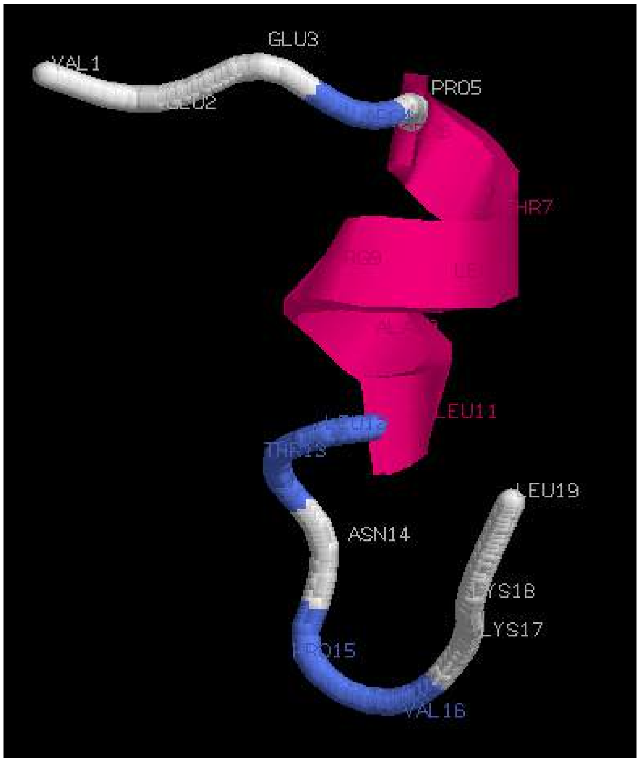
Cluster No. (dihedral PCA)	**Frames (out of 11,087,250)**	**Cluster No. (cartesian PCA for backbone)**	**Frames**	**Cluster No. (cartesian PCA for non-hydrogen)**	**Frames**
1	475,196 (4.3%)	**1**	247,619 (52%) (out of 475,196)	**1**	114,016 (46%) (out of 247,619)
3	653,715 (5.9%) **(most populated)**	**1**	384,398 (59%) (out of 653,715)	**1**	125,811 (32.9%)(out of 382,398)

Table 1: statistical analysis of the most populated clusters produced by PCA for all the frames.



In figure 14 only residues 6-8 built an α-helix conformation, while the rest of the peptide is forming turns and coils.

Figure 14: Structure produced by the most populated cluster.



The second most populated cluster reveals that residues 6-11 participate in an α-helix conformation, while the rest of the peptide is forming turns and coiled coils.

Figure 15: Structure produced by PCA analysis of the second most populated cluster.

4.4.b PCA for frames produced in temperature range less than 320K

It has been created and used a DCD file in which there were included only the frames that have been produced in temperature range less than 320K during simulation. The same procedure was followed and the continuous rounds of different PCAs were practised. There were produced in total 40 clusters by dihedral PCA and the three most populated (the first, third and fifth from forty) were further analyzed. The new results are presented in the Table 2.

Cluster No. (dihedral PCA)	**Frames (out of 11,087,250)**	**Cluster No. (cartesian PCA for backbone)**	**Frames**	**Cluster No. (cartesian PCA for non-hydrogen)**	**Frames**
1	365,435 (6.5%)	**1**	215,176 (58.9%) (our of 365,435)	**1**	101,225 (47%) (out of 215,176)
3	393,559 (7%) **(most populated)**	**1**	243,511 (61.9%) out of (393,559)	**1**	38,637 (15.9%) (out of 243,511)
5	275,759 (4.9%)	**1**	104,028 (37.7%) out of (275,759)	**1**	28,146 (27%) (out of 104,028)

Table 2: Statistical analysis of the most populated clusters produced by PCA for frames produced in temperature range less than 320K.

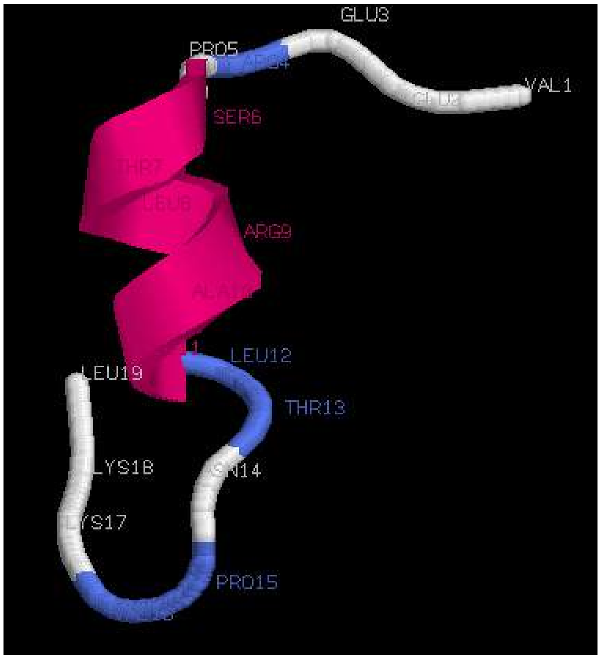


Figure 16: Structure produced by continuous PCAs for the first cluster of dihedral PCA.

In figure 16 it is depicted a PDB structure produced by the three-round PCA analysis for the first cluster of dihedral PCA. An α-helix is formed for residues 6-11.

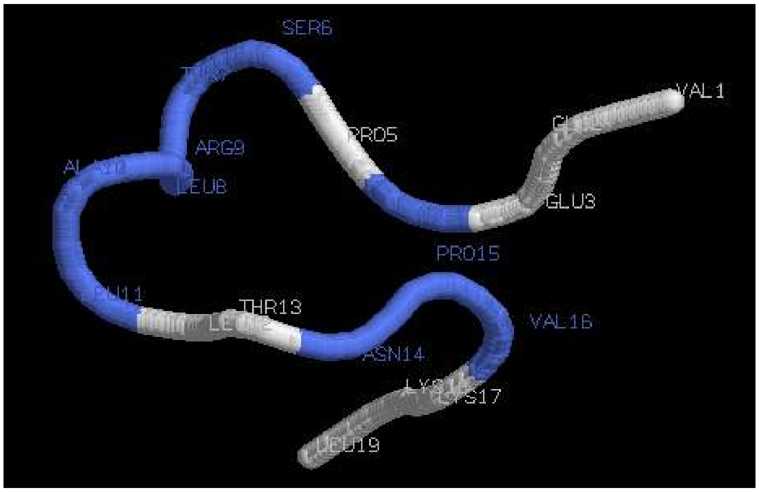


Figure 17: Structure produced by continuous PCAs for the third cluster (most populated) of dihedral PCA.

In figure 17 it is depicted a PDB structure produced by continuous PCAs for the third cluster of dihedral PCA. The third cluster in the most populated cluster. According to this clustering, our peptide does not form any specific conformation.

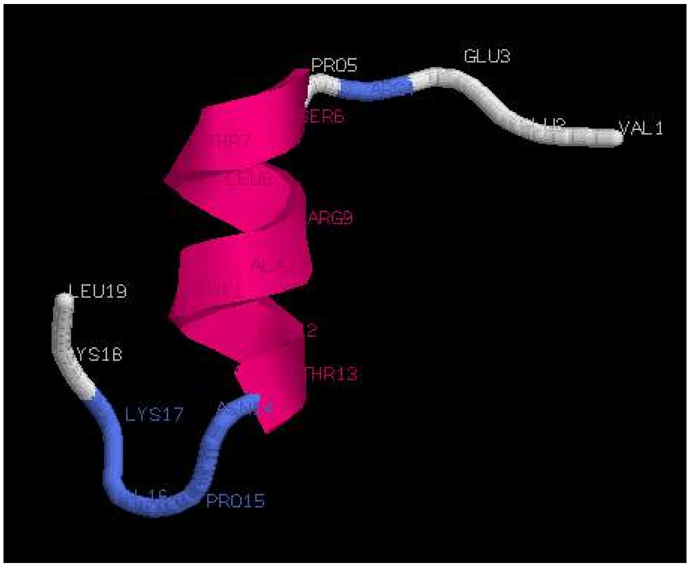


Figure 18: Structure produced by continuous PCAs for the fifth cluster of dihedral PCA.

In figure 18 it is depicted a PDB screen shot from the fifth cluster of dihedral PCA, after the three-round PCAs. Residues 6-13 participate in the formation of α-helix.

4.4.c PCA for frames produced in temperature range less than 300K

Cluster No. (dihedral PCA)	**Frames (out of 11,087,250)**	**Cluster No. (cartesian PCA for backbone)**	**Frames**	**Cluster No. (cartesian PCA for non-hydrogen)**	**Frames**
1	151,198 (4.8%) **(most populated)**	**1**	69,428 (45.9%) (out of 151,198)	**1**	18,389 (26.5%) (out of 69,428)

Table 3: Statistical analysis of the most populated clusters produced by PCA for frames produced in temperature range less than 300K.

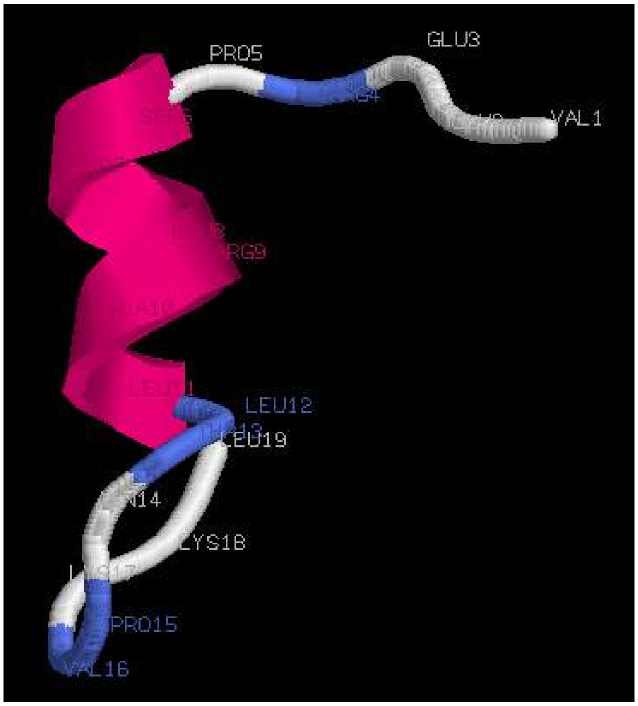


Image 19: Structure produced by continuous PCAs for the first (most populated) cluster of dihedral PCA.

4.4.d PCA for residue selection 6-14

Cluster No. (dihedral PCA)	**Frames (out of 11,087,250)**	**Cluster No. (cartesian PCA for backbone)**	**Frames**	**Cluster No. (cartesian PCA for non-hydrogen)**	**Frames**
1	1,422,039 (12.8%) **(most populated)**	**1**	571,914 (40.2%) (out of 1,422,039)	**1**	325,911 (57%) (out of 571,914)

Table 4: Statistical analysis of the most populated clusters produced by PCA for residue selection (6-14 residues)

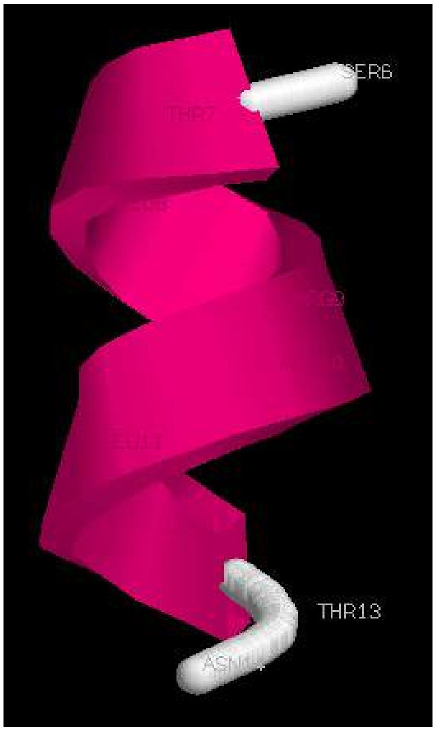


Figure 20: Structure produced by continuous PCAs for the first (most populated) cluster of dihedral PCA.

4.5 Comparison with experimental data

Nuclear Magnetic Resonance is one of the main experimental methods that are used to identify the structure of a molecule and is concerned with the magnetic properties of certain nuclei and more specifically with their spin. These nuclei act like tiny magnets due to that spinning, which generates a magnetic moment along the axis of the spin. One such nucleus is the proton, the nucleus of hydrogen ^{1}H. In general, radiation of steady frequency is applied through the substance. While the strength of the magnetic field is changing, at some point absorption is occured and a signal is produced. This is the nuclear magnetic resonance spectrum. There are many other phenomena caused by NMR, such as NOE (Nuclear Overhauser Effect), J-couplings and chemical shifts.

4.5.a NOEs

The NOE is based in Nuclear Overhauser Effect and it is useful in NMR spectroscopy. NOE occurs through space, not through chemical bonds. Thus, atoms that are close spatially, can give NOE signal. The inter-atomic distances derived from the observed NOE can often help to confirm the three-dimensional structure of a molecule. NOE allows us to study groups that may be separated by many bonds, but they are relatively close to each other in space. Nuclear Overhauser Effect is the alteration in the absorption intensity of one nuclei A due to the radiation of another nuclei B which is close to nuclei A, as an attempt of the system to reach it's equilibrium again. This alteration is caused when the population of spins of nuclei B changes. It should be noted that there is no observable NOE signal for nuclei distance larger than 5.5 Angstrom.

NOE signals are strictly dependent on the distance between the two nuclei and they can be expressed as:

$$\mathbf{NOE = 1/r^{6} f(t_{c}) \quad (11)}$$

Where r states for the distance between two nuclei A and B and t_c is the time needed for a full rotation for 1 rad. [43] Distance restraints that are derived from NOEs are essential for the identification of the secondary structure of a protein. NOE signals, and therefore distance restraints, are depended on the $(r)^{-6}$ distance of the nuclei. In many NMR experiments $(r)^{-3}$ is preferred for the calculations. Calculation of $(r)^{-6}$ is preferred for smaller peptides, like FTZpep, and $(r)^{-3}$ are preferred for bigger peptides. [44]

NOEs derived from the simulation have been calculated and compared with the experimental data. A list with all the possible proton pairs has been created through a Perl script, prep_proton.pl, based on the PSF file of the simulation. This list has been modified later, according to the pairs of protons that have been observed by the NMR experiment. More specifically, six new proton lists have been created, depending on the atoms of the peptide that take part in the signal and including only proton pairs that produced NOE signal in the experiment.

- **dαN(i,i+1):** NOE intensity classification for the observed Hα (i) to Hn (i+1) NOE.
- **dNN(i,i+2):** NOE intensity classification for the observed Hn (i) to Hn (i+2) NOE.
- **dNN(i,i+1):** NOE intensity classification for the observed Hn (i) to Hn (i+1) NOE.
- **dαN(i,i+2):** NOE intensity classification for the observed Hα (i) to Hn (i+2) NOE.
- **dαN(i,i+3):** NOE intensity classification for the observed Hα (i) to Hn (i+3) NOE.
- **dβN(i,i+1):** NOE intensity classification for the observed Hβ (i) to Hn (i+1) NOE.

Signals have been categorized as:

- strong (1.8 – 2.7 Angstrom)
- medium (2.7 – 3.3 Angstrom)
- weak (3.3 – 5.0 Angstrom)

The $(r)^{-6}$ and $(r)^{-3}$ average distances have been calculated by a C program, noe_averaging.c [51]:

$$\mathbf{R = (< R_{ij}^{-6} >)^{-1/6}} \text{ and } \mathbf{R = (< R_{ij}^{-3} >)^{-1/3}}$$

NOEs signals have been calculated for:

1. all the frames that were produced by the simulation
2. frames that were produced in temperature range less than 320K
3. frames that were produced in temperature range less than 300K

in order to examine the behavior of the peptide according to temperature differences. It should be noted that NMR experiments were conducted at 298K.

As a method of validation between experimental and simulation derived results, it has been used the *upper bound violation* process. [45] An upper bound violation identifies an inconsistency between a restraint and a structure. A restraint is not considered to be an upper bound violation as long as (r) $^{-6}$ value is lower than NOE upper bound value. For example, the lower bound for strong signal is 1.8 and the upper bound is 2.7 Angstrom. If the signal is defined experimentally as strong but the (r) $^{-6}$ value is greater than 2.7 (p.e. 2.9), then this restraint is considered to be an upper bound violation and it should be calculated in the average upper bound violation value.

$$\mathbf{v(i,j) = r^{-6} – nmr(i,j) \ (12)}$$

where v(i,j) is the violation between two protons i, j and nmr(i,j) is the experimental upper bound value. In our example, v(i,j) = 2.9 – 2.7 = 0.2

Angstrom. The average upper bound violation is calculated by the sum of all the upper bound violations observed, divided by the total number of proton pairs. [44, 45] The tables that are following present the results from the simulation and the experimental classification of the NOEs signals. S is for strong, M for medium, W for weak and O for overlapping. The number of proton pairs that have been included into the calculations is 44.

daN (i, i+1)

pair	**$(r)^{-6}$ (300K)**	**Upper bound violation (300K)**	**$(r)^{-6}$ (320K)**	**Upper bound violation (320K)**	**$(r)^{-6}$ (all)**	**Upper bound violation (all)**	**Proton No.**	**Residue number & experimental classification**
1	2.241339 S		2.233715 S		2.242464 S		6-20	1V-2E W
2	2.256559 S		2.255621 S		2.278608 S		22-35	2E-3E M
3	2.960644 M		2.916426 M		2.854506 M		37-50	3E-4R M
4	2.300573 S		2.300797 S		2.295347 S		84-88	5P-6S M
5	3.378978 W		3.365568 W		3.277516 M		90-99	6S-7T W
6	3.218781 M		3.196329 M		3.163602 M		101-113	7T-8L W
7	3.169373 M		3.138844 M		3.086221 M		115-132	8L-9R W
8	2.818224 M	**0,118224**	2.847682 M	**0,16694**	2.840357 M	**0,140357**	134-156	9R-10A S
9	2.859265 M		2.86694 M		2.816129 M		158-166	10A-11L M
10	2.704108 M	**0,004108**	2.750996 M		2.745855 M		168-185	11L-12L W
11	2.463704 S		2.464122 S		2.486882 S		187-204	12L-13T W
12	2.590176 S		2.600706 S		2.622371 S		206-218	13T-14N O
13	2.689788 S		2.683320 S		2.692376 S		242-246	15P-16V S
14	2.719207 M		2.700131 M		2.694515 S		248-262	16V-17K M
15	2.387494 S		2.395202 S		2.400014 S		264-284	17K-18K S
16	2.316668 S		2.320116 S		2.325811 S		286-306	18K-19L S

dNN (i, i+2)

pair	$(r)^{-6}$ (300K)	Upper bound violation (300K)	$(r)^{-6}$ (320K)	Upper bound violation (320K)	$(r)^{-6}$ (all)	Upper bound violation (all)	Proton No.	Residue number & experimental classification
1	4.427110 W		4.400797W		4.397888 W		246-284	16V-18K W

dNN (i, i+1)

pair	$(r)^{-6}$ (300K)	Upper bound violation (300K)	$(r)^{-6}$ (320K)	Upper bound violation (320K)	$(r)^{-6}$ (all)	Upper bound violation (all)	Proton No.	Residue number & experimental classification
1	2.072932 S		2.086154 S		2.130339 S		35-50	3E-4R W
2	2.909250 M		2.90365 M		2.894426 M		88-99	6S-7T W
3	2.589192 S		2.595493 S		2.594189 S		99-113	7T-8L W
4	2.245346 S		2.252443 S		2.279503 S		113-132	8L-9R W
5	2.780326 M		2.76246 M		2.722105 M		132-156	9R-10A O
6	2.476873 S		2.472350 S		2.515948 S		156-166	10A-11L W
7	2.448551 S		2.432662 S		2.440888 S		166-185	11L-12L O
8	2.796214 M		2.80072 M		2.768924 M		185-204	12L-13T M
9	2.504755 S		2.475301 S		2.466145 S		204-218	13T-14N M
10	2.298336 S		2.308209 S		2.325863 S		246-262	16V-17K W

daN (i, i+2)

pair	**(r) $^{-6}$ (300K)**	**Upper bound violation (300K)**	**(r) $^{-6}$ (320K)**	**Upper bound violation (320K)**	**(r) $^{-6}$ (all)**	**Upper bound violation (all)**	**Proton No.**	**Residue number & experimental classification**
1	4.054661 W	**0,754661**	4.04844 W	**0,74844**	4.077923 W	**0,777923**	134-166	9R-11L M
2	4.339273 W	**1,039273**	4.33728 W	**1,03728**	4.301064 W	**1,001064**	158-185	10A-12L M
3	4.380828 W		4.37998 W		4.367947 W		220-246	14N-16V W
4	4.389874 W		4.38433 W		4.352303 W		242-262	15P-17K W

daN (i, i+3)

pair	**(r) $^{-6}$ (300)K**	**Upper bound violation (300K)**	**(r) $^{-6}$ (320K)**	**Upper bound violation (320K)**	**(r) $^{-6}$ (all)**	**Upper bound violation (all)**	**Proton No.**	**Residue number & experimental classification**
1	4.281650 W		4.29272 W		4.221236 W		101-156	7T-10A O
2	3.602238 W		3.58682 W		3.609497 W		115-166	8L-11L O
3	4.923632 W		3.89841 W		3.891693 W		134-185	9R-12L O
4	4.441164 W		4.46464 W		4.438529 W		168-218	11L-14N W
5	5.292905 W		5.27468 W		5.093577 W		242-284	15P-18K W

dβN (i, i+1)

pair	$(r)^{-6}$ (300K)	Upper bound violation (300K)	$(r)^{-6}$ (320K)	Upper bound violation (320K)	$(r)^{-6}$ (all)	Upper bound violation (all)	Proton No.	Residue number & experimental classification
1	3,606738W		3,578669W		3,528181W		39,40-50	3E-4R W
2	3,448726W		3,453233W		3,451229W		92,93-99	6S-7T W
3	3,136871M		3,153859M		3,142528M		136,137-156	9R-10A W
4	3,398513W	**0,098513**	3,39466 W	**0,09466**	3,37219 W	**0,07219**	160,161,162-166	10A-11L M
5	3,249211M		3,247284M		3,245031M		189,190-204	12L-13T W
6	2,950030M		2.987206M		3,022802M		208-218	13T-14N M
7	3,122237M		3,115741M		3,107458M		250-262	16V-17K W
8	2,996287M		3,000627M		3,016553M		288,289-306	18K-19L W

Average violation $(r)^{-6}$ (300K)	Number of violations	Average violation $(r)^{-6}$ (320K)	Number of violations	Average violation $(r)^{-6}$ (all)	Number of violations
0,0457904	5	0,04653	4	0,0452621	4

It is obvious that the average upper bound violation for all the three cases does not exceed 0.05 Angstrom. This indicated that the simulation is coming to a big agreement with the experimental data.

NOE signals for the residues 6-14 have been also measured. There has not been taken under consideration a temperature cut off in this case. The upper bound violation is presented below:

Average violation r) $^{-6}$ **(all)**	**Number of violations**
0,0738	4

The number of protons that have been used for the average upper bound calculation is 27 out of 44 and the average upper bound violation for the residues 6-14 is 0.07 Angstrom.

4.5.b J-couplings

J-couplings is a through-bond interaction in which the spin of one nucleus is polarized itself and polarizes the spins of electrons that surround the nucleus. As a consequence, the energy levels of the nuclei that are in close proximity with the initial nucleus are perturbated, causing increase or decrease of their energy, in dependence with the spin of the polarized nucleus. J-couplings are calculated in Hz and are independent of the applied field. J-couplings remain the same towards the changes of the applied field. They are also mutual (Jax = Jxa, where x,a is a pair of protons). They are affected by the number of bonds that might separate two nuclei and the more the bonds, the less J-coupling phenomena are observed. Coupling constants refer to the distance between two peaks in the NMR spectra.

Backbone vicinal coupling constants have been observed ($^3J_{HN\text{-}H\alpha}$). Backbone $^3J_{HN\text{-}H\alpha}$ coupling constants have been used extensively for the characterization of φ torsion angle and in order to distinguish α-helix from β-sheet structure. [46] Differences between experimental and theoretically predicted ^{3}J couplings can provide useful information regarding fluctuations of motions of torsion angles, especially for a wide range of time scales (fs to ms) that are difficult to be observed by NMR experiments.

In Karplus equation it is described the correlation between ^{3}J-coupling constants and dihedral torsion angles in NMR spectroscopy and it has been used in the measurement of J-couplings derived from data produced by molecular dynamics simulation experiments. [47]

$$\mathbf{J(\varphi) = \mathit{C} \cos 2\varphi + \mathit{B} \cos\varphi + \mathit{A} \ (13)}$$

where J is the ^{3}J coupling constant, φ is the dihedral angle and *A, B, C* are parameters measured empirically and their values depend on the atoms. The number 3 in front of the symbol “J” indicates that a proton is coupled to another proton three bonds away, per example H-C-C-H. J-coupling is very valuable in identifying backbone torsion angles in NMR experiments.

Density Functional Theory (DFT) is a computational quantum mechanical modeling method that is used to study the electronic structure of many-body systems. Is has been extensively used to calculate A, B, C scalar couplings in Karplus equation. [46] It is important to be mentioned that the choice of Karplus equation parameters has a great impact on the results. [48] We used results that occurred by the DFT1 parameter set. According to a force field validation experiment applied on hepta-alanine [49], the DFT1 parameter set always leads to better agreement with the experimental data, independently of the applied force field or the error data set. Moreover, AMBER is the most suitable force field for DFT1 set of parameters. [49]

In the present study only $^3J_{HN\text{-}H\alpha}$ coupling constants have been measured, as they are very sufficient to discriminate the presence -or not- of α-helix conformation. Yun et.al. mention that they categorized the signals according to their magnitude in dependence with 6 Hz (if their magnitude is lower or bigger than 6Hz). They also observed unambiguous signals.

Phi and psi angles have been calculated through a Perl script *phi_psi_indeces.pl* and J-couplings have been calculated through another Perl script *calc_Jcouplings.pl* [52]. We used only the data that occurred through the DFT1 set of parameters and the Karplus equation used to calculate them is:

$$J(\varphi) = 9.44 * \cos(\varphi - pi / 3.00)^2 - 1.53 * \cos(\varphi - pi / 3.00) - 0.07 \quad (14)$$

The table below presents the theoretically predicted J-couplings. We also calculated J-couplings for different temperature range, like NOEs signals. The Std is Standard Deviation and the Exper. is the experimental measurement. The measurements that match the experimental data are marked with green color.

residue	3J (HN, HA) 300K	Std 300K	3J (HN, HA) 320K	Std 320K	3J (HN, HA) all	Std all	Exper.
2	7.570	2.537	7.599	2.537	7.651	2.532	< 6
3	9.030	2.079	8.962	2.125	8.677	2.271	> 6
4	6.501	2.687	6.582	2,728	6.821	2.823	< 6
5	-	-	-	-	-	-	Pro
6	3.189	1.887	3.260	1.954	3.444	2.135	< 6
7	5.405	2.478	5.414	2.507	5.373	2.559	*
8	8.968	2.376	8.884	2.420	8.649	2.518	< 6
9	4.669	2.641	4.688	2.680	4.906	2.832	*
10	6.782	2.825	6.752	2.839	6.624	2.850	*
11	7.473	2.747	7.486	2.751	7.442	2.760	> 6
12	7.546	2.640	7.590	2.628	7.575	2.613	> 6
13	7.990	2.536	7.986	2.543	7.995	2.546	> 6
14	6.628	2.912	6.671	2.922	6.770	2.947	> 6
15	-	-	-	-	-	-	Pro
16	9.225	2.136	9.152	2.184	8.988	2.297	> 6
17	7.770	2.425	7.781	2.412	7.751	2.399	> 6
18	7.598	2.718	7.592	2.709	7.627	2.699	< 6
19	8.162	2.563	8.194	2.560	8.288	2.531	> 6

Residues 5 and 15 have not been included in the calculation because they represent proline residues which does not participate in $^3J_{HN\text{-}H\alpha}$ because it looses the two H of the -NH2 group while the peptide bond is built.

If we will not take under consideration the unresolved values and the proline residues, 13 residues remain for the calculation of J-couplings. Nine out of thirteen J-coupling values come to agreement with the experimental data, irrespectively of the temperature cut off, which leads to a 69.2% agreement between the theoretically predicted J-couplings and the experimental data.

J-couplings that are referred to the residues 6-14, are in total 9 and 5 out of 9 theoretically predicted values agree with the experimental, leading to a 55.6% match.

Chapter 5: Discussion

The present study aimed to validate the accuracy of Molecular Dynamics method in identifying a protein’s structure and folding process. The peptide FTZpep that has been used in the present thesis acquires a dynamic behavior in aqueous solution and it acquires helical structure for specific residues. However, simulation results mostly agree with the experimental data. RMSD matrix for the whole peptide indicates that the peptide shows indeed a dynamic behavior and is mostly disordered. The color that prevails in the matrix is yellow, which means that the RMSD magnitude lies in the medium RMSD magnitude range. The RMSD matrix for residues 6-14 undeniably shows more stability for this part of the peptide but it still shows that it has a dynamic behavior in aqueous solution.

Secondary structure graph implies the presence of helical conformation in combination with turns for residues 6-14 implying that the part of the peptide is indeed more stable than the whole peptide but still dynamic in water solution. Peptide’s ends are completely disordered as they do not form any specific conformation, besides turns and coils. Weblogo graph reinforces this statement, as it shows that our peptide forms α-helix for residues 6-13, showing higher preference for α-helix for residues 6-11. The rest of the peptide remains in turns and coils.

Principal Component Analysis has revealed information regarding FTZpep behavior along temperature range. The structure that has been extracted from the most populated cluster from all frames, shows a great instability and the formation of α-helix for residues 6-8 only. While the temperature goes lower, more residues are included in the formation of α-helix structure. More specifically, when the PCA has been conducted for frames produced in temperature range less than 320K, residues 6-13 form α-helix. The most populated cluster from this temperature cut-off is completely disordered. The representative structure that has been extracted from continuous PCAs for temperature cut-off of 300K and lower implies that the residues 6-11 take part in the formation of α-helix. It must be mentioned that the temperature of NMR experiments was 298K. The PCA for the residues 6-14 implies that this part of the peptide is mainly helical with residues 7-12 forming α-helix.

NMR theoretically predicted data also agree with the experimental data. The average upper bound violation for the whole peptide in NOEs calculation is only 0.05 Angstrom while for residues 6-14 is 0.07 Angstrom. Theoretically predicted J-couplings for the whole peptide come to an agreement of 69.2% with the experimental data, while the predicted J-couplings for residues 6-14 come to an agreement of 55.6%

We ascertained that FTZpep has a dynamic behavior in water solution with a tendency to form helical conformations. Residues 6-14 shows a more stable behavior and a higher tendency to form α-helix.

Literature

1. Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th edition. New York: W H Freeman; 2002.
2. Anfinsen, C B. “The Formation and Stabilization of Protein Structure.” *Biochemical Journal* 128.4 (1972): 737–749. Print.
3. Anfinsen, C. B. et al. “THE KINETICS OF FORMATION OF NATIVE RIBONUCLEASE DURING OXIDATION OF THE REDUCED POLYPEPTIDE CHAIN.” *Proceedings of the National Academy of Sciences of the United States of America* 47.9 (1961): 1309–1314. Print.
4. Levinthal, Cyrus. "ARE THERE PATHWAYS FOR PROTEIN FOLDING ?". *Extrait du Journal de Chimie Physique* 65.1 (1968): 44. Print.
5. Onuchic, José Nelson, Zaida Luthey-Schulten, and Peter G. Wolynes. "THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective". *Annual Review of Physical Chemistry* 48 (1997): 545-600. Print.
6. Ruhong, Zhou et al. "Hydrophobic Collapse In Multidomain Protein Folding". *Science* 305.5690 (2004): 1605-1609. Print.
7. Karplus, M., and D. L. Weaver. “Protein Folding Dynamics: The Diffusion-Collision Model and Experimental Data.” *Protein Science : A Publication of the Protein Society* 3.4 (1994): 650–668. Print.

8. Fersht, A R. "Optimization Of Rates Of Protein Folding: The Nucleation-Condensation Mechanism And Its Implications". *PNAS* 92.24 (1995): 10869-10873. Print.
9. Honig, Barry and An-Suei Yang. "Free Energy Balance In Protein Folding". Advances in Protein Chemistry 46 (1995): 27-58. Print.
10. Bryngelson, Joseph D. et al. "Funnels, Pathways, And The Energy Landscape Of Protein Folding: A Synthesis". Proteins: Structure, Function, and Genetics 21.3 (1995): 167-195. Print.
11. Zwanzig, R, A Szabo, and B Bagchi. "Levinthal's Paradox." Proceedings of the National Academy of Sciences of the United States of America 89.1 (1992): 20–22. Print.
12. Hafen, Ernst, Atsushi Kuroiwa, and Walter J. Gehring. "Spatial Distribution Of Transcripts From The Segmentation Gene Fushi Tarazu During Drosophila Embryonic Development". *Cell* 37.3 (1984): 833-841.
13. Yun, Ji-Hye et al. "Solution Structure Of LXXLL-Related Cofactor Peptide Of Orphan Nuclear Receptor FTZ-F1". *Bulletin of the Korean Chemical Society* 33.2 (2012): 583-588. Print.
14. Allen, Michael P. "Introduction To Molecular Dynamics Simulation". *Computational Soft Matter: From Synthetic Polymers to Proteins, Lecture Notes* 23 (2004): 1-28. Print.
15. Zhang, Jiapu. "The Hybrid Idea Of (Energy Minimization) Optimization Methods Applied To Study Prion Protein Structures Fo- Cusing On The B2-A2 Loop". *Biochemical Pharmacology* 4.4 (2015): 1-23. Print.
16. AMBER force field official page, http://ambermd.org/

17. CHARMM force field official page, https://www.charmm.org/charmm/?CFID=dad41029-cdcd-407f-b5aa-523aaaa367b3&CFTOKEN=0
18. GROMOS force field official page, http://www.gromacs.org/Documentation/Terminology/Force_Fields/GROMOS
19. Jorgensen, William L., David S. Maxwell, and Julian Tirado-Rives. "Development And Testing Of The OPLS All-Atom Force Field On Conformational Energetics And Properties Of Organic Liquids". *Journal of the American Chemical Society* 118.45 (1996): 11225-11236. Web. 27 Jan. 2017.
20. Bizzarri, Anna Rita and Salvatore Cannistraro. "Molecular Dynamics Of Water At The Protein−Solvent Interface". *The Journal of Physical Chemistry B* 106.26 (2002): 6617-6633. Web. 27 Jan. 2017.
21. Tomasi, Jacopo, Benedetta Mennucci, and Roberto Cammi. "Quantum Mechanical Continuum Solvation Models". *Chemical Reviews* 105.8 (2005): 2999-3094. Web. 30 May 2017.
22. Yuet, Pak K. and Daniel Blankschtein. "Molecular Dynamics Simulation Study Of Water Surfaces: Comparison Of Flexible Water Models". *The Journal of Physical Chemistry B* 114.43 (2010): 13786-13795. Web. 27 Jan. 2017.
23. Nelson, M. T. et al. "NAMD: A Parallel, Object-Oriented Molecular Dynamics Program". *International Journal of High Performance Computing Applications* 10.4 (1996): 251-268. Web.
24. The Norma computer cluster, http://norma.mbg.duth.gr/

25. Structural and Computational Biology group of Molecular Biology and Genetics Department in Alexandroupolis, Greece. https://utopia.duth.gr/glykos/people.html

26. Phillips, James C. et al. "Scalable Molecular Dynamics With NAMD". *Journal of Computational Chemistry* 26.16 (2005): 1781-1802. Web. 5 Feb. 2017.

27. Schlick, Tamar. *Molecular Modeling And Simulation: An Interdisciplinary Guide*. 1st ed. New York, NY: Springer Science+Business Media, LLC, 2010. Print.

28. Leach, Andrew R. *Molecular Modelling*. 1st ed. Harlow, England: Prentice Hall, 2001. Print.

29. Glykos, Nicholas M. "Software News And Updates Carma: A Molecular Dynamics Analysis Program". *Journal of Computational Chemistry* 27.14 (2006): 1765-1768. Web. 6 Feb. 2017.

30. Koukos, Panagiotis I. and Nicholas M. Glykos. "Grcarma: A Fully Automated Task-Oriented Interface For The Analysis Of Molecular Dynamics Trajectories". *Journal of Computational Chemistry* 34.26 (2013): 2310-2312. Web. 6 Feb. 2017.

31. Heinig, M. and D. Frishman. "STRIDE: A Web Server For Secondary Structure Assignment From Known Atomic Coordinates Of Proteins". *Nucleic Acids Research* 32.Web Server (2004): W500-W502. Web. 6 Feb. 2017.

32. Price, Daniel J., and Charles L. Brooks. "A Modified TIP3P Water Potential For Simulation With Ewald Summation". *The Journal of Chemical Physics* 121.20 (2004): 10096-10103. Web. 30 May 2017.

33. Zhou, Ruhong. "Free Energy Landscape Of Protein Folding In Water: Explicit Vs. Implicit Solvent". *Proteins: Structure, Function, and Genetics* 53.2 (2003): 148-161. Web. 30 May 2017.

34. Hsin, Jen et al. "Using VMD: An Introductory Tutorial". *Current Protocols in Bioinformatics* (2008): n. pag. Web. 31 May 2017.

35. Knapp, B. et al. "Is An Intuitive Convergence Definition Of Molecular Dynamics Simulations Solely Based On The Root Mean Square Deviation Possible?". *Journal of Computational Biology* 18.8 (2011): 997-1005. Web. 14 June 2017.

36. Schneider, T D, and R M Stephens. “Sequence Logos: A New Way to Display Consensus Sequences.” *Nucleic Acids Research* 18.20 (1990): 6097–6100.

37. Crooks, G. E. "Weblogo: A Sequence Logo Generator". *Genome Research* 14.6 (2004): 1188-1190. Web. 31 May 2017.

38. David, Charles C., and Donald J. Jacobs. "Principal Component Analysis: A Method For Determining The Essential Dynamics Of Proteins". *Protein Dynamics* (2013): 193-226. Web. 1 June 2017.

39. Mu, Yuguang, Phuong H. Nguyen, and Gerhard Stock. "Energy Landscape Of A Small Peptide Revealed By Dihedral Angle Principal Component Analysis". *Proteins: Structure, Function, and Bioinformatics* 58.1 (2004): 45-52. Web. 1 June 2017.

40. Baltzis, Athanasios S., and Nicholas M. Glykos. "Characterizing A Partially Ordered Miniprotein Through Folding Molecular Dynamics Simulations: Comparison With The Experimental Data". *Protein Science* 25.3 (2015): 587-596. Web. 2 June 2017.

41. Sayle, R. "RASMOL: Biomolecular Graphics For All". *Trends in Biochemical Sciences* 20.9 (1995): 374-376. Web. 2 June 2017.

42. Altis, Alexandros et al. "Construction Of The Free Energy Landscape Of Biomolecules Via Dihedral Angle Principal Component Analysis". *The Journal of Chemical Physics* 128.24 (2008): 245102. Web. 4 June 2017.

43. http://www1.udel.edu/chem/bahnson/chem645/NMR-vs-Crystal.pdf

44. Zagrovic, Bojan, and Wilfred F. van Gunsteren. "Comparing Atomistic Simulation Data With The NMR Experiment: How Much Can Noes Actually Tell Us?". *Proteins: Structure, Function, and Bioinformatics* 63.1 (2006): 210-218. Web. 4 June 2017.

45. Patapati, Kalliopi K., and Nicholas M. Glykos. "Three Force Fields' Views Of The 310 Helix". *Biophysical Journal* 101.7 (2011): 1766-1771. Web. 4 June 2017.

46. Case, David A., Christoph Scheurer, and Rafael Brüschweiler. "Static And Dynamic Effects On Vicinal Scalarjcouplings In Proteins And Peptides: A MD/DFT Analysis". *Journal of the American Chemical Society* 122.42 (2000): 10390-10397. Web. 4 June 2017.

47. Coxon, Bruce. "Chapter 3 Developments In The Karplus Equation As They Relate To The NMR Coupling Constants Of Carbohydrates". *Advances in Carbohydrate Chemistry and Biochemistry* (2009): 17-82. Web. 4 June 2017.

48. Best, Robert B., Nicolae-Viorel Buchete, and Gerhard Hummer. "Are Current Molecular Dynamics Force Fields Too Helical?". *Biophysical Journal* 95.1 (2008): L07-L09. Web. 4 June 2017.

49. Georgoulia, Panagiota S., and Nicholas M. Glykos. "Usingj-Coupling Constants For Force Field Validation: Application To Hepta-Alanine". *The Journal of Physical Chemistry B* 115.51 (2011): 15221-15227. Web. 4 June 2017.

50. Lindorff-Larsen, Kresten et al. “Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field.” *Proteins* 78.8 (2010): 1950–1958. *PMC*. Web. 12 June 2017.

51. http://norma.mbg.duth.gr/index.php?id=research:howto:md_and_nmr_calculation_of_noes

52. http://norma.mbg.duth.gr/index.php?id=research:howto:md_and_nmr_calculation_of_j-couplings

53. Lawrence, Peter A. et al. "Borders Of Parasegments In Drosophila Embryos Are Delimited By The Fushi Tarazu And Even-Skipped Genes". *Nature* 328.6129 (1987): 440-442.