Alexandroupolis 2019

DEPARTMENT OF MOLECULAR BIOLOGY & GENETICS DUTH

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ

DEMOCRITUS UNIVERSITY OF THRACE

UNDERGRADUATE THESIS

[MOLECULAR DYNAMICS SIMULATION OF A VAMMIN-DERIVED MUTANT PEPTIDE]

Evangelia Valaroutsou

Supervisor: Nicholaos M. Glykos

Αλεξανδρούπολη 2019

ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ & ΓΕΝΕΤΙΚΗΣ ΔΠΘ

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

[ΠΡΟΣΟΜΟΙΩΣΗ ΜΟΡΙΑΚΗΣ ΔΥΝΑΜΙΚΗΣ ΕΝΟΣ ΜΕΤΑΛΛΑΓΜΕΝΟΥ ΠΕΠΤΙΔΙΟΥ ΠΡΟΕΡΧΟΜΕΝΟΥ ΑΠΟ ΤΗΝ ΠΡΩΤΕΪΝΗ VAMMIN]

Ευαγγελία Βαλαρούτσου

Επιβλέπων Καθηγητής: Νικόλαος Μ. Γλυκός

Acknowledgments

Firstly, I wish to thank my professor and supervisor, Dr. Nicholaos M. Glykos, for all the help that he provided and for his unending patience. I would also like to thank my colleagues at Democritus University and especially in NMG group for making this team a great environment to work in. Furthermore, I wish to thank my close friends because they filled my life with beautiful memories, as well as my special one who helped me stand when things were most difficult. Finally, I'm sending my immense gratitude to my family for believing in me all these years. Especially to my brother who is always there for me, and to my mother who does everything to support me: thank you from the bottom of my heart.

Table of Contents

Abstract

Molecular Dynamics Simulations is a revolutionary technique used for the determination and study of molecular structures and motions. In present project, this technique is used to study the folding of W2W11, a Vammin-derived mutant peptide, and the results are compared with NMR studies in order to examine the accuracy of Molecular Dynamics Simulations. Vammin is a Vascular Endothelial Growth Factor found in snake venom. Loop-3 of Vammin, which adopts a β-hairpin structural conformation, appears to be a possible anti-angiogenic candidate. W2W11 is a mutant peptide that contains the sequence of Vammin's loop-3, but with a Trp-Trp pair inserted in a hydrogen-bonded site to act as a possible stabilizer for the peptide's folded structure in water solution. W2W11 has a total of 12 amino acids with the following sequence: MWVNPRTQSSWM. NMR experiments proved that when placed at a hydrogen-bonded site, the Trp side chains do not interact with each other. The peptide is mostly disordered, and no native-like β-hairpin is formed. Simulation of W2W11 is conducted using Amber-ff99SB-ILDN force field and TIP3P water model, and produces 5.868.600 frames over 4.69μs. The results derived from Q-T diagrams, RMSD matrix, Secondary Structure analysis and Principal Components analysis come in strong agreement with the experimental findings. NOE distance restraints calculated for the simulation have certain differences with the NMR results. Simulation chemical shifts show a significant level of agreement with the experimental results.

Περίληψη

Οι Προσομοιώσεις Μοριακής Δυναμικής αποτελούν μια επαναστατική μέθοδο για τον προσδιορισμό και την μελέτη της δομής και της κίνησης μορίων. Στην παρούσα εργασία, η τεχνική αυτή χρησιμοποιείται για την μελέτη της αναδίπλωσης του πεπτιδίου W2W11, ένα μεταλλαγμένο πεπτίδιο προερχόμενο από την πρωτεΐνη Vammin, και τα αποτελέσματα συγκρίνονται με πειράματα NMR ώστε να διαπιστωθεί η ακρίβεια της μεθόδου προσομοιώσεων. Η πρωτεΐνη Vammin είναι ένας αγγειακός ενδοθηλιακός αυξητικός παράγοντας που συναντάται στο δηλητήριο του φιδιού. Ο βρόχος 3 της πρωτεΐνης, που έχει δομή β-φουρκέτας, αποτελεί πιθανό αντι-αγγειογόνο στόχο. Το W2W11 είναι ένα μεταλλαγμένο πεπτίδιο που περιέχει την αλληλουχία του βρόχου 3 της Vammin, με ένα ζευγάρι τρυπτοφανών εισαγμένο σε θέση υδρογονικού δεσμού για να λειτουργήσει σαν πιθανός σταθεροποιητής της αναδιπλωμένης δομής του πεπτιδίου σε υδατικό διάλυμα. Το W2W11 αποτελείται από 12 αμινοξέα με την εξής αλληλουχία: MWVNPRTQSSWM. Πειράματα NMR απέδειξαν ότι όταν τοποθετούνται σε θέση υδρογονικού δεσμού οι τρυπτοφάνες δεν αλληλεπιδρούν. Το πεπτίδιο είναι κυρίως αποδιατεταγμένο, και η σταθερή δομή β-φουρκέτας δεν παρατηρείται. Τα αποτελέσματα που λαμβάνονται από τα διαγράμματα Q-T, το πλέγμα RMSD και τις αναλύσεις δευτεροταγούς δομής και ομαδοποίησης βάση κύριων συνιστωσών έρχονται σε συμφωνία με τα πειραματικά ευρήματα. Οι περιορισμοί αποστάσεων (NOEs) που υπολογίστηκαν για την προσομοίωση έχουν ορισμένες διαφορές με αυτούς που εξήχθησαν από το φάσμα NOE των πειραμάτων NMR. Οι χημικές μετατοπίσεις της προσομοίωσης έχουν σημαντικό επίπεδο συμφωνίας με τα πειραματικά αποτελέσματα.

Introduction

1.1 Life is proteins

Proteins are the most abundant biomolecule in nature; they exist in every part of every cell and are very heterogeneous molecules. Thousands of different protein types and sizes can be found within a single cell, each having a unique role and faring its own responsibilities. The proper cooperation of all results in a well-functioning living organism. The source of such vast protein diversity are the amino acids; simple monomer subunits, combined in multiple ways specified by genes to create thousands of different polypeptide sequences. The process from sequence to a full-fledged protein follows several steps, referred to as the primary, secondary and tertiary structures, followed sometimes by the quaternary structure; the creation of protein complexes. Each amino acid residue is bound to its neighbor residue with a special type of covalent bond.[1] A peptide bond is created between the -NH_2 of one amino acid and the -COOH of the other, and the linkage of the two is accompanied by the loss of a molecule of water. This creates a specific amino acid order for each sequence; the primary structure of a polypeptide chain. A polypeptide chain has polarity because its ends are different, with a free α-amino group at one end and a free α-carboxyl group at the other. Furthermore, it consists of a repeating part, called the main chain or backbone, and a variable part, that contains the distinctive side chains.[2]

Hydrogen bonds that form between atoms of the backbone create the secondary structure. These bonds can form between a partially negative oxygen atom and a partially positive nitrogen atom. Most proteins have parts of their polypeptide chains that take the shape of either coiled or folded patterns, something that contributes to the protein's structure. Many of these coils and folds appear often in nature, and thus have been given names. Two very common examples are the α-helix and the β-sheet.[3]

The tertiary structure, also known as native state of the protein, is the overall three-dimensional shape of the protein, formed by interactions of the R groups; the side chains of the various amino acids. These interactions can be polar, nonpolar, or charged. The polar and charged amino acids are hydrophilic and as such can dissolve in water, while the nonpolar amino acids are hydrophobic and cannot dissolve in water. Multiple different secondary structure domains can be present at this point.[4]

The tertiary structure, and thus the protein's function, is determined by the primary structure. The most important proof for this comes from Christian Anfinsen's thermodynamic hypothesis. From his experiments he concluded, "These results suggest that the native molecule is the most stable configuration, thermodynamically speaking, and that the major force in the correct pairing of sulfhydryl groups in disulfide linkage is the concerted interaction of side-chain functional groups distributed along the primary sequence.” Further investigations on the reversible denaturation of several additional proteins helped verify Anfinsen's hypothesis.[5][6] Certain proteins denatured by heat, extreme pH, or denaturing reagents will regain their original structure and function when conditions return to the ones in which the native state of the protein was stable.[4] Anfinsen’s work enabled a large research enterprise of in vitro protein folding that has come to understand native structures by experiments inside test tubes rather than inside cells.[7]

Throughout the years, countless of researches into the field of protein structure gave birth to a very important question; how does a protein’s amino acid sequence dictate its three-dimensional atomic structure? This issue, called the Protein Folding Problem, is still being researched, and understanding it is the key to discovering more about the roles and functions of all proteins.

1.2 The enigma of protein folding

As mentioned already, a major milestone in protein science was the thermodynamic hypothesis of Christian Anfinsen. Through his experiments, Anfinsen claimed that the native structure of a protein is the lowest free-energy thermodynamically stable conformation, which depends only on the amino acid sequence and the solution conditions. The folding problem can be broken down into three basic steps: the folding code, the folding process and the structure prediction.

The folding code question is whether there is one dominant factor that indicates why different proteins will have different native structures. Truth is, not one but many different small interactions, such as hydrogen bonds, ion pairs, van der Waals attractions, or hydrophobic interactions are considered part of the folding code. Some of them, such as electrostatic interactions, affect protein folding in a less dominant way. Others can play a major role in the process, the most evident example being the hydrophobic interactions. It is generally accepted that the folding code has to be written less in the backbone and more in the side chains, since that is where proteins differ from one another.[7]

A very important development on the question of the folding process took place in 1968, when Cyrus Levinthal made the argument that "there are too many possible conformations for the unfolded protein to find the native state in conformational space by random searching". In its description of the 125 most important unsolved problems in science, *Science* magazine framed the problem this way: "Can we predict how proteins will fold? Out of a near infinitude of possible ways to fold, a protein picks one in just tens of microseconds. The same task takes 30 years of computer time".[7][8][10] This was named Levinthal's Paradox and it inspired scientists to search for ways in which the different polypeptides are guided through specific folding patterns. Different models were proposed to explain this process, some of which will now be briefly presented in chronological order.

Framework model

Also known as the sequential or hierarchical model, it was proposed by Ptitsyn in 1973. This model suggests that folding takes place in a step-by-step orderly fashion, starting with the quick formation of native secondary structural elements and followed by interactions between them which result in the formation of an advanced intermediate. The latter part of the process is relatively slow and includes the formation of the tertiary structure by diffusion and collision and also the construction of the quaternary structure in the case that there exist multiple protein strands. According to this model, each step stabilizes the previous step, which suggests the creation of several intermediates, and the local interactions play a dominant role in guiding the formation of the different elements. Although accurate in some cases, this model fails to explain the fast kinetics of protein folding, and has not been proven experimentally for a large number of proteins.[8][9][11]

Diffusion collision model

A 1976 suggestion by Karplus and Weaver, the diffusion collision model treats the protein as an assembly of microdomains; unstable portions of the developing secondary structure. To gain stability, two microdomains diffuse, collide and coalesce into a more stable formation, which then repeats the process with a third microdomain, and so on. Like an extension to the framework model, this model also suggests a gradually increasing stability on each step of the process towards the native structure. The different properties and interactions of the microdomains, rather than those of individual amino acids, are of great importance. They ensure that each process that leads to a different protein has a unique order of steps, and are the leads to our understanding of the kinetic event that describes protein folding.[8][12]

Hydrophobic collapse model

As mentioned before, proteins consist of multiple amino acids with different properties according to their chemistry. Amino acids with polar side chains are hydrophilic and able to dissolve in water, while amino acids with non-polar side chains are hydrophobic and avoid water. It was observed since the beginning of protein folding studies that, in a water-rich solution, proteins prefer to hide their non-polar amino acids towards the core of the structure, while water friendly polar amino acids would be exposed on the protein surface. According to the hydrophobic collapse model, first described by Kauzmann in 1959, proteins react to this hydrophobic force field by folding into a state that reminds us of an “oil-drop”, a state that has lower free energy than the unfolded state and so is preferred, but still has higher free energy than the protein’s native structure.[8][13][14]

Nucleation condensation model

The nucleation model suggests a process similar to crystallization. As protein folds, it has a tendency to form a nucleus, which is then used as a guiding point for the creation of the protein’s structure. There exist two versions of the nucleation model. Nucleation-propagation describes the formation of a strong local nucleus, followed by rapid propagation of the structure. Nucleation-condensation suggests the initial formation of a weak local nucleus that consists of several locally folded regions and is then stabilized by interactions between them. Once a more stable intermediate with lower overall free energy is formed, a transition state is reached. Only then does the nucleus build up begin. Key to the condensation model is the fact that the formation of nucleus is accompanied by the formation of secondary and tertiary structures.[8][15]

A schematic overview of the classical approach to folding mechanisms is presented below.

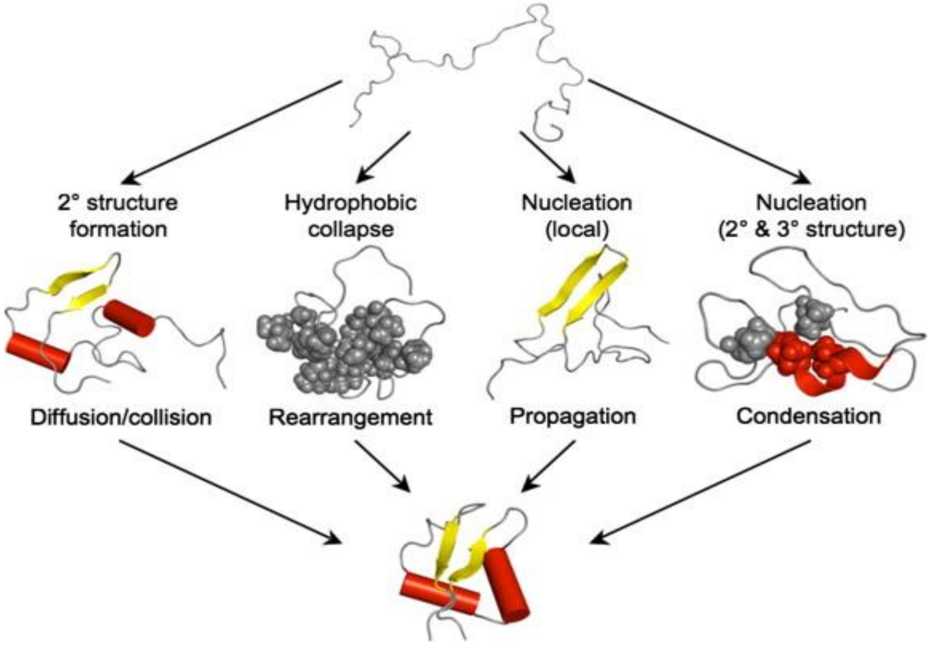


Figure 1: A description of the four 'classical' folding mechanisms. Reproduced without permission from Adrian A Nickson, Jane Clarke. "What lessons can be learned from studying the folding of homologous proteins?" Methods. 52(1):38-50. 2010 [16]

Energy landscape and protein folding funnel

Introduced by Bryngelson et al in 1995 the energy landscape theory is a modern approach to the folding problem that describes protein folding as a downhill process. According to the second law of thermodynamics in a system with constant pressure and temperature, Bryngelson used Gibbs free energy as a function of protein conformation to describe the protein-solvent system.[17] This description was depicted as a funnel. The folding funnel may have many local minima that can trap folding proteins into non-native states. The funnel's deep minimum corresponds to a single well-defined structure, the protein's native form, a state in which Gibbs free energy reaches the absolute minimum. The depth of the funnel represents the protein's energy stabilization, and the width represents the conformational entropy of the system.[18]

Gibbs free energy can be determined through the terms of entropy and enthalpy. During the first stages of protein folding, hydrophobic reactions are dominant and drive the protein into the formation of subsequent favored states with lower free energy each time (Figure 2, A). Throughout this process, due to the gradual loss of its degrees of freedom, the protein's entropy is reduced, which favors the unfolded state. At the same time, an overall increase in enthalpy happens with the formation of favorable hydrogen bonds, electrostatic interactions and van der Waals attractions, both inside the protein itself and between protein and solvent, which results in even further decrease in free energy and also favors the folded state (Figure 2, B). Both entropy and enthalpy contribute in the protein-solvent system reaching Gibbs free energy minimum which corresponds to the native protein state.

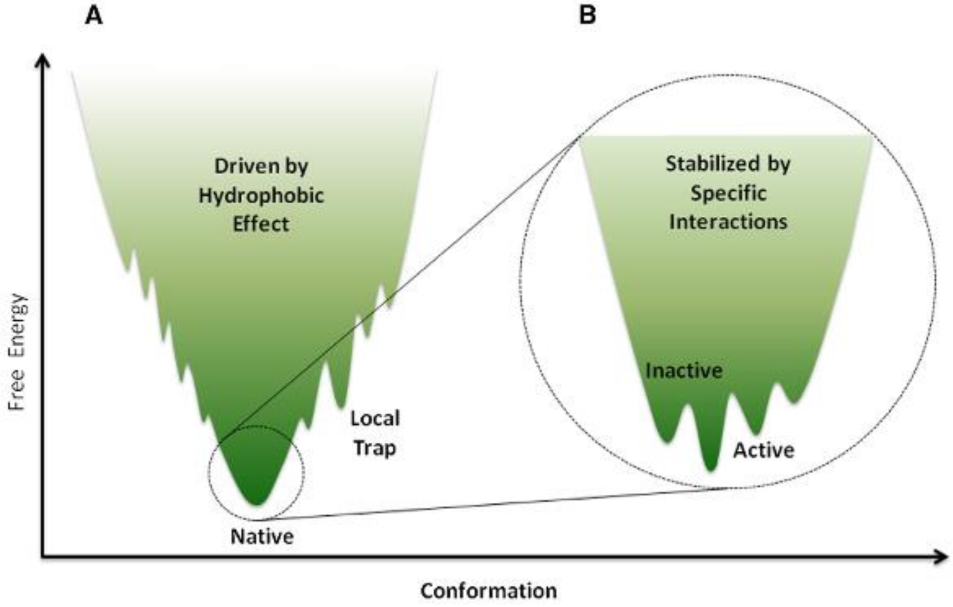


Figure 2: Protein folding depicted by free energy as a function of conformations. Reproduced without permission from Ruth Nussinov, Chung-Jung Tsai. "Free Energy Diagrams for Protein Function." Cell Press, Chemistry & Biology 21(3): 311-318. 2014. [18]

The relationship between Gibbs free energy, entropy and enthalpy can be expressed through the following formula:

$$\boldsymbol{\Delta G = \Delta H - T \, \Delta S}$$

where ΔG is the change in Gibbs free energy, ΔH is the change in enthalpy, ΔS is the change in entropy and T is the temperature in Kelvin.

The combined effects of the changes between entropy and enthalpy drive the protein to fold at any stage of protein folding, from high energy unstable states (funnel peaks), through lower energy more stable non-native conformations (local minima) and towards the lowest free energy native structure (deep minimum). [19]

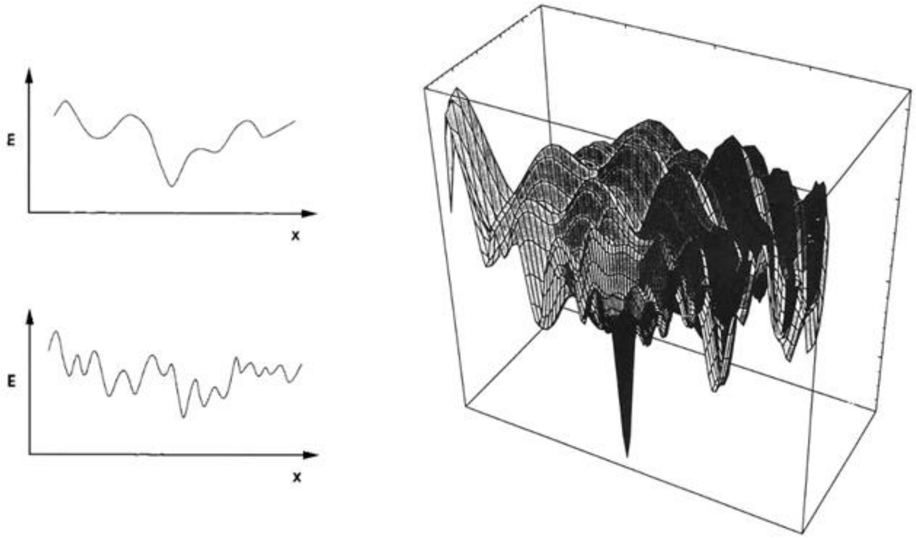


Figure 3: Energy landscapes as depicted by Bryngelson et al. in the initial study on protein folding funnels and energy landscapes. Reproduced without permission from Joseph D. Bryngelson, Jose Nelson Onuchic, Nicholas D. Socci, and Peter G. Wolynes. “Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis.” PROTEINS: Structure, Function, and Genetics 21:167-195. 1995 [17]

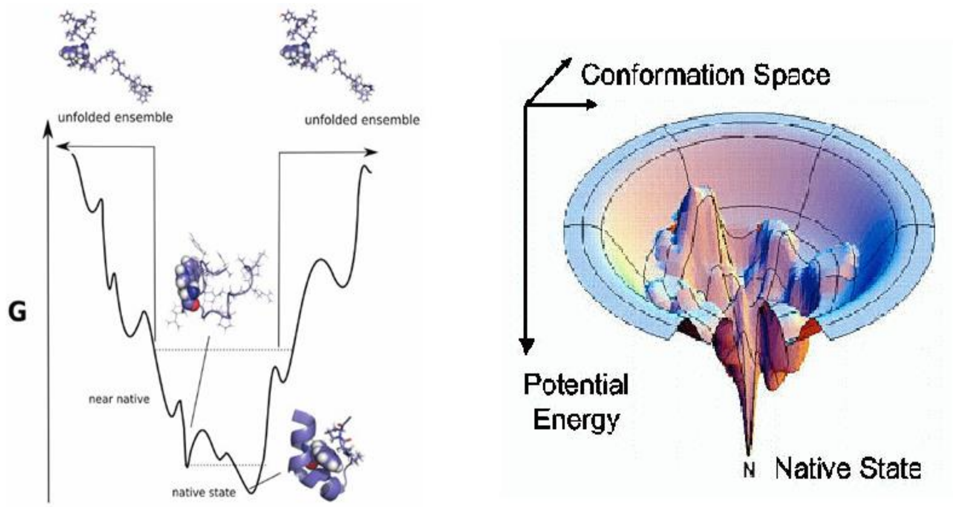


Figure 4: Left: Example of a small peptide’s energy landscape. Reproduced without permission from Emanuel Karl Peter. “Enhanced Sampling Techniques for Protein Folding Simulations.” Bunsen-Magazin 18(1). 2016 [20] Right: 3D image of an energy landscape. Reproduced without permission from Sadi Carnot. “Energy landscape: Landscape types” (source: http://www.eoht.info/page/Energy+landscape last update Dec 23 2015).

1.3 Protein analysis in the lab

Due to the technological revolution of our century and the creation of free and vast molecular databases, the study of protein structure and behavior is now more accessible than ever before. The increasing amount of data, the new evolutionary methods and experimental information, as well as the use of powerful machines has given scientists the necessary means to analyze biomolecules on an atomic level. Online molecular databases are now filled with valuable information and provide common access to all researchers. These databases have been enriched with the work and results of years of lab experiments and are now a great tool for even further analysis. Among other information, they provide 3D shapes of proteins, nucleic acids, and complex assemblies to be used widely in the fields of research and education. The two most popular methods for determining molecular structures in the lab are the following:

X-ray Crystallography

This method is used to determine atomic structures in high resolution and with great detail. A very brief and possibly inaccurate description to this technique would be as follows: proteins are crystallized and then hit with an intense X-ray beam. This beam is diffracted by the crystallized proteins and collected on a film. The crystal is then rotated and the process is repeated several times. Analyzing the film provides detailed atomic information and a magnified image of the protein can be recreated. It is important that the proteins are rigid and aligned in a nice order and orientation inside the crystal to be able to get precise results. Serial femtosecond crystallography, an evolution from classical X-ray crystallography, a method that utilizes X-ray Free Electron Lasers (XFEL) uses very short and bright pulses of radiation on tiny crystals to create thousands of individual diffraction patterns. This method enables the analysis of proteins' structure and behavior within very short time periods (femtoseconds to nanoseconds).[21]

Nuclear Magnetic Resonance Spectroscopy

NMR is a very common technique for analyzing molecular structures today. It uses strong magnetic fields and radio frequency waves to investigate how different atoms resonate according to their nuclei spin properties. This way, the location of each atom and interactions between different atoms can be determined. NMR is used to study proteins in solutions and as such, unlike crystallography, can be applied to structural studies of flexible proteins. [23]

Two other very effective techniques that are worth mentioning are:

Small-Angle Scattering

Like the name suggests this method investigates structures by studying the intensity of coherent scattering of radiation at small angles. Two types of radiation are most popular; X-ray radiation (SAXS) and neutron radiation (SANS), though other short wavelength particle beams are also available.[22]

Transmission Electron Microscopy

A big step forward from simple light microscopy, Transmission Electron Microscopy (TEM) also known as 3D Electron Microscopy (3DEM) uses a beam of electrons on the sample to produce high resolution 2D images that are then combined into 3D assemblies. The most common variation of this method today is called Cryo-Electron Microscopy and uses cell samples preserved in non-crystalline ice. This allows for the particles to be observed in their native environment, and as such is a very important tool for understanding the nature of their structure and behavior. [21]

1.4 Computational approach to protein analysis

With the use of methods like the ones mentioned above, thousands of molecular structures have already been determined. Today, molecular graphics software can be utilized to visually represent and study these structures (Figure 5). Furthermore, a new generation of mathematical and computational formulas able to simulate models for proteins' structure and dynamic behavior, using mere sets of data for atomic and desired solvent characteristics, are receiving increasing scientific interest. Weeks' worth of laboratory work is now reanimated on a computer screen in just a few hours, with gradually more reliable results and higher success rates.

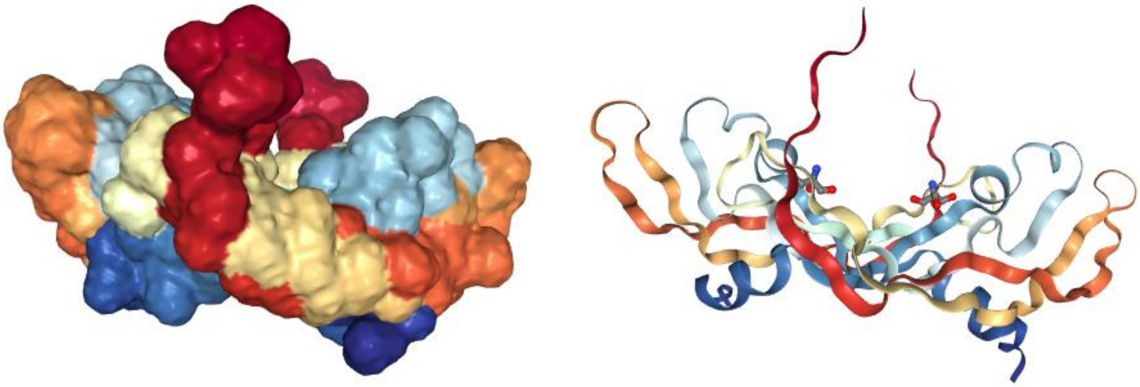


Figure 5: Visual representation of Vammin using surface style (left) and cartoon style (right). PDB entry: 1WQ8. Source: RCSB PDB.

This revolutionary technique, termed Molecular Dynamics Simulations, will be used in current project to analyze the structural stability of a mutant peptide.

1.5Vammin and the loop-3 peptide

Vascular endothelial growth factors (VEGF) play an important role in angiogenesis. As such, they appear to be promising candidates for angiogenic or anti-angiogenic responses in the treatment of numerous vascular related pathogenies.[24] Several VEGF groups have been discovered so far, with mammalian VEGF-A being the first one. It has been identified that VEGF can bind to four receptors, three of which are tyrosine kinases; Flt-1, KDR and

Flt-4, and one is a non-tyrosine kinase co-receptor for certain subtypes; neuropilin-1.

Vammin (PDB: 1WQ8, Figure 5) is a Vascular Endothelial Growth Factor found in snake venom and is a member of VEGF-F group alongside VR-1 (PDB: 1WQ9). It has 110 residues and shows very strong receptor selectivity for KDR.[25]

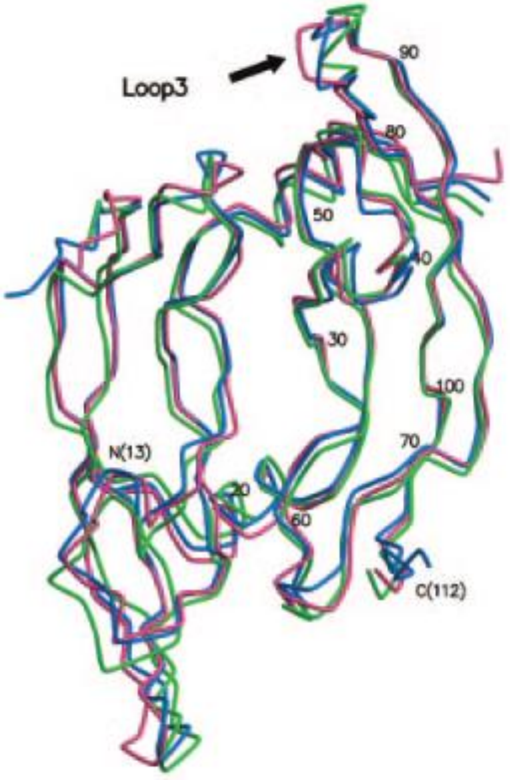


Figure 6: Backbone superposition models of vammin (blue), VR-1 (green) and VEGF-A (magenta). N (13) and C (112) indicate N- and C-terminal residues of vammin. Arrow indicates loop 3. Reproduced without permission from Kyoko Suto, Yasuo Yamazaki, Takashi Morita and Hiroshi Mizuno. “Crystal Structures of Novel Vascular Endothelial Growth Factors (VEGF) from Snake Venoms”. The Journal of Biological Chemistry. 280(3): 2126–2131. 2005 [25]

Amino-acid sequences for VEGF-A, vammin and VR-1 are presented in Figure 7. Reports have shown that the insertion of threonine residue 86 in loop 3 for vammin and VR-1 is particularly important for their KDR specific binding.[26]

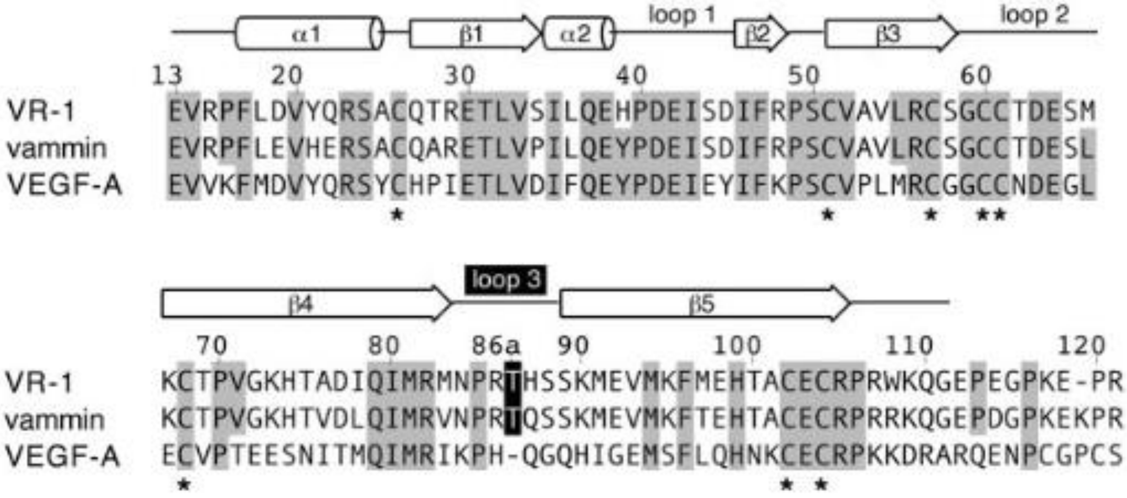

α1
β1
α2
loop 1
β2
β3
loop 2
13
20
30
40
50
60
VR-1 EVRPFLDVYQRSACQTRETLVSILQEHPDEISDIFRPSCVAVLRCSGCCTDESM
vammin EVRPFLEVHERSACQARETLVPILQEYPDEISDIFRPSCVAVLRCSGCCTDESL
VEGF-A EVVKFMDVYQRSYCHPIETLVDIFQEYPDEIEYIFKPSCVPLMRCGGCCNDEGL
β4
loop 3
β5
70
80
86a
90
100
110
120
VR-1 KCTPVGKHTADIQIMRMNPRTHSSKMEVMKFMEHTACECRPRWKQGEPEGPKE-PR
vammin KCTPVGKHTVDLQIMRVNPRTQSSKMEVMKFTEHTACECRPRRKQGEPDGPKEKPR
VEGF-A ECVPTEESNITMQIMRIKPH-QGQHIGEMSFLQHNKCECRPKKDRARQENPCGPCS

Figure 7: Structure-based sequence alignments of vammin, VR-1 and VEGF-A. Primary structures of the receptor-binding domains of VEGFs are aligned based on their crystal structures, and identical residues are shaded. The numbering at the top refers to residue number of VEGF-A. The inserted threonine residue in loop 3 of the two VEGF-Fs is highlighted in black and marked with 86a. The secondary structural elements are shown as arrows for β-strands and cylinders for α-helices, and loops are labeled. Conserved cysteine residues are marked with asterisks at the bottom. Reproduced without permission from Kyoko Suto, Yasuo Yamazaki, Takashi Morita and Hiroshi Mizuno. "Crystal Structures of Novel Vascular Endothelial Growth Factors (VEGF) from Snake Venoms" The Journal of Biological Chemistry. 280(3): 2126–2131. 2005 [25]

When it is part of the protein, loop 3 of vammin appears as a well-defined antiparallel 4:6 β-hairpin with a non-Gly β-bulge and overlapping β turns of type IV and I at the loop region.[27] In order to study it as a possible anti-angiogenic candidate, its isolation in solution is necessary. To achieve this, mutant peptides with stabilizing elements inserted at positions that are not important for the KDR interaction can be constructed, and whether they take stable native-like forms can be determined with the methods described above.

1.6Trp-Trp pair stabilizers: hydrogen bonded sites vs non-hydrogen bonded sites

Trp-Trp pairs can be used as stabilizers when installed in β-hairpin peptide conformations. In order to test this, Jiménez's group [27][28] has created, among others, two specific mutant peptides derived from loop 3 of vammin, W3W10 and W2W11. W3W10 mutant peptide holds the Trp-Trp pair in a non-hydrogen bonded site, as it replaced the native's non-hydrogen bonded residues V71 and S78. W2W11 holds the Trp-Trp pair in a hydrogen-bonded site, replacing residues R70 and K79.

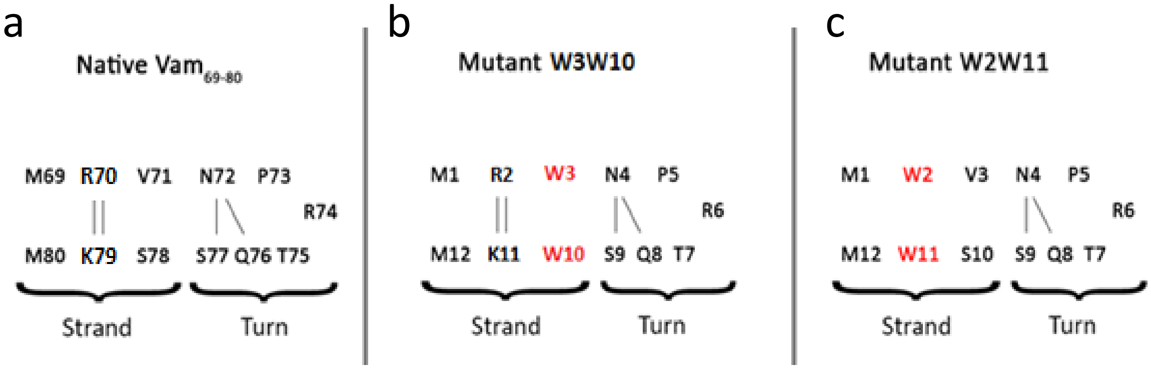


Figure 8: Sequences for a) loop 3 of native vammin, residues 69-80, b) W3W10 mutant peptide and c) W2W11 mutant peptide.

Jiménez's NMR spectroscopy experiments have shown that the Trp-Trp pair placed in non-hydrogen bonded site works as a perfect stabilizer of the native-like form that W3W10 takes in solution.[27] Furthermore, Koukos' Folding Molecular Dynamics Simulation approach [29] for the determination of

W3W10's structural stability has proven to be in excellent agreement with the experimental NMR results.

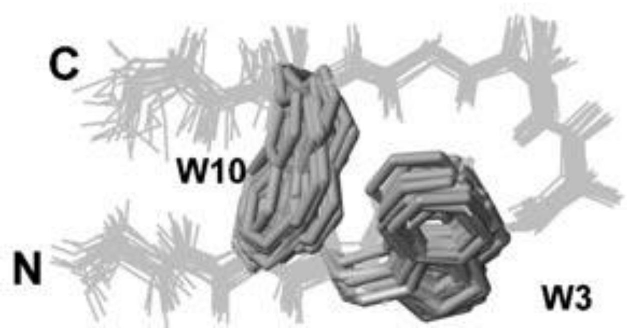


Figure 9: Superposition of the 20 lowest target function structures calculated for peptide W3W10. [27] Mirassou et al. Disulfide Bonds versus Trp···Trp Pairs in Irregular β-Hairpins: NMR Structure of Vammin Loop 3-Derived Peptides as a Case Study. ChemBioChem 10: 902-910. 2009

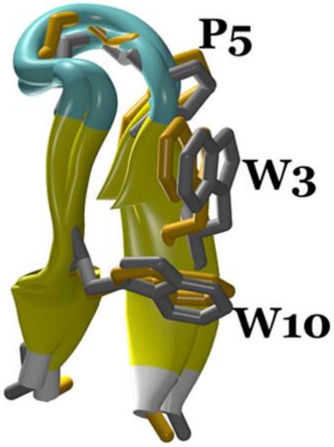


Figure 10: Structural comparison between W3W10 representative molecular-dynamics-derived structure (colored orange) with the experimentally (NMR) determined one (colored gray). [29] Koukos et al. Folding Molecular Dynamics Simulations Accurately Predict the Effect of Mutations on the Stability and Structure of a Vammin-Derived Peptide. J. Phys. Chem. B 118, 10076–10084. 2014

In present project, an MD simulation analysis for mutant peptide W2W11 was held, in order to test the stabilizing potency of the Trp-Trp pair in a hydrogen-bonded site. The technique will be explained and methods and results will be presented, compared with NMR experimental data and discussed.

Molecular Dynamics Simulations

2.1 Introduction

Scientists who study proteins and other biomolecules often wish to find a relation between their three-dimensional structure and their biological functions. Molecular dynamics simulations are considered a great tool to achieve that. They analyze the dynamic properties of molecules as a function of time, provided the scientist inputs initial positions of the particles (usually taken from crystallography or NMR results). MD simulations can answer questions about the attributes of the molecule-solvent system in an easier way than actual experiments on such system. Experiments play an important role in validating the simulation’s results through comparisons between the data so as to achieve accurate calculations, since the appearance of computational error always poses a threat. However, the scientist has complete control over the system’s properties, thus is able to test multiple alterations within a short amount of time in search for the ones that best simulate not only the native state, but also any other possible variation they wish to examine. In protein research today, the use of faster and cheaper supercomputers has enabled the successful folding of small peptides or small proteins.[30]

Molecular dynamics are also a great complement to other structure analysis tools, such as Monte Carlo simulations, Poisson-Boltzmann analyses, energy minimization, Brownian dynamics or enhanced sampling methods.

2.2Walk down history road

In 1687 Isaac Newton's second law described the basics behind force and motion. According to it, "*a body's acceleration equals the net force divided by its mass*". This concept is the foundation stone for molecular dynamics.

About a century later, Laplace thought about designing appropriate analysis tools that include mathematical equations of forces and motions, basically a vision of our modern molecular simulations.

Another century passes, and the figures of Van Der Waals and Boltzmann pose the initial ideas on how this method could actually work.

Traveling a few decades in time back to the 1950's, the birth years of molecular dynamics, one will surely be intimidated by names like Alder and Wainwright, Stillinger and Rahman, McCammon, Gelin and Karplus, all great men of modern science that have given their insight, built models, shaped and nourished the method of MD simulations towards its current glory.

2.3What can we learn?

Today, applications of simulation methods enable scientists to determine structures or refine low quality ones already obtained from experimental procedures. Furthermore, the development of molecular systems is represented and observed over time, a feature of significant importance compared to classical experiments. The results that we acquire through Molecular Dynamics simulations consist of a time series of conformations, i.e. a trajectory followed by each atom in accordance with Newton’s laws of motion. Analysis of this trajectory offers valuable information about a molecule’s structure and behavior; atomic mean-square fluctuations, local fluctuations like the creation of bonds or interactions with the solvent, particle motions, configurational changes, binding sites, free energies etc.

2.4Basic principles of MD simulation

The basics behind MD simulations will now be presented. The explanations given here are mainly based on Dr. Stote’s *Theory of Molecular Dynamics Simulations* [32] which analyzes the method in a very straightforward and understandable way. For even more information, Tamar Schlick [31] in her guide for *Molecular Modeling and Simulation* illustrates the method with great detail.

As mentioned before, the method is based on Newton's second law for force and motion. In a system with N number of atoms, each having its own Cartesian vector (x, y and z coordinates) and velocity vector, we have:

$$\boldsymbol{F} = \boldsymbol{m}\,\boldsymbol{a} \text{ (1)}$$

where F is the force exerted on the atom, m its mass and a its acceleration.

Another equation for force expresses it through the gradual change of potential energy:

$$\boldsymbol{F} = -\frac{\boldsymbol{dV}}{\boldsymbol{dr}} \text{ (2)}$$

where V is the potential energy and r the atom's position.

The important for MD simulation principle of time is added through the equation for each atom's acceleration. At a specific time t, the equation can be expressed as:

$$\boldsymbol{a} = \frac{\boldsymbol{d^2r}}{\boldsymbol{dt^2}} \text{ (3)}$$

Combining equations (1), (2) and (3), for each atom we have:

$$-\frac{\boldsymbol{dV}}{\boldsymbol{dr}} = \boldsymbol{m}\,\frac{\boldsymbol{d^2r}}{\boldsymbol{dt^2}} \quad \text{and} \quad \boldsymbol{a} = -\frac{\boldsymbol{dV}}{\boldsymbol{mdr}}$$

Looking at these two formulas it’s apparent that one can successfully calculate the atom’s trajectory as long as one knows the atom’s Cartesian vector and velocity vector, i.e. position, initial velocity and acceleration.

Atom positions in the system are typically known through data from crystallographic or NMR classical experiments, though *build-up* techniques can be used to construct a structure when data isn’t available (homology modeling).

Initial velocities are chosen randomly using a Maxwell-Boltzmann distribution:

$$\boldsymbol{p(v)} = \left(\frac{\boldsymbol{m}}{\boldsymbol{2\pi\, k_B\, T}}\right)^{\boldsymbol{1/2}} \boldsymbol{exp}\left(\frac{\boldsymbol{-m\, v^2}}{\boldsymbol{2\, k_B\, T}}\right)$$

where v is the velocity, k_B is Boltzmann’s constant and T is the temperature of the system.

2.5 Integration algorithms

Calculating atoms' accelerations is a tricky and demanding process. It is done by using force fields to calculate potential energies. Because of its level of difficulty, the use of algorithms is the generally preferred approach. When choosing an algorithm, one must pay attention to several factors: it must have the ability to conserve energy and momentum, it must have high computational efficiency and it must allow a long-time step for integration. This way, the results can be as close to reality as possible, though a level of inaccuracy will always be present. Some well-known algorithms are:

- Verlet algorithm
- Leap-frog algorithm
- Velocity Verlet algorithm
- Beeman's algorithm

The majority of algorithms work based on Taylor's series for the expansion of a function. For atomic positions, velocities and accelerations we have:

$$r(t + dt) = r(t) + v(t)\,dt + ½\,a(t)\,dt2 + \ldots$$

$$v(t + dt) = v(t) + a(t)\,dt + ½\,b(t)\,dt2 + \ldots$$

$$a(t + dt) = a(t) + b(t)\,dt + \ldots$$

where r is the position, v the velocity (the first derivative with respect to time) and a the acceleration (the second derivative with respect to time).

2.6Force fields

A force field is a collection of empirical equations and associated constants designed to calculate potential energies by reproducing molecular geometries and selected properties of the simulated structure. It describes the time evolution of bond lengths, bond angles and torsions, as well as the non-bonding van der Waals and electrostatic interactions between atoms by calculating vibrational frequencies, heats of formation, intermolecular energies, and more.

The total energy V of the system as a function of all the atomic positions R can be expressed as:

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

where E_{bonded} is the internal energy that describes bond stretch, angle and rotation and $E_{non-bonded}$ is the external energy that describes van der Waals and electrostatic interactions between non-bonded atoms.

E_{bonded} can be expressed as a sum of three elements:

$$E_{bonded} = E_{bond-stretch} + E_{angle-bend} + E_{rotate-along-bond}$$

$E_{non-bonded}$ can be describes as a sum of two elements:

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

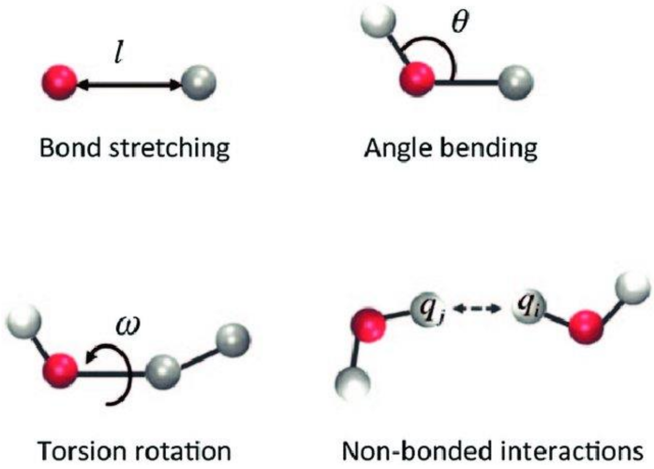


Figure 11: Representation of atoms as charged spheres, which have bonded (bond stretch, angle bend and torsional angle rotation) and nonbonded interactions (van der Waals and electrostatics). Reproduced without permission from Patricia Saenz-Méndez, Samuel Genheden, Anna Reymer, Leif A. Eriksson. Computational chemistry and molecular modeling basics. Computational Tools for Chemical Biology. 2017 [33].

A force field's parameterization process is a difficult task. One must make important decisions regarding the functional form and numerical values for the parameters. There exist endless combinations of parameters, even when one already has structure and energy data to work on. Much freedom and manipulation are possible in constructing empirical energy surfaces. Only if constructed and parameterized correctly will the energy model generate reliable structural predictions.

Several general issues still remain unresolved and/or need improvement:

Determination of partial charges; Improvement of electrostatic potentials; Methods for solvent representation; Interpretation of results in the absence of solvent; Cartesian vs. torsion space representation; Interpretation of conflicting results by different models and potentials.

Among the force fields used today, four are the most popular:

- AMBER (Assisted Model Building with Energy Refinement) [34]
- CHARMM (Chemistry at Harvard Macromolecular Mechanics) [35]
- GROMOS (Groningen Molecular Simulation) [36]
- OPLS (Optimized Potentials for Liquid Simulations) [37]

Improvement of potential energy functions of force fields has been an ongoing venture. The currently in use "*second-generation*" molecular mechanics and dynamics force fields appear to be a lot more sophisticated than the 1960s and 1970s originals. The better understanding and use of quantum mechanics today is significantly affecting the creation of better force fields, already leading to the emersion of "t*hird generation*" more accurate ones.

2.7System solvent

The environment in which a protein naturally abides plays a major role in its behavior and function. This environment, the solvent, which usually is water, has many effects that influence protein folding, dynamics and thermodynamics parameters. In the same way, when performing a simulation of said protein, having the proper solvent model included in the force field is also of great importance. In molecular simulations, the correct screening of electrostatic interactions in the system is one of the most important effects one must pay attention to when designing or choosing a solvent model. Solvent simulation can be approached in two ways: by adding or not adding water molecules.

Implicit solvent approach

In this type of solvent, water molecules are absent, but their effects are simulated by adding certain properties to the potential energy function. For example, a dielectric constant is used to properly screen the electrostatic interactions. The most usual and simplest method used is to add a distance-depended dielectric constant that calculates electrostatics as described by the Poisson Boltzmann equation. This method is termed Generalized Born Implicit Solvent (GBIS).[38] There also exist models that base the electrostatic screening on the molecule's accessible surface.

Explicit solvent approach

Here, molecules of water are explicitly simulated by all-atom force field models. The electrostatic interaction is modeled using Coulomb's law, and the dispersion and repulsion forces using the Lennard-Jones potential. It's apparent that the volume of complexity required is greater, but so is the level of detail. For that reason, certain limitations must also be included. The various explicit solvent models that exist today are able to represent different water properties such as radial distribution function, diffusivity, density anomaly and others, but no model can represent all of these properties simultaneously. Some well-known models are the simple point charge model (SPC), extended SPC/E, and TIP3P and TIP5P models. [39]

Methods

3.1 Software

To perform Molecular Dynamics simulation of W2W11 mutant peptide derived from Vammin, the NAMD software was used.[40][41] NAMD is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems, developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign. It uses molecular graphics program VMD for simulation setup and trajectory analysis, and is compatible with AMBER and CHARMM force fields.

3.2 Computing cluster

Molecular Dynamics simulation is a computationally demanding procedure. In order to increase the computational power and achieve better simulation performance, multiple computers can be connected in a parallel way to create a computing cluster. This way, the simulation is processed simultaneously by multiple cores.

For present project, the Norma computing cluster was used. [42] The cluster is located at the Department of Molecular Biology and Genetics of Democritus University of Thrace in Alexandroupolis, Greece.

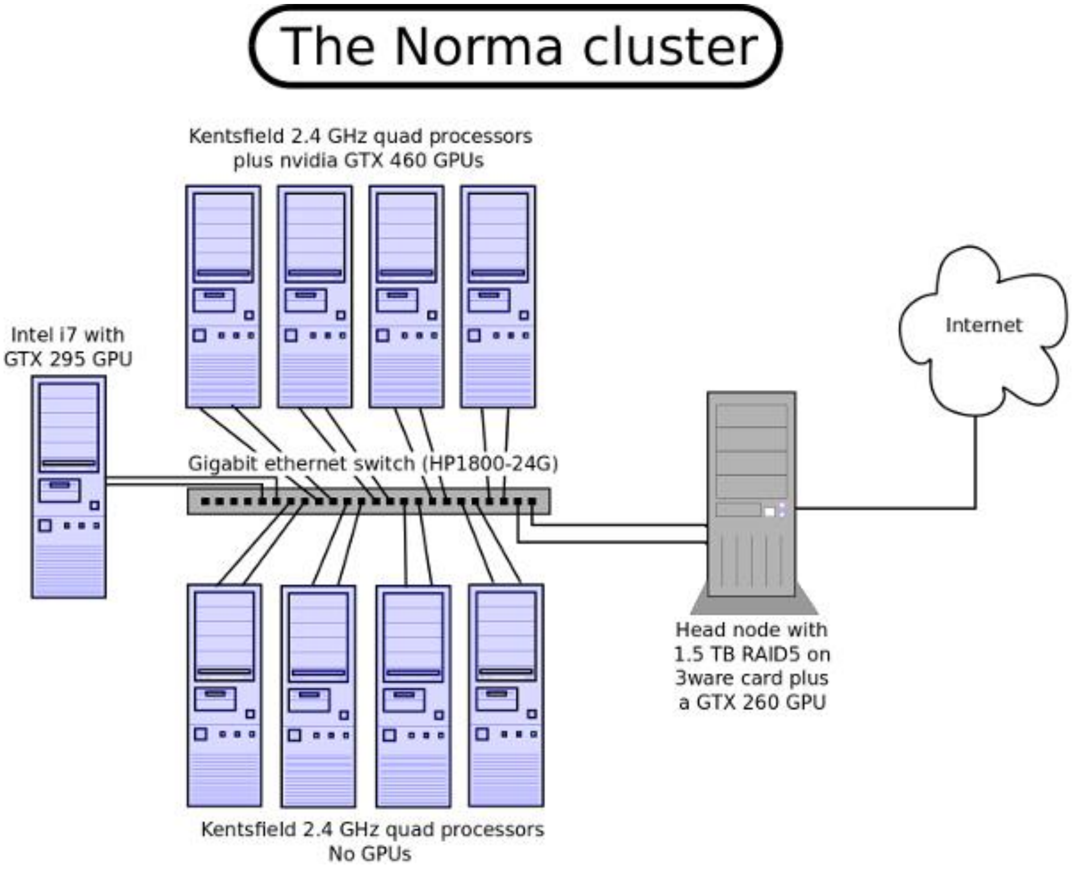

The Norma cluster
Kentsfield 2.4 GHz quad processors plus nvidia GTX 460 GPUs
Intel i7 with GTX 295 GPU
Gigabit ethernet switch (HP1800-24G)
Internet
Head node with 1.5 TB RAID5 on 3ware card plus a GTX 260 GPU
Kentsfield 2.4 GHz quad processors No GPUs

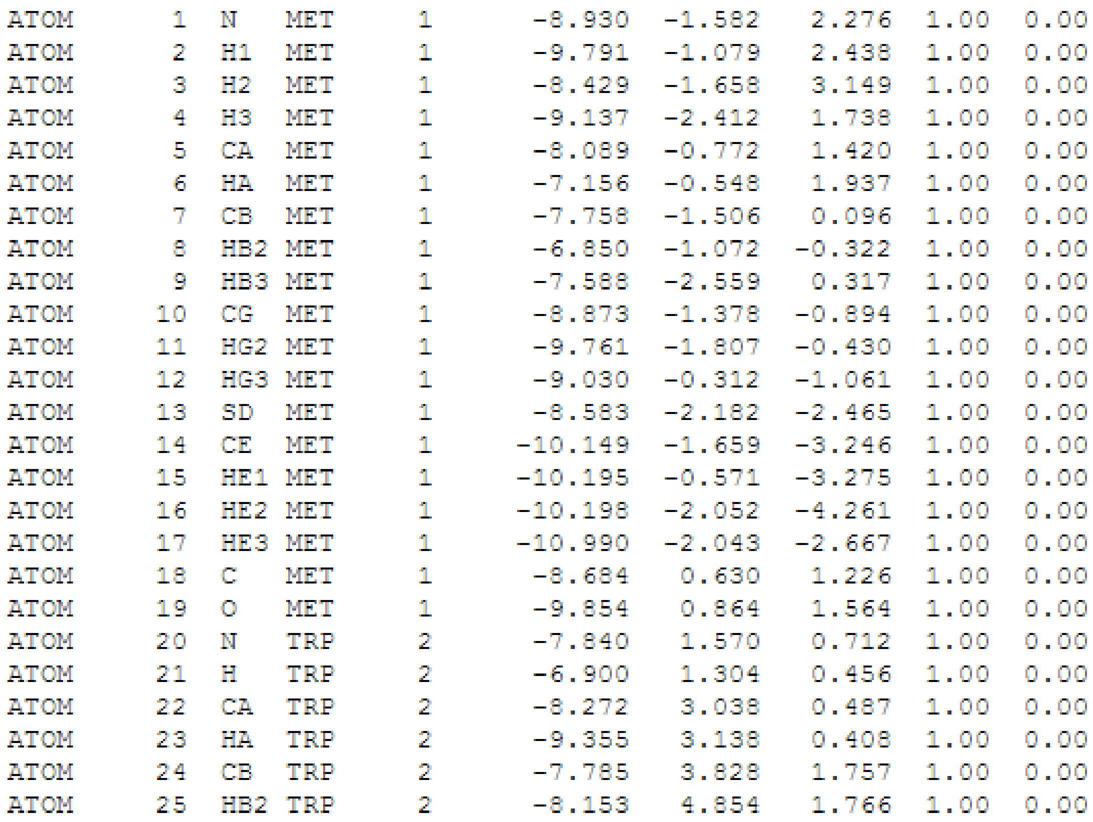
Figure 12: The Norma cluster characteristics: Norma comprises 40 CPU cores, 46 Gbytes of physical memory and 6 GPGPUs distributed over 10 nodes. The nodes are based on Intel's Q6600 Kentsfield 2.4 GHz quad processors and are connected via a dedicated HP ProCurve 1800-24G Gigabit ethernet switch. Each of the nine diskless (compute-only) nodes offers four cores, four Gbytes of physical memory and two (gigabit) network interfaces, with the exception of one node based on Intel's i7 965 extreme which 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. Norma is presently used almost exclusively for computational biology and crystallography projects of the structural and computational biology group. [42]

3.3 System customization

Several files are required to perform a Molecular Dynamics simulation using the NAMD software.

Protein Data Bank file

Atomic coordinates and/or velocities for the system are included in a .pdb file which can either be created or accessed and downloaded through the Protein Data Bank platform. This file has the following form:



```
ATOM      1  N    MET     1      -8.930  -1.582   2.276  1.00   0.00
ATOM      2  H1   MET     1      -9.791  -1.079   2.438  1.00   0.00
ATOM      3  H2   MET     1      -8.429  -1.658   3.149  1.00   0.00
ATOM      4  H3   MET     1      -9.137  -2.412   1.738  1.00   0.00
ATOM      5  CA   MET     1      -8.089  -0.772   1.420  1.00   0.00
ATOM      6  HA   MET     1      -7.156  -0.548   1.937  1.00   0.00
ATOM      7  CB   MET     1      -7.758  -1.506   0.096  1.00   0.00
ATOM      8  HB2  MET     1      -6.850  -1.072  -0.322  1.00   0.00
ATOM      9  HB3  MET     1      -7.588  -2.559   0.317  1.00   0.00
ATOM     10  CG   MET     1      -8.873  -1.378  -0.894  1.00   0.00
ATOM     11  HG2  MET     1      -9.761  -1.807  -0.430  1.00   0.00
ATOM     12  HG3  MET     1      -9.030  -0.312  -1.061  1.00   0.00
ATOM     13  SD   MET     1      -8.583  -2.182  -2.465  1.00   0.00
ATOM     14  CE   MET     1     -10.149  -1.659  -3.246  1.00   0.00
ATOM     15  HE1  MET     1     -10.195  -0.571  -3.275  1.00   0.00
ATOM     16  HE2  MET     1     -10.198  -2.052  -4.261  1.00   0.00
ATOM     17  HE3  MET     1     -10.990  -2.043  -2.667  1.00   0.00
ATOM     18  C    MET     1      -8.684   0.630   1.226  1.00   0.00
ATOM     19  O    MET     1      -9.854   0.864   1.564  1.00   0.00
ATOM     20  N    TRP     2      -7.840   1.570   0.712  1.00   0.00
ATOM     21  H    TRP     2      -6.900   1.304   0.456  1.00   0.00
ATOM     22  CA   TRP     2      -8.272   3.038   0.487  1.00   0.00
ATOM     23  HA   TRP     2      -9.355   3.138   0.408  1.00   0.00
ATOM     24  CB   TRP     2      -7.785   3.828   1.757  1.00   0.00
ATOM     25  HB2  TRP     2      -8.153   4.854   1.766  1.00   0.00
```

Figure 13: Preview of W2W11 mutant peptide .pdb file. From left to right the columns indicate: record type, atom ID, atom name, residue name, residue ID, x, y, and z coordinates, occupancy, and temperature factor.

Force field parameters file

As mentioned in chapter 2.6, the force field parameters file (.prm) contains all of the equations and constants needed to evaluate forces and energies. It defines bond strengths, equilibrium lengths, etc. For this simulation, Amber-ff99SB-ILDN [46] force field was used. This is a preview of the parameters file:

```
%VERSION  VERSION_STAMP = V0001.000  DATE = 10/08/14  13:02:43
%FLAG TITLE
%FORMAT(20a4)
default_name
%FLAG POINTERS
%FORMAT(10I8)
    4938      16    4831     110     224     150     439     353       0
0
    7430    1590     110     150     353      35      70      49      26
1
       0       0       0       0       0       0       0       1      24
0
       0
%FLAG ATOM_NAME
%FORMAT(20a4)
N   H1  H2  H3  CA  HA  CB  HB2 HB3 CG  HG2 HG3 SD  CE  HE1 HE2 HE3 C   O
N
. . .
%FLAG CHARGE
%FORMAT(5E16.8)
  2.90099016E+00  3.61530432E+00  3.61530432E+00  3.61530432E+00
4.02712830E-01
. . .
%FLAG ATOMIC_NUMBER
%FORMAT(10I8)
       7       1       1       1       6       1       6       1       1
6
. . .
%FLAG MASS
%FORMAT(5E16.8)
  1.40100000E+01  1.00800000E+00  1.00800000E+00  1.00800000E+00
1.20100000E+01 . . .
%FLAG ATOM_TYPE_INDEX
%FORMAT(10I8)
       1       2       2       2       3       4       3       5       5
3
. . .
%FLAG NONBONDED_PARM_INDEX
```

```
%FORMAT(10I8)
       1       2       4       7      11      16      22      29      37
46
. . .

%FLAG RESIDUE_LABEL
%FORMAT(20a4)
MET TRP VAL ASN PRO ARG THR GLN SER SER TRP MET Cl- WAT WAT WAT WAT WAT
WAT
. . .
%FLAG RESIDUE_POINTER
%FORMAT(10I8)
       1      20      44      60      74      88     112     126     143
154
. . .
%FLAG BOND_FORCE_CONSTANT
%FORMAT(5E16.8)
  5.70000000E+02  4.90000000E+02  3.40000000E+02  2.27000000E+02
3.40000000E+02
. . .
%FLAG BOND_EQUIL_VALUE
%FORMAT(5E16.8)
  1.22900000E+00  1.33500000E+00  1.09000000E+00  1.81000000E+00
1.09000000E+00
. . .
%FLAG ANGLE_FORCE_CONSTANT
%FORMAT(5E16.8)
  8.00000000E+01  5.00000000E+01  5.00000000E+01  3.50000000E+01
5.00000000E+01
. . .
%FLAG ANGLE_EQUIL_VALUE
%FORMAT(5E16.8)
  2.14501057E+00  2.09439600E+00  2.12755727E+00  1.91113635E+00
1.91113635E+00
. . .
%FLAG DIHEDRAL_FORCE_CONSTANT
%FORMAT(5E16.8)
  2.00000000E+00  2.50000000E+00  0.00000000E+00  2.00000000E+00
4.00000000E-01
. . .
%FLAG DIHEDRAL_PERIODICITY
%FORMAT(5E16.8)
  1.00000000E+00  2.00000000E+00  2.00000000E+00  2.00000000E+00
3.00000000E+00
 . . .
%FLAG DIHEDRAL_PHASE
%FORMAT(5E16.8)
  0.00000000E+00  3.14159400E+00  0.00000000E+00  0.00000000E+00
0.00000000E+00
. . .

. . .
```

Configuration file

This file is needed for the user to specify the options that NAMD should adopt while running the simulation. The configuration options used for this simulation are the following:

```
# Input files
#
amber                   on
readexclusions          yes
parmfile                vammin_2W.prmtop
coordinates             heat_out.coor
velocities              heat_out.vel
extendedSystem          heat_out.xsc


#
# Adaptive ...
#
adaptTempMD             on
adaptTempTmin           300
adaptTempTmax           500
adaptTempBins           1000
adaptTempRestartFile    output/restart.tempering
adaptTempRestartFreq    10000
adaptTempLangevin       on
adaptTempRescaling      off
adaptTempOutFreq        400


#
# 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            10
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            36                      # <===== CHANGE ME
PmeGridsizeY            36                      # <===== CHANGE ME
PmeGridsizeZ            36                      # <===== 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           30000                  # <===== CHANGE ME
run                     500000000              # <===== CHANGE ME
```

3.4System setup

A sequence of steps is performed in order to produce a simulation of the structure's trajectory.[32] A schematic representation of these steps is shown below.

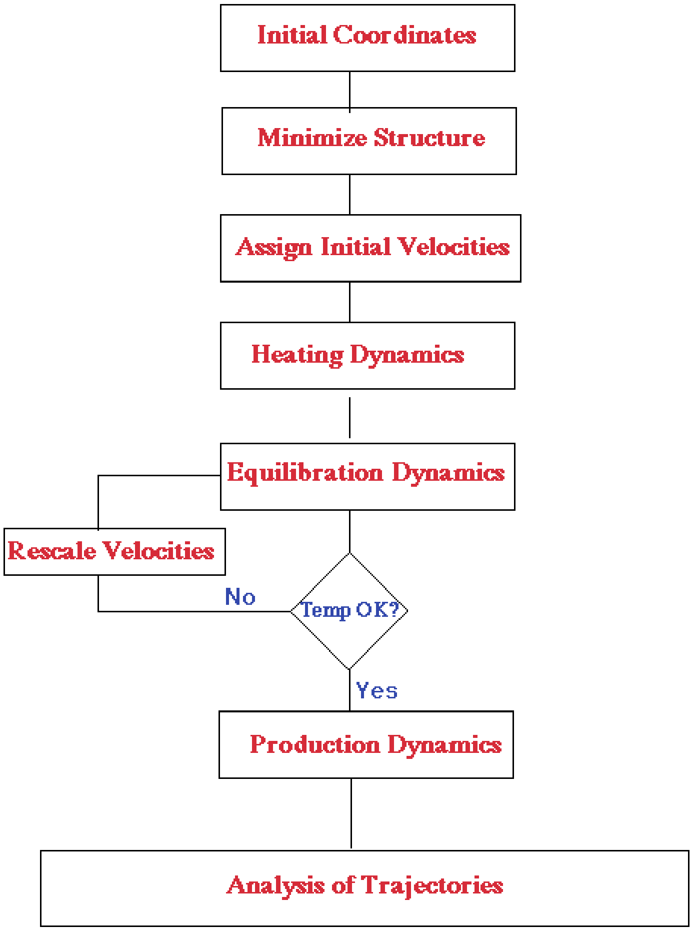


Figure 14: Schematic representation of the steps performed to produce a Molecular Dynamics simulation.[32]

Step 1 is the addition of the system's *Initial Coordinates.* As mentioned before in chapter 2.4, the initial configuration is usually known from crystallographic or NMR classical experiments, or when such data isn't available can also be constructed with techniques like homology modeling. For this project, initial structure for W2W11 was in fully extended conformation.

Step 2 comprises an *Energy Minimization of the Structure.* This process is done in order to remove strong van der Waals interactions that may lead to local distortion of the structure. At this point, explicit water molecules were added to the system using TIP3P water model, followed by another energy minimization so that the water molecules can adjust to the protein.

Steps 3 and 4 include the *Assignment of Initial Velocities* at low temperatures at the start of the simulation and the *Heating* performed throughout the simulation. As the structure evolves over time, new velocities are assigned at a slightly higher temperature and the simulation continues. This step is repeated several times until the desired temperature is reached. For this simulation the system was heated to an initial temperature of 300K

Step 5 begins once the desired temperature is reached. During this step the system undergoes a process of *Equilibration*; the simulation continues while structure, pressure, temperature and energy are monitored until they become stable overtime. If any significant temperature changes take place, the *Velocities are Rescaled* in order to return to equilibrium.

During the simulation process, electrostatics interactions were calculated using the Particle Mesh Ewald method (PME) and all bonds involving hydrogen atoms were restrained using the SHAKE algorithm.

Step 6 is the final step of the simulation. This is the stage where further heating was applied with NAMD's adaptive tempering method through the Langevin thermostat; when the potential energy for a given structure is lower than the average energy calculated thus far, the temperature is lowered, and when the current energy is higher than the average energy, the temperature is raised. The method was first described by Zhang and Ma.[43] Adaptive tempering temperature range for this simulation was 300 – 500K. At this point we acquire a *Produced Result*, that is a simulation of the system's trajectory over several hundred ps to ns or more. Our simulation produced 5.868.600 frames through runtime of 4.69μs.

Results

Analysis of the trajectory was performed using the program *CARMA*[44] and specifically through its graphic interface GRCARMA[45]. The files required for the analysis are a DCD file and a PSF file. The .psf file (protein structure file) contains structural information and can be generated by the user with psfgen, VMD, X-PLOR etc. using the initial pdb and topology files. The columns of information included in this file indicate values for each atom's name, type, ID, charge and mass, as well as the name and ID of the residue to which it belongs. The .dcd file is the trajectory file that was produced by the simulation; it consists of all the sets of atomic coordinates, each set corresponding to one frame.

To study the folding properties of mutant peptide W2W11 and test the stabilizing potency of the addition of a Trp-Trp pair in the hydrogen-bonded site, we performed the following analyses: fraction of native contacts vs adaptive tempering temperatures diagram, root mean square deviation matrix analysis (RMSD), secondary structure analysis and principal components analysis, as well as NOE averaged distances and chemical shifts for comparison with the NMR experimental findings.

4.1 Fraction of native contacts

Native contacts are non-sequential residues that interact and are able to guide the folding process of a protein. Native contacts play a major role in the determination of a peptide's folding mechanism during a simulation [50]. The fraction of native contacts (Q) is used to measure a simulated structure's deviation from the native state [51]. Values range between zero and 1.0; numbers closer to 1.0 indicate a structural conformation similar to that of the native state of a protein, while numbers closer to zero indicate disordered conformations. $Q - T$ diagrams depict the Q values throughout a simulation for which adaptive tempering was used, where T represents the different temperatures. In such diagram, yellow areas indicate low populations of conformations, while red areas indicate highly populated conformations.

GRCARMA was used to calculate the Q values from W2W11's simulation, using only CA atoms and selecting as reference the .pdb file of the native state. This file contains 20 experimentally determined structures. Backbone conformations of these structures are similar and resulted in diagrams with insignificant differences. The one shown below corresponds to the 3rd structure in said file. The $Q - T$ grid was created with script *2matrix* (can be found on Norma) using the Q values and the temperatures used for the simulation's adaptive tempering process. The grid was then plotted and the result is shown below (Figure 15) [52]. We can see a vertical line distribution at low Q values indicating a preference for disordered states throughout the simulation, and that temperatures lower than 320K are highly populated.

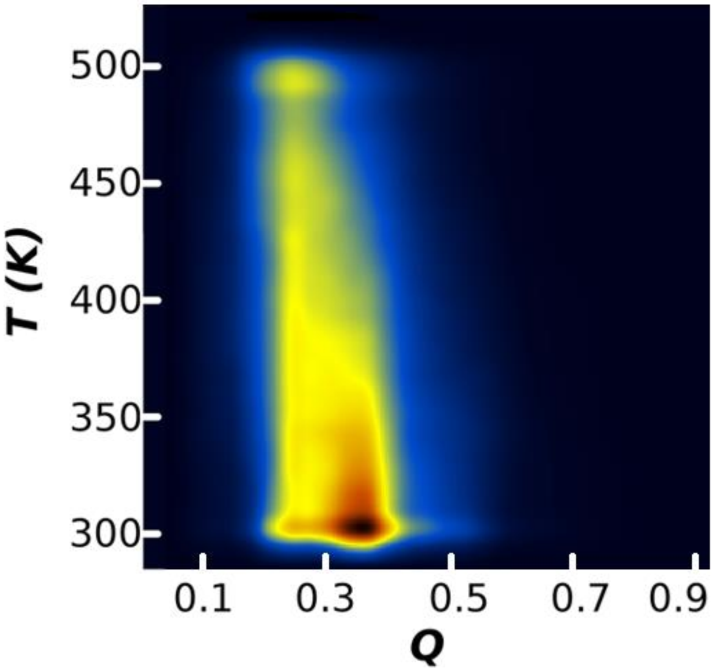


Figure 15: $Q - T$ diagram. Yellow indicates low populations. Red indicates higher populations. Q values range between zero and 1.0; numbers closer to 1.0 indicate a structural conformation similar to that of the native state, while numbers closer to zero indicate disordered conformations.

Figure 16 shows a comparison between $Q - T$ diagrams derived from molecular dynamics simulations for the Native loop 3 peptide, W3W10 mutant peptide and W2W11 mutant peptide. There exists an evident similarity between the behavior of the Native loop 3 peptide and W2W11 mutant peptide; both adopt disordered conformations (low Q distribution), while W3W10 mutant has a high Q – low T population and as such appears to have stability during its folding process.

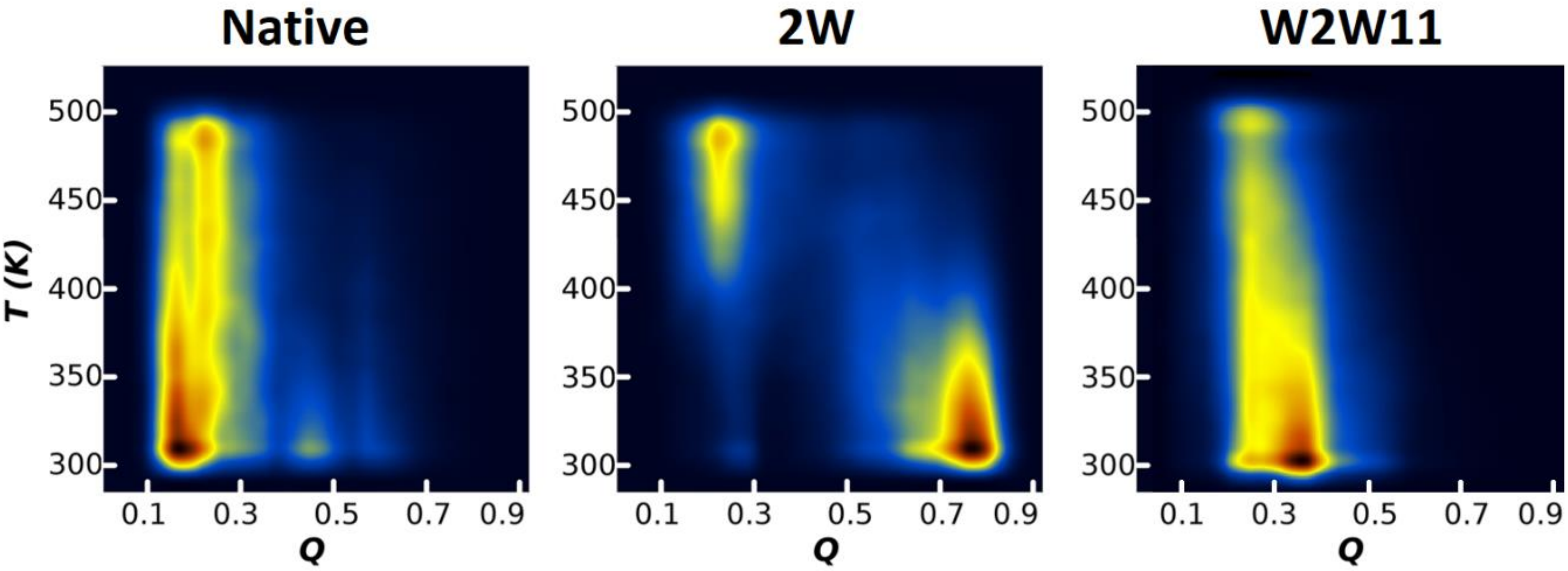


Figure 16: Comparison of Q – T diagrams acquired through molecular dynamics simulations for Native loop 3 peptide, 2W (W3W10) mutant peptide and W2W11 mutant peptide. Data for native loop 3 and 2W is the work of Koukos et al. [29]

4.2 Root Mean Square Deviation Matrix

Generation of a Root Mean Square Deviation (RMSD) matrix is used to calculate distances between atoms of superimposed structural conformations. The equation used for RMSD calculations is the following:

$$\mathbf{RMSD} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{N}}\sum(\mathbf{x_i} - \mathbf{x_{ref}})^{\mathbf{2}}}$$

where $\mathrm{x_i}$ is the atomic coordinates at a specific time, $\mathrm{x_{ref}}$ is the atomic coordinates of the reference conformation and N is the number of atoms.

Atomic distances are calculated in Å and are depicted in a color-coded way, with a range that varies from 0.00Å (dark blue), through yellow, up to 10.6Å (dark red). Blue color (shorter distances) corresponds to conformations that are generally preferred and maintain stability. Red color (bigger distances) corresponds to unstable conformations. Yellow corresponds to random medium states. The main diagonal is depicted as a black line and corresponds to null values, since those are the RMSD values for each conformation superimposed with itself. Blue regions on the diagonal indicate stable conformations of the structure for a period of time proportional to the length of the region, while blue regions away from the diagonal indicate similar structures that appeared at different times during the simulation. The following RMSD matrix for mutant peptide W2W11 was calculated with GRCARMA using Ca atoms of all residues (bottom) and heavy atoms (top).

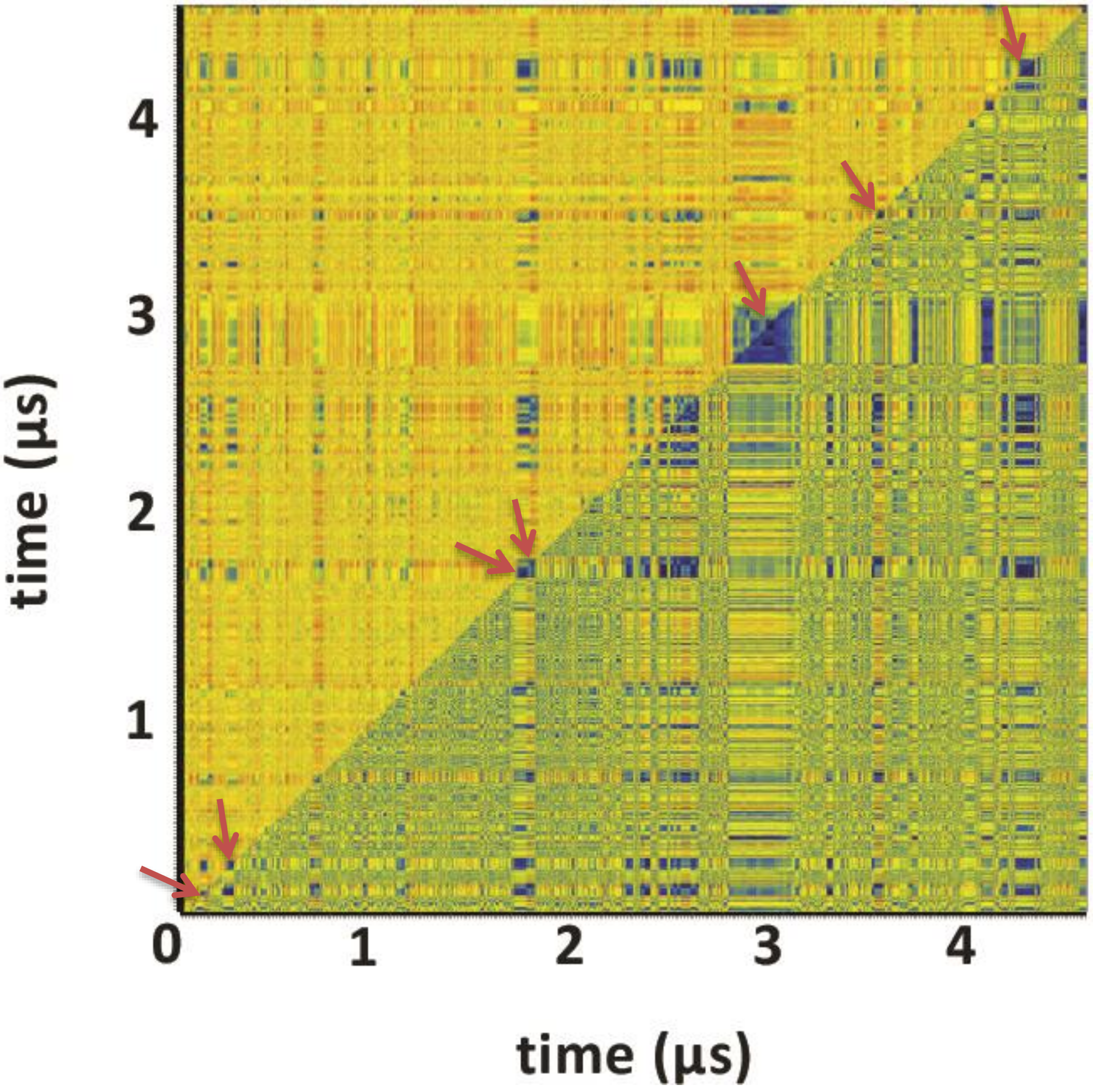


Figure 17: RMSD matrix of 5.868.600 frames. The color gradient for the postscript image ranges from 0.00Å (dark blue), through yellow, to 10.6Å (dark red). Bottom half represents RMSD for backbone atoms and top half represents RMSD for heavy atoms. Red arrows indicate stable conformations.

The RMSD matrix shows that the mutant peptide is mostly unstable throughout the simulation (big number of yellow regions), taking mainly random conformations. There exist only a few small periods of time (short blue regions) where the peptide prefers more stable conformations. Some examples are at 0.1μs, 0.25μ, 1.7μs, 1.8μs, 2.9μs – 3μs, 3.6μs and 4.3μs (red arrows), while for 2.4 μs – 2.7μs we can see the peptide rapidly changing between stable and unstable state, visiting several of the different stable conformations. The following conformations correspond to the timestamps mentioned above. The images were produced by RasMol.[49]

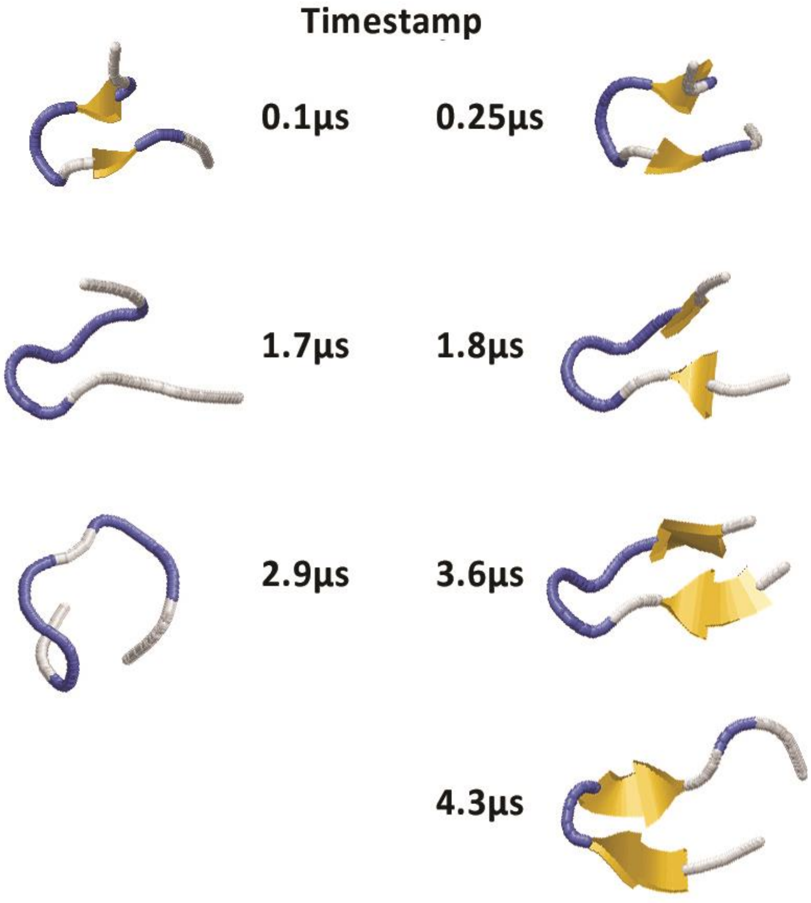


Figure 18: More stable conformations that W2W11 mutant peptide takes throughout the simulation of 4.69μs according to the RMSD matrix. Structure conformation is color coded: Blue corresponds to turn, yellow to β-sheet and white to random coil.

4.3 Secondary Structure Analysis

Schematics for the peptide’s secondary structures can be calculated with the secondary structure analysis. Studying the secondary structure schematic can give information about how the peptide behaves and folds throughout the simulation process. GRCARMA uses STRIDE (STRuctural IDEntification) [47] to produce a text file with secondary structure information for frames according to indicated step, as well as a schematic representation of this information using a color-coded depiction. The following is the schematic produced for our analysis with step set to 200 frames.

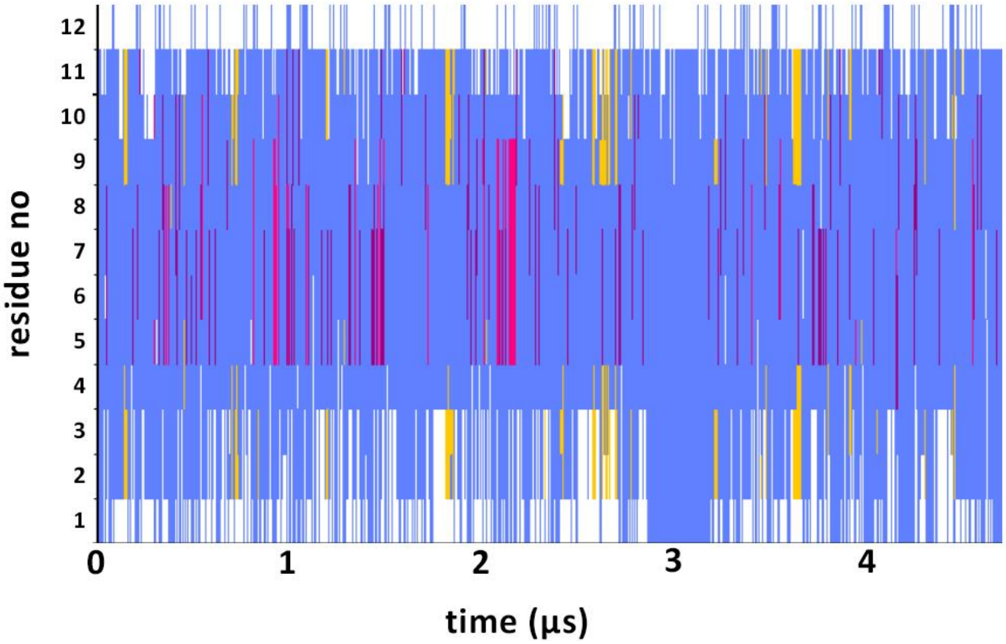
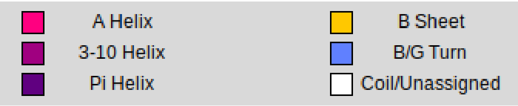


Figure 19: STRIDE schematic for W2W11 mutant peptide throughout a simulation of 4.69μs. Colors are assigned according to the following table:



🟥	A Helix	🟨	B Sheet
🟪	3-10 Helix	🟦	B/G Turn
🟪	Pi Helix	□	Coil/Unassigned

In addition to STRIDE, GRCARMA uses WebLogo [48] to produce a graphical representation of the secondary structure for the 12 residues sequence. The structure is depicted with letters indicating different conformations preferred by each residue stacked together. A preferred structural conformation that appears more frequently is indicated by a larger letter stacked at the top. This is the WebLogo produced for simulation of W2W11 mutant peptide.

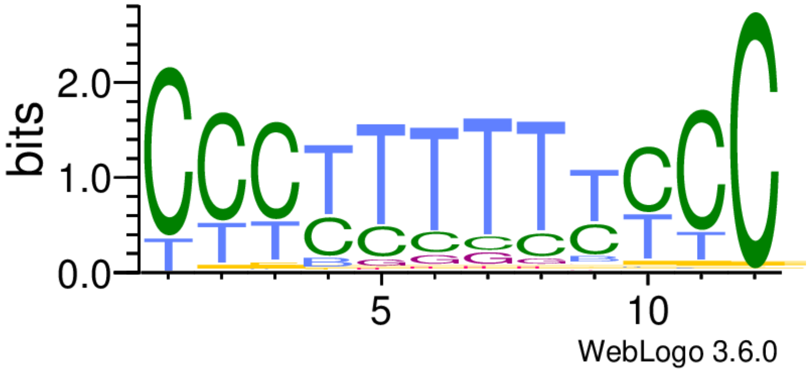


Figure 20: WebLogo graphical representation of W2W11 mutant peptide. The letters indicate: H: α-helix, G: 3_{10} helix, I: π-helix, E: β-sheet, B: β-bridge, T: turn, C: random coil/unassigned.

Both STRIDE (Figure 19) and WebLogo (Figure 20) results agree that residues 4-9 prefer adopting the conformation of a turn, while residues 1-3 and 10-12 behave mainly as random coils that occasionally come together and interact creating a β-sheet structure and allowing the peptide to take more stable structural conformations (fairly similar to the β-hairpin). These stable conformations can be seen in figure 18.

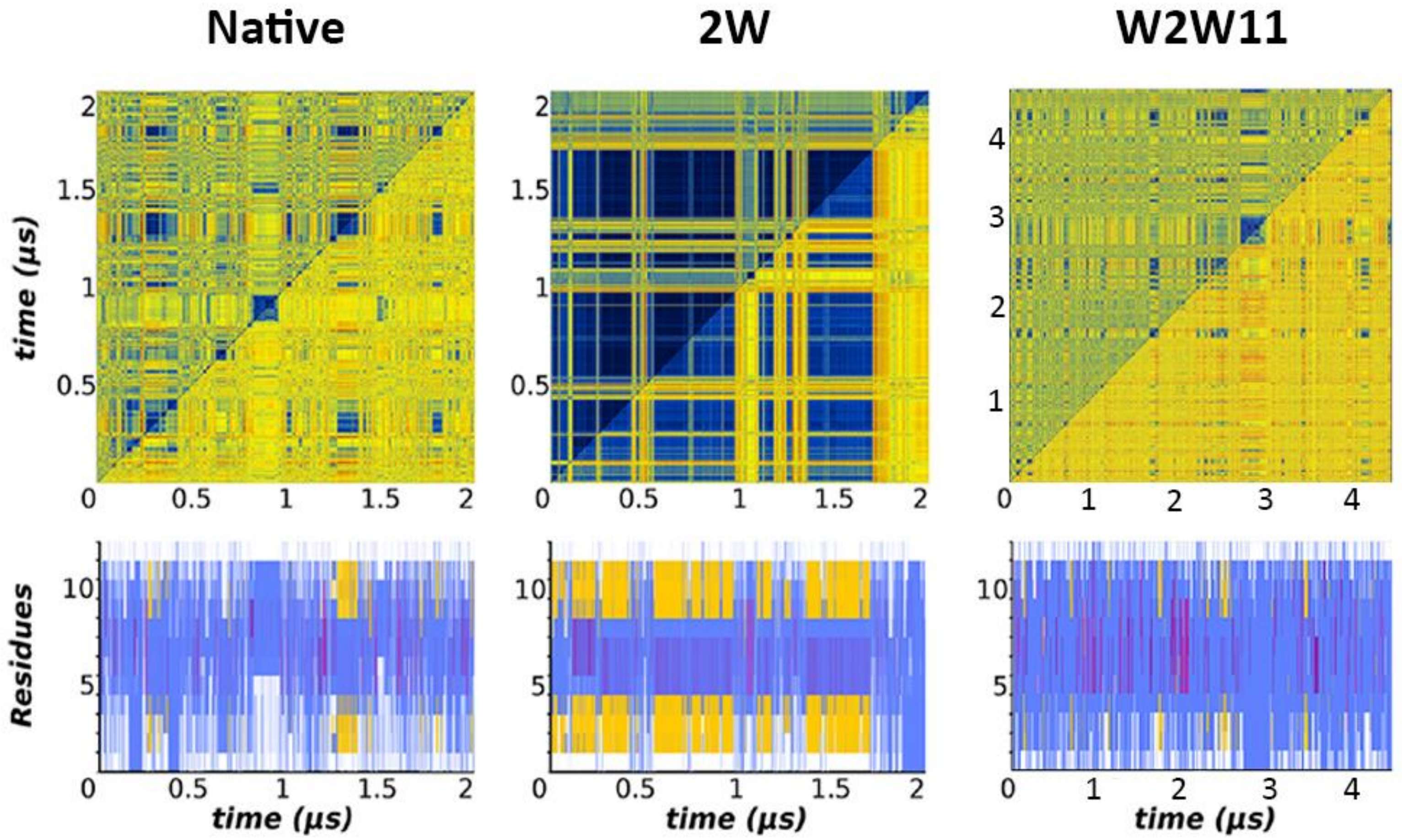


Figure 21: Comparison of RMSD matrices and STRIDE schematics acquired through molecular dynamics simulations for Native loop 3 peptide, 2W (W3W10) mutant peptide and W2W11 mutant peptide. Data for native loop 3 and 2W is the work of Koukos et al. [29]

4.4Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique for creating groups (clusters) of data according to patterns of similarities and differences from large amounts of information. PCA is used for complex systems with high dimensionality; it helps reduce the number of dimensions without severe loss of information.[52]

In Molecular Dynamics simulations, PCA (also called quasiharmonic analysis or essential dynamics method) is a very useful tool because of its ability to filter observed motions. It creates clusters using a covariance matrix or a correlation matrix (C-matrix) constructed from system variables that describe the accessible degrees of freedom (DOF) of the protein. An eigenvalue decomposition of the C-matrix gives a set of eigenvectors, each with a corresponding eigenvalue that characterizes a portion of the motion. When the original data is projected onto an eigenvector, the result is called a principal component (PC). Ultimately, the description of protein dynamics can be done in terms of only a few principal components.

There are mainly two variables for describing protein motion that are used in Molecular Dynamics Principal Components Analysis: Cartesian coordinates and internal coordinates. Cartesian PCA (cPCA) uses atomic coordinates in Cartesian space, but sometimes mixes internal and overall motion and creates artifacts. On the other hand, use of internal coordinates can separate internal and overall dynamics correctly. Dihedral PCA (dPCA) uses internal dihedral angles (ϕ, ψ) of the backbone. [53][54] Combining both methods can increase the overall accuracy of the results especially for complicated trajectories.

For this project, dihedral PCA was performed through CARMA using GRCARMA's interface [44][45]. The program was set to calculate 5 principal components and isolate up to a maximum of 10 clusters. A total of 5 clusters were produced. Temperature was set to 298K. Following figure is a schematic representation of the 5 clusters produced by dPCA.

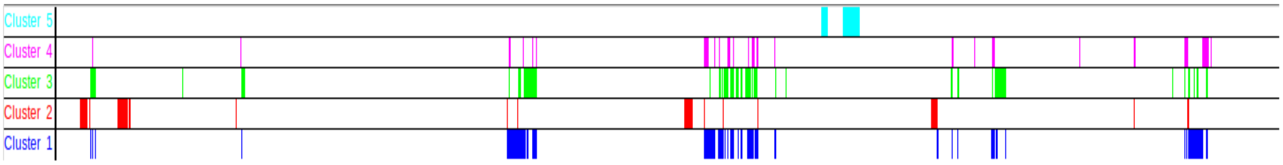


Figure 22: The 5 clusters that were produced by dPCA for W2W11 mutant peptide simulation trajectory of 5.868.600 frames. Blue: cluster 1, red: cluster 2, green: cluster 3, purple: cluster 4 and cyan: cluster 5.

Following table consists of the number of frames that were isolated in each cluster:

Cluster No	Frames (out of 5.868.600)
1	167.689
2	56.267
3	67.332
4	12.468
5	22.351

A total of 326.107 out of 5.868.600 frames were isolated into clusters, about 5.6% of the trajectory. Percentages that correspond to each cluster are:

2.9% for cluster 1 (the most populated cluster),

1% for cluster 2,

1.1% for cluster 3,

0.2% for cluster 4,

0.4% for cluster 5.

Superimposed and representative structures for each cluster were also produced by GRCARMA and presented via RasMol [49] in the figures below.

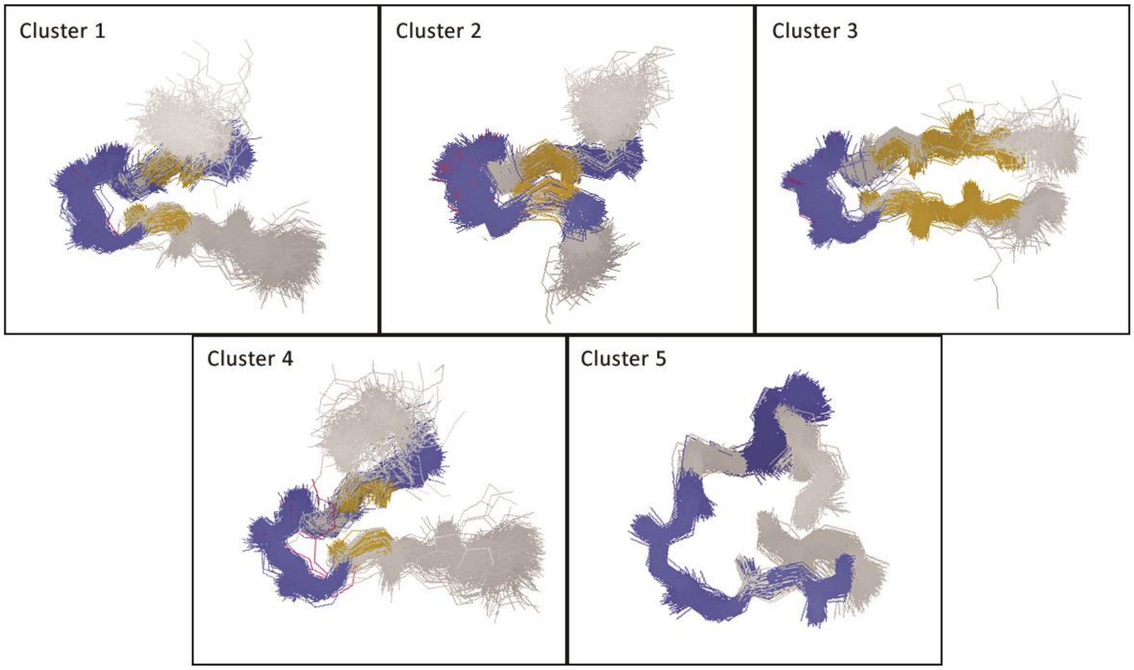


Figure 23: Superimposed structures for the 5 clusters produced by grcarma and represented with RasMol. Structural conformation is color coded: blue color represents turn, yellow represents β-sheet and white represents random coil.

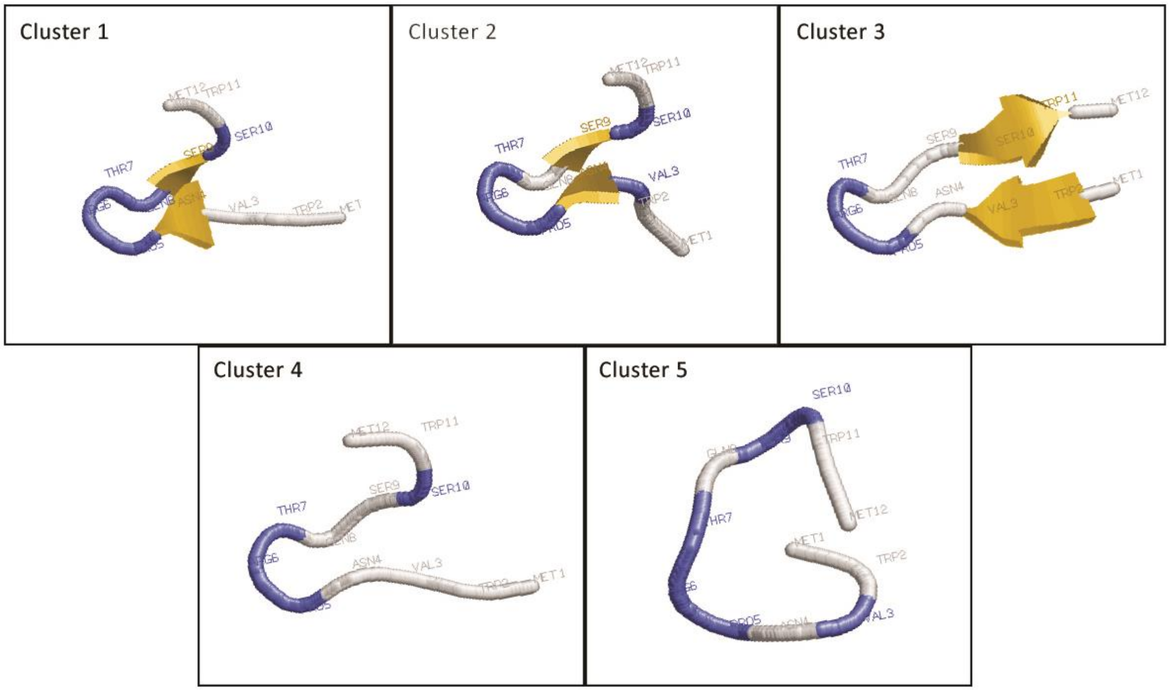


Figure 24: Representative structures for the 5 clusters produced by grcarma and represented with RasMol. Structural conformation is color coded: blue color represents turn, yellow represents β-sheet and white represents random coil. Residues are labeled.

It is already apparent, like shown before through the RMSD matrix, that the peptide takes mostly random unstable conformations throughout the simulation (94.4% of the trajectory), and only 5.6% of the trajectory contains stable structural conformations. Figures 23 and 24 show superimposed and representative stable conformations adopted by the peptide for clusters 1 to 5. For clusters 1 to 4, residues 5, 6 and 7 (PRO, ARG, THR) successfully create a well-defined turn. The structures that correspond to the most populated group (Cluster 1) and the third most populated one (Cluster 3) have ASN 4 and SER 9 close together, creating a β-sheet conformation, while the N- and C- ends are extended in random coil conformations and do not seem to interact. The second most populated group (Cluster 3) contains structures that have a well-defined β-hairpin conformation. Clusters 4 and 5, the least populated ones, contain conformations of mostly turns and random coils. Especially the conformations in Cluster 5 have multiple turns (residues 3, 5-7, 9-10) creating an almost circular structure. These secondary structures adopted by the peptide can be seen clearly in the following STRIDE schematics:

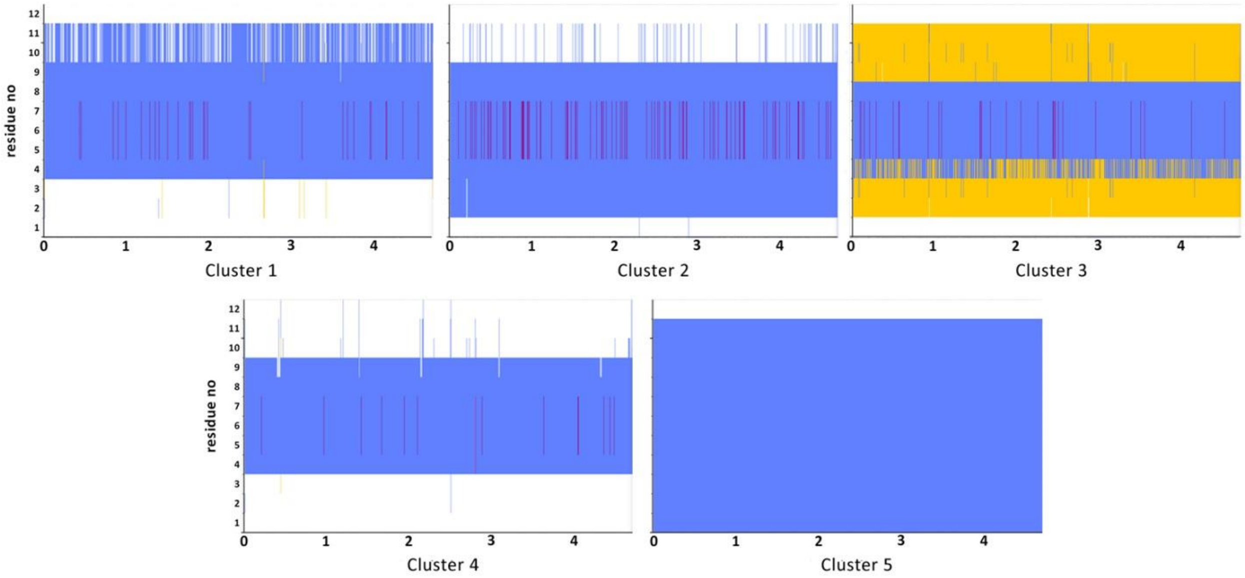


Figure 25: STRIDE schematics for the representation of secondary structures for the 5 clusters produced by grcarma. Colors are assigned according to the following table:

🟥	A Helix	🟨	B Sheet
🟪	3-10 Helix	🟦	B/G Turn
🟪	Pi Helix	□	Coil/Unassigned

In order to have a complete picture of the Molecular Dynamics simulation results for W2W11 mutant peptide's structural study, comparison with those taken from Santiveri et al NMR classical experiments [28] is necessary. NMR, short for Nuclear Magnetic Resonance, as mentioned already in chapter 1.3, is an experimental technique used to study molecule structures according to the spin properties of their atoms' nuclei, and specifically how these atoms resonate if an external magnetic field is applied. There exist several phenomena that occur because of the nuclei's spin. Such phenomena are the Nuclear Overhauser Effect (NOE), Chemical shifts or J-couplings. Hydrogen (^{1}H) is the atom generally preferred for NMR studies because of the simplicity of its nucleus (1 proton), though Carbon (^{13}C) is also frequently used. Both are atoms abundant in all molecules. For this project, ^{1}H NOEs and Chemical shifts were calculated for the simulated trajectory and compared with the experimental results.

4.5 NOE averaged distances

The Nuclear Overhauser Effect (NOE) was discovered by Albert Overhauser in 1953. The NOE process can be summarized as a cross relaxation from one spin state to another spin state; when a specific nucleus is magnetically excited and its neighbor is at equilibrium, relaxation occurs between the two nuclei. The dipolar interaction that happens causes the neighbor nucleus' spin to intensify. These interactions are the direct magnetic coupling (the dipolar coupling) between the two nuclei and there is a strong $1/r^6$ distance dependence of the relaxation; NOE signals become insignificant when distance between nuclei increases. In other words, NOE happens through spatial distance and sequential distance does not matter. For that reason, studying them allows us to calculate intermolecular distances and molecular motion. [55][57]

The NOE signal can be expressed through the following equation:

$$\mathbf{NOE\ =\ 1/r_6\,f(t_c)}$$

where r is the distance between the two nuclei and t_c is the time required for a full rotation of 1 rad.

What we receive is an NOE spectrum (NOE Spectroscopy, NOESY), and from it we can extract a list of the distance restraints between the nuclei pairs; NOE enhancement is dependent on the $1/r^6$ distance.

In order to compare simulation with experiment, r^{-6} averaging of the internuclear distance is used to obtain predicted average distances from the simulation, which are then compared with the distance restraints derived from NMR experiment. The process was done as follows:

A list with all possible proton pairs based on the protein structure file for W2W11, (PSF) which contains atomic information, was created using perl script *prep_proton.pl*. The list was edited to remove redundant information. Ultimately, only the pairs that were studied in the experiment were included, a total of 20 pairs. C script *noe_averaging* was then used to calculate r^{-6} distances for listed pairs. Three sets of calculations were conducted; one for all frames of the simulation (table 1), one for the frames that correspond to temperatures lower than 320K (table 2), and one for frames that correspond to temperatures lower than 300K (table 3).[56]

In order to compare these results with the NMR-derived NOEs, calculations of upper bound violations were also conducted. In NMR experiments, lower bound for NOE signals is typically set equal to the closest possible distance (van der Waals contact or approximately 2 Å), while the upper bound is set to 3, 4, 5 or 6 Å depending on whether the NOE is classified as strong, medium or weak. This classification is based on the first NMR protein structure determination done by Williamson, Havel and Wüthrich in 1985. Violation occurs when the r^{-6} value is greater than the experimental NOE upper bound value, and is calculated through the following equation:

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

where $v(i,j)$ is the upper bound violation for proton pair i/j and $nmr(i,j)$ is the experimental upper bound value.

NMR-derived NOEs (at temperature of 278K), as well as r^{-6} values calculated from the simulation are presented in the tables below. The Q stands for pseudo atoms; certain atoms having been joined together to create virtual atoms in order to calculate the structure. QD represents δ2-amido H atoms, QB represents β-methylene H atoms and QG represents γ-methylene H atoms.

Residues i/j	Proton i	Proton j	NMR upper bound values & classification	r^{-6} (all) & classification	Upper bound violation
Trp 2 / Asn 4	$C_{\varepsilon1}H$	$C_{\alpha}H$	5,50 **W**	4,868130 **M**	-
	$C_{\zeta3}$	$C_{\alpha}H$	5,50 **W**	4,227607 **M**	-
	$C_{\zeta3}$	$C_{\beta}H$	5,85 **W**	4,718050 **M**	-
	$C_{\zeta2}$	$C_{\alpha}H$	4,38 **M**	4,050517 **M**	-
	$C_{\eta2}$	$C_{\alpha}H$	4,76 **M**	3,914567 **S**	-
	$C_{\eta2}$	$C_{\beta}H$	6,32 **W**	4,988475 **M**	-
Trp 2 / Pro 5	$C_{\zeta2}$	$C_{\gamma\gamma'}H$	6,38 **W**	4,515324 **M**	-
	$C_{\zeta2}$	$C_{\delta\delta'}H$	5,35 **M**	4,185821 **M**	-
Val 3 / Pro 5	$C_{\gamma}H_3$	$C_{\alpha}H$	8,09 **W**	5,772530 **W**	-
	$C_{\gamma}H_3$	$C_{\beta'}H$	8,13 **W**	8,229871 **W**	0,099871
Asn 4 / Arg 6	QD2	QB	6,31 **W**	4,103769 **M**	-
Asn 4 / Thr 7	$N_{\delta'}H$	$C_{\gamma}H_3$	7,40 **W**	4,330974 **M**	-
	$C_{\beta'}H$	HN	5,14 **W**	3,728186 **S**	-
Asn 4 / Gln 8	HN	QB	3,80 **S**	4,454235 **M**	0,654235
	HN	QG	6,38 **S**	5,366504 **W**	-
Arg 6 / Gln 8	HN	QB	3,80 **S**	6,495701 **W**	2,695701
	HN	QG	6,38 **S**	6,180353 **W**	-

Table 1: NOEs for all frames of the simulation. Columns indicate (left to right): Residue pair (i/j), proton of residue i, proton of residue j, upper bound distance restraints derived from NMR and used for structure calculation of peptide W2W11 and experimental classification, r^{-6} for all frames of the simulation and classification and upper bound violation values.

Residues i/j	Proton i	Proton j	NMR upper bound values & classification	r^{-6} (320K) & classification	Upper bound violation
Trp 2 / Asn 4	$C_{\varepsilon1}H$	$C_{\alpha}H$	5,50 **W**	4,710386 **M**	-
	$C_{\zeta3}$	$C_{\alpha}H$	5,50 **W**	3,862881 **S**	-
	$C_{\zeta3}$	$C_{\beta}H$	5,85 **W**	4,534957 **M**	-
	$C_{\zeta2}$	$C_{\alpha}H$	4,38 **M**	3,596455 **S**	-
	$C_{\eta2}$	$C_{\alpha}H$	4,76 **M**	3,450356 **S**	-
	$C_{\eta2}$	$C_{\beta}H$	6,32 **W**	4,592774 **M**	-
Trp 2 / Pro 5	$C_{\zeta2}$	$C_{\gamma\gamma'}H$	6,38 **W**	4,125062 **M**	-
	$C_{\zeta2}$	$C_{\delta\delta'}H$	5,35 **M**	3,726879 **S**	-
Val 3 / Pro 5	$C_{\gamma}H_3$	$C_{\alpha}H$	8,09 **W**	5,616470 **W**	-
	$C_{\gamma}H_3$	$C_{\beta'}H$	8,13 **W**	8,487907 **W**	0,357907
Asn 4 / Arg 6	QD2	QB	6,31 **W**	3,890221 **S**	-
Asn 4 / Thr 7	$N_{\delta'}H$	$C_{\gamma}H_3$	7,40 **W**	4,061490 **M**	-
	$C_{\beta'}H$	HN	5,14 **W**	3,491515 **S**	-
Asn 4 / Gln 8	HN	QB	3,80 **S**	4,463096 **M**	0,663096
	HN	QG	6,38 **S**	5,285152 **W**	-
Arg 6 / Gln 8	HN	QB	3,80 **S**	6,488513 **W**	2,688513
	HN	QG	6,38 **S**	6,058747 **W**	-

Table 2: NOEs for simulation frames that correspond to T less than 320K. Columns indicate (left to right): Residue pair (i/j), proton of residue i, proton of residue j, upper bound distance restraints derived from NMR and used for structure calculation of peptide W2W11 and experimental classification, r^{-6} for frames that correspond to T less than 320K and classification and upper bound violation values.

Residues i/j	Proton i	Proton j	NMR upper bound values & classification	r^{-6} (300K) & classification	Upper bound violation
Trp 2 / Asn 4	$C_{\varepsilon1}H$	$C_{\alpha}H$	5,50 W	4,716235 M	-
	$C_{\zeta3}$	$C_{\alpha}H$	5,50 W	3,836434 S	-
	$C_{\zeta3}$	$C_{\beta}H$	5,85 W	4,488259 M	-
	$C_{\zeta2}$	$C_{\alpha}H$	4,38 M	3,555507 S	-
	$C_{\eta2}$	$C_{\alpha}H$	4,76 M	3,408413 S	-
	$C_{\eta2}$	$C_{\beta}H$	6,32 W	4,535264 M	-
Trp 2 / Pro 5	$C_{\zeta2}$	$C_{\gamma\gamma'}H$	6,38 W	4,107531 M	-
	$C_{\zeta2}$	$C_{\delta\delta'}H$	5,35 M	3,696960 S	-
Val 3 / Pro 5	$C_{\gamma}H_3$	$C_{\alpha}H$	8,09 W	5,605573 W	-
	$C_{\gamma}H_3$	$C_{\beta'}H$	8,13 W	8,489692 W	0,359692
Asn 4 / Arg 6	QD2	QB	6,31 W	3,850298 S	-
Asn 4 / Thr 7	$N_{\delta'}H$	$C_{\gamma}H_3$	7,40 W	4,024441 M	-
	$C_{\beta'}H$	HN	5,14 W	3,478327 S	-
Asn 4 / Gln 8	HN	QB	3,80 S	4,485261 M	0,685261
	HN	QG	6,38 S	5,228597 W	-
Arg 6 / Gln 8	HN	QB	3,80 S	6,49785 W	2,69785
	HN	QG	6,38 S	6,030803 W	-

Table 3: NOEs for simulation frames that correspond to T less than 300K. Columns indicate (left to right): Residue pair (i/j), proton of residue i, proton of residue j, upper bound distance restraints derived from NMR and used for structure calculation of peptide W2W11 and experimental classification, r^{-6} for frames that correspond to T less than 300K and classification and upper bound violation values.

Great upper bound violation can specifically be seen for protons (i-j) HN-QB of pairs Asn 4 / Gln 8 and Arg 6 / Gln 8, and $C_{\gamma}H_{3}$-$C_{\beta'}H$ of pair Val 3 / Pro 5 in all cases. However no further upper bound violation can be seen in the rest of the results. It's apparent that results for the N- terminal region mostly come in agreement with the experiment, while differences can be seen for the C- terminal region.

Average values for upper bound violations were calculated:

Average upper bound violation for r^{-6} (all)	**Number of violations**
0.202	3
Average upper bound violation for r^{-6} (320K)	Number of violations
0.218	3
Average upper bound violation for r^{-6} (300K)	Number of violations
0.220	3

Upper bound violation average values in all three cases greatly exceed the threshold of 0,05 Å which indicates that the overall simulation results do not come in agreement with the experimental results.

4.6 Chemical shifts

As it was mentioned before, one of the phenomena that occurs because of the nuclei’s spin is called chemical shifts. When nuclei undergo a spin flip they absorb energy. This energy can be measured against the irradiation frequency in Hz. The proportional frequency change is on the order of a number of Hz versus an operating frequency in MHz. This ratio is expressed as a *chemical shift (δ)* from a standard compound's frequency in parts per million (ppm). Chemical shifts provide atomic distance information for the molecule being studied; acquiring backbone shifts is the perfect first step for deciphering a molecular structure.

For decades now chemical shifts have been considered the mileposts of NMR. They were observed in 1950 by Proctor and Yu based on ^{14}N NMR studies and in 1957 by Arnold et al on ^{1}H studies. Today they are one of the most reliable tools for determining biomolecular structures. Furthermore, they provide detailed information about hydrogen bonding interactions, ionization and oxidation states, the ring current influence of aromatic residues, or the nature of hydrogen exchange dynamics.[58]

In 1983 in his studies on peptide C_αH protons Dalgarno defined *secondary shifts* as the difference between the chemical shifts observed in the experiment and the chemical shifts that correspond to random coil conformations.[59] Secondary shifts for $C_{\alpha}H$ and NH protons are expressed through the following formulas:

$$\boldsymbol{\Delta\delta_{C\alpha H} = \delta_{C\alpha H}^{OBS} - \delta_{C\alpha H}^{RC}}$$

$$\boldsymbol{\Delta\delta_{NH} = \delta_{NH}^{OBS} - \delta_{NH}^{RC}}$$

where $\Delta\delta$ is the secondary shift, δ^{OBS} is the observed chemical shift and δ^{RC} is the random coil chemical shift.

Positive values correspond to upfield shifts that indicate α-helix conformations. Higher negative values correspond to large downfield shifts that indicate β-sheet conformations, while lower negative values correspond to smaller downfield shifts that indicate residues that are closer to the β-sheet edges.

In 1995 Wishart et al listed the random coil NMR chemical shifts for multiple nuclei of the protected linear hexapeptide Gly-Gly-X-Y-Gly-Gly (where X and Y are any of the 20 common amino acids). Among the different peptides that were measured, Gly-Gly-X-Ala-GIy-Gly provided internally consistent random coil chemical shift values for nuclei of residue X. [60] These were the values used for present project’s calculations.

Chemical shifts of the simulated peptide's nuclei were calculated using SPARTA+ [61] through perl script calc_shifts.pl [62]. Chemical shifts as well as secondary shifts for NH and C$_\alpha$H of the experiment and the simulation are presented in the following tables.

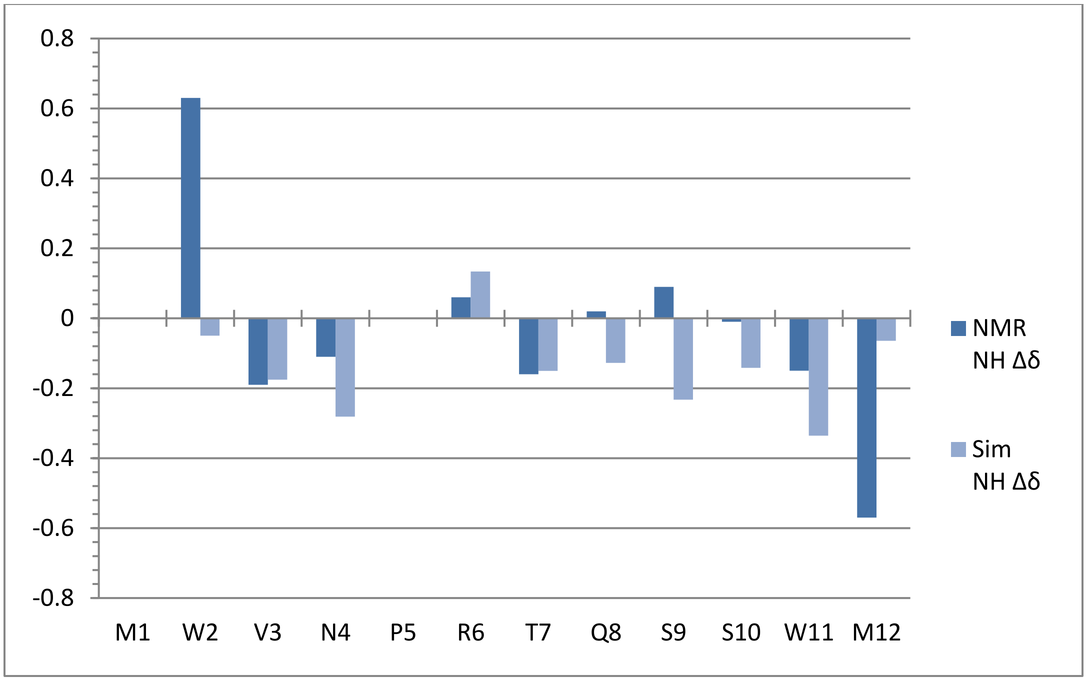
Residue No	Residue	NMR NH δ (ppm)	Simulation NH δ (ppm)	Random coil NH δ (ppm)	NMR NH Δδ	Simulation NH Δδ
1	**Met**					
2	**Trp**	**8,88**	**8,2004**	**8,25**	**0,63**	**-0,0496**
3	**Val**	**7,84**	**7,8543**	**8,03**	**-0,19**	**-0,1757**
4	**Asn**	**8,29**	**8,1190**	**8,40**	**-0,11**	**-0,2810**
5	**Pro**					
6	**Arg**	**8,29**	**8,3637**	**8,23**	**0,06**	**0,1337**
7	**Thr**	**7,99**	**7,9995**	**8,15**	**-0,16**	**-0,1505**
8	**Gln**	**8,34**	**8,1924**	**8,32**	**0,02**	**-0,1276**
9	**Ser**	**8,40**	**8,0771**	**8,31**	**0,09**	**-0,2329**
10	**Ser**	**8,30**	**8,1683**	**8,31**	**-0,01**	**-0,1417**
11	**Trp**	**8,10**	**7,9143**	**8,25**	**-0,15**	**-0,3357**
12	**Met**	**7,71**	**8,2155**	**8,28**	**-0,57**	**-0,0645**

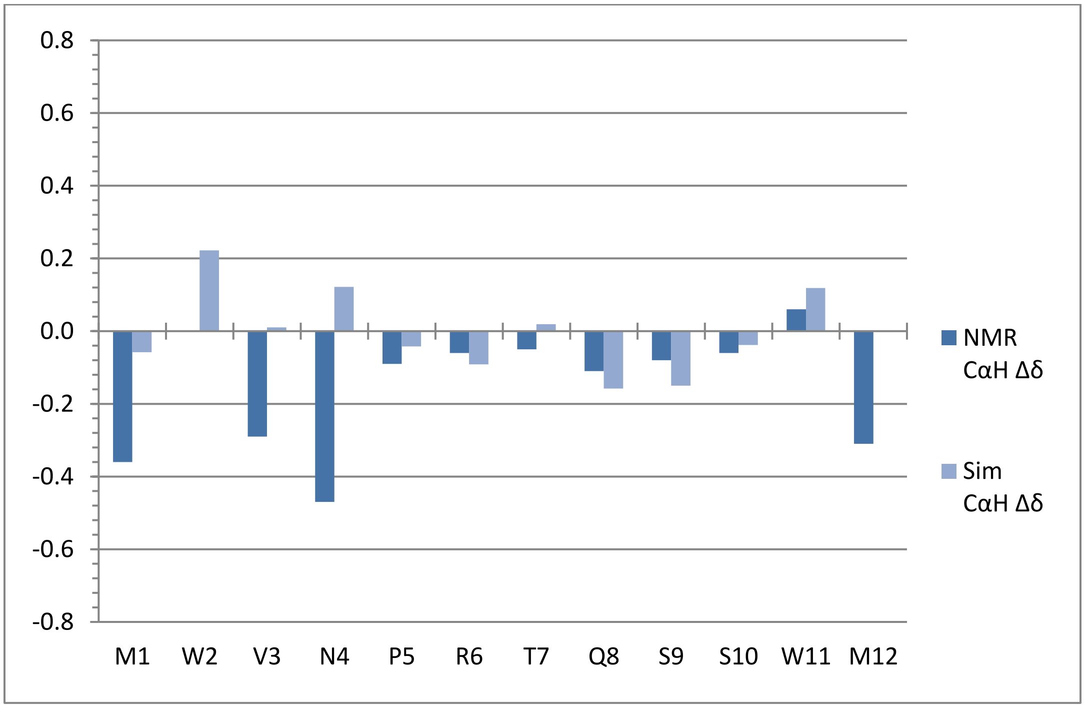
Table 4: NH chemical shifts. From left to right: residue number, residue, experiment shift value, simulation shift value, random coil shift value, experiment secondary shift, simulation secondary shift.

Residue No	Residue	NMR $C_{\alpha}H$ δ (ppm)	Simulation $C_{\alpha}H$ δ (ppm)	Random coil $C_{\alpha}H$ δ (ppm)	NMR $C_{\alpha}H$ $\Delta\delta$	Simulation $C_{\alpha}H$ $\Delta\delta$
1	**Met**	**4,12**	**4,4222**	**4,48**	**-0.36**	**-0.0578**
2	**Trp**	**4,66**	**4,8820**	**4,66**	**0.00**	**0.2220**
3	**Val**	**3,83**	**4,1303**	**4,12**	**-0.29**	**0.0103**
4	**Asn**	**4,27**	**4,8617**	**4,74**	**-0.47**	**0.1217**
5	**Pro**	**4,33**	**4,3778**	**4,42**	**-0.09**	**-0.0422**
6	**Arg**	**4,28**	**4,2486**	**4,34**	**-0.06**	**-0.0914**
7	**Thr**	**4,30**	**4,3690**	**4,35**	**-0.05**	**0.0190**
8	**Gln**	**4,23**	**4,1820**	**4,34**	**-0.11**	**-0.1580**
9	**Ser**	**4,39**	**4,3198**	**4,47**	**-0.08**	**-0.1502**
10	**Ser**	**4,41**	**4,4320**	**4,47**	**-0.06**	**-0.0380**
11	**Trp**	**4,72**	**4,7788**	**4,66**	**0.06**	**0.1188**
12	**Met**	**4,17**	**4,4779**	**4,48**	**-0.31**	**-0.0021**

Table 5: $C_{\alpha}H$ chemical shifts. From left to right: residue number, residue, experiment shift value, simulation shift value, random coil shift value, experiment secondary shift, simulation secondary shift.

Schematic diagrams for the comparison of experiment and simulation NH and $C_{\alpha}H$ secondary shifts were also created and depicted below.





Figures 26 and 27: Schematic diagrams for experiment (dark blue) and simulation (light blue) secondary shifts. Top: NH shifts. Bottom: $C_\alpha H$ shifts. Y axis indicates Δδ value and X axis indicates residues.

Certain similarities and differences can be seen between the chemical shifts calculated for the experiment and the simulation. More specifically:

In the list of NH chemical shifts, shift values for residues 3, 6 and 7 (Val, Arg, Thr) are significantly similar. However, the rest of the list presents differences in values. Average NH shift difference is 0.22 ppm.

In the list of $C_{\alpha}H$ chemical shifts, a fair similarity for shifts of residues 5 - 11 (Pro, Arg, Thr, Gln, Ser, Ser, Trp) can be seen, but that is not the case for residues 1 – 4 and 12 (Met, Trp, Val, Asn and Met). Average $C_{\alpha}H$ shift difference is 0.17 ppm.

These comparisons can also be seen in the secondary shift schematic diagrams. It should also be noted that the N- and C-terminal residues are affected by their charged ends.

In order to evaluate the results and measure the strength of the linear association between the experimental and the simulation derived chemical shifts, the Pearson correlation coefficient (PCC), also referred to as r, was calculated for $C_{\alpha}H$ shifts for the whole peptide, $C_{\alpha}H$ shifts excluding N- and C-terminal residues, NH shifts for the whole peptide, and NH shifts excluding N- and C-terminal residues.

The PCC between two variables (x and y) can be calculated with the following formula:

$$\boldsymbol{r_{xy}} = \frac{\sum_{i=1}^{n}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})(\boldsymbol{y_i} - \overline{\boldsymbol{y}})}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2}\sqrt{\sum_{i=1}^{n}(\boldsymbol{y_i} - \overline{\boldsymbol{y}})^2}}$$

where x_i and y_i are the first and second variables respectfully for position i, $\bar{x}$ is the mean of x variable and $\bar{y}$ is the mean of y variable.

Values of r range between 1 and -1, with 1 meaning perfect positive correlation and -1 meaning perfect negative correlation. The correlation becomes weaker when closer to zero.

The r values between experimental and simulation results for $C_\alpha H$ shifts and NH shifts are as follow:

	$C_\alpha H$ shifts (all)	**$C_\alpha H$ shifts excluding terminals**	**NH shifts (all)**	**NH shifts excluding terminals**
r	0.3957	0.6433	0.6719	0.7062

We can see a positive correlation for all results, although weaker for $C_\alpha H$ shifts for the whole peptide and moderate for the rest.

Discussion

The goal of present project was to use the technique of Molecular Dynamics Simulation to study the folding of W2W11, a Vammin-derived mutant peptide, and to compare the results with those taken from Santiveri et al NMR studies in order to examine the accuracy of Molecular Dynamics Simulations in determining molecular structures. According to the NMR results: "*W2W11 peptide does not fold into a native-like β-hairpin structure, but instead adopts a non-random structure involving residues 2–8. The ordered structure is not highly populated and coexists with random coil conformations*". Analyses that were performed on the simulation trajectory were: $Q - T$ diagram, RMSD matrix, Secondary Structure and Principal Components Analysis, as well as simulation derived NOEs and Chemical shifts. A significant level of agreement can be seen in most of these results.

The $Q - T$ diagram for W2W11, in relation to the 3rd conformation of the experimentally known structures .pdb file, showed a vertical line distribution at low Q values, something that indicates that the mutant peptide adopts disordered conformations throughout the simulation. This result is very similar to that derived from the simulation of the isolated loop 3 peptide, which also shows low Q values. This is the first indication that W2W11 does not fold into a stable native-like β-hairpin conformation.

The RMSD matrix showed mostly medium values (yellow color) with a few blue regions. These results indicate a dynamic behavior where the peptide is mostly unstable throughout the simulation adopting random coil conformations. However there exist several periods of time when the peptide assumes stable non-random conformations.

Secondary structure analysis showed a preference to blue and white colors in the STRIDE schematic, something that indicates that turn and coil conformations dominate the structure. Specifically, the N- and C- ends (residues 1 – 3 and 10 – 12) are mostly random coils, with a few yellow regions indicating occasional β-sheet conformations. The center-most residues 4 – 9 show high preference for turn. These results indicate a dynamic behavior. It is apparent that the mutant peptide successfully forms the turn that is also adopted by the native protein's loop 3. However, the Trp side chains fail to stabilize the peptide ends into a β-sheet conformation, since those residues seem to highly prefer to extend into random coils. WebLogo schematic agrees with these results.

Principal Components Analysis was performed in order to study the trajectory as a set of clusters of data with similar information. The clusters derived from the PCA corresponded to only 5.6% of the trajectory, indicating that 94.4% of the trajectory showed an unstable behavior. A total of five clusters were calculated, with only one of them that corresponded to 1.1% of the trajectory showing a β-hairpin conformation, with Trp 2, Val 3, Ser 10 and

Trp 11 creating a β-sheet and Pro 5, Arg 6 and Thr 7 forming the turn. The rest of the conformations in the four remaining clusters were non-random stable structures forming mostly turns and coils, with no interaction between the Trp pair. These results come in agreement with the experimental findings.

NOE averaged distances analysis showed upper bound violation for protons (i-j) HN-QB of pairs Asn 4 / Gln 8 and Arg 6 / Gln 8, and $C_{\gamma}H_3$-$C_{\beta'}H$ of pair Val 3 / Pro 5, but no further violation for the rest of the results. Results concerning the N- terminal region seem to mostly agree with the experiment, while differences can be seen for the C- terminal region. Upper bound violation average values were also calculated and greatly exceed the threshold of 0,05 Å, something that indicates that the overall simulation NOE results do not come in agreement with the experimental results.

Chemical shifts analysis showed NH chemical shift values for residues 3, 6 and 7 (Val, Arg, Thr) between simulation and experiment to be fairly similar, but also showed differences for the rest of the values. $C_{\alpha}H$ chemical shifts gave similarities for residues 5 - 11 (Pro, Arg, Thr, Gln, Ser, Ser, Trp), which isn't the case for residues 1 – 4 and 12 (Met, Trp, Val, Asn and Met). Some of these differences are due to that fact that the N- and C-terminal residues are affected by their charged ends. Secondary shift diagrams gave a schematic representation of these observations. Calculation of the Pearson correlation coefficient to measure the association between the experimental and the simulation derived chemical shifts resulted in positive values for all pairs; weaker for $C_{\alpha}H$ shifts for the whole peptide and moderate for $C_{\alpha}H$ shifts excluding N- and C-terminal residues, NH shifts for the whole peptide, and NH shifts excluding N- and C-terminal residues. This indicates a significant level of agreement between experiment and simulation.

To conclude, the insertion of a Trp-Trp pair in a hydrogen-bonded site does not seem to be able to guide W2W11 vammin-derived mutant peptide through a stable native-like folding process. However, the main goal of this project was to examine the accuracy of Molecular Dynamics Simulations at determining molecular structures. The simulation performed here, or rather the parameters used for this simulation, managed to accurately predict the mostly disordered dynamic behavior of W2W11 as previously shown through NMR experiments, something that proves once more that this revolutionary technique is an excellent tool in the hands of anyone who wishes to study molecular structures and folding behaviors.

Literature

[1] David L. Nelson, Michael M. Cox. Lehninger Principles of Biochemistry 4th Edition. New York: W. H. Freeman and Company, 2005

[2] Berg JM, Tymoczko JL, Stryer L. Biochemistry 5th edition. New York: W H Freeman, 2002.

[3] Ibraheem Rehman, Salome Botelho. Biochemistry, Secondary Protein Structure. Treasure Island (FL): StatPearls Publishing, 2019

[4] Ibraheem Rehman, Salome Botelho. Biochemistry, Tertiary Structure, Protein. Treasure Island (FL): StatPearls Publishing, 2019

[5] Nicole Kresge, Robert D. Simoni, Robert L. Hill. The Thermodynamic Hypothesis of Protein Folding: the Work of Christian Anfinsen. *The Journal of Biological Chemistry,* 281, e11. 2006

[6] Edgar Haber, Christian B. Anfinsen. Side-chain Interactions Governing the Pairing of Half-cystine Residues in Ribonuclease. The Journal of Biological Chemistry, 237:1839-1844. 1962

[7] Ken A. Dill, S. Banu Ozkan, M. Scott Shell, Thomas R. Weikl. The Protein Folding Problem. Annu Rev Biophys. 37:289–316. 2008

[8] Unnati Ahluwalia, Nidhi Katyal, Shashank Deep. Models of Protein Folding. JOURNAL OF PROTEINS AND PROTEOMICS. 3(2):85-93, 2012

[9] Ptitsyn O. B. Stages in the mechanism of self-organization of protein molecules. Dokl Akad NaukSSSR 210:1213-1215. 1973

[10] So much more to know... Science. Vol. 309, Issue 5731, pp. 78-102. 2005

[11] ChemTube3D. Protein Folding Kinetics. University of Liverpool
http://www.chemtube3d.com

[12] Martin Karplus, David L. Weaver. Protein folding dynamics: The diffusion-collision model and experimental data. Protein Science, Cambridge University Press. 3:650-668. 1994

[13] Go N. The consistency principle in protein structure and pathways of folding. Adv Biophys 18:149-164. 1984

[14] Michal Brylinski, Leszek Konieczny, Irena Roterman. Hydrophobic collapse in (*in silico*) protein folding. Computational Biology and Chemistry 30:255–267. 2006

[15] A R Fersht. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. PNAS 92 (24):10869-10873. 1995

[16] Adrian A Nickson, Jane Clarke. What lessons can be learned from studying the folding of homologous proteins? Methods. 52(1):38-50. 2010

[17] Joseph D. Bryngelson, Jose Nelson Onuchic, Nicholas D. Socci, and Peter G. Wolynes. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. PROTEINS: Structure, Function, and Genetics 21:167-195. 1995

[18] Ruth Nussinov, Chung-Jung Tsai. Free Energy Diagrams for Protein Function. Cell Press, Chemistry & Biology 21(3): 311-318. 2014

[19] Li-Quan Yang, Xing-Lai Ji, Shu-Qun Liu. The free energy landscape of protein folding and dynamics: a global view. Journal of Biomolecular Structure and Dynamics 31(9):982-992. 2013

[20] Emanuel Karl Peter. Enhanced Sampling Techniques for Protein Folding Simulations. Bunsen-Magazin 18(1). 2016

[21] Methods for Determining Atomic Structures. PDB-101 Education portal of RSCB PDB. https://pdb101.rcsb.org

[22] Sebastian Jaksch. Small-Angle Scattering. Cornell University. 2019

[23] William Reusch. Nuclear Magnetic Resonance Spectroscopy. Michigan State University. 2013

[24] Napoleone Ferrara, Terri Davis-Smyth. The Biology of Vascular Endothelial Growth Factor. Endocrine Reviews. Vol. 18, No. 1. Genentech Inc. 1997

[25] Kyoko Suto, Yasuo Yamazaki, Takashi Morita, Hiroshi Mizuno. Crystal Structures of Novel Vascular Endothelial Growth Factors (VEGF) from Snake Venoms. The Journal of Biological Chemistry. 280(3): 2126–2131. 2005

[26] Bruce A. Keyt, Hung V. Nguyen, Lea T. Berleau, Carlos M. Duarte, Jeanie Park, Helen Chen, Napoleone Ferrara. Identification of Vascular Endothelial Growth Factor Determinants for Binding KDR and FLT-1 Receptors. The Journal of Biological Chemistry. Vol. 271, No. 10, pp. 5638–5646, March 1996

[27] Yasmina Mirassou, Clara M. Santiveri, M. Jesús Pérez de Vega, Rosario González-Muñiz, M. Angeles Jiménez. Disulfide Bonds versus Trp···Trp Pairs in Irregular β-Hairpins: NMR Structure of Vammin Loop 3-Derived Peptides as a Case Study. ChemBioChem 10: 902-910. 2009

[28] Clara M. Santiveri, María Jesús Pérez de Vega, Rosario González-Muñiz, M. Angeles Jiménez. Trp-Trp pairs as β-hairpin stabilisers: Hydrogen-bonded versus non-hydrogen-bonded sites. Org. Biomol. Chem. 9, 5487. 2011

[29] Panagiotis I. Koukos, Nicholas M. Glykos. Folding Molecular Dynamics Simulations Accurately Predict the Effect of Mutations on the Stability and Structure of a Vammin-Derived Peptide. J. Phys. Chem. B 118, 10076−10084. 2014

[30] Martin Karplus, J. Andrew McCammon. Molecular dynamics simulations of biomolecules. Nature Structural Biology vol 9 no 9. 2002

[31] Tamar Schlick. Molecular Modeling and Simulation; An Interdisciplinary Guide 2nd edition. Springer Verlag, New York 2010

[32] Roland Stote, Annick Dejaegere, Dmitry Kuznetsov, Laurent Falquet. Theory of Molecular Dynamics Simulations. Version 1.0. 1999
https://embnet.vital-it.ch/MD_tutorial/

[33] Patricia Saenz-Méndez, Samuel Genheden, Anna Reymer, Leif A. Eriksson. Computational chemistry and molecular modeling basics. Computational Tools for Chemical Biology. 2017

[34] The Amber Project. http://ambermd.org/AmberModels.php

[35] CHARMM – Chemistry at HARvard Macromolecular Mechanics https://www.charmm.org/charmm/

[36] GROMACS – Fast. Flexible. Free. http://www.gromacs.org/Documentation/Terminology/Force_Fields/GROMOS

[37] William L. Jorgensen, David S. Maxwell, Julian Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. Journal of American Chemical Society. 118(45):11225-11236. 1996

[38] Generalized Born Implicit Solvent. https://www.ks.uiuc.edu/Research/namd/2.10b1/ug/node28.html

[39] Sikandar Y. Mashayak, David E. Tanner. Comparing Solvent Models for Molecular Dynamics of Protein. 2011

[40] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kale, Klaus Schulten. Scalable molecular dynamics with NAMD. Journal of Computational Chemistry. 26:1781-1802. 2005

[41] NAMD – Scalable Molecular Dynamics. https://www.ks.uiuc.edu/Research/namd/

[42] The Norma computing cluster. http://norma.mbg.duth.gr/

[43] Cheng Zhang, Jianpeng Ma. Enhanced sampling and applications in protein folding in explicit solvent. The Journal of Chemical Physics. 132(24): 244101. 2010

[44] N. M. Glykos. Carma: a molecular dynamics analysis program. *J. Comput. Chem.*, 27, 1765-1768. 2006

[45] P.I. Koukos, N. M. Glykos. grcarma: A Fully Automated Task-Oriented Interface for the Analysis of Molecular Dynamics Trajectories. *J. Comput. Chem.*, 34, 2310-2312. 2013

[46] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins. 15;65(3):712-25. 2006

[47] Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res. 32. (Web Server issue):W500-2. 2004

[48] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 14(6):1188-90. 2004

[49] Roger Sayle, E. James Milner-White. "RasMol: Biomolecular graphics for all". Trends in Biochemical Sciences (TIBS). 20(9): 374. 1995

[50] Robert B. Best, Gerhard Hummer, William A. Eaton. Native contacts determine protein folding mechanisms in atomistic simulations. PNAS. 110(44): 17874–17879. 2013

[51] Samuel S. Cho, Yaakov Levy, Peter G. Wolynes. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. PNAS. 103(3): 586 – 591. 2006

[52] Norma. Adaptive tempering: (T vs Q) and other diagrams. http://norma.mbg.duth.gr/

[53] Lindsay I. Smith. A tutorial on principal components analysis. Cornell University, USA. 2002

[54] Charles C. David, Donald J. Jacobs. Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. Methods Mol Biol. 1084: 193–226. 2014

[55] Alexandros Altis, Moritz Otten, Phuong H. Nguyen, Rainer Hegger, Gerhard Stock. Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. The Journal of Chemical Physics 128, 245102. 2008

[56] Nuclear Overhauser Effect. UCL Chemistry NMR Lectures. https://www.ucl.ac.uk/nmr/NMR_lecture_notes

[57] Norma. MD and NMR, calculation of (r)^-3 and (r)^-6 averaged distances. http://norma.mbg.duth.gr/

[58] Mike P. Williamson. Applications of the NOE in Molecular Biology. Annual Reports on NMR Spectroscopy. Vol 65. ISSN 0066-4103. 2009

[59] Steven P. Mielke, V.V. Krishnan. Characterization of protein secondary structure from NMR chemical shifts. Prog Nucl Magn Reson Spectrosc. 54(3-4): 141–165. 2009

[60] D. C. Dalgarno, B. A. Levine, R. J. P. Williams. Structural information from NMR secondary chemical shifts of peptide α C-H protons in proteins. Bioscience Reports 3, 443-452. 1983

[61] D. S. Wishart, C. G. Bigam, A. Holm, R. S. Hodges, B. D. Sykes. 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. Journal of Biomolecular NMR. 5(1): 67–81. 1995

[62] Yang Shen, Ad Bax. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. J Biomol NMR. 48(1): 13–22. 2010

[63] Norma. Calculation of chemical shifts from a molecular dynamics trajectory https://norma.mbg.duth.gr/