Conformation of Peptides in the Unfolded State:

A Coarse-grained Model Representation

by

Özge Engin

A Thesis Submitted to the

Graduate School of Engineering

in Partial Fulfillment of the Requirements for

the Degree of

Master of Science

in

Computational Science and Engineering

Koç University

September 2007

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Özge Engin

and have found that it is complete and satisfactory in all respects,

and that any and all revisions required by the final

examining committee have been made.

Committee Members:

_____

Burak Erman, Ph. D. (Advisor)

_____

Mehmet Sayar, Ph. D. (Co-advisor)

_____

Özlem Keskin, Ph. D.

_____

Atilla Gürsoy, Ph. D.

_____

Alkan Kabakçıoğlu, Ph. D.

Date: _____

To my grandfathers, Seyfettin and Basri

ABSTRACT

The native structure of proteins is stabilized by both local and non-local interactions. The available phase space is highly reduced in size due to the local interactions, but non-local interactions determine the final, physiologically active native structure. Although it is well-known that these two types of interactions are the key factors in determining the tertiary structure, their relative contributions is open for debate. This study will lay the groundwork for the investigation of relative contributions of these interactions. The contribution of both local and non-local neighbors to the Ramachandran map of the residue in question is examined by using statistical weight matrices (U) constructed according to the Markov assumption. An efficient matrix multiplication scheme based on rotational isomeric states model is introduced for studying realistic conformations of homotripeptides of all-alanine, tryptophan, valine, and tyrosine and AXA tripeptides, where X represents alanine, valine, tryptophan, and tyrosine in the unfolded state. This scheme is based on U's obtained from mono and dipeptide molecular dynamics simulations. By using these matrices one can obtain the Ramachandran map of the central residue of longer sequences, such as tripeptides. Comparison of explicit tripeptide simulations with the Markov model shows that the Markov assumption fails to capture interactions specific to the tripeptide. Here, a systematic correction is proposed for efficient calculation of realistic protein conformations. Preliminary results suggest that the Markov assumption can be improved significantly by adding the contributions from hydrogen bonds, which are only present in the tripeptide sequences. Such a coarse-grained model, Modified Markov model, will help elucidate the protein folding problem and improve secondary structure prediction algorithms.

ÖZET

Proteinler, doğal hallerine birincil yapılarında birbirine zincir üzerinde yakın ve yakın olmayan amino asitler arasında kurulan etkilesimler yardımıyla sahip olur. Yakın komşu etkileşimleri, proteinlerin bulunabileceği konformasyonların sayısını sınırlar; bununla beraber proteinlerin fizyolojik olarak aktif halde bulundukları doğal hallerine kavuşmaları birbirine yakın komşu olmayan amino asitler arasında kurulan etkileşimlerin de yardımıyla gerçekleşir. Proteinlerin üçüncül yapılarının belirlenmesinde, yukarda da belirtildiği gibi, birincil yapıda birbirine yakın ve yakın olmayan amino asitler arasındaki etkileşimler etkili olmaktadır; fakat kesin bir sekilde aydınlatılamayan nokta bu iki tip etkileşimin proteinlerin doğal hallerinin kazanılmasına hangi oranda katkı sağladığıdır. Bu çalışma, bu iki tip etkileşimin oranının aydınlatılmasına bir zemin hazırlayacaktır. Bu çalışmada, Markov varsayımına göre oluşturulan istatistiksel ağırlık matrisleri kullanılarak, bu iki etkileşimin, verilen bir peptit dizisinin ortasındaki amino asitin konformasyon tercihine ne ölçüde katkı sağladığı incelendi. Dönme açılarını değişik dönme izomerleri olarak gruplandıran (Rotational isomeric state model) bir modelin temel alınarak oluşturulduğu ve içinde bir dizi seri matris çarpımlarının yapıldığı bir yöntem kullanılarak, bütün amino asit bileşenleri alanine, valine,tryptophan ve tyrosine olan üçlü peptit dizileri ile birinci ve üçüncü bileşeni alanine, ortadaki bileşeni alanine, valine, trpytophan ve tyrosine olan AXA şeklindeki üçlü peptit dizilerinin ortasında bulunan amino asitin konformasyon tercihleri, uygun tekli ve ikili peptit dizilerine ait istatistiksel ağırlık matrisleri kullanılarak incelendi. Bu calışmada kullanılan tüm istatistiksel ağırlık matrisleri, moleküler dinamik simülasyonu ile elde edildi. Üçlü dizilerin simülasyonu sonucu elde edilen Ramachandran haritaları, Markov modelinden elde edilenlerle karşılaştırıldığında, Markov modelinin üçlü diziye ait etkileşimleri yakalamada başarılı olamadığı gözlemlendi. Bu amaçla, eksik etkileşim terimlerinin eklenmesine olanak sağlayacak sistematik bir yöntem geliştirildi. Elde edilen

sonuçlar, Markov modelinin, üçlü diziye ait hidrojen bağ etkileşimlerinin de eklenmesiyle, üçlü diziye ait etkileşimleri de büyük ölçüde yakalayabildiğini gösterdi. Bu şekilde tasarlanan kabataslak bir model, protein katlanma probleminin aydınlatılmasına ve ikincil yapı tayin algoritmalarının daha verimli bir şekilde çalışır hale getirilmesine olanak sağlayacaktır.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Proteins are the most abundant biological polymers found in all cells as well as in all compartments within those cells. All proteins, independent of their complexity, are constructed from the same ubiquitous set of twenty different amino acids linked to each other with the help of a peptide bond as shown in Figure 1.1. Peptide bond formation is a dehydration reaction in which a water molecule is released as a result of the interaction between the hydroxyl (–OH) of carboxylic acid group of one amino acid residue and the hydrogen (–H) atom of the amino group of the succeeding amino acid residue.
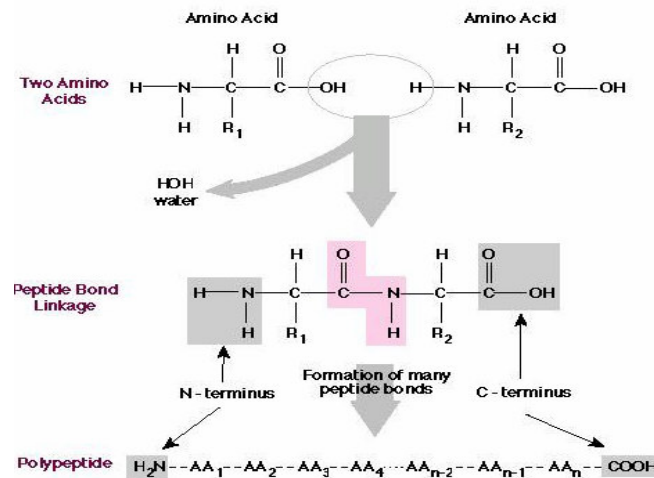
Figure 1.1 Peptide bond formation [1]

Under physiological conditions, peptide bond shows the characteristics of a double bond with a ratio of approximately 40% [2]. Due to having a partial double bond character, the peptide bonds have a large dipole moment which leads to fixed rotational angle around the peptide bond to 180^0 as shown in Figure 1.2 and also plays a role in the formation of secondary structures such as alpha helix and beta sheets. The fixation of the ω angle, the dihedral angle of a peptide bond, results in N, H, C, and O atoms to lay on a rigid plane, the amide plane, as shown in Figure 1.2 [3]. Although the partial double bond limits the rotation around the peptide bond, bonds located at each side have still have the ability of rotating around those bonds between the HN and C_α and C_α and C=O. The rotational angle around the first bond is called ϕ whereas that of the second is called ψ.
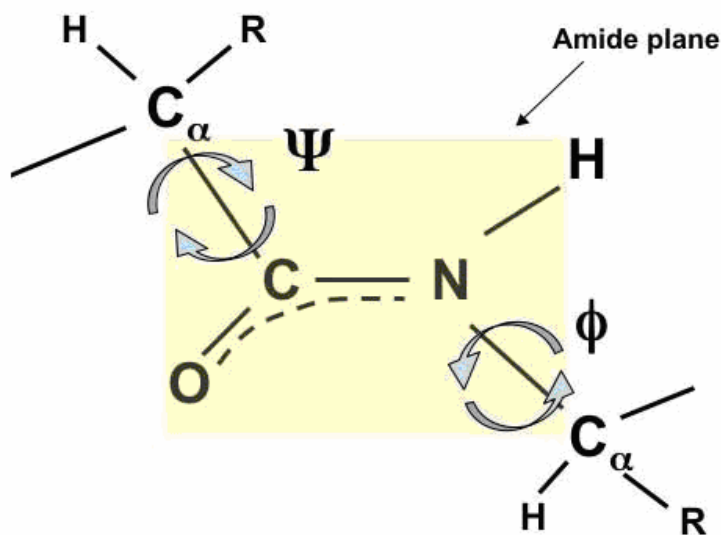
Figure 1.2 The planarity of peptide bond together with other rotational angles ϕ and ψ [3]

Calculation of a torsion angle between atoms i and i+1 requires the atoms i-1,i ,i+1, and i+2. For instance, if the torsion angle ϕ_i is calculated, then, C_{i-1}, N_i, $C_{\alpha i}$, and C_i atoms must be included, and in the same manner, if the torsion angle ψ_i is calculated, then, N_i, $C_{\alpha i}$, C_i,

and N_{i+1} atoms must be included. In this respect, the first ϕ, and last ψ torsion angles of a protein will remain undefined.

Ramachandran maps, bounded by the angle interval of $(-180^0)/(180^0)$, show the $\phi-\psi$ angle distribution of amino acid residues in which ϕ angle values are shown on the x axis, and ψ angle values are shown on the y axis. For most amino acids, the dihedral angles are confined to well-defined regions of the Ramachandran map which are called as allowed regions. These regions not only satisfy excluded volume constraints, but also carry signatures of specific conformations of a residue [4]. The boundaries of those allowed regions of the Ramachandran map are identified via $\phi-\psi$ angle pairs of protein structures with known three-dimensional structures from the Protein Data Bank (PDB) [5]. PDB serves as a repository for 3-D structures obtained via mainly X-ray crystallography or NMR spectroscopy.

The primary structure of a protein is its linear amino acid sequence. The secondary structure, specifically $\alpha-$helices, and $\beta-$sheets, is fomed via a fixed bonding pattern which is not covalent, but weaker nonbonded interactions, especially hydrogen bonds that are formed between the corresponding backbone amide and carboxyl groups. Specific secondary structures can be identified from the Ramachandran map as shown in Figure 1.3. The tertiary structure is formed via interactions formed among the amino acid residues located far apart in the primary structure of the protein whereas the quaternary structure of the protein is formed by the formation of interactions among different polypeptide chains.
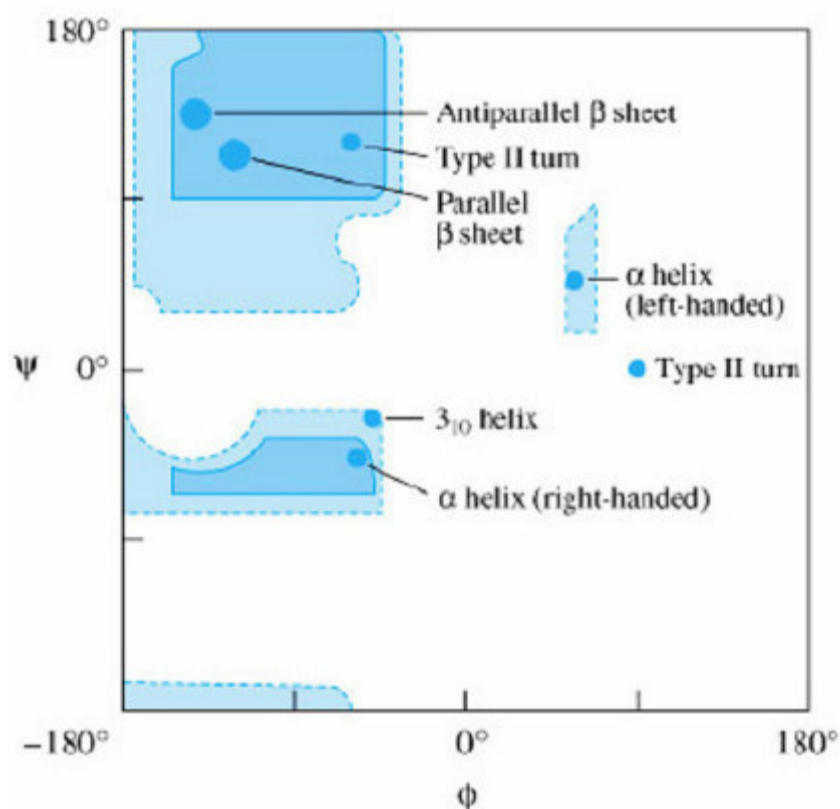
Figure 1.3 Ramachandran map showing the corresponding regions of secondary structures[6]

The conformational $\phi-\psi$ preference and secondary structure propensity of amino acid residues is determined by many factors, which include the hydrophobic contacts, side-chain interactions, and steric effects, and the types of nearest neighbors in their native state [7]; however, it has been thought the $\phi-\psi$ preference of amino acids are not influenced by those factors in the unfolded state which is also known as random coil. Definitions for the structural characteristics of random coil vary. In 1970, Tanford defined the random coil as a state in which each bond could rotate freely [8]whereas Shortle described random coil state as one in which no side-chain-side-chain interactions are present [9]. Smith *et al.* stated that $\phi-\psi$ angles of each residue are independent of $\phi-\psi$ angles of other residues in

random coil [22]; however, experimental and theoretical studies have provided evidence for the notion that well defined conformations can emerge locally even in unfolded peptides [11] which motivated researchers to reevaluate the Flory's isolated pair hypothesis in which it is stated that each $\phi-\psi$ pair in the peptide is insensitive to the values of its neighbors [12,13,14]. In addition, it has been shown that polyproline-II helix, which is a α_L which is composed of three proline residues per turn [15], has been the most dominant structural motif of unfolded peptides. It exists only in water and reflects the local conformational preference of a given residue in contrast to classical secondary structures.

Conformational preference of an amino acid residue can be obtained experimentally, by the analysis of PDB data, by performing molecular dynamics (MD) or Monte Carlo (MC) simulations. MD, as explained in more detail in Chapter 3, numerically solves Newton's equations of motions of a molecular system to get information about its time-dependent properties in which the interactions between objects are described by a force-field, quantum mechanical model, or the combination of both [16]. It has gained popularity in material science, theoretical physics, and since the 1970s also in biochemistry and biophysics; however, simulation of dense macromolecular systems is impossible if all degrees of freedom and interactions of a chemically realistic chain are taken into account because of insufficient computational power [17]. In this respect, coarse-grained models have been constructed to make computer simulations a tractable approach.

Contribution

In this study, our goal is to construct a coarse-grained model to represent the conformational preferences of unfolded peptides via MD simulations. We made a detailed comparison of several available force-fields in order to identify a reliable candidate for sampling the conformations of unfolded peptides in both aqueous and hydrophobic medium. Selection of an appropriate force-field is a crucial step since almost all force-

fields are parameterized, and tuned to represent the native states of peptides. We used Gromos force-field with parameter set 53A6 [18], which gives similar results to experimental studies, to represent the unfolded states of peptides.

In other studies, the validity of Flory's isolated-pair hypothesis was checked for extended, polyproline-II helix, α_R conformations; however, we included two more basins in our study namely as α_L, and γ–turn. In this respect, we could investigate the validity of the hypothesis for the five well-defined Ramachandran basins.

The relative contributions of short and long-range interactions to the conformational preferences of unfolded tripeptides are elucidated depending on the results obtained via our coarse-grained model.

Outline

Chapter 2 summarizes the related work made on the structural preferences of unfolded peptides from experimental studies, statistical analysis of PDB data, and computer simulations, and the comparison of those conformational preferences with coil library data, the comparison of different force-fields, MD simulations that are performed in different media. In chapter 3 detailed information about the methods, algorithms, simulation details and main ideas used in this study are given. Chapter 4 gives a detailed investigation of the conformational probability distribution of peptides obtained via MD simulations that were performed in both aqueous and hydrophobic media. In that chapter, detailed information about the performance of the model on different peptides is also given. In addition, corrections that were incorporated into the model are analyzed in detail. The last chapter finalizes the study by both giving a summary of the overall results, and some of the application areas in which it may be used successfully.

Chapter 2

RELATED WORK

In order to have a deeper understanding of the protein folding problem, studies have been concentrated on the structural characteristics of unfolded states of short peptides. The conformational space of peptides is considered to have a rugged energy landscape which is composed of an exorbitant number of local minima [19,20]. Under suitable conditions, the peptide can visit that energy landscape spontaneously, and find the global minimum, namely the native state [19]. Fowler and Guggenheim state that peptides in their unfolded states wander the whole range of conceivable stationary states randomly [21]. In contrast to these ideas, experimental and theoretical evidence have been provided for the notion that well defined conformations can emerge locally even in unfolded peptides. In unfolded states of proteins and peptides, although torsion angles undergo some fluctuations, there are some preferred conformational states [22,23]. The prevalence of the structural biases in unfolded peptides provide a motivation for the reevaluation of Flory's isolated-pair hypothesis in which it was shown that each ϕ,ψ pair in the peptide backbone is independent of its neighbors.

Theoretical calculations of Pappu and coworkers state that Flory's isolated pair hypothesis is valid for unfolded peptides, provided the residues are found in the upper left quadrant of the Ramachandran map [19]. Since Flory's isolated pair hypothesis states that each $\phi-\psi$ pair is sterically independent, local conformation of an unfolded peptide reflects

the intrinsic structural propensities of the corresponding amino acid residue [24,19,25,26,27].

In the literature, numerous studies have been performed on intrinsic structural propensities of amino acids by using data extracted from coil library [28], and making host-guest studies on short peptide sequences [29,30]. In 1995, Swindells and coworkers found that the coil library, which can be treated as a working model of the backbone conformations of residues in unfolded peptides, give intrinsic preferences that are related to Chou-Fasman statistical frequencies of amino acids in secondary structures such as α−helices and β−sheets [31]. During this period, the extended-β and polyproline-II helix conformations were analyzed as a whole basin within the upper left quadrant of Ramachandran map [29]. After 2000, it has been observed that PP-II helix conformation of short peptides is the most dominant structural motif of unfolded peptides [32,33,30,34]. It exists only in water and reflects the local conformational preference of a given residue in contrast to classical secondary structures . In 2004, Eker and coworkers combined a variety of experimental techniques to obtain the dihedral angles of the central residues of AXA peptides where A represents alanine, and X represents glycine, valine, methionine, histidine, serine, proline, lysine, leucine, tryptophan, tyrosine, and phenylalanine as in host-guest studies. In this respect, they obtained the structural propensities of mentioned amino acid residues, and used them to predict possible conformations of the monomeric amyloid β peptide $A\beta_{1\text{-}42}$ in aqueous solution. Finally, they demonstrated that the unfolded state of peptides can be understood in terms of the intrinsic structural propensities of their amino acid residues, in agreement with Flory's isolated pair hypothesis as opposed to random coil model [35]. However, Avbelj and Baldwin stated that the neighboring residue effect is observed in the coil library. Furthermore they demonstrate that when a neighboring residue (i+1, i-1) belongs to class L which consists of aromatic and β−branched amino acids, then the ϕ of residue i is more negative for essentially all amino acid residues [36].

In a study of Hermans and coworkers, the unfolded state of proteins was investigated by using two different approaches: database statistics and molecular dynamics simulations. They stated that the two models give different distributions, and highlighted the difficulty in representing an accurate behavior of the unfolded state of proteins [37]. A similar problem arises upon using different force-fields for representing the unfolded states of proteins. Avbelj and Baldwin, they stated that for making a force-field succeed with unfolded peptides, the solvation free energies of amides should be reproduced accurately [38]. Some force-fields fail to do this; however, opls-aa and Gromos force-field with a parameter set of 53A6 [18,36] , used in this study, has been adjusted to solve this problem.

The ability of a peptide to fold into a defined structure is governed not only by information encoded in its primary sequence, but also by the chemical environment in which it resides [39]. In this respect, peptides as well as proteins have been simulated in different environments. However, much effort has been invested in studying the structural characterization of peptides in polar, aqueous environment. In a number of studies, it is shown that apolar solvents, such as trifluroethanol and hexafluoroisopropanol have been shown to stabilize secondary structure elements whereas different micellar states of sodium dodecyl sulfate (SDS) promote either β−sheet and α−helix formation [40].

Chapter 3

MATERIALS AND METHODS

The present study is composed of three main parts. In the first part, conformational space of capped alanine (A), valine (V), tryptophan (W), and tyrosine (Y) residues were analyzed. In order to indicate the constituent amino acids of a given peptide, single letter codes of each amino acid were used as shown in parenthesis above. For instance, a capped dipeptide that was composed of alanine and tryptophan residues was shown as AW. After analyzing the conformational space of capped residues, in order to understand the influence of neighboring amino acid residues, sequence of length one, three, and five were tested for the following capped tripeptides and pentapeptides: WWW, AWA, AVA, AAA, VVV, YYY, AYA, and AAAAA in both aqueous and hydrophobic media in detail via MD simulations. Moreover, following dipeptides were simulated to extract necessary information for our model system: AA, AV, VA, YY, AY, YA, WW, AW, WA, VV.

AXA tripeptides where X represents the amino acids mentioned above were simulated to get intrinsic conformational preferences of the central residues whereas homopeptides were simulated to investigate the nearest-neighbor effect on the central residue of given tripeptides or, in other words, to test the applicability of the Flory's isolated-pair hypothesis on short peptides.

In the second part, a box of capped A and W monopeptides were simulated to obtain the radial distribution function. In addition, potential of mean forces among those two residues were obtained via their radial distribution functions.

In the third part, detailed calculations and analysis were performed by means of the configurational probabilities obtained in the first and second part.

This chapter is composed of two parts. In the first part, general information about MD simulations is given in order to make the reader familiar with the methods used. In the second part, a summary of the key points necessary to carry out an MD simulation is given. The results obtained from the second part are given in Section 4.1.

3.1 General Information about Molecular Dynamics simulations

The aim of carrying out computer simulations is the investigation of the properties of molecules in terms of their structure and microscopic interactions. They serve as a complement to experiments because hidden details of bulk properties of systems can be revealed. Two main types of simulation techniques are molecular dynamics (MD) and Monte Carlo (MC). There are also a range of hybrid techniques which use features of those two techniques. By performing MD simulations under different conditions to represent different ensembles, one can collect time averaged statistical data from the system under investigation, which can be used to learn about the thermodynamic equilibrium state of the system.

The MD simulation method is based on Newton's second law for the equation of motion [41]:

$$\vec{F}_i = m_i \vec{a}_i$$

(3.1)

where F_i is the force exerted on the particle, m_i is the mass of that particle, and a_i is its acceleration. The acceleration of each atom can be determined by the help of knowledge of

force on each atom. The trajectory of the system, which describes the positions, velocities, and accelerations as they vary with time, is obtained by the integration of equations of motion. From this trajectory, the time average values can be determined [41]. The method is deterministic: By knowing the positions and velocities of each atom, the state of the system can be predicted at any time in the future or the past.

The force can also be expressed as the gradient of the potential energy.

$$\overrightarrow{F_i} = -\overrightarrow{\nabla_i} V$$

(3.2)

Combination of these two equations yields the following equation:

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

(3.3)

where V is the potential energy of the system in question [41].

3.1. 1 Numerical Integration Algorithms for solving equations of motions

The potential energy function is composed of atomic coordinates of all the atoms in a system. Due to its complex nature and large number of degrees of freedom in a typical MD system, it can not be solved analytically. In this respect, different types of numerical integration algorithms have been developed for solving equations of motion. Independent of type of the algorithm, following criteria should be considered. 1) Conservation of energy and momentum, 2) Computational efficiency, 3) Permission of a long time step for integration. Some of those integration algorithms are Verlet, Leap-frog, Velocity Verlet, and Beeman's algorithm [41]. The Verlet algorithm uses positions and accelerations at time t and the positions from time t-δt in order to calculate new positions at time t+δt [41].

We used Verlet algorithm because it is straightforward, and the storage requirements are modest; however, it has a disadvantage of having moderate precision.

3.1.2 Different Ensembles used in MD simulations

MD simulations generate a sequence of points in phase space as a function of time all of which belong to the same ensemble: different conformations of the system with their respective momenta. Several different ensembles are used in MD simulations depending on the system in question. These are microcanonical (NVE) which corresponds to an adiabatic process, canonical (NVT), isothermal-isobaric (NPT), and grand-canonical (μVT) ensembles where N, E, T, V, P, and μ refer to fixed number of atoms, a fixed energy, temperature, volume, pressure, and chemical potential, respectively [41]. In each ensemble, those letters in parenthesis are kept constant during the simulation. However, most MD simulations are performed by using NPT ensemble to mimic the conditions at which many experiments are performed. This can be done by coupling to system to an external temperature and pressure bath. A variety of methods are required to add or remove energy from the system in a realistic way. Some of the popular techniques to maintain temperature include Berendsen, Nose-Hoover [41,42] and Langevin dynamics [41,42]. For pressure coupling, Parrinello-Rahman and Berendsen algorithms are used [42]. The frequency for coupling to an external bath can be changed depending on the state of the system. If the equilibrium is not reached, then the system should be coupled to an external bath more frequently; however, if the equilibrium is reached, then, there is no need to couple frequently, the time step for coupling can be increased.

3.1.3 Different sampling techniques

The potential energy surface of a molecular system is a very complex landscape which is composed of a number of local minima, at which all derivatives of the potential energy function with respect to the coordinates are zero and second derivatives are nonnegative and one deepest local minimum, named as global minimum. The number of local minima on the energy surface is so high, that it is impossible to sample the space sufficiently. If one wants to find other local minima and to discover the global minimum, an advanced MD simulation should be performed such as simulated annealing, umbrella sampling, replica-exchange method etc [42]. Simulated annealing technique involves cyclic heating and cooling of the system, which enables the atoms to overcome energy barriers on the complex energy surface. The atoms are prevented from becoming unstuck from their initial positions and they can wander through states of higher energy by heating whereas slow cooling gives a chance to find lower energy configurations compared to the initial one.

3.1.4 Periodic Boundary Conditions

Finite size system effects can be reduced by using periodic boundary conditions. The atoms of the system are put into a space-filling box, which is surrounded by translated copies of that central box. Thus, there are no boundaries of the system. Simulation packages use periodic boundary conditions with the minimum image convention in which only one nearest image of each particle is considered for the calculation of short-range nonbonded forces whereas for long-range electrostatic interactions, this does not yield accurate results, hence one has to use a method to sum over all periodic images of the atoms such as, Ewald sum, Particle Mesh Ewald (PME), and Particle-Particle-Particle-Mesh (PPPM) [41].

3.1.5 Force fields

A force-field, potential function, is composed of two distinct components: the set of equations used to generate potential energy and forces, and the parameters used in this set of equations. Within an equation, various sets of parameters can be used providing they are consistent. A typical potential function in an MD simulation is composed of three parts: nonbonded, bonded, and special terms. Nonbonded interactions generally include Lennard-Jones and Coulomb interactions [18,42]. They are computed on the basis of a neighbor list composed of nonbonded atoms within a certain radius. Nonbonded interactions in Gromacs are pair-additive and centro-symmetric [42]. They contain repulsion, dispersion, and a Coulomb term. The former two are combined in the Lennard-Jones potential. In addition, charged atoms are treated via Coulomb potential [18,42]. Bonded interactions include covalent bond-stretching (2-body), angle bending (3-body), improper dihedrals (4-body), and proper dihedrals (4-body). They are computed on the basis of fixed atom lists [42] and as indicated above, they are not pair interactions because they include 3- and 4- body interactions as well. Improper dihedral interaction is used to force atoms to remain in a plane.

3.2 Methods

In this study, three different MD simulation methods were used which are named as Method1, Method2, and Method3. Method1 was used for simulations in an aqueous medium whereas Method2 was used in simulations performed in hydrophobic medium. Method3 was used for simulations of a box of capped alanine and capped tryptophan residue.

3.2.1 Method 1

Each peptide was solvated with water, such that the desired density at 310 K and 1 atm was reached and there were approximately 250 water molecules per residue. Water molecules were modeled with the SPC216 model [42].

The minimization step was performed in order to remove overlapping Van der Waals contacts and to start with a good initial conformation. In this respect, no constraints were used for both bond angle and lengths. –DFLEX option [42] was used to make water molecules flexible which helped the system minimized further. As shown in Figure 3.1, a 1000 kJ/mol*nm convergence criterion was used which meant that when the maximum force acting on the system was smaller than that value, the minimization was converged. Steepest descent minimization algorithm was used because it is faster than conjugate gradient algorithm in the early stages of minimization. As will be mentioned in the following sections, smaller convergence criterion was used to perform a more stringent energy minimization step.

In NVT equilibration step, the temperature of the system was maintained at 310 K via Berendsen temperature coupling algorithm [43]. Time constant for coupling (tau_t) was taken as 0.1 ps both for peptide and solvent molecules which meant that the system was coupled to the bath every 0.1 ps. Initial velocities were assigned from a Maxwell-Boltzmann distribution, corresponding to a temperature of 310 K. The nonbonded pair forces need to be calculated only for those i, j pairs for which the distance between i and its nearest image j is less than a given cut-off distance. A pair list, which contains those particle pairs, is generated. Nonbonded interactions must be calculated for those pairs. The pair list is composed of atom i, a displacement vector for that atom, and all particles that are within the cutoff radius [42]. That list was (nslist) updated every 10 steps as indicated in Figure 3.1. Short-range interactions were calculated for those atoms within 0.9 nm which

was indicated via r_list parameter. Van der Waals interactions were calculated via cut-off method whereas electrostatic interactions were calculated via PME method due to the fact that computational complexity in the calculation of Van der Waals interaction relative to that electrostatic interaction. Single-range cut-off was used, so that no additional forces were evaluated during the neighbor list generation.

In NPT equilibration step, Berendsen coupling algorithm [43] was used with isotropic pressure coupling constant of 0.5 ps which meant that the system was coupled to an external pressure bath every 0.5 ps.

In the main MD simulation we applied the simulated annealing method with periodic boundary conditions. NVT ensemble was chosen to prevent the explosion of the system that may occur as a result of application of a high temperature. Each simulated annealing cycle composed of 80 ps. The system temperature was increased from 310 K to 1010 K in 20 ps, followed by a cooling period where the temperature was decreased to 310 K in 20 ps. During the last 40 ps, the temperature was kept constant.

For simulations in which the simulated annealing technique was performed, data points were collected from the last 20 ps of each cycle which corresponded to number of 37125, for using in further analysis.

ENERGY MINIMIZATION STEP
Integrator =Steepest Descent
Constraint =None
dt =0.002(ps)   -DFLEX option
nslist =10   and SPC water model
ns_type = grid
rlist = 0.9(nm)
vdw_type = Cut-off
coulomb_type = PME
rcoulomb = 0.9 (nm)
rvdw = 0.9 (nm)
emtol = 1000.0 (kj/mol*nm) with emstep = 0.01

NVT EQUILIBRATION STEP
Integrator=MD
Constraint=All bonds
dt =0.002(ps)
nslist =10
ns_type = grid   SPC water model
rlist = 0.9(nm)
vdw_type = Cut-off
coulomb_type = PME
rcoulomb = 0.9(nm)
rvdw = 0.9 (nm)
T_coupl=Berendsen
tau_t=0.1 (ps)
ref_t=310 (K)

NPT EQUILIBRATION STEP
Integrator=MD
Constraint=All bonds
dt =0.002(ps)
nslist =10
ns_type = grid
rlist = 0.9(nm)   SPC water model
vdw_type = Cut-off
coulomb_type = PME
rcoulomb = 0.9 (nm)
rvdw = 0.9 (nm)
T_coupl=Berendsen
tau_t=0.1 (ps)
ref_t=310 (K)
P_coupl=Berendsen
P_coupltype=Isotropic
tau_p =0.5 (ps)
ref_p=1.0 (bar)
compressibility =4.5 e-5/ bar (@ 310 K and 0.983 bar)

MAIN MD SIMULATION
Integrator=MD   rcoulomb = 0.9(nm)
Constraint=All bonds   rvdw = 0.9(nm)
dt =0.002(ps)   T_coupl=Berendsen
nslist =10   tau_t=0.1 (ps)
ns_type = grid   ref_t=310 (K)
rlist = 0.9(nm)   Simulated annealing
vdw_type = Cut-off   (using with periodic boundary
coulomb_type = PME   conditions for both solvent and solute)

Figure 3.1 Minimization, pre-equilibration, and main MD steps with their corresponding parameters for Method 1

3.2.2 Method 2

   Method 2 was used to perform simulations in a hydrophobic environment which composed of cyclohexane molecules. Although an equilibrated box of water has been defined in the Gromos force field with a parameter set of 53A6 [18], an equilibrated box of

cyclohexane has not been defined. In this respect, after a cycle of energy minimization, and NPT equilibration, the density close to the desired value, which was approximately 0.720 g/cm^3 at 310 K and 1 atm, was reached. The topology file of the cyclohexane molecule was obtained via Dundee PRODRG server [44] and given in Appendix A.2 in detail. We performed only NPT equilibration step to maintain a density value closer to real one before starting to the main simulation.

Single-range cut-off was used, so that no additional forces were evaluated during the neighbor list generation.

Cyclohexane molecules have a ring structure, which is composed of six carbon atoms. Due to the fact that the force-fields are parameterized to reproduce strong ring stacking, larger values of r_list, r_coulomb, and r_vdw parameters (1.4 nm) increase the possibility of formation of aggregates within the system. We used a large number of cyclohexane molecules in order to get the desired density. In this respect, for both the prevention of the formation of aggregates within the system and reduction of the computational time, we decided to use a value of 0.9 nm for the parameters mentioned above.

For simulations in which the simulated annealing technique was performed, data points were collected from the last 20 ps of each cycle which corresponded to number of 37125, for using in further analysis.

ENERGY MINIMIZATION STEP

Integrator = Steepest Descent
Constraint = None
dt = 0.002 (ps)
nslist = 10
ns_type = grid
r_list = 0.9 (nm)
vdw_type = Cut-off
coulomb_type = PME
r_coulomb = 0.9 (nm)
r_vdw = 0.9 (nm)
em_tol= 100 kJ/mol*nm   emstep=0.01

MAIN MD SIMULATION

Integrator = MD
Constraint = All-bonds
dt = 0.002 (ps)
nslist = 10
ns_type = grid
r_list = 0.9 (nm)
vdw_type = Cut-off
coulomb_type = PME
r_coulomb = 0.9 (nm)
r_vdw = 0.9 (nm)
P_coupl= Parrinello-Rahman
tau_t=0.5 (ps)
ref_p=1.0 (bar)
Simulated annealing
(using with periodic boundary
conditions for both solvent and solute)

NPT EQUILIBRATION STEP

Integrator = MD
Constraint = All-bonds
dt = 0.002 (ps)
nslist = 10
ns_type = grid
r_list = 0.9 (nm)
vdw_type = Cut-off
coulomb_type = PME
r_coulomb = 0.9 (nm)
r_vdw = 0.9 (nm)
P_coupl= Parrinello-Rahman
tau_t=0.5(ps)   P_coupl=Berendsen
ref_p=1.0 (bar)   tau_t=0.1(ps)
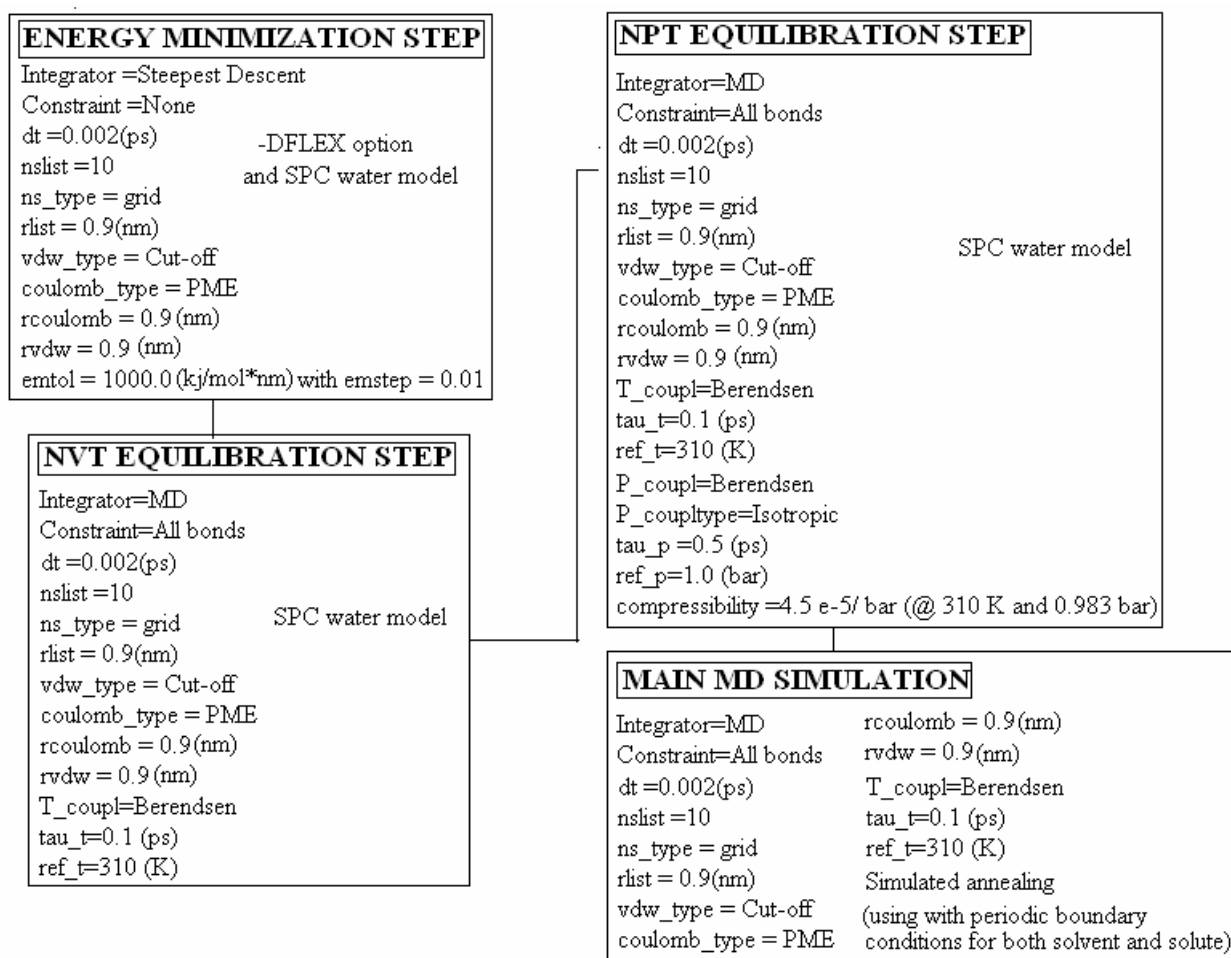   ref_t=310 (K)

Cyclohexane environment

Figure 3.2 Minimization, pre-equilibration, and main MD steps with their corresponding parameters for Method 2

3.2.3 Method 3

Method 3 was used to perform simulations in a box of capped alanine and tryptophan residues. This type of simulation was different from the other two. Minimization step was more important for this type of simulation because probability of having overlapping molecules was higher, especially for tryptophan molecules. In this respect, more stringent convergence criterion, 10 kJ/mol*nm, was used in the energy minimization step. After that, an NPT equilibration step with Parrinello-Rahman coupling algorithm was performed to fill the holes among the molecules by changing the volume of the box. After the equilibrium was reached, NPT simulation was stopped and the output of that step was given to main MD simulation as input. The main MD simulation was performed by using an NVT ensemble in order to obtain a constant value of volume which is important for further analysis.

Our system contained many atoms, approximately 46656, so we used 0.8/1.4 nm of twin-range cut-off method. In this method, interactions within 0, 8 nm (r_list) were calculated at each time step from a pair list that was generated every ten steps. At this time points, interactions between 0,8 and 1,4 nm were also calculated and kept constant between updates and a reaction-field contribution was added to the electrostatic interactions and forces to maintain a homogeneous medium outside the long-range cutoff. Since calculating of all forces were not done at every step, energy was not conserved exactly. This was an artifact, but we had to compromise between accuracy and computational cost. In addition, errors introduced by using the twin-range cut-off were negligible. MD simulation data obtained via Method 3 were used to construct the radial distribution function (RDF) of the simulated systems. In order to obtain reliable RDFs, we included interactions among the atoms located not only at short distances but also at long distances, so we used larger r_coulomb and r_vdw values compared to those used in Method 1 and Method 2. Due to

having a large number of atoms, we decided to use the twin-range cut-off method for preventing the calculation of long-range interactions at each step. The twin-range cut-off method is not used with PME due to a restriction in the implementation algorithm of PME. In this respect, we used the Reaction-Field method for the calculation of electrostatic interactions.

ENERGY MINIMIZATION STEP

Integrator = Steepest Descent
Constraint = None
dt = 0.002 (ps)
nslist = 10
ns_type = grid
r_list = 0.8 (nm)
vdw_type = Cut-off
coulomb_type = Reaction-field
r_coulomb = 1.4 (nm)
r_vdw = 1.4 (nm)
emtol=10.0 kJ/mol*nm   emstep=0.01

MAIN MD SIMULATION

Integrator = MD
Constraint = All-bonds
dt = 0.002 (ps)
nslist = 10
ns_type = grid
r_list = 0.8 (nm)
vdw_type = Cut-off
coulomb_type = Reaction-field
r_coulomb = 1.4 (nm)
r_vdw = 1.4 (nm)
T_coupl=Berendsen
tau_t=0.1(ps)
ref_t=310 (K)

NPT EQUILIBRATION STEP

Integrator = MD
Constraint = All-bonds
dt = 0.002 (ps)
nslist = 10
ns_type = grid
r_list = 0.8 (nm)
vdw_type = Cut-off
coulomb_type = Reaction-field
r_coulomb = 1.4 (nm)   P_coupl = Parrinello-Rahman
r_vdw = 1.4 (nm)   tau_p = 0.5 (ps)
T_coupl=Berendsen   ref_p = 1.0 (bar)
tau_t=0.1(ps)   compressibility = 4.5e-5/bar
ref_t=310 (K)   (@ 310 K and 0.983 bar)

Figure 3.3 Minimization, pre-equilibration, and main MD steps with their corresponding parameters

3.3 Key points for the pre-equilibration steps and main MD simulations

There were some key points that should be taken into consideration before passing to the main MD simulation steps.

3.3.1 Necessity of putting caps to termini of peptides

Advantages of using caps to obtain a reliable conformational distribution have been discussed in previous studies [45]. Without the terminal groups, the peptides will be charged, and the interaction of the positive and negative charges will strongly bias the configurational space. The Ramachandran map of an uncapped monoalanine is shown in Section 4.1. In addition, these terminal groups mimic the extension of the peptide backbone, so that the amino acid under investigation behaves as if it is part of a long chain which enables us to compare the MD distributions of backbone conformation to proteins whose structure is accurately known [45]. In this respect, the peptides were terminally blocked with an acetyl group (ACE) for their N and a methyl amide group (NME) for their C termini. Although ACE group was defined in the Gromos force-field with parameter set 53A6, NME was not defined. The parameters for this end group were extracted from the other residues' information in the force-field, and they are defined in Appendix A.1. The biased configurational space of uncapped monoalanine is shown in Figure 3.4. Due to the interaction between positive and negative termini, uncapped monoalanine visited similar basins to those of glycine.

RAMACHANDRAN MAP OF
UNCAPPED MONO ALANINE
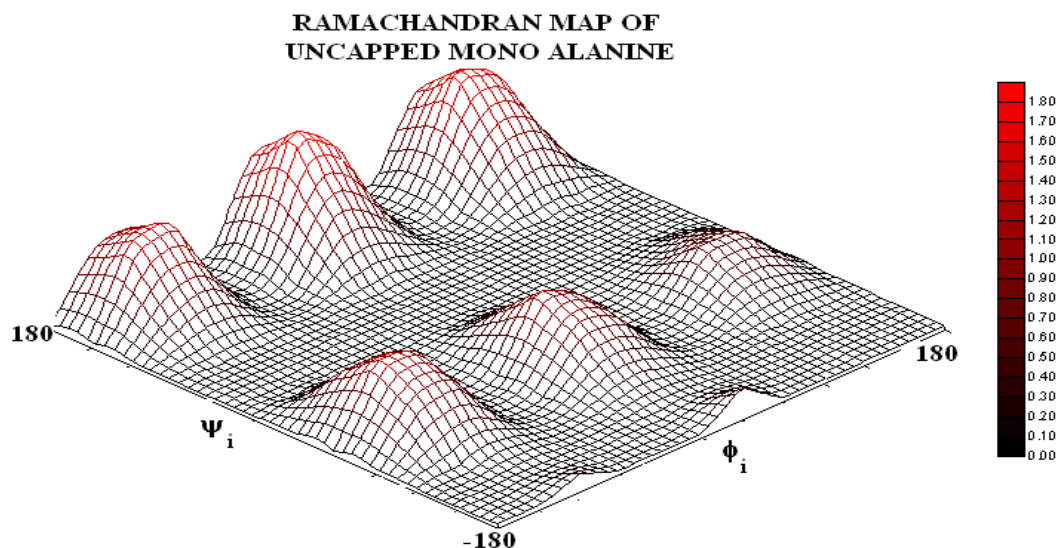
Figure 3.4 Ramachandran map of uncapped monoalanine

3.3.2 Force-field in molecular dynamics simulation in the unfolded state

Nearly all force-fields have been tuned to represent the characteristics of proteins in their native states. Therefore we tested a variety of force-fields to find an appropriate one for proper representation of the unfolded state of peptides. In order to test the validity of the force-fields, the statistical data obtained from MD simulations were compared with both statistical data and results obtained via other force-fields such as AMBER99 [46], OPLS-AA/L all atom [47], charmm22 [48], Gromos force-fields with parameter set 45A3 [49] and 53A6 [18]. The resultant conformational distributions were similar to those obtained by Hermans and coworkers, and they compared a variety of molecular mechanics force-fields to investigate the problem of modeling the unfolded peptide backbone [18]. In a statistical study of Swindells and coworkers, the intrinsic $\phi-\psi$ propensities of amino acids derived from the coil regions, structures not in $\alpha-$helix and not in $\beta-$strand, of known structures were given in Table 3.2. As mentioned above, for selecting an appropriate force-field, both

experimental and statistical data were taken into account which led to using of Gromos force field with parameter set 53A6 [18] which was based on the free enthalpy of hydration and salvation. However, peptides were mostly trapped around the extended basin of the Ramachandran map not visiting other basins with a significant probability. In order to overcome that problem, we used the simulated annealing method.

Table 3.1 Conformational distribution of capped mono alanine in simulations of different force-fields [18]

| Force fields | Amber | Charmm 22 | Gromos | Opls-aa |
Basins				
Beta (PP-II & extended)	0.16	0.50	0.82	0.86
α_R	0.84	0.50	0.13	0.135
α_L	-	-	0.04	0.004
γ–turn	-	-	0.0005	0.0006

Table 3.2 Intrinsic phi-psi propensities of amino acid residues derived from coil regions of known structures [31]

Amino acid type	Intrinsic $\phi-\psi$ propensities derived from coil region		
	α_R /coil	B/coil (PP-II, extended)	$(\alpha_L, \gamma-\text{turn})$ /coil
Alanine (A)	1.23	1.09	0.39
Valine (V)	0.89	1.33	0.36
Tyrosine (Y)	0.85	1.26	0.59
Tryptophan (W)	1.17	1.09	0.48

3.3.3 Initial configuration of peptides

It is well-known that the initial configuration plays a crucial role on the conformational preferences of simulated molecules in MD simulations. In this respect, we set the initial structures to different conformations to investigate whether the conformational preferences were dependent on the initial configurations; however, we observed that it had a negligible effect on the equilibrium $\phi-\psi$ distributions of peptides when the simulated annealing technique is used. So, the simulations were started with extended conformations, in which both ϕ and ψ angles were set to 180^0.

3.4 Rotational Isomeric States Approximation

The interdependence of bond rotation of the ith bond on the i-1th and the i+1th bonds necessitates some simplifications in order to represent the potential of chain molecules reliably [50,51]. In this respect, the continuous space of dihedral angles is replaced by discrete values. This approach is known as the rotational isomeric state approximation (RIS) in which each bond is allowed to be in a finite number of different rotational states. By using distinct rotational isomeric forms, description of conformations of simple molecules can be made in a convenient way [50,51].

Both ϕ and ψ dihedral angles can take values in the interval of $-\pi$ to $+\pi$. In this study, the discrete state formalism was used for torsion angles. The Ramachandran map was divided into a 12x12 grid each of which represented a 30^0 interval for both ϕ and ψ dihedral angles. Some of these grids were combined to construct some defined regions, basins, on Ramachandran map. The names and boundaries of those basins are described in detail in Section 3.5.

3.5 Definition of Ramachandran Basins

The investigation of the distribution of data for the Ramachandran map was concentrated on basins as depicted in Figure 3.5.
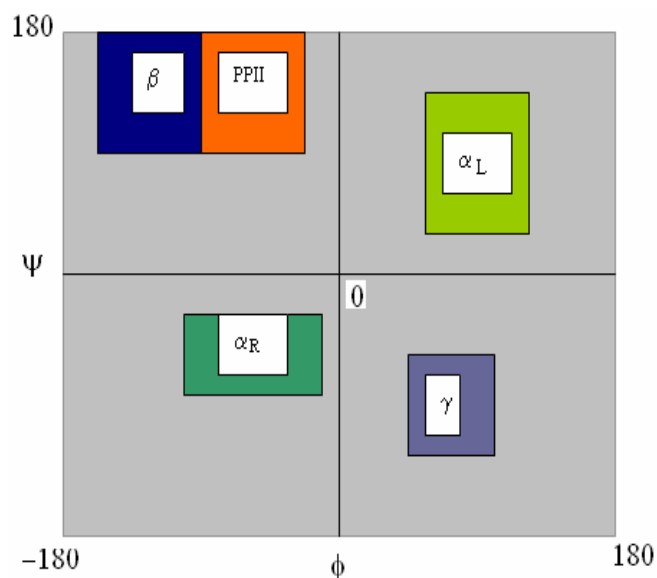
Figure 3.5 Ramachandran basins

The boundaries of the basins are shown in Table 3.3 together with their full names. The boundaries of the basins were chosen to fulfill two conditions : to get both the smallest basins in the Ramachandran map and maximum number of data points within that regions.

Table 3.3 Boundaries of Ramachandran basins with their full names

Ramachandran Basins	Full name	ϕ interval	ψ interval
PPII	Polyproline-II helix region	$-90<\phi<-30$	$60<\psi<180$
β	Extended (β)region	$-150<\phi<-90$	$60<\psi<180$
α_R	Right-handed helix region	$-120<\phi<0$	$-90<\psi<0$
α_L	Left-handed helix region	$0<\phi<120$	$0<\psi<90$
γ	Gama turn region	$0<\phi<180$	$-180<\psi<0$

3.6 Statistical weight matrices with interdependent bonds

As Flory stated, a configuration (ϕ) of a chain molecule composed of *n* bonds can be represented via rotational isomeric state scheme in a *v*-digit system, *v* being the number of discrete rotational isomeric states chosen to describe the rotational potential of a bond. For demonstration, by choosing *v*=2, a configuration of a 4- bond system can be specified by the series of digits as:

$$0\ 0\ 1\ 0$$

It is supposed that the rotational potential affecting any given bond i depends exclusively on the configuration of i-1 [th], i [th], and i+1 [th] bonds. Interactions among the longer range residues, residues beyond the first nearest neighbors, are not included. Then, the total configurational energy can be expressed as the sum of energies of the first-nearest neighbors. In this respect, the total configurational energy can be expressed as follows [50]:

$$E(\phi) = E_0 + E_{00} + E_{01} + E_{10} \tag{3.4}$$

The first term on the right hand side carries a single index since it has no predecessor in the given sequence. The other residues have double index, first of which represents the latter index of its predecessor whereas the second index represents its own index. The equation for total energy can be generalized as follows [50]:

$$E(\phi) = \sum_{i=1}^{n-1} E_i(\phi_{i-1}, \phi_i) = \sum_{i=1}^{n-1} E_{\eta\zeta;i} \tag{3.5}$$

Where η denotes the configuration for bond i-1 and ς for that of i. $E_{\eta\varsigma i}$ describes the energy contribution of ϕ_{i-1}–ϕ_i bond pair to the overall rotational potential of chain molecule while bond i being in state ς and the bond i-1 in η [50].

The conformational probability of residue Y to be in a specified state is calculated according to the following formula [52]:

$$P_Y(\phi_i, \varphi_i) = N_Y(\phi_i, \varphi_i) / \sum N_Y \tag{3.6}$$

where P_Y (Φ_i, Ψ_i) indicates the ratio of the indicated values of the argument to the total number of observations [52].

The conformational energy of a residue X along the primary sequence of the protein is defined as :

$$E_Y(\phi_i, \varphi_Y) = -RT * \ln\left[\frac{P_Y(\phi_i, \varphi_i)}{P_Y{}^0(\phi_i) P_Y{}^0(\varphi_i)}\right] \tag{3.7}$$

where $P^0{}_Y$ (Φ_i) and $P^0{}_Y$ (ψ_i) are the uniform distribution probabilities. In isomeric state formalism, they are equal to angular intervals of the states (1/12) [52].

The statistical weights can be calculated by using the conformational energies. The relation between the statistical weight for bonds i-1 and i in the state $\eta\varsigma$ and the corresponding energy is given as [50,51]:

$$U_i = u_{\eta\varsigma i} = \exp\left(-\frac{E_{\eta\varsigma, i-1, i}}{R * T}\right) \tag{3.8}$$

The $\eta\varsigma$ ' th element of U_i represents the statistical weight for bond i when it is in state ς whereas bond i–1 is in state η [50].

The partition function, Z, of the chain is given as follows:

$$Z = J * \left[\prod_{i=2}^{2n} U_i \right] J \qquad (3.9)$$

Where $J* = [1\ 0\ 0\ \ldots]$; and J=column $[1\ \ldots\ 1]$ [50].

The probability $P_{\varsigma;i}$ that bond i will be in state ξ is given by the following equation[52]:

$$p_{\varsigma;i} = Z^{-1} J * \left[\prod_{m=2}^{i-1} U_m \right] U'_{\varsigma;i} \left[\prod_{m=i+1}^{2n} U_m \right] J \qquad (3.10)$$

$U'_{\varsigma;i}$ is the matrix obtained by equating all entries of all its columns to zero except those columns of ζ. One can obtain ϕ or ψ dihedral angle distributions separately for a given bond by using equation 3.10.

The joint probability $P_{\eta\varsigma;i-1,i}$ that bond $i–1$ is in state η and bond i is in state ς is given by the following equation [52]:

$$p_{\eta\varsigma;i-1,i} = Z^{-1} J * \left[\prod_{m=2}^{i-1} U_m \right] U'_{\eta\varsigma;i} \left[\prod_{m=i+1}^{2n} U_m \right] J \qquad (3.11)$$

Statistical weight matrices were multiplied in the serial matrix multiplication scheme by using Equation 3.11.

3.7 The geometric criteria used for different hydrogen (H) bond types

Different geometric criteria were used for the investigation of different types of H-bonds formed only within peptides. H-bonds that were formed with water atoms were not taken into consideration. They were examined in two groups as shown in Table 3.4.

Table 3.4 H-bond types with both their corresponding atom groups and geometric criteria

H bond types	Atom groups included in the formation of corresponding H bond		Geometric criteria	
			Length	Angle
Backbone-Backbone	The backbone carbonyl (-CO)& amide (–NH) groups		$r \leq r_{HB}=0.35$nm	$0 \leq \alpha_{HB}=30^0$
Backbone-Side chain	For tyrosine	-OH group of the side chain & -CO or –NH groups of the backbone	$r \leq r_{HB}=0.35$nm	$70 \leq \alpha_{HB} \leq 150$
	For tryptophan	-NH group of the side chain & -CO or –NH groups of the backbone	$r \leq r_{HB}=0.35$nm	$80 \leq \alpha_{HB} \leq 160$

In the above table, r represents the distance between the acceptor and the donor group whereas r_{HB} represents the distance criterion for H-bond formation. In the same manner, α represents the angle between the bond vectors of DH and DA where D is a donor, H is hydrogen and A is an acceptor atom. An imaginary bond vector is assumed to exist among the donor and acceptor atoms.

H-bonds that are formed between the backbone-side chain and side chain-side chain were included only for peptides containing tyrosine or tryptophan amino acids as their

residues. The geometric criteria used for hydrogen bonds that were made between the side-chain and backbone atom groups of tryptophan and tyrosine are taken from a study in which a detailed database survey of corresponding H-bond type was made [53]. As shown in Table 3.4, the 0.35 nm was used as length criterion for all types of hydrogen bonds [52,54]. We used such a relatively big value to cover both weak and strong hydrogen bonds since both of them are important contributors for our model.

3.8 Definition of types of order of interactions

As mentioned in section 3.7, different types of H-bonds were formed between different types of donor and acceptor groups. In the first stage of the categorization, the atoms were labeled as in the Gromos force field with parameter set 53A6 [18] in which only heavy and polar hydrogen atoms were taken into consideration. In the second stage, probable H-bonds were identified. In the third stage, each H-bond was represented with the corresponding notation of X/Y/Z in which each letter represents the corresponding atom number of donor, acceptor, and hydrogen atoms of the given peptide. It is also important to categorize them according to the types of the order of interactions. Then, they were classified depending on the type of order of interactions according to the following scheme: Second order H-bonds were those having two rotatable bonds between the H bonding atoms whereas third order H-bonds were those having three rotatable bonds between the H-bonding atoms where other dihedral angles were kept in trans conformation, and so on. As shown in Figure 3. **6**, although some H-bond types were categorized under the same category as second order, they could be divided further into different groups within the same group depending on the differences in the values $\phi-\psi$ pair of the central residues of given peptides. In the following figure, the two regions on the map shown as A and B indicate regions of highest occurrence of the second order H-bonds. The limits of the regions were those beyond

which the H-bond disappeared. Through the text, unless otherwise specified, H-bonds were those formed between the corresponding backbone atom groups.
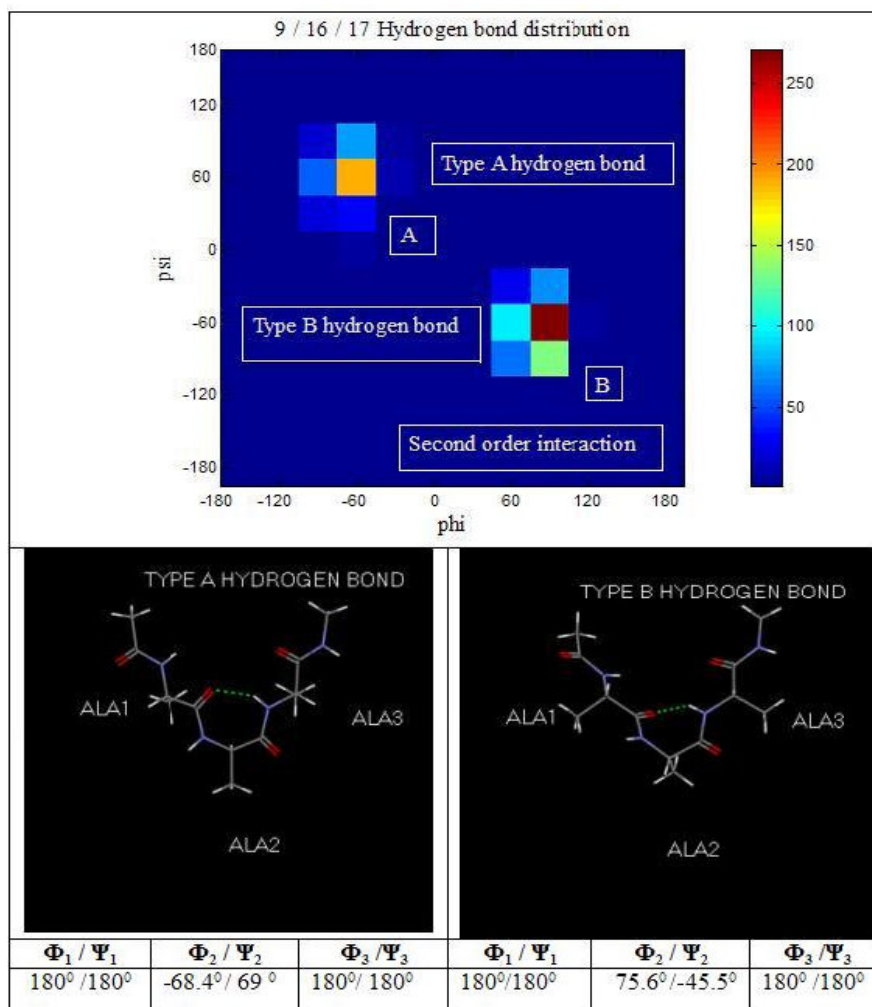

Figure 3. 6 9/16/17 H-bond distribution map in which color distribution represents the number of corresponding H-bond in each grid in accordance with the color bar located at right side of the map. Two lower figures represent the conformation of the core.

3.9 Construction of the coarse-grained model

The following assumptions were made in our coarse-grained model:

- State of a given bond depends on the state of the preceding bond along the peptide backbone (Flory's isolated pair hypothesis which is based on Markov assumption.) In this respect, we assumed that Ramachandran map of a central residue located within an n-length peptide (maximum of n was taken as five in this study) can be represented by the help of information coming from ϕ_i -ψ_i and $\psi_{i-1} - \phi_i$ statistical weight matrices of mono and dipeptides, respectively.

- Each peptide bond prefers to be found in trans conformation.

- Water molecules influence the conformational preferences of capped mono-, di-, and tripeptides in a similar way.

In order to check the applicability of our coarse-grained model on denatured peptides, we simulated corresponding capped mono-, di-, and tripeptides with the same simulation procedure. Each Ramachandran map was divided into grids as described in Section 3.4 to be used as statistical weight matrices in the model. The serial matrix multiplication scheme of the RIS model was suitably carried out using those $\phi_i - \psi_i$ and $\psi_{i-1} - \phi_i$ statistical weight matrices.

3.10 Validation of the model

For demonstration, suppose we want to test the applicability of our model on a capped AVA tripeptide.
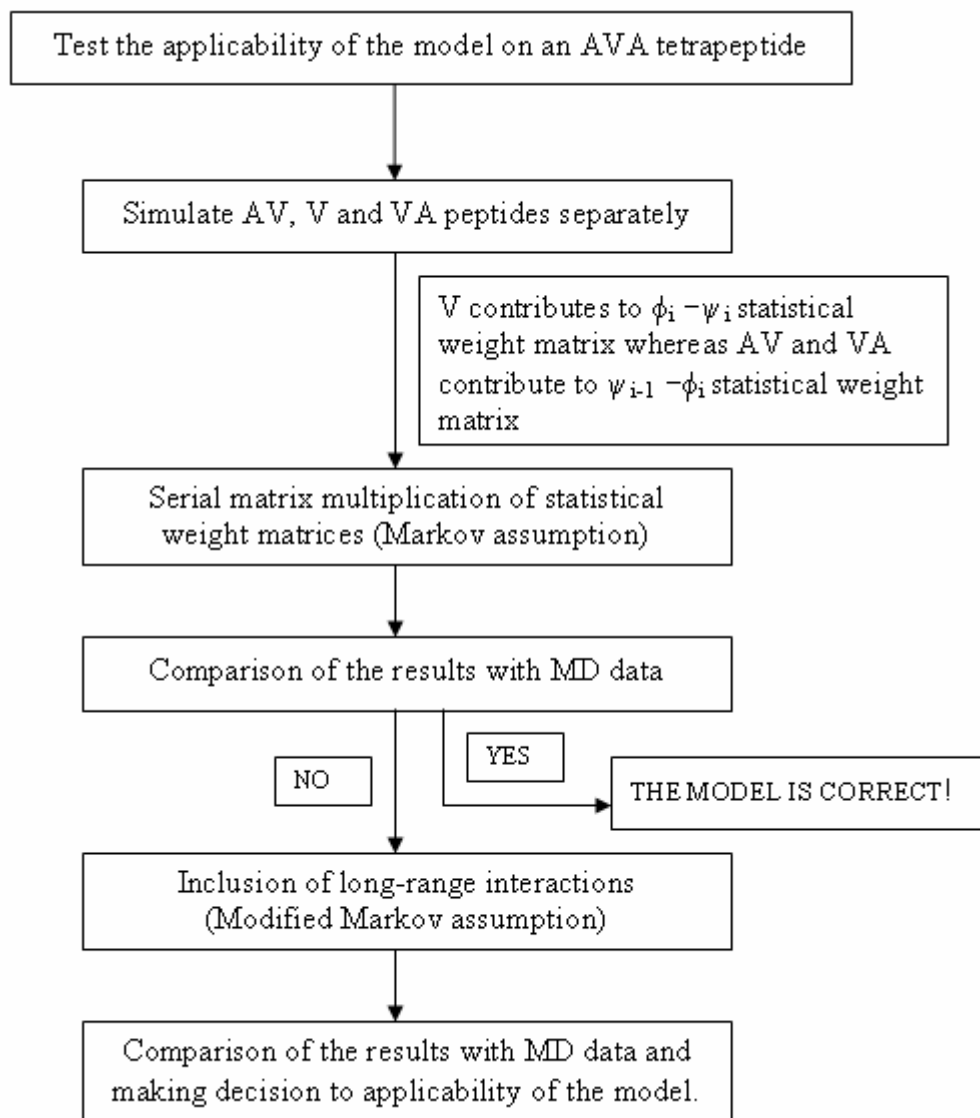

Figure 3.7 Flow chart for testing our model

Chapter 4

RESULTS AND DISCUSSION

As described in Chapter 3, our goal is to develop and test a coarse-grained model based on the Markov assumption for predicting the conformational preference of unfolded short peptide sequences. For demonstration we have looked at the Ramachnadran map of the central amino acid residue of a variety of different tripeptides. This Ramachandran map will be referred to as the target map for any given peptide sequence. In order to obtain the predicted map, serial matrix multiplication scheme of RIS model was carried out by using statistical weight matrices of corresponding mono-, and dipeptides using Equation 3.11. This enables one to check the validity of Flory's isolated pair hypothesis for unfolded peptides. In addition to this, the relative contribution of short- and long-range interactions on the conformational preference of the central residue of tripeptides can be determined.

In this study, we aimed at constructing a coarse-grained model of which assumptions were given in Section 3.9, and since it is based on Markov assumption, it will be referred to as the Markov model in the remainder of the text. The Markov assumption in the context of the current study implies that, the probability distribution of any given dihedral is dependent only on the preceeding dihedral, and is independent of all the other bonds. Therefore, the information conveyed in the statistical weight matrices of mono and dipeptides was used to represent the Ramachandran map of the central residue of tripeptides. We checked the validity of the Markov model on unfolded peptides which gives a chance of determining the relative contributions of short and long-range

interactions on the conformational preference of the central residues of given tripeptides. The Markov model could not represent the whole target map, so we looked for the missing interaction terms. In the Markov model, short-range interaction terms, are conveyed in $\phi_i-\psi_i$ and $\psi_{i-1}-\phi_i$ statistical weight matrices of mono and dipeptides, respectively, were included; however, the interaction terms that were specific to tripeptides, which are named as long-range interaction terms, were not included. Since we aim at representing the target map of central residue of tripeptides, it is thought that incorporation of the long-range interactions to the model may improve the Markov model.

Statistical weight matrices are given as Ramachandran maps through the text, in accordance with their order in the serial matrix multiplication scheme, for both a better visualization and investigation of the extent of information conveyed by them. There were two types of statistical weight matrices: $\phi_i-\psi_i$ and $\psi_{i-1}-\phi_i$. The information related to intra-residue interactions is conveyed in $\phi_i-\psi_i$ matrix, whereas the information related to inter-residue interactions is conveyed in $\psi_{i-1}-\phi_i$ matrix.

We tested the model on the following representative tripeptides, AAA, VVV, WWW, YYY, AVA, AYA, and AWA.

4.1 Statistical weight matrices of capped mono-, di- and tripeptides obtained via MD simulations in explicit water

The target map of AAA tripeptide and statistical weight matrices in the form of Ramachandran maps of mono and dialanine that was used to obtain the target map are shown in Figure 4.1 and Figure 4.2, respectively. Statistical weight matrices and target maps of other tripeptides in the form of Ramachandran map are given in Appendix A.3.
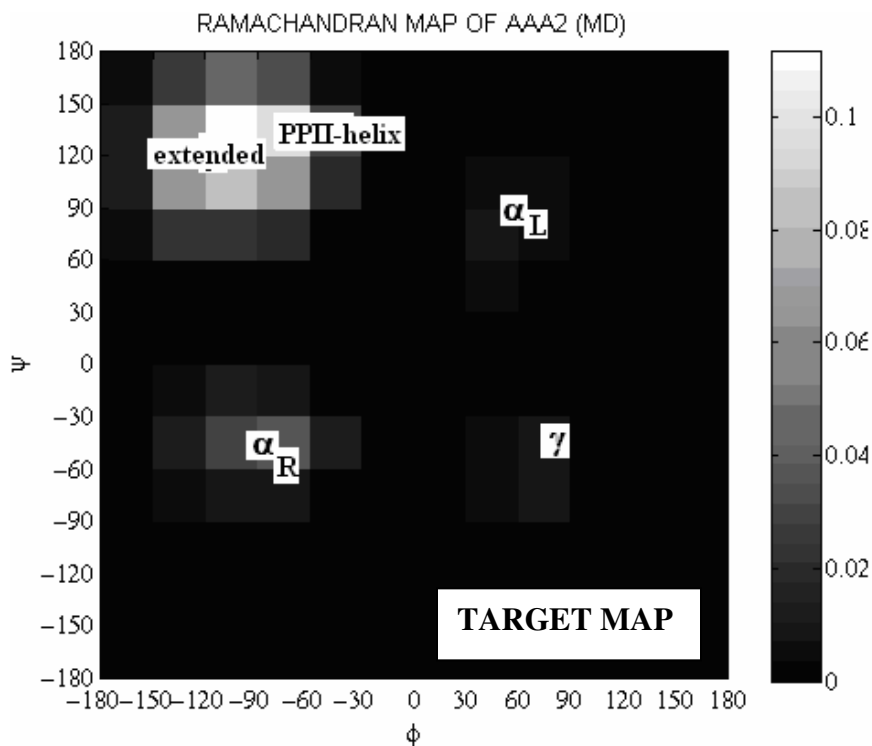
Figure 4.1 The target map of AAA tripeptide (The Ramachandran map of the central alanine of AAA tripeptide) Color represent the probabilities in accordance with the color bar located at the right side of the map

We obtained the Ramachandran map of the central alanine residue of AAA tripeptide, which is named as target map via MD simulations. As seen from Figure 4.1, extended (β) and polyproline-II helix conformations were dominated. The contribution of α_R and γ–turn conformations to the conformational space was smaller. It is well-known from analysis of coil library data that [36],   alanine has a high secondary structure propensity for α_R conformation; however, this could not be reflected by AAA tripeptide in aqueous environment as shown in Figure 4.1 as a result of absence of (i)-(i+4) hydrogen bonds.
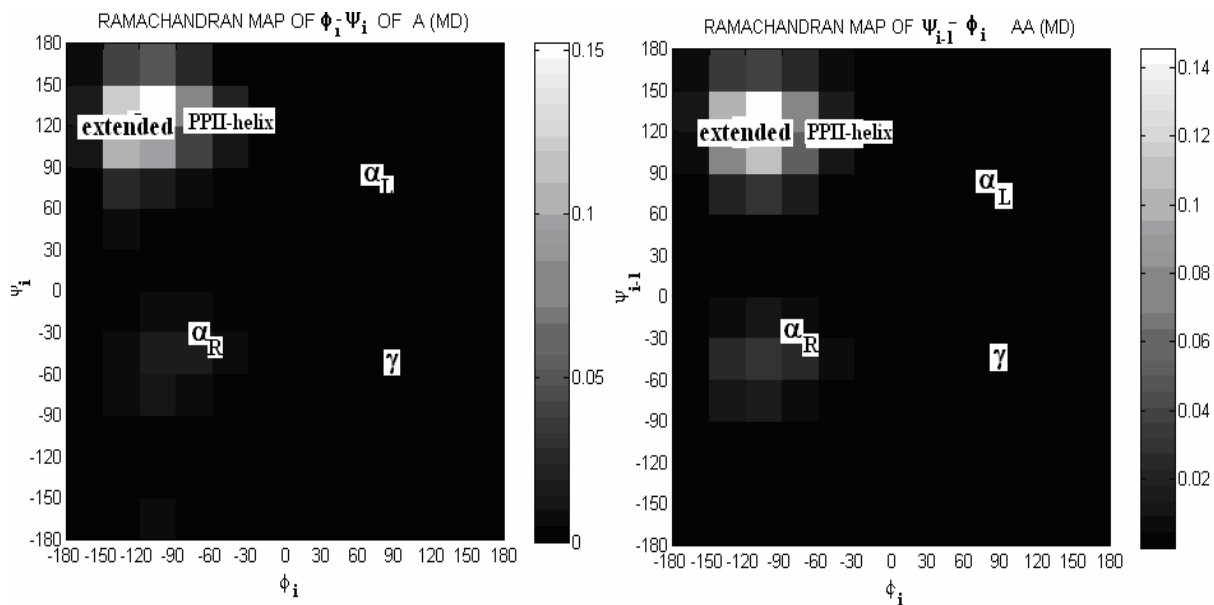
Figure 4.2 Statistical weight matrices of mono and dialanine used for the construction of the target map. Colors represent the probabilities in accordance with the color bar located at the right side of the map.

The Ramachandran maps of mono and dialanine were obtained via MD simulations. Similar to the central residue of AAA tripeptide, mono and dialanine dominated extended and polyproline-II helix conformations. The conformational probability of α_R was smaller compared to that of the central alanine of AAA tripeptide.

MD simulation data showed that there were five different local minima on the Ramachandran maps of peptides; namely: polyproline-II helix, extended, α_R, α_L, and γ–turn; however, they had not contributed to the conformational space of different types of peptides equally. For monopeptides, in general, the extended basin was populated with 65% probability, polyproline-II helix basin was populated with a 20%, α_R basin was populated with 7% probability, γ–turn basin was populated with a 2% probability in explicit water MD simulations. For alanine and valine, γ–turn basin was not dominated. As can be seen from the conformational probability distribution, α_R conformation is not favored in short peptides. This is mostly due to the unfavorable backbone electrostatic energy. Formation of helix favoring are necessary to make α_R conformation favorable; however, water improves the stability of that conformation to some extent [55]. This effect was also seen in the conformational preference of AAAAA pentapeptide in water; the conformational probability of the α_R of central alanine increased relative to those of mono alanine and the central alanine of AAA tripeptide as a result of increase in the number of both (i) - (i+3) and (i)-(i+4) hydrogen bond types.

Among the Ramachandran maps of monopeptides, a difference was observed in their preference for the γ–turn and α_L basin. Even though it is small, tryptophan and tyrosine amino acid residues, which having aromatic side-chains, populated the γ–turn basin with a probability of 2%. On the other hand, no conformational preference was observed for alanine and valine amino acid residues. This observation is in agreement in correspondence with the results of a statistical study of Swindells and coworkers on coil library [36], in which it is stated that amino acid residues those having aromatic side-chains have a larger γ -turn preference.

Conformational probabilities obtained via MD simulations were compared to those obtained via experimental studies of Eker and coworkers [11] in which the dihedral angles

of the central amino acid residues of alanyl-X-alanine tripeptides, where X represents glycine, valine, methionine, histidine, serine, proline, lysine, leucine, tryptophan, tyrosine, and phenylalanine, were investigated in ^2H_2O by using a combined Fourier transform IR, polarized Raman spectroscopy, and vibrational CD measurements of amide I band profiles of mentioned peptides. They showed that valine, phenylalanine, tryptophan, histidine, and serine predominantly adopt an extended β–strand conformation. On the other hand, alanine, methionine, glycine, and leucine populate polyproline-II helix and extended β–strand conformations with comparable probability [11]. In our case, the relative distribution of the two local minima among the capped AXA tripeptides was the same as that of experimental data and shown in Figure 4. **3**. The ratio of the polyproline-II helix to that of extended conformation was higher for AAA tripeptide compared to the other AXA tripeptides. It should be kept in mind that experimental results give information about the most probable state of a given peptide instead of the complete conformational space. Depending on the experimental results, it can be said that, the most probable conformations of the central residues of AXA tripeptides were polyproline-II, and extended. In addition, we obtained the same two structures as dominant conformations via MD simulations. This shows that the selected force-field was able to successfully represent peptides in their unfolded states. However, the relative distribution of conformational probabilities of other basins is important as well, and this requires further experimental evidence.
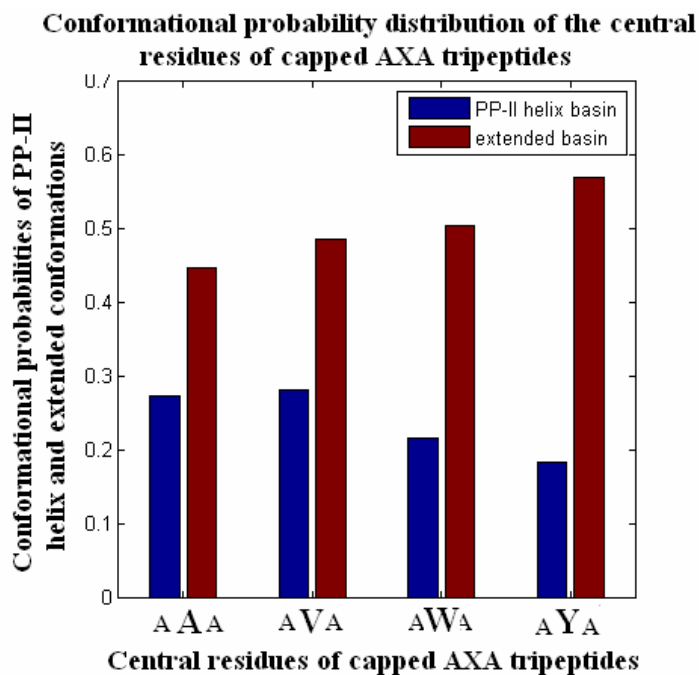
Figure 4. 3 Polyproline-II helix and extended basin probability distribution of the central residues of AXA tripeptides, where X represents alanine, valine, tryptophan, and tyrosine.

We analyzed the differences among the Ramachandran maps of the central residues of homo-peptides. General trend observed in the conformational preferences of these residues was their relative preference for γ–turn basin. As the size of the side-chains of their first nearest neighbors got bigger, the γ–turn basin probability increased. However, γ–turn preferences of monopeptides were smaller than those of the central residues of corresponding tripeptides; AXA and homotripeptides. In addition, the difference was bigger in homo-tripeptides compared to AXA tripeptides. The biggest difference was observed for the tryptophan series among homopeptides. Upon placing a tryptophan residue in the middle of a WWW peptide, the probability of possible nonbonded interactions such as Van der Waals or electrostatic among the atoms of the first and third

residues increase. This leads to the central residue of tripeptides to adopt a twisted (more compact) conformation rather than being in the extended conformation.

In a study by Nemethy and Printz on γ–turns [56], it is argued that in the γ–turn conformation, side-chains of the first nearest neighbors of the central residues extend on the same side of the backbone. This leads to further stabilization of γ–turns for amino acid residues with larger side chains. This observation is in agreement with our results, as described above.

As mentioned above, γ–turn conformation is favored by both electrostatic and Van der Waals interactions, both of which are named as nonbonded interactions. We investigated the relative contributions of those interactions to the γ–turn basin preference of the central residue of tripeptides by both analyzing the distances between the side-chains of amino acid residues and hydrogen bonding profiles of tripeptides. It was observed that when the central residue was in the γ–turn conformation, the distance between the centers of the side-chains populated as shown in smaller distance intervals as shown for WWW tripeptide in Figure 4.4. The distance distribution between the side-chains of the second and the third residues of WWW tripeptide is similar to the distance distribution between the side-chains of the first and second residues, so it is not shown here.
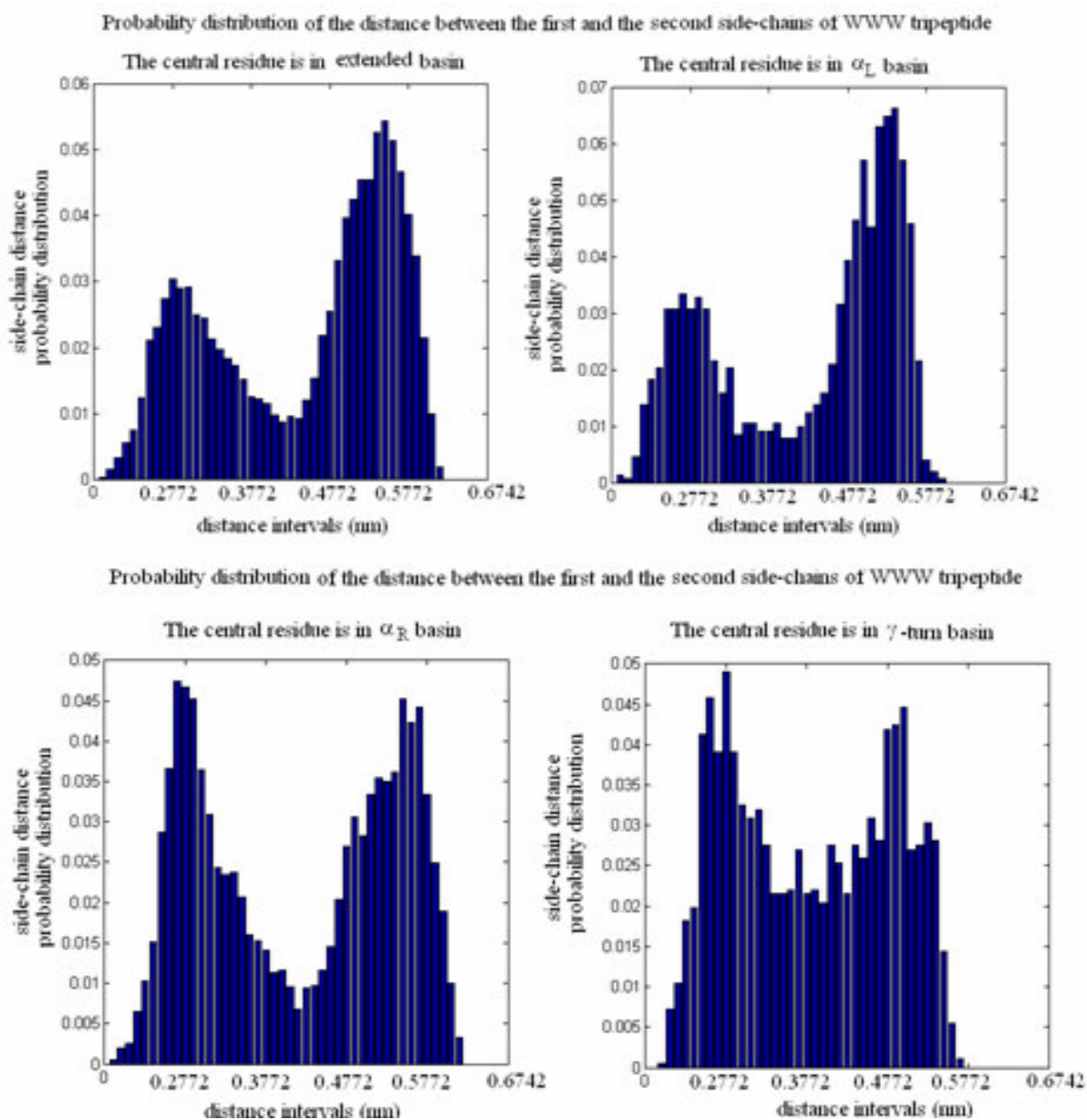
Figure 4.4 Probability distribution of the distance between the side-chains of the first and the second residue of WWW tripeptide

On the other hand, for the other three conformations, the distance between the side-chains was on average larger for homo and AXA tripeptides. We have also observed that, the number of hydrogen bonds of homo-peptides formed when the central residues are in γ–turn conformation is smaller than that of AXA tripeptides. Therefore, one can conclude that the dominant contributor to the γ–turn preference for homo-peptides was hydrophobic interactions. For AXA tripeptides, both hydrophobic interactions, and hydrogen bonding play a role.

Side-chains of valine (V), tyrosine (Y), and tryptophan (W) amino acids.
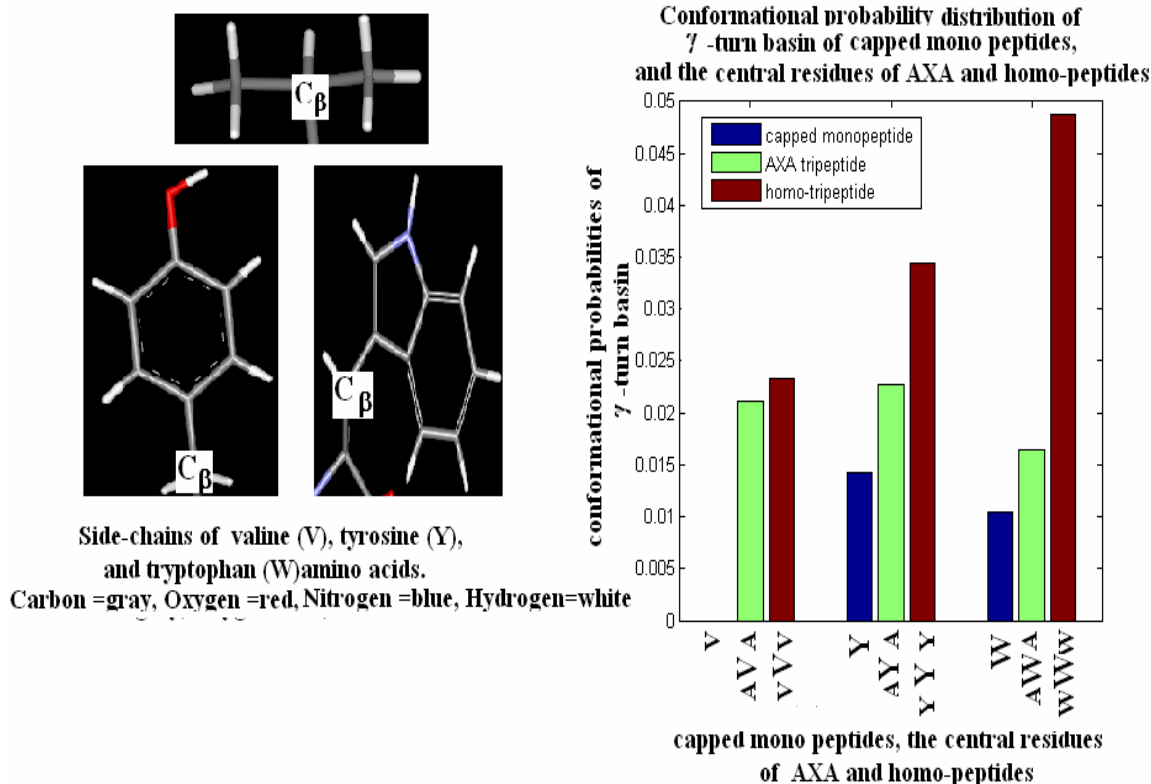Carbon =gray, Oxygen =red, Nitrogen =blue, Hydrogen=white

Figure 4.5 γ-turn basin probability distribution of capped monopeptides, the central residues of AXA tripeptides, where X represents valine, tyrosine, and tryptophan, and homo-peptides

In a study of Baldwin and coworkers [3], they show that the neighboring residue effect on peptide backbone conformation is present in coil library. They express that when a neighboring residue (i-1 or i+1) belongs to class L (aromatic and β−branched amino acids, FHITVWY), rather than class S (other residues, G and P excluded), the ϕ angle of the residue i takes more negative values. Moreover, they investigate the origin of that shifting in the ϕ dihedral angle values, and it is concluded that upon having first nearest neighbors of class L, the central amino acids are shielded from solvent water molecules. This results in a slightly higher preference for the extended conformation compared to the polyproline-II helix conformation.

In order to check the validity of that phenomenon for our tripeptides, we analyzed the ϕ distribution of the central residues of both homo- and AXA peptides. The central residues of tripeptides did not take more negative values, upon having neighbors with aromatic or β-branched side chains. The relative distribution was approximately the same for valine and tryptophan series. For tyrosine the probability of extended conformation decreased approximately 2% upon having neighbors mentioned above. This may be due to the presence of a –OH group on its side-chain. The tyrosine neighbors of the central tyrosine may not shield the central tyrosine efficiently, and interaction with water molecules are favored due to their –OH groups. As a result of the inefficient shielding of the central tyrosine by its first nearest neighbors, this residue may be directed to be in the polyproline-II helix conformation. The probable source of that discrepancy may be the comparison of the coil library data with those obtained via MD simulations. The coil library is constructed by averaging over different environments of amino acid residues within proteins. On the contrary, the conformational preference of amino acid residues in MD simulations is investigated not by averaging over different environments, but when amino acid residues are fully exposed to water molecules.

Another point that should be taken into consideration was that the domination of the polyproline-II helix conformation upon passage from monopeptides to tripeptides. In a study of Jha and coworkers, it was stated that the geometry of the extended conformation places the amide –NH and carbonyl –CO groups in the same plane whereas the geometry of the polyproline-II helix places those groups out of plane [20]. In this respect, upon having neighbors, the central residues start to make interactions with their neighbors leading to an increase in their polyproline-II helix preference. Indeed, we have observed that the number of inter-residue hydrogen bonds of the central amino acid residues of tripeptides increases.

4.2 Testing the validity of Markov model on tripeptides

As shown in Figure 4.6, after the application of Markov model, two of the local minima, extended and polyproline-II helix conformation, were dominated with an approximately 80%, 10% probability, respectively. The contribution of the other three minima, α_R, α_L, and γ–turn, to the conformational space diminished. In addition, as shown in Figure 4.6, the model worked in a similar manner for different types of tripeptides. As explained in detail in Section 4.3, due to exclusion of long-range interactions in the Markov model, which are specific to tripeptides, we could not represent the whole Ramachandran map of the central residues of tripeptides.

In this section, we looked for the missing terms in Markov model, and included them to the corresponding statistical weight matrices to improve the Markov model.
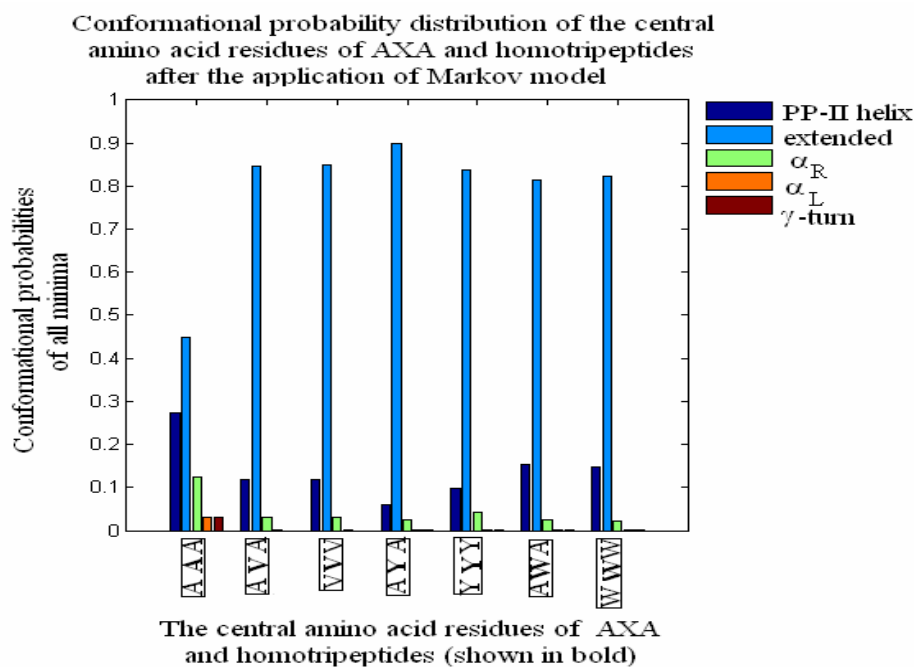
Figure 4.6 Distribution of five local minima after the application of Markov model

4.3 Inclusion of long-range interaction terms to Markov Model

By using information conveyed by the statistical weight matrices of mono and dipeptides, we could not represent the whole target map. Even though extended and polyproline-II helix conformation probabilities are represented successfully, α_R, α_L, and γ–turn local minima are missing from the computed Ramachandran map. The contributions for these three minima to the statistical weight matrices are missing in the mono and dipeptide matrices. Therefore, the final Ramachandran map for a tripeptide obtained via Markov model does not contain a significant probability for these regions. On the other hand, the Ramachandran map for the tripeptide obtained via direct MD simulation clearly contains a finite probability for these three regions. One possible explanation for this major

difference between the Markov results and the actual Ramachandran map, could be the contributions that are not included in the statistical weight matrices of mono and dipeptides. These contributions are long-range interactions which are only present in the tripeptide sequence, but are absent in mono and dipeptides.

We investigated the hydrogen bonding profiles of mono, di, and tripeptides. We observed that the central residues of tripeptides were stabilized in one of these three local minima via hydrogen bonds that were specific to tripeptides. In this respect, one could in principal construct an improved model by incorporating the contributions from these hydrogen bonds to the corresponding statistical weight matrices of mono and dipeptides.

The procedure for inclusion of long-range interaction terms to statistical weight matrices of mono and dialanine of AAA tripeptide is shown in Figure 4.7. Procedures for other peptides are given in Appendix A.4. Ramachandran basins on the energy surface of peptides are shown only for AAA tripeptide. As a first stage in the procedure, hydrogen bonding profiles of mono, and dipeptides were extracted, and compared to those of the corresponding tripeptide. The contributions from interactions, which are only present in tripeptides, have to be explicitly included into the statistical weight matrices. Conformational probabilities were converted to conformational energies by using Boltzmann relation as shown in Figure 4.7. New statistical weight matrices, now, included the contributions from the long-range interaction terms of tripeptides. These modified matrices were used in the serial multiplication scheme of RIS model by using Equation 3.11. The new model which is composed of new statistical weight matrices is named as Modified Markov model.
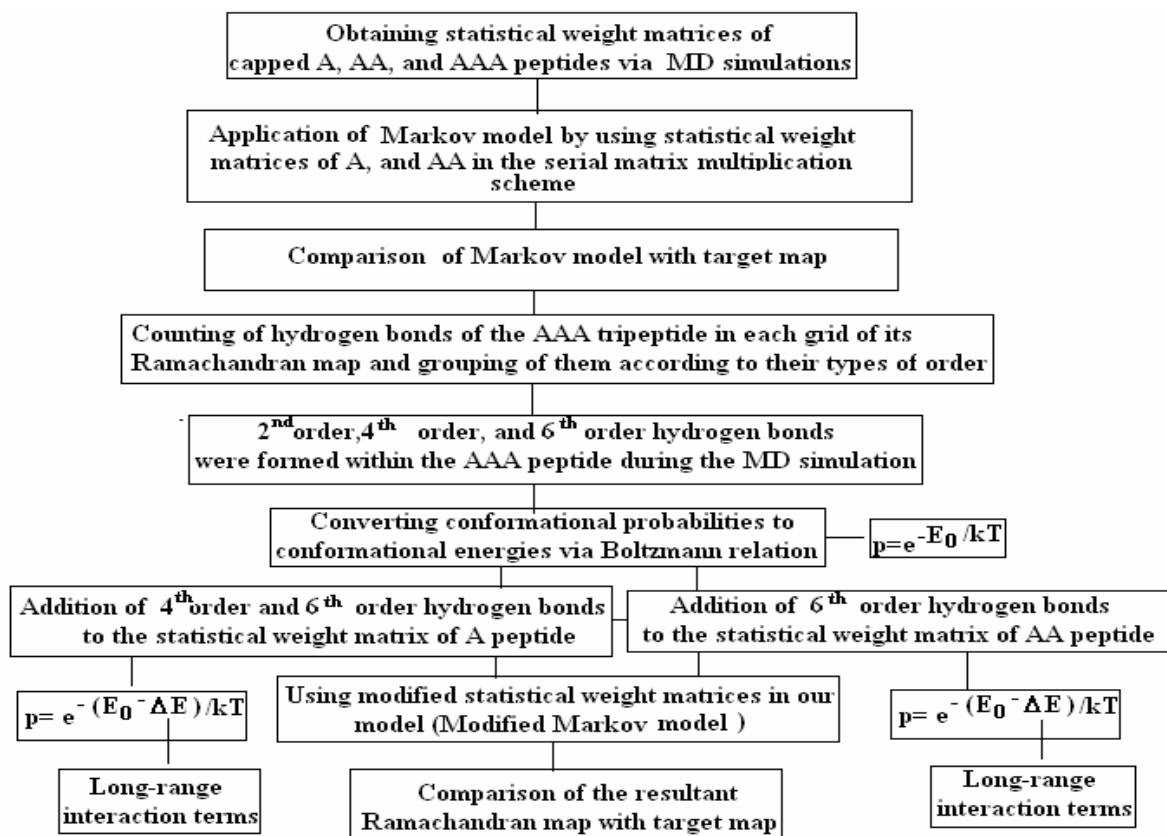
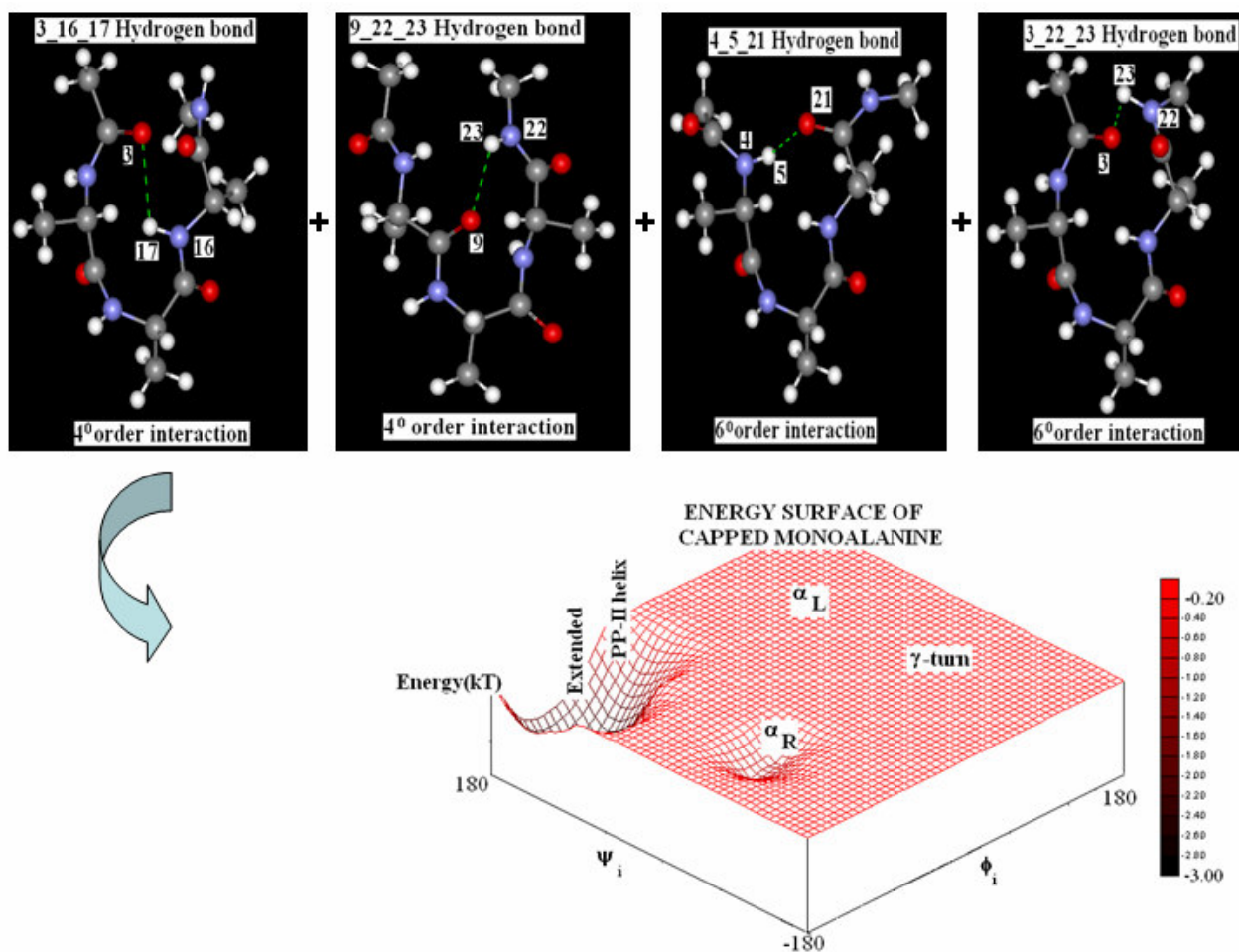Figure 4.7 The flow chart of the procedure followed for the addition of corrections to the Markov model

Figure 4. 8 Inclusion of the corresponding hydrogen bonding terms to the statistical weight matrix of monoalanine (Hydrogen bonds were shown as dotted green lines)

As shown in Figure 4. **8**, 4^0 and 6^0 order hydrogen bonds were specific to AAA tripeptide. In these types of hydrogen bonds, atoms that participated to the formation of the corresponding bond were separated from each other with 4 and 6 rotatable bonds, respectively, so they were not formed within monoalanine whereas 4^0 order hydrogen bonding terms were present in dialanine. 4^0 and 6^0 order hydrogen bonding terms were

added to the statistical weight matrix of monoalanine as shown in Figure 4. 9, whereas only 6^0 order hydrogen bonding terms were added to the statistical weight matrix of dialanine as shown in Figure 4. 9. 4^0 order hydrogen bonds were made via 3_16_17 and 9_22_23 atoms groups whereas 6^0 order hydrogen bonds were made via 3_22_23, and 4_5_21 atoms groups of AAA tripeptide.
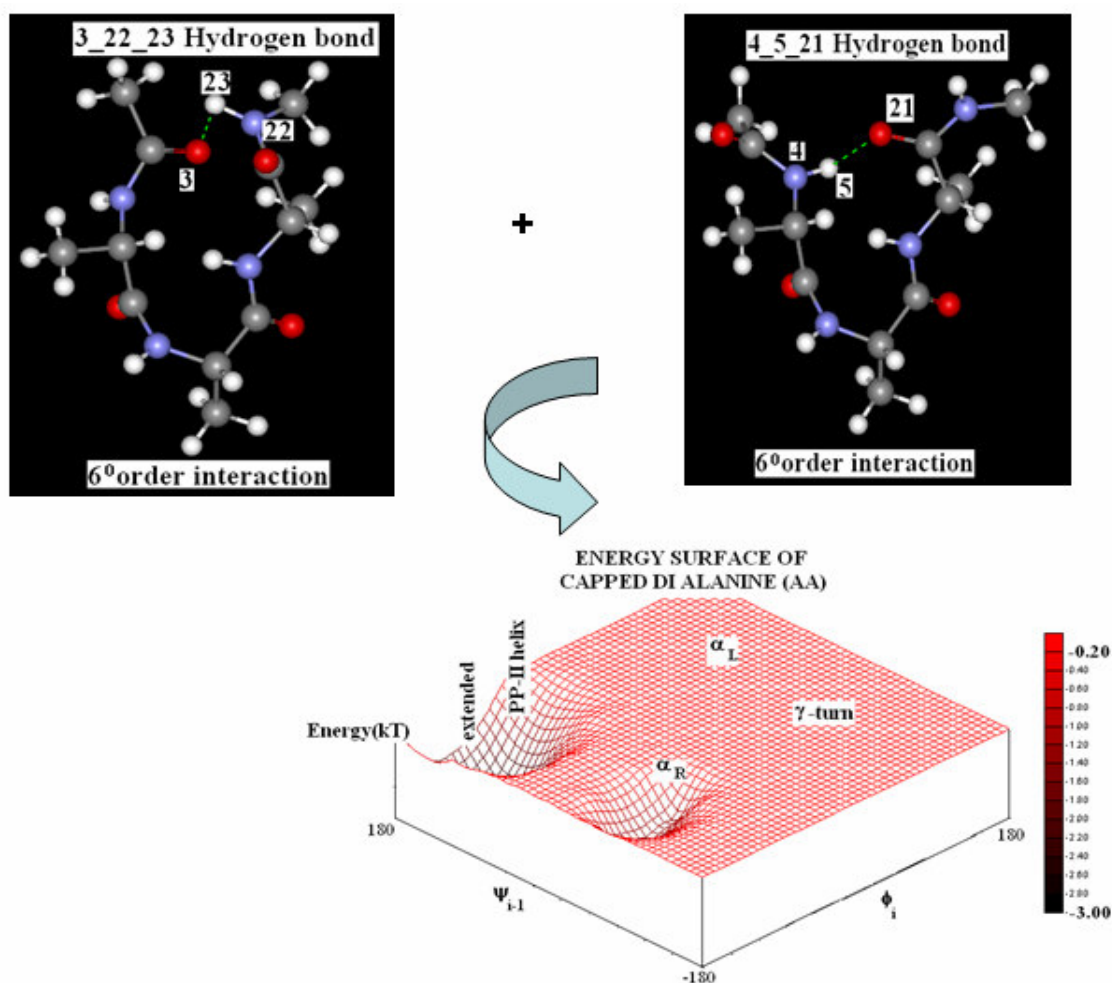

Figure 4. 9 Inclusion of the corresponding hydrogen bonding terms to the statistical weight matrix of dialanine (Hydrogen bonds were shown as dotted green lines)

For AVA tripeptide, 4^0 and 6^0 order hydrogen bonding terms were added to statistical weight matrices of monoalanine and monovaline; however, only 6^0 order hydrogen bonding terms were added to statistical weight matrices of AV and di VA dipeptides since 4^0 order hydrogen bonds could be made via atom groups of those dipeptides.

For VVV tripeptide, 4^0 and 6^0 order hydrogen bonding terms were added to statistical weight matrix of monovaline; however, only 6^0 order hydrogen bonding terms were added to statistical weight matrix of valine dipeptide since 4^0 order hydrogen bonds could be made via atom groups of that dipeptide as in AV and VA dipeptides.

For AYA tripeptide, 4^0 and 6^0 order hydrogen bonding terms were added to statistical weight matrix of monoalanine whereas only 6^0 order hydrogen bonding terms were added to statistical weight matrix of monotyrosine. AYA could make 4^0 order hydrogen bonds via its backbone atom groups. In this case, no additional terms were added to statistical weight matrices of AY and YA dipeptides since they could form both types of hydrogen bonds.

For YYY tripeptide, 6^0 and 8^0 order hydrogen bonding terms were added to statistical weight matrix of monotyrosine whereas only 8^0 order hydrogen bonding terms were added to that of tyrosine dipeptide.

For AWA tripeptide, 3^0, 4^0, and 5^0 order of hydrogen bonding terms were added to statistical weight matrix of monoalanine whereas 4^0 and 5^0 were added to that of monotryptophan; however, only 5^0 order of hydrogen bonding terms were added to those of WA and AW dipeptides.

For WWW tripeptide, 4^0, 5^0, 6^0, and 7^0 order of hydrogen bonding terms were added to statistical weight matrix of monotryptophan whereas 6^0, and 7^0 order of hydrogen bonding terms were added to that of WW dipeptide.

In the following figures, the comparison of MD data with Markov and Modified Markov model is shown for all tripeptides.

As can be seen, before the inclusion of long-range interaction terms to Markov model, the extended conformation was dominated with an amount of approximately 85% probability for all types of tripeptides. In Markov model, to represent the target map, we used information conveyed by the statistical weight matrices of $\phi_i-\psi_i$ and $\psi_{i-1}-\phi_i$ of mono and dipeptides, respectively. In those matrices, the highest contribution to conformational space came from both extended basin. On the other hand, α_L, γ-turn basins were populated with a small probability value which explained the absence of those basins via Markov model. Mono and dipeptides populated α_R basin with a probability that was close to that of the central residues of tripeptides; however, we could not represent that basin via Markov model. The statistical weight matrices of mono and dipeptides could not transmit the information of α_R conformation to the model since the central residues of tripeptides were directed to be found in α_R conformation via long-range hydrogen bonds. Extended conformation was successfully represented via Markov model, because it was formed via short-range interactions, which places the amino and carboxyl group of amino acid residues in same plane [20].
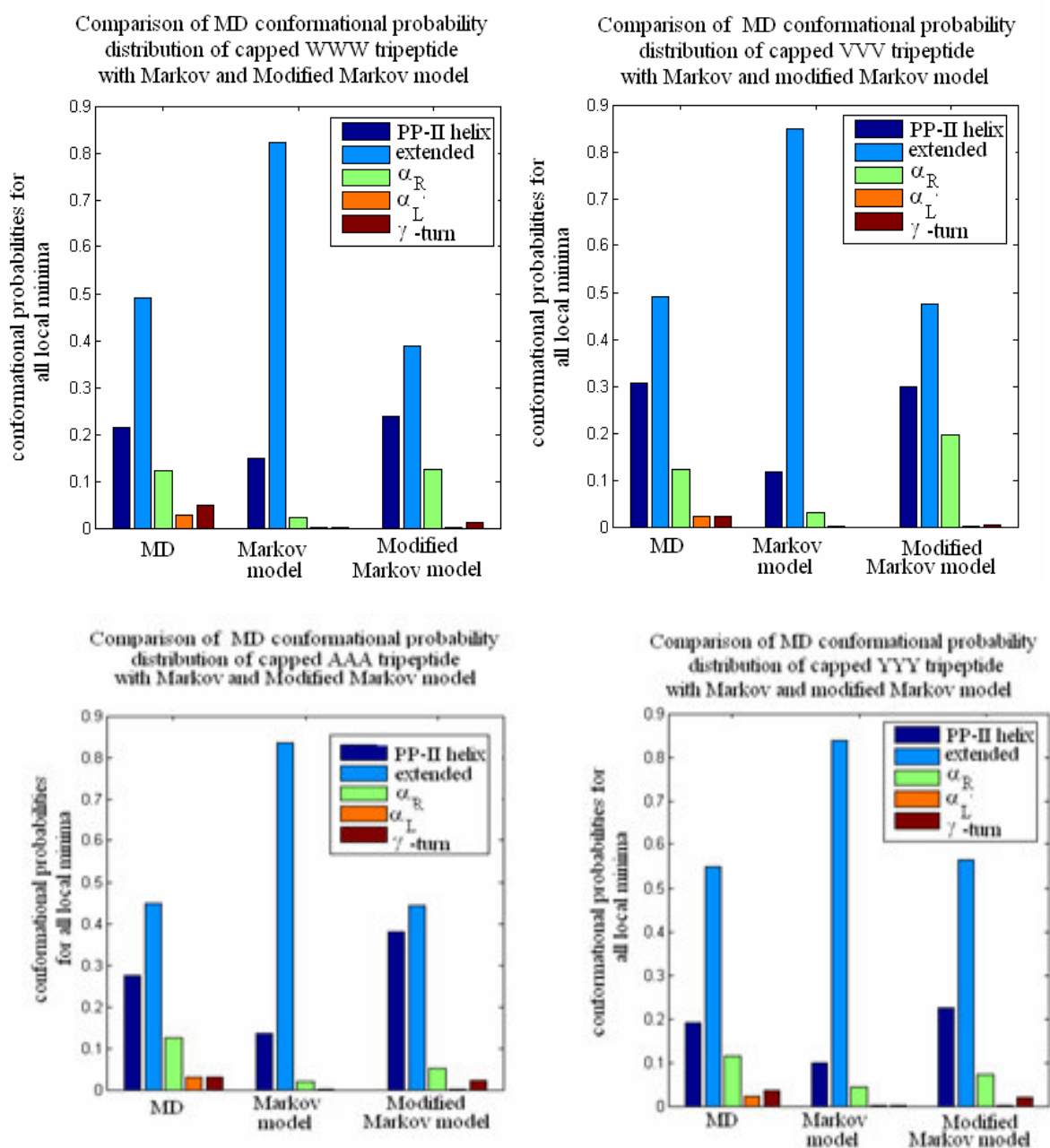
Figure 4.10 Comparison of MD conformational probability distribution of homotripeptides with Markov and Modified Markov model
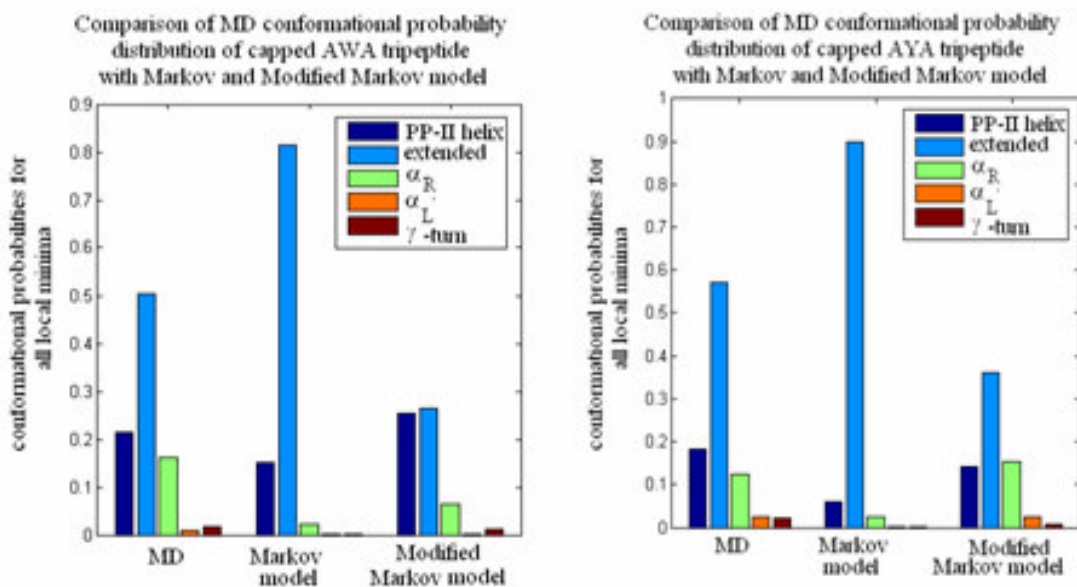
Figure 4.11 Comparison of MD conformational probability distribution of AXA tripeptides with Markov and Modified Markov model

After the inclusion of long-range interaction terms, which were specific to tripeptides, we were able to obtain a map that was similar to the target map. It should be kept in mind that we included only contributions from hydrogen bonding terms, but not other nonbonded interactions, such as hydrophobic interaction terms. Influence of the exclusion of hydrophobic interaction terms was reflected in the conformational probability of γ–turn of tripeptides. That basin could be represented with a similar probability to that of target maps for the central residues of AXA tripeptides. However, it could not be represented with a similar probability for the central residues of homopeptides because γ–turn preference of the central residues of homo-tripeptides was constructed dominantly by hydrophobic interaction terms rather than hydrogen bonding terms [56].

4.4 Solvent influences the conformational preferences of amino acid residues

In order to investigate the changes on the conformational preferences of the central residues of homo and AXA tripeptides, we performed simulations in hydrophobic medium which is composed of cyclohexane molecules. In general, the number of intra- and inter-residue hydrogen bonds of mono and both types of tripeptides increased in hydrophobic medium due to the replacement of water–peptide hydrogen bonds with intra-peptide hydrogen bonds in such a hydrophobic environment. In addition, we investigated the hydrogen bonding patterns of peptides formed in both media, and compared the relative numbers of the 4^0 and 6^0 order hydrogen bonds. It was observed that the former was favored with a higher probability in hydrophobic medium whereas the latter was favored in aqueous medium.

For monoalanine and the central alanine of trialanine, passage from aqueous environment to hydrophobic one resulted with an increase of the probability of the extended conformation. The change is more drastic for monoalanine, since it has very limited hydrogen bonding capabilities on its own. In this respect, monoalanine made intra-residue hydrogen bonds which are possible in the extended conformation, since this places amino and carboxyl groups in the same plane. However, the probability of polyproline-II helix conformation of the central alanine of pentaalanine was increased since AAAAA has significantly more hydrogen bond candidates. the due to having more neighbors. In this case, pentaalanine made inter-residue hydrogen bonds at a higher rate which was favored by polyproline-II helix since it places amino and carboxyl groups out of plane, giving a chance for making inter-residue hydrogen bonds. As was the case in aqueous medium, the conformational probability of α_R of alanine was higher in pentaalanine since (i)-(i+4) hydrogen bonds could be made. As shown in Figure 4.12, the γ–turn probability of mono, and the central alanine of tri- and pentaalanine was higher in hydrophobic medium

compared to that obtained in aqueous medium. In this case, as indicated above, the main contribution to the formation of that conformation came from hydrogen bonding terms as was for AXA peptides in aqueous medium.
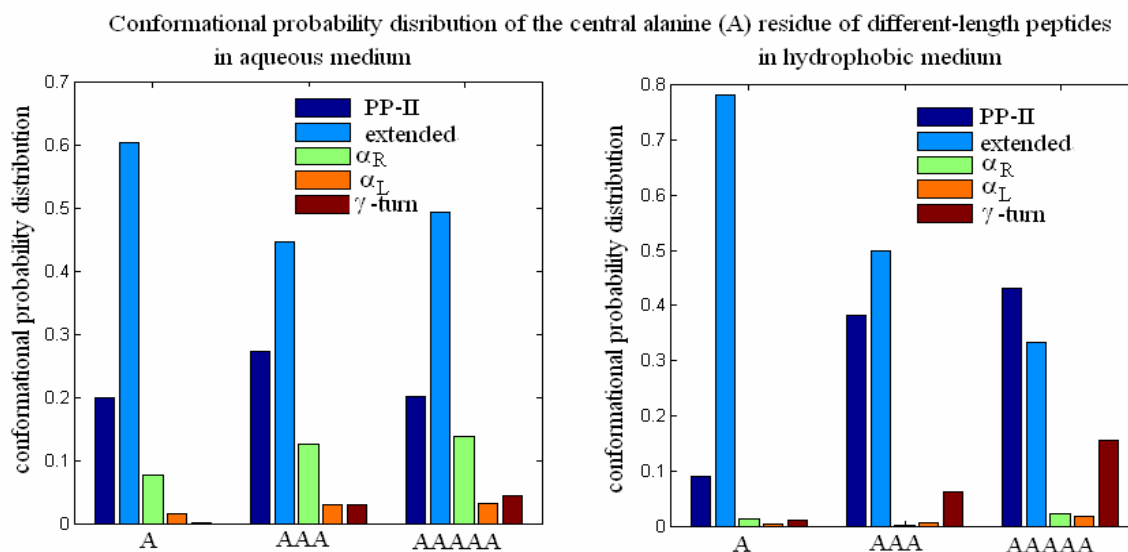

Figure 4.12 Conformational probability distribution of the central alanine of different-length peptides in both aqueous (left) and hydrophobic medium (right)

For other types of both homo and AXA tripeptides, the pattern of conformation probability distribution was similar to that observed in aqueous medium. The two dominant conformations were extended and polyproline-II helix which indicated the dominance of both intra- and inter-residue hydrogen bonds as in the case of alanine series. In addition, the contributions of those two conformations to the conformational space increased whereas the contribution of other three conformations decreased in hydrophobic medium.

4.5 Radial distribution function of alanine and tryptophan amino acid residues

Another important contribution to peptide contribution comes from interactions among amino acids which are separated along the chain but come into contact during

conformational search. Interactions among residues which belong to different chains can also be considered under the same category. In this section, our preliminary results on such interactions are presented.

One possible representation for such contacts in the context of a coarse-grained model, is to obtain the potential of mean force (PMF) among different residues. PMF can be obtained by Boltzman inverting the radial distribution function of the target molecules in a medium. In this study, we constructed the the radial distribution function (RDF) of alanine and tryptophan amino acid residues relying on the distance distribution data obtained via MD simulation. These simulations were performed with a box full of capped alanine (tryptophan) residues, where the details of the simulation were described in detail in Section 3. In Figure 4.13, the RDFs for the C_α atoms of alanine (tryptophan) are shown. The amino acid residues alanine and tryptophan are highly asymmetric objects. The side-chain breaks the symmetry of the molecules around the backbone axis. Therefore, a full description of the relative conformational preference of these residues would require a much more detailed selection for the representative degrees of freedom. Here, we will use the simplest representation (i.e, C_α atoms), which we hope will provide a basic understanding of the nature of such interactions.

The RDF shows a peak at approximately 0.57 nm for both alanine and tryptophan residues. This marks the thickness of the first coordination shell for both residues. RDF for alanine shows a small peak at roughly double the distance, which is likely to be with the second coordination shell. For tryptophan a second peak is observed at 0.73 nm, which is completely absent from the RDF of alanine. Alanine has a much smaller side-chain compared to tryptophan. Due to the small size of alanine side-chain, the symmetry of the molecule around the backbone is only slightly perturbed. However, the presence of side-chain does not lead to a second peak in the RDF. On the other hand, the bulky side-chain of tryptophan clearly distorts the symmetry, and leads to the second peak in RDF.

As explained above, the PMF can be obtained by Boltzmann inverting the RDF. Therefore, the height and the sharpness of the RDF peaks, determine the potential energy well's depth and shape for the corresponding residue. The RDF peak of alanine which is relatively higher compared to tryptophan, indicates that alanine residues are bound to each other more strongly compared to tryptophan residues. The well defined nature of the alanine peak, also suggests that, alanine residues have a more well-defined characteristics distance among each other, compared to tryptophan residues.
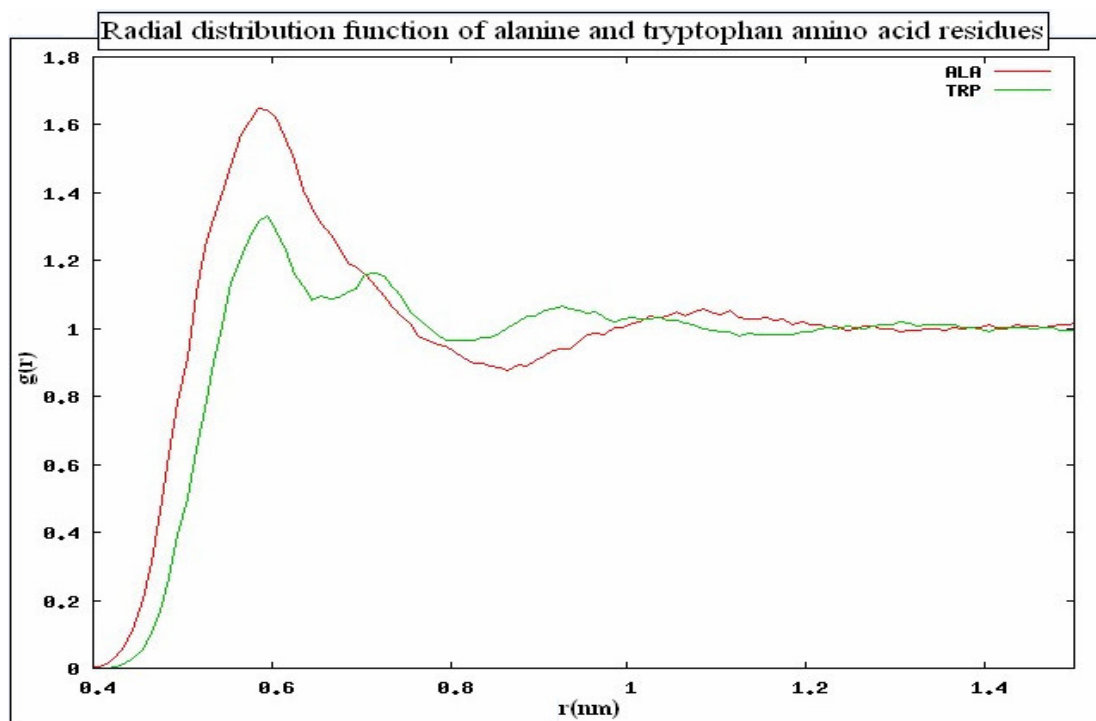
Figure 4.13 Radial distribution function of alanine and tryptophan amino acid residues constructed by including Cα atoms.

Chapter 5

CONCLUSION

Protein folding problem is a reversible disorder-order transition; consequently the conformational preferences of amino acid residues in both unfolded and folded states of peptides must be described precisely. In this respect, we investigated the conformational preference of alanine, valine, tryptophan, and tyrosine in homo and AXA tripeptides, where X represents alanine, valine, tryptophan, and tyrosine amino acid residues, via MD simulations. Since almost all force-fields have been tuned to represent the native state of proteins, the selection of an appropriate force-field to represent the unfolded state was a crucial step in our study. In a study of Avbelj and Baldwin, it was stated that in order to make a force-field succeed with unfolded peptides, the solvation free energies of amides should be reproduced accurately [18]. We used Gromos force-field with a parameter set of 53A6 since it has been adjusted to take account of our problem, and gave similar results to those obtained in an experimental study of Eker and coworkers [11]. However, in order to wander the conformational space efficiently, we used the simulated annealing method.

MD simulation results in explicit water demonstrated that alanine, valine, tryptophan, and tyrosine residues, which are collectively named as monopeptides, adopted dominantly to extended and polyproline-II helix conformations. The difference was in their γ–turn preferences. Tryptophan and tyrosine residues, which have aromatic side-chains, showed a higher preference for γ–turn conformation. On the other hand, alanine and valine did not

show any preference for γ–turn. These results were consistent with data extracted from coil library due to the placement of tripeptides in an environment, in which the turn structures found in native proteins; on the surfaces of proteins. The difference among the conformational preferences of the central residues of AXA and the conformational preferences of the central residues of homopeptides was in γ–turn conformation. There were two contributors that governed the formation of such conformations; hydrophobic and electrostatic interactions. For the central residue of AXA peptides, both hydrophobic and electrostatic interactions played role for the formation of γ–turn conformation. For the central residues of homopeptides hydrophobic interactions mostly governed the formation of the γ–turn conformation, since the first and the third residues had larger side-chains relative to alanine's.

Avbelj and Baldwin stated that the neighboring residue effect is observed in coil library when a neighboring residue (i+1, i-1) belongs to class L, which consists of aromatic and β–branched amino acids, then the ϕ of residue i is more negative for essentially all amino acid residues. This effect was only observed when AVA tripeptide was altered to VVV tripeptide, but it was not observed for tryptophan and tyrosine series. It could be due to the comparison of data that reflected the characteristics of different environments. In coil library, such sequences can be found on the surface or interior parts of proteins which lead to the averaging out of different environmental effects whereas the peptides were exposed to water molecules in our MD simulations. Comparison of coil data with those of MD gives reliable results, provided that data were collected for residues with similar environments. However, our results were consistent with experimental data of Eker and coworkers, in which the conformational preferences of AXA tripeptides were investigated experimentally, since peptides were exposed to similar environments. In this respect, having similar results with those of experiments indicated the success of the force-field for the representation of unfolded peptides.

We also performed MD simulations in hydrophobic media to observe the influence of environment on the conformational preferences of amino acid residues of short peptides. The conformational distribution was similar to that obtained in aqueous medium; however, the contribution of both extended and polyproline-II helix conformations increased whereas the contribution of other three decreased indicating that in hydrophobic medium short-range interactions were favored relative to long-range interactions.

We observed that changing of the first nearest neighbors of the central amino acid residues of AXA tripeptides did not influence the conformational preferences of the central residues. Therefore, absence of the nearest neighbor effect on short unfolded peptides motivated us to think as Flory's isolated-pair hypothesis is valid for short peptides, and we thought that it was possible to represent the Ramachandran map of the central residues of tripeptides by using information of mono and dipeptides. In this respect, we constructed a model based on Markov assumption. This model revealed that the extended conformation , in which the intra-residue interactions are dominant, is overrepresented. α_R conformation was not represented by Markov model although that conformation was populated in the statistical weight matrices of mono and dipeptides. The statistical weight matrices of mono and dipeptides could not transmit the information of α_R conformation during the serial matrix multiplication since the central residues of tripeptides were directed to be found in α_R conformation via long-range hydrogen bonds. However, they were directed to be found in that conformation via short-range interactions in mono and dipeptides. In this respect, we included the contributions from the long-range interaction terms, which were specific to tripeptides, to corresponding statistical weight matrices of mono and dipeptides. This Modified Markov model includes not only short-range interactions, but also the long-range interaction terms, and is capable of representing also the α_R conformation. However, the Modified Markov model did not show the same performance for the representation of γ–turn preferences of AXA and homopeptides. We believed that this is due to the different

origin of the driving forces for these basins. Hydrophobic and electrostatic interactions governed the conformational preference of AXA tripeptides whereas mostly hydrophobic interactions governed the conformational preference for homotripeptides.

The Modified Markov model still excludes the nonbonded interactions formed among the water molecules and peptides. Moreover, other types of nonbonded interaction terms such as hydrophobic and van der Vaals interactions, which were specific to tripeptides, were not included in the model. By incorporation of those criteria to the model, we may represent the conformational preferences more reliably, and those statistical weight matrices can be used to represent the Ramachandran map of central residues of longer peptides.

In conclusion, our results are in agreement with of the statistical coil model, rather than the random coil model. Well-defined conformations emerge locally in unfolded peptides. In agreement with Flory's isolated-pair hypothesis providing that unfolded peptides are found in extended or polyproline-II helix conformation. However, for other three basins, α_R, α_L, and γ–turn, the hypothesis is not valid since they are formed via long-range interactions that are specific to tripeptides.

APPENDIX
A.1 TOPOLOGY FILE OF –NME CAP

[NME]
[atoms]

  N   N -0.280   0

  H   H 0.28000 0

  CA CH3 0.00000 1

[bonds]

  N   CA gb_20

  N   H   gb_2

[angles]

 -C   N   H ga_31

  H   N   CA ga_17

 -C   N   CA ga_30

[impropers]

  N -C CA   H gi_1

[dihedrals]

 -CA -C   N   CA gd_4

A.2 TOPOLOGY FILE OF CYCLOHEXANE MOLECULE

;include "ffG53a6.itp"

[moleculetype]
; name nrexcl
CHE   3

[atoms]
; nr type   resnr residu atom   cgnr   charge   mass
1   CH2   1   CHE   C1   1   0.0 14.027
2   CH2   1   CHE   C2   1   0.0 14.027
3   CH2   1   CHE   C3   1   0.0 14.027
4   CH2   1   CHE   C4   1   0.0 14.027
5   CH2   1   CHE   C5   1   0.0 14.027
6   CH2   1   CHE   C6   1   0.0 14.027

[bonds]
; ai aj funct   c0   c1
1   2 2   gb_26
2   3 2   gb_26
3   4 2   gb_26
4   5 2   gb_26
5   6 2   gb_26
6   1 2   gb_26

[pairs]
1 4
2 5
3 6

[angles]
; ai   aj   ak   funct c0   c1
  1   2   3   2   ga_14
  2   3   4   2   ga_14
  3   4   5   2   ga_14
  4   5   6   2   ga_14
  5   6   1   2   ga_14
  6   1   2   2   ga_14

[dihedrals]

1	2	3	4	1	gd_17
2	3	4	5	1	gd_17
3	4	5	6	1	gd_17
4	5	6	1	1	gd_17
5	6	1	2	1	gd_17
6	1	2	3	1	gd_17

A.3 Statistical weight matrices of mono, di, and tripeptides obtained via MD simulations in explicit water

Φ_i - ψ_i and ψ_{i-1} - ϕ_i statistical weight matrices in the form of Ramachandran
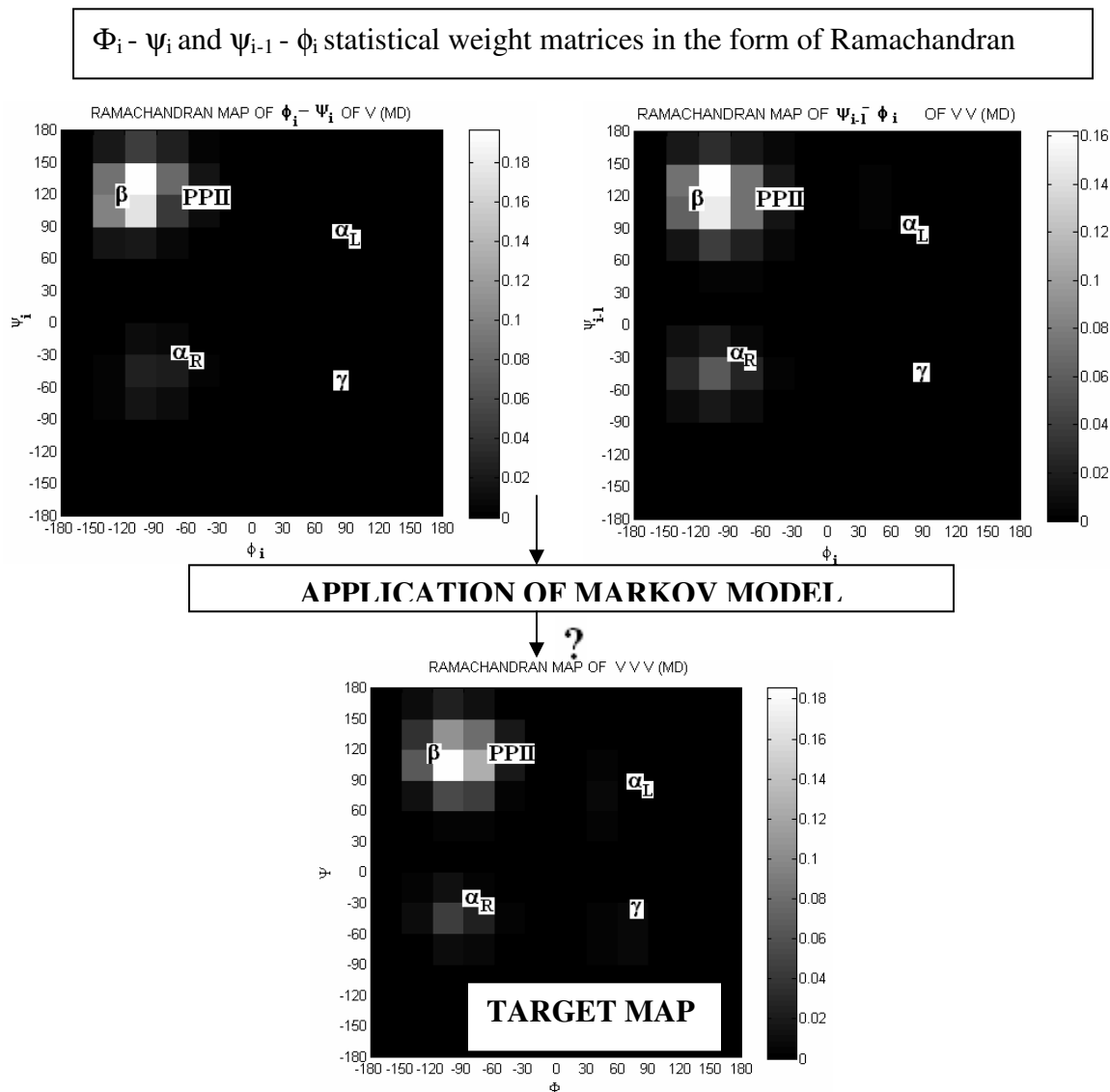
Figure A.3 2 Statistical weight matrices used for the construction of the target map and the target map of the central valine. Colors represent the probabilities in accordance with the color bar located at the right side of the map.

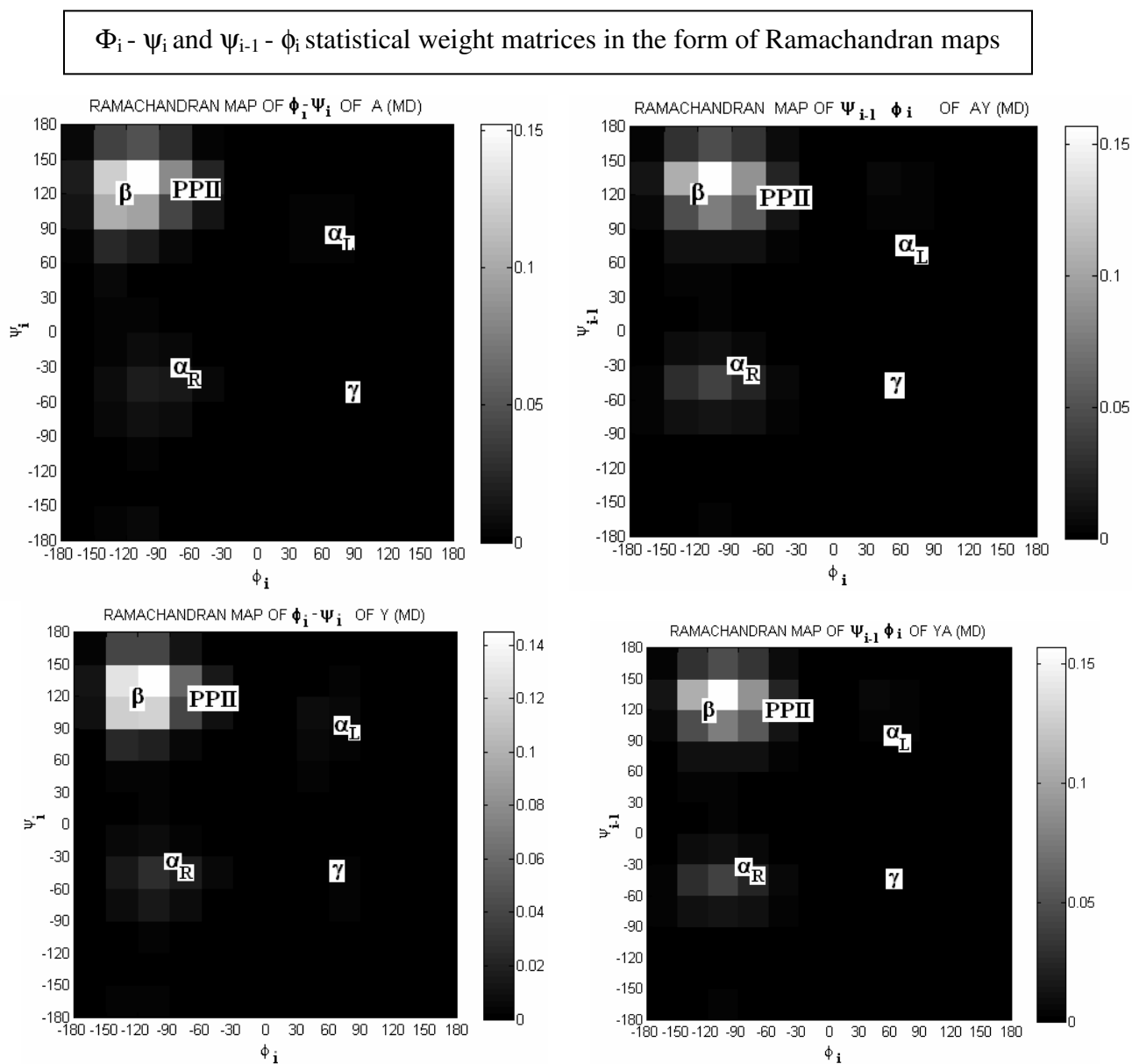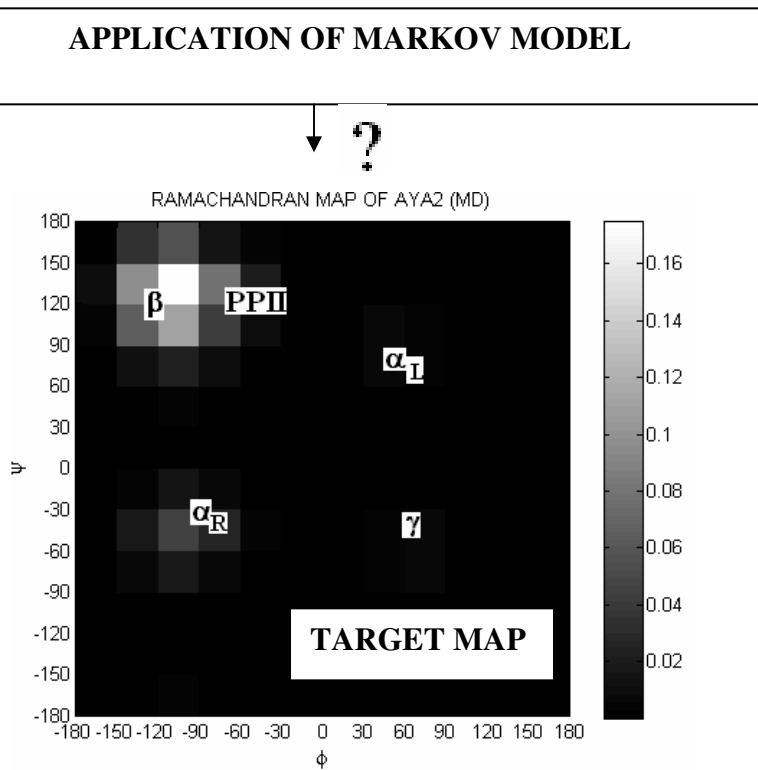Φ_i - ψ_i and ψ_{i-1} - ϕ_i statistical weight matrices in the form of Ramachandran maps
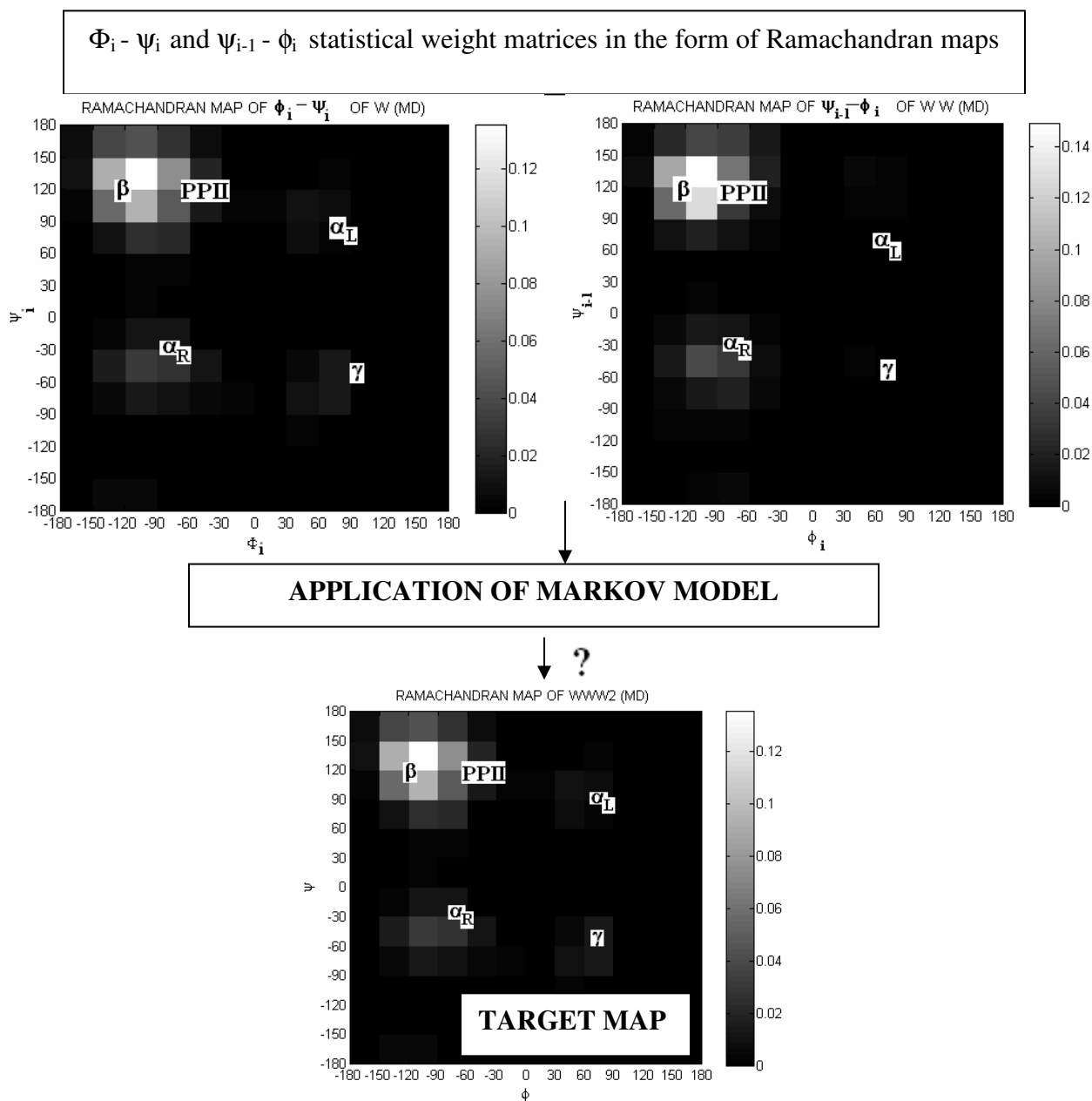
Figure A.3 3 Statistical weight matrices used for the construction of the target map and the target map of the central tryptophan. Colors represent the probabilities in accordance with the color bar located at the right side of the map.

Φ_i - ψ_i and ψ_{i-1} - ϕ_i statistical weight matrices in the form of Ramachandran maps
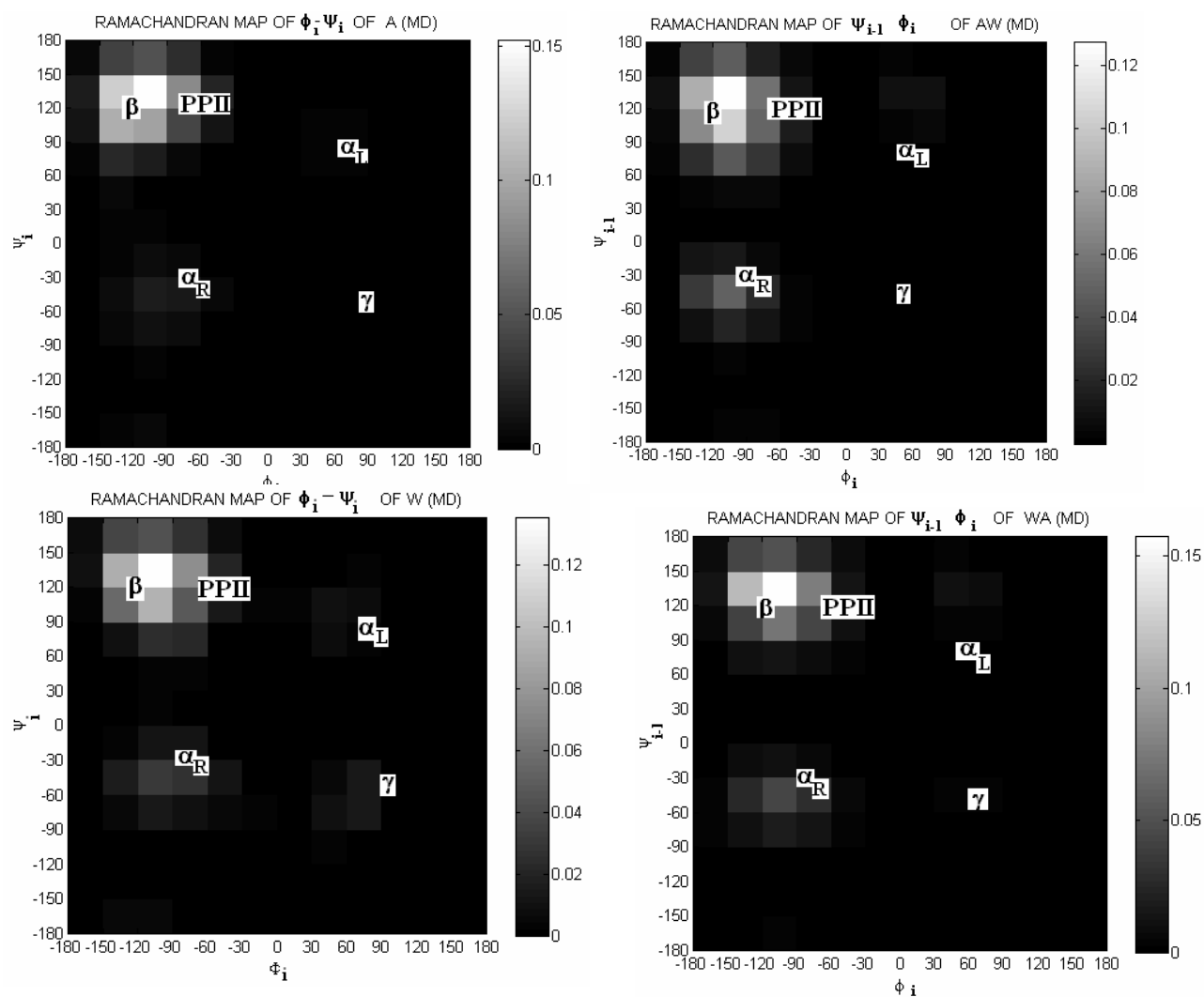
APPLICATION OF MARKOV MODEL
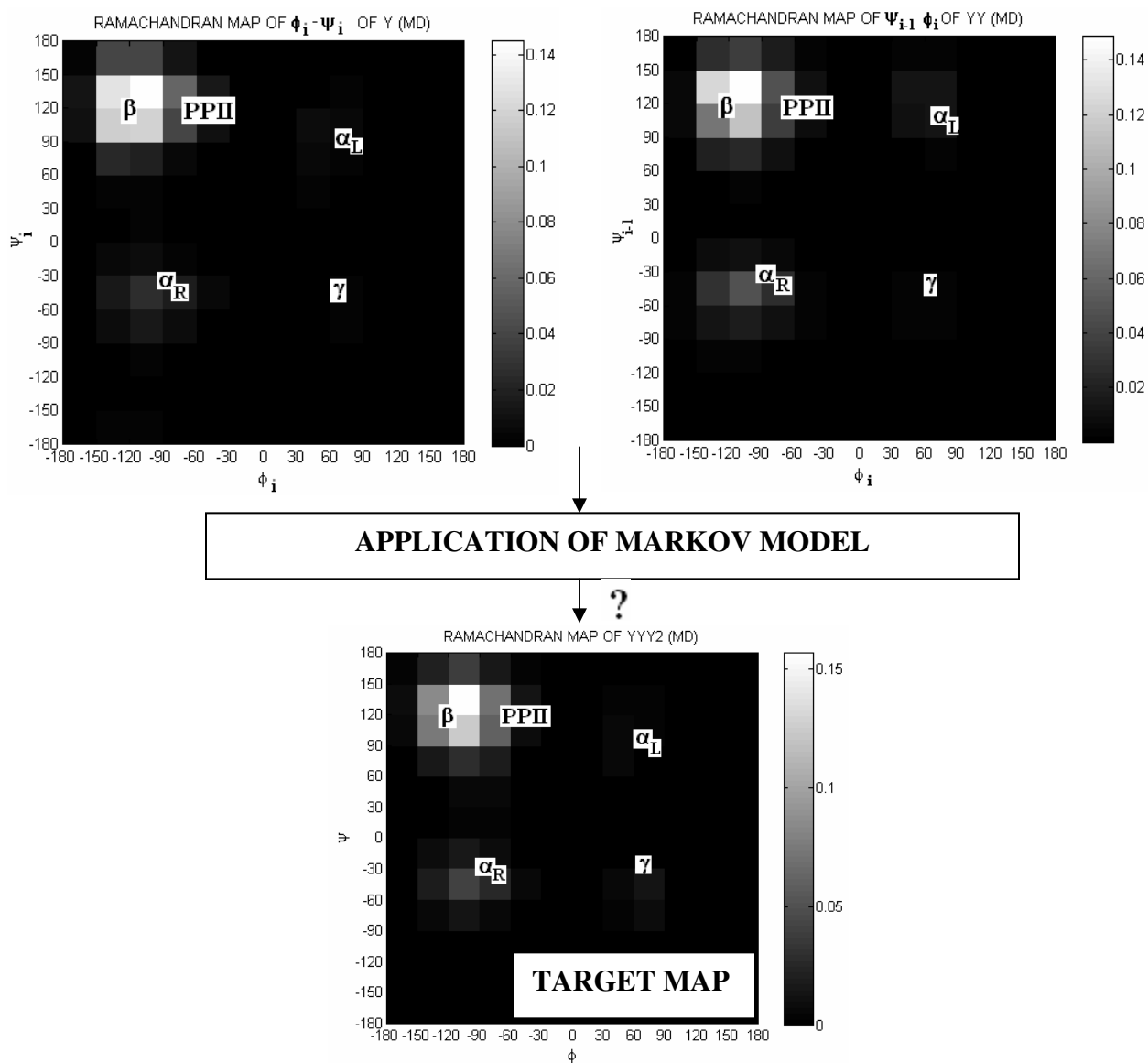
TARGET MAP

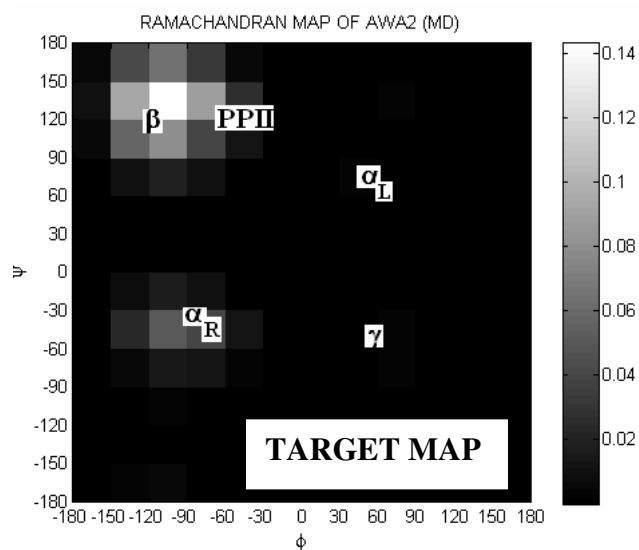Figure A.3 4 Statistical weight matrices used for the construction of the target map and target map of the central tyrosine. Colors represent the probabilities in accordance with the color bar located at the right side of the map.

Φ_i - ψ_i and ψ_{i-1} - ϕ_i statistical weight matrices in the form of Ramachandran maps
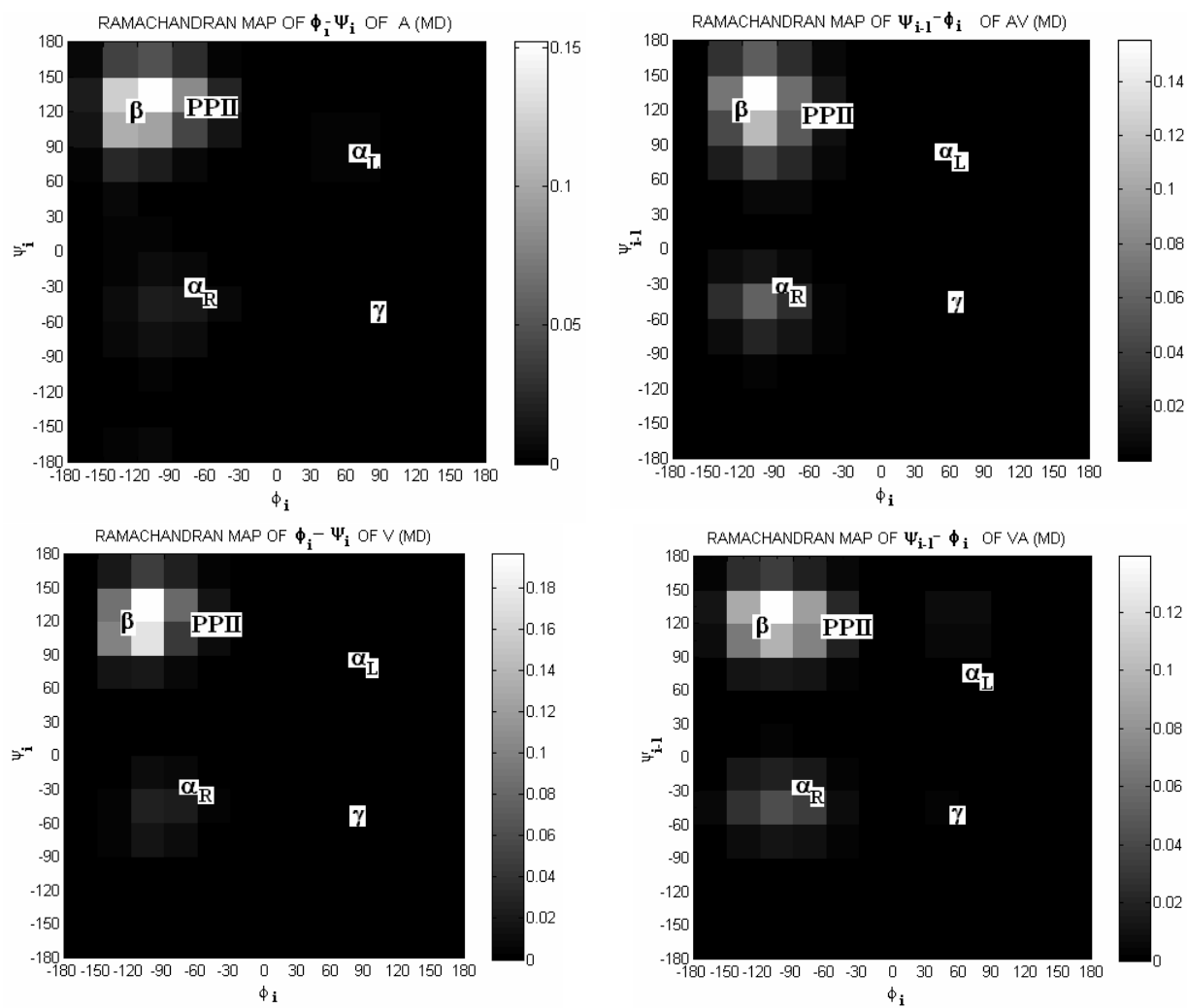
Figure A.3 4 Statistical weight matrices used for the construction of the target map and target map of the central valine. Colors represent the probabilities in accordance with the color bar located at the right side of the map.
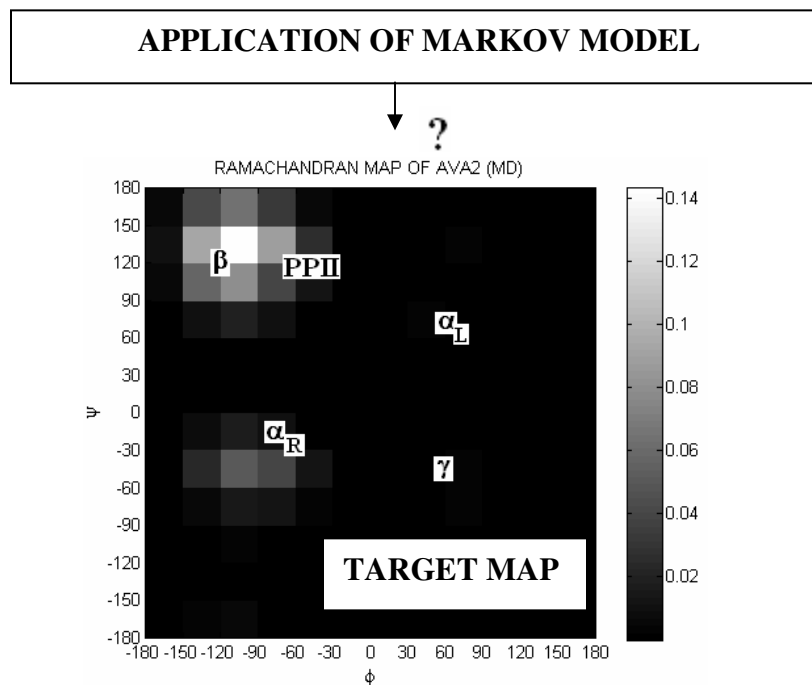
Figure A.3 5 (cont'd) Target Ramachandran map of the central valine. Colors represent the probabilities in accordance with the color bar located at the right side of the map.

Figure A.3 5 Statistical weight matrices used for the construction of the target map and target map of the central tyrosine. Colors represent the probabilities in accordance with the color bar located at the right side of the map.

Figure A.3 6 (cont'd) Target Ramachandran map of the central tyrosine. Colors represent the probabilities in accordance with the color bar located at the right side of the map.

Φ_i - ψ_i and ψ_{i-1} - ϕ_i statistical weight matrices in the form of Ramachandran maps

Figure A.3 6 Statistical weight matrices used for the construction of the target map and target map of the central tryptophan. Colors represent the probabilities in accordance with the color bar located at the right side of the map.

Figure A.3 7 (cont'd) Target Ramachandran map of the central tryptophan. Colors represent the probabilities in accordance with the color bar located at the right side of the map.

**A.4 ADDITION OF CORRESPONDING LONG-RANGE INTERACTION TERMS
TO THE STATISTICAL WEIGHT MATRICES OF MONO AND DIPEPTIDES**
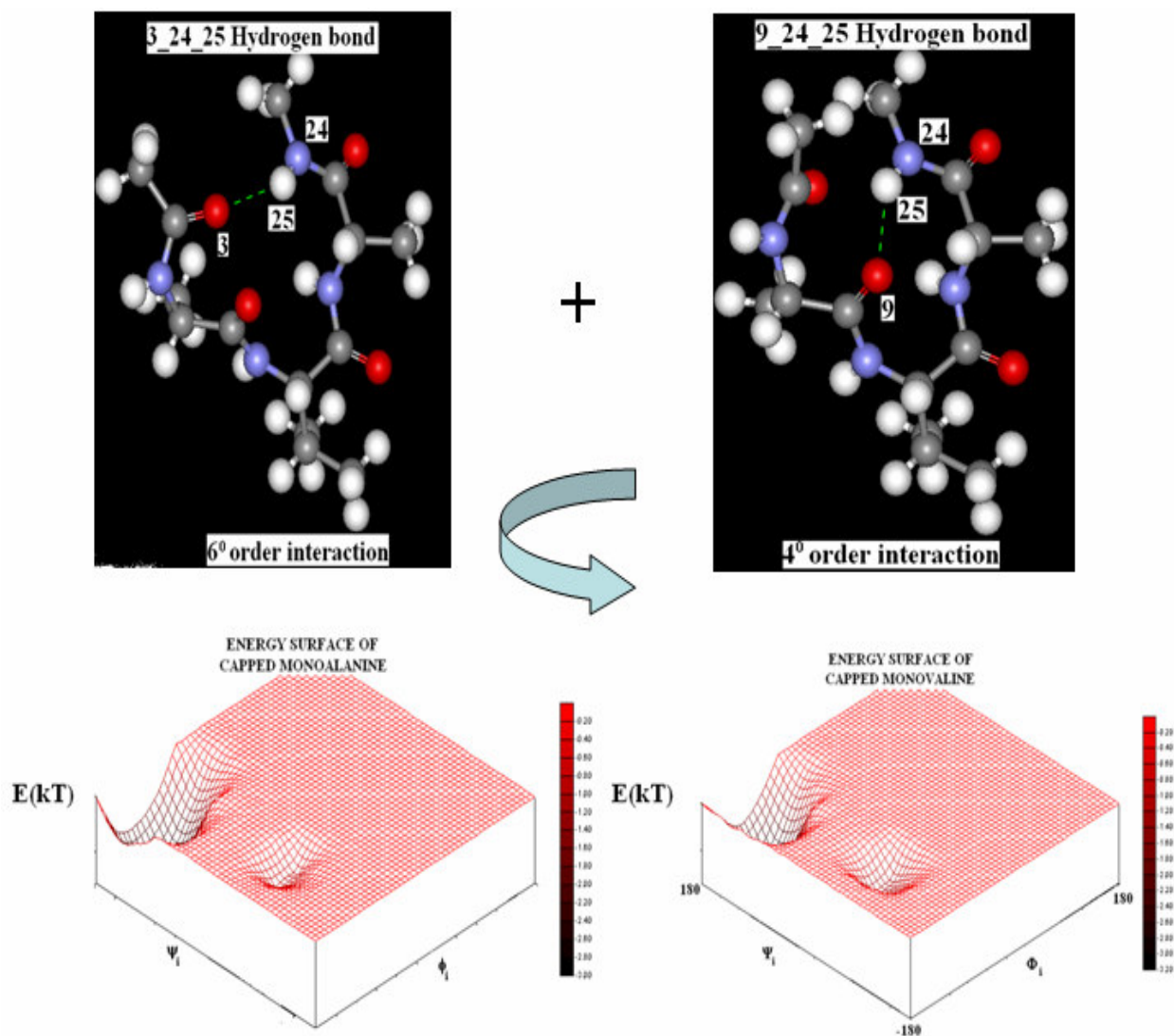
Figure A.4. 2 Addition of the corresponding hydrogen bonding terms to the statistical weight matrices of monoalanine and valine (Hydrogen bonds were shown as dotted green lines).
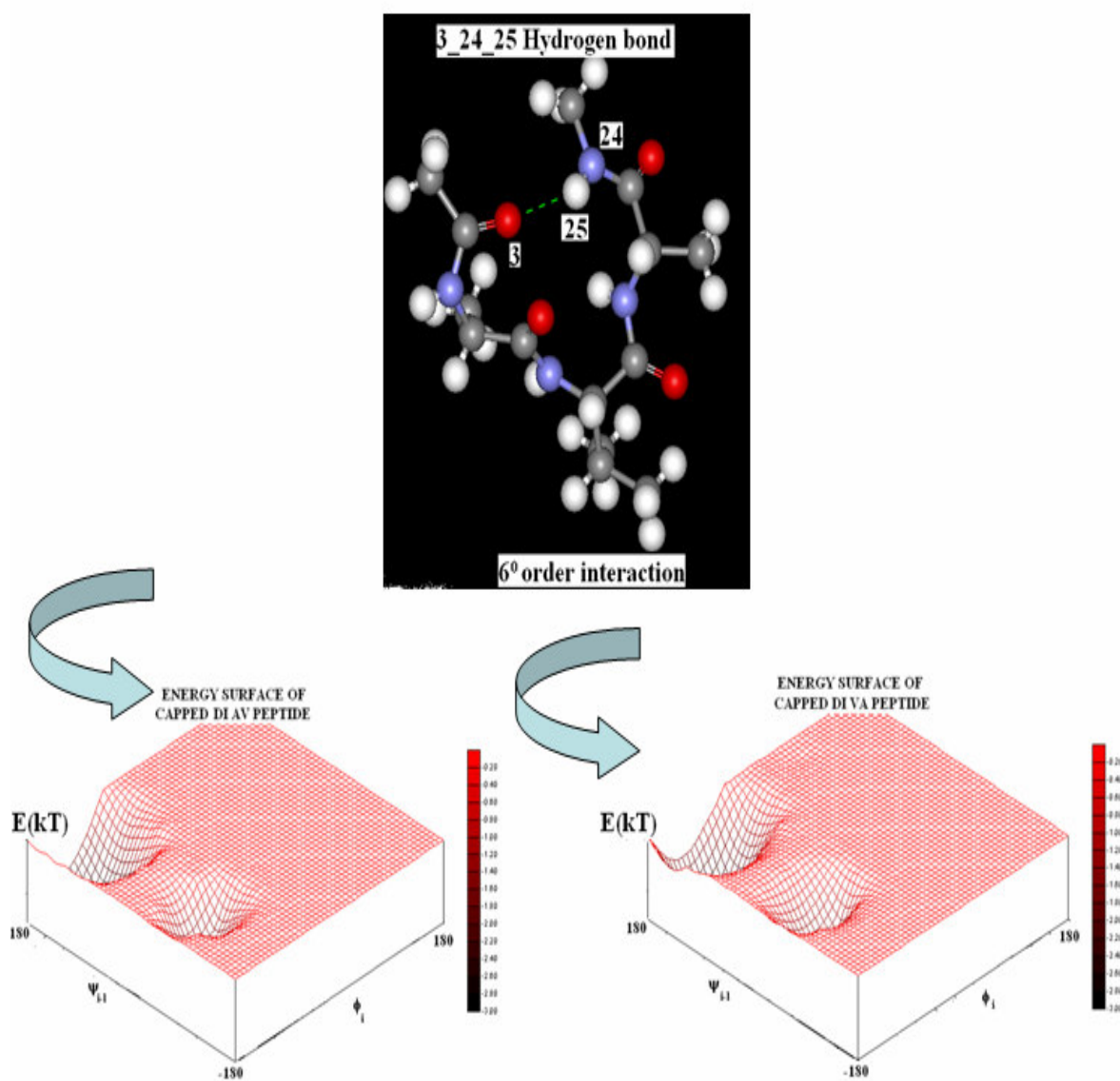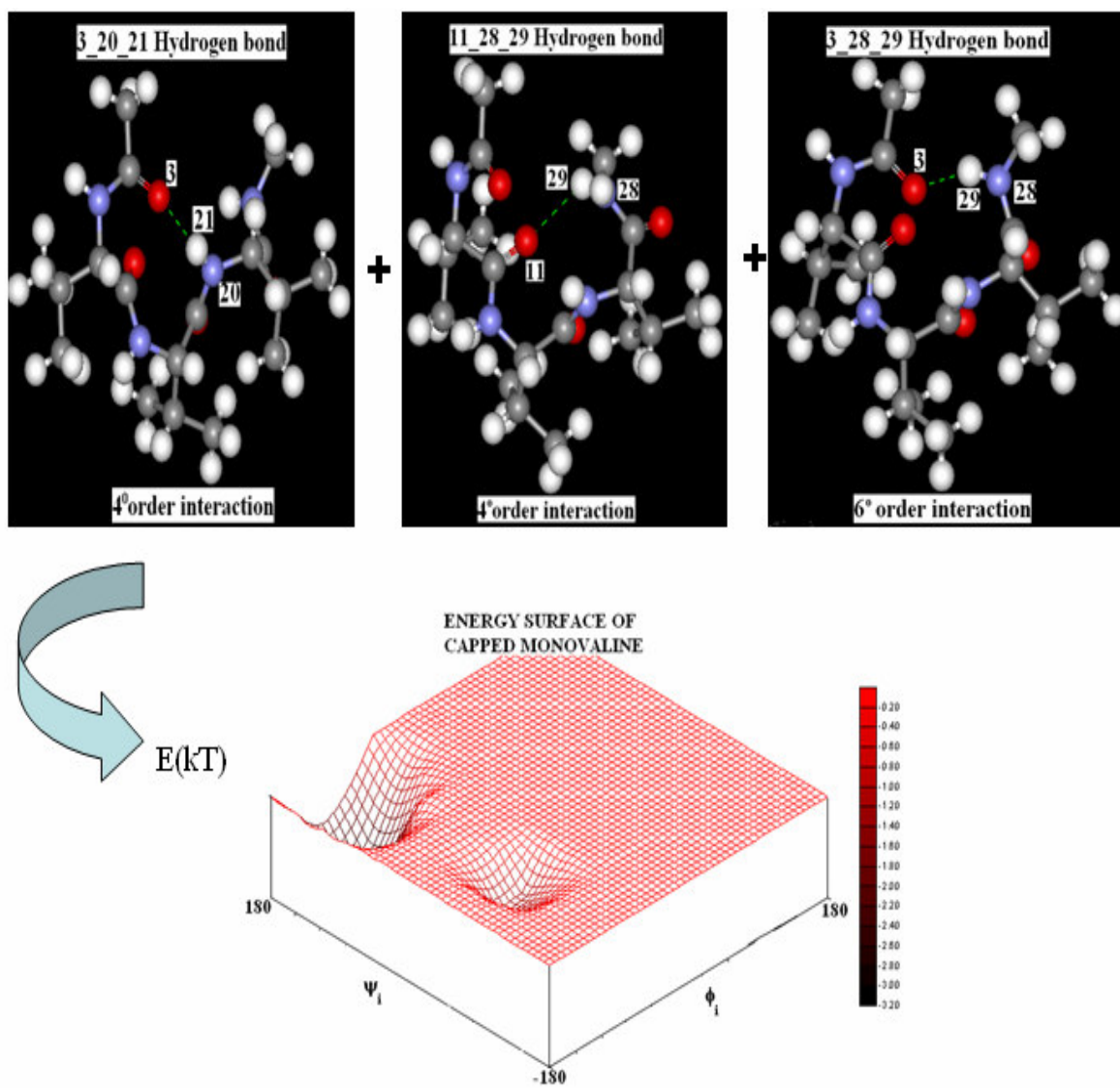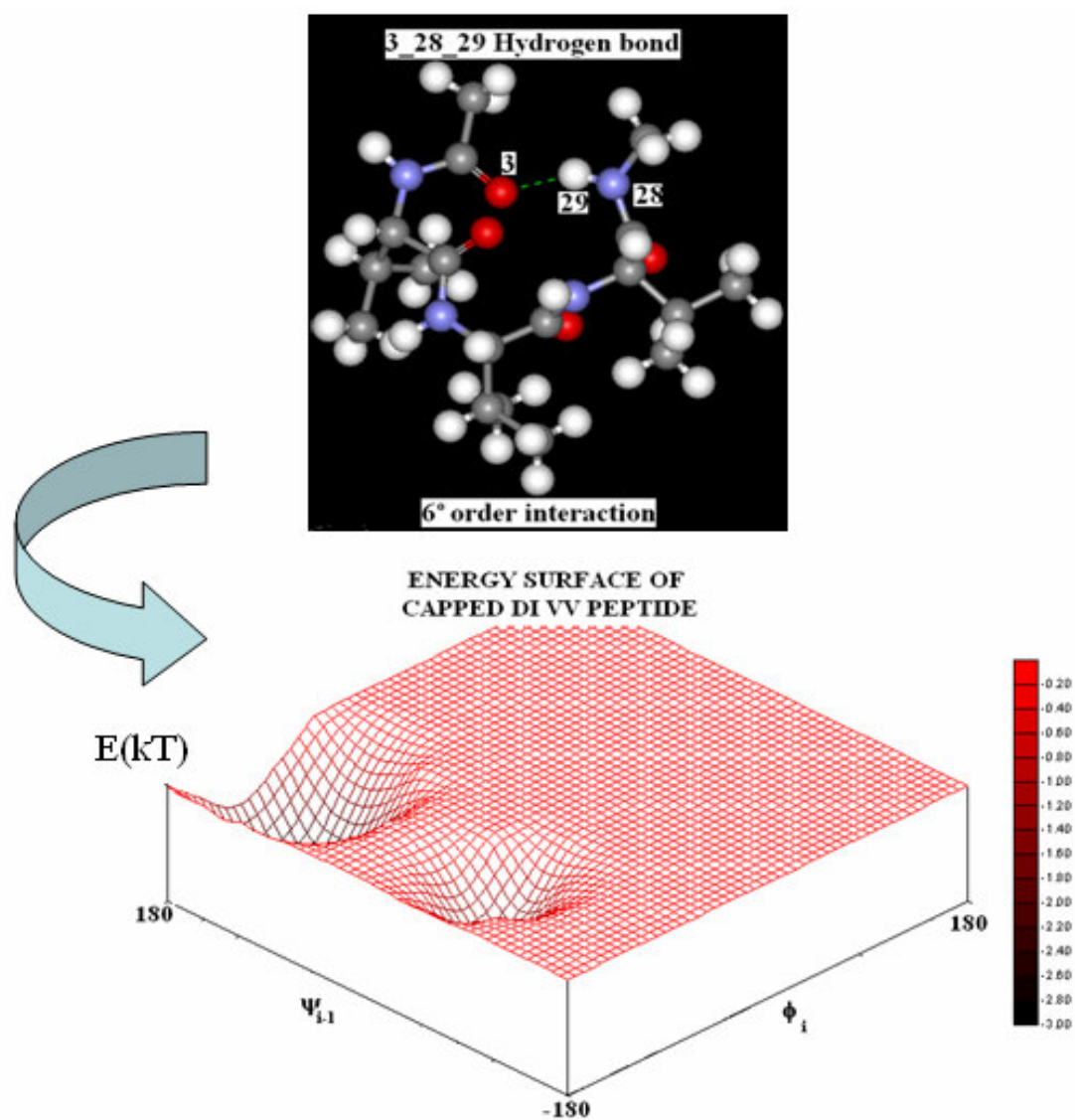
Figure A.4. 3 Addition of the corresponding hydrogen bonding terms to the statistical weight matrices of AV and VA dipeptide(Hydrogen bonds were shown as dotted green lines).
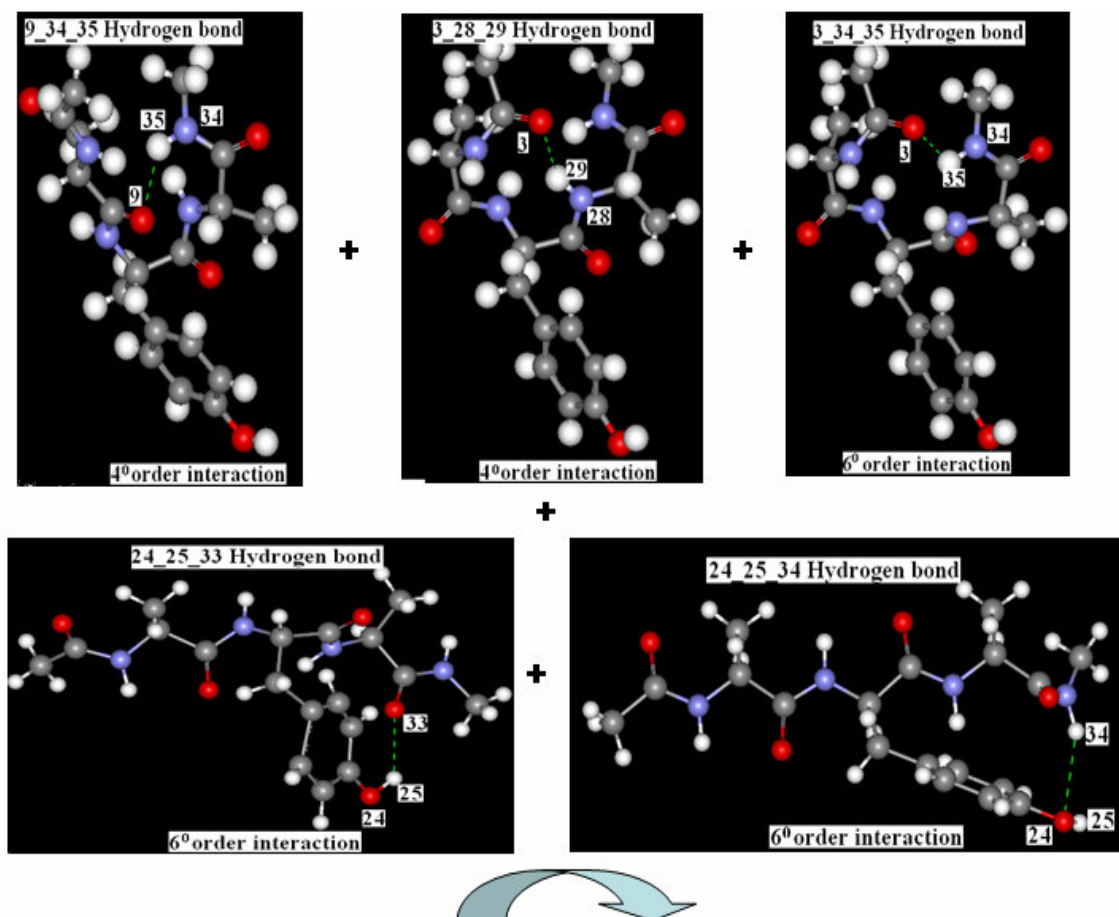
Figure A.4. 4 Addition of the corresponding hydrogen bonding terms to the statistical weight matrix of monovaline(Hydrogen bonds were shown as dotted green lines).
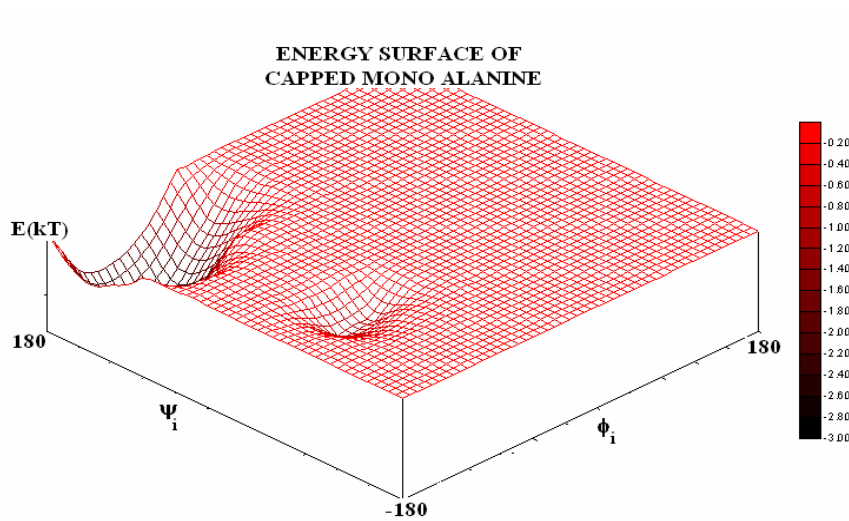
Figure A.4. Addition of the corresponding hydrogen bonding terms to the statistical weight matrices of valine dipeptide (Hydrogen bonds were shown as dotted green lines).

Figure A.4. 5 Addition of the corresponding hydrogen bonding terms to the statistical weight matrix of monoalanine (Hydrogen bonds were shown as dotted green lines).

Figure A.4. 5 (cont'd)Addition of the corresponding hydrogen bonding terms to the statistical weight matrices of monoalanine(Hydrogen bonds were shown as dotted green lines).
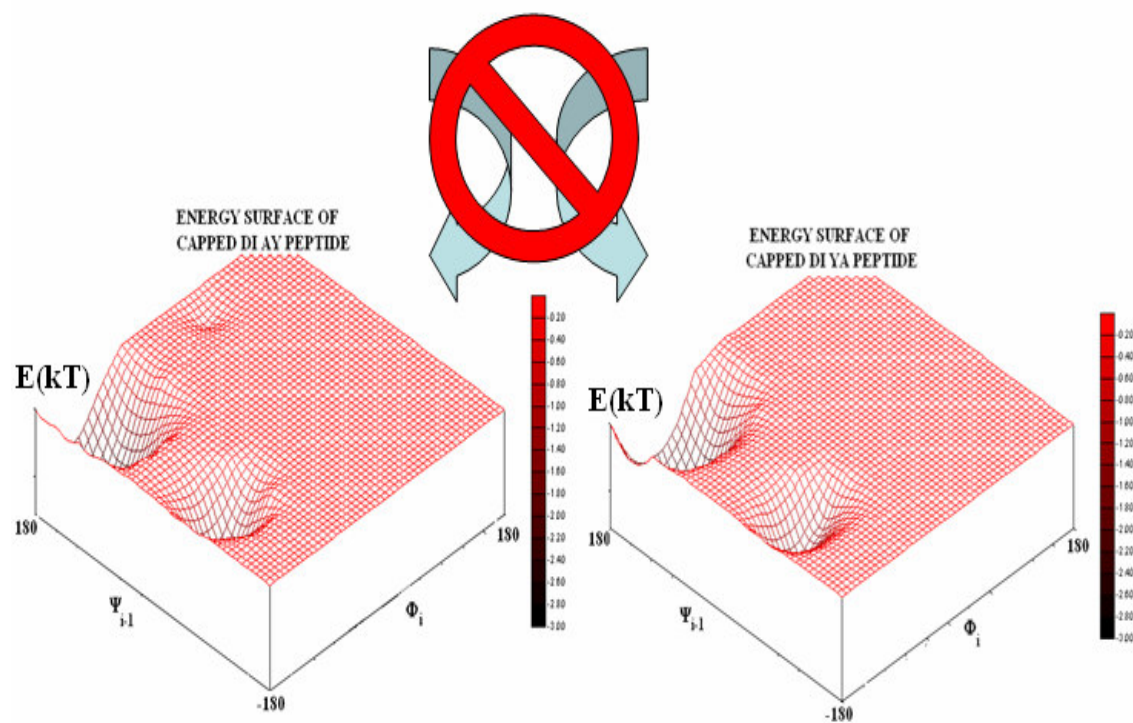
Figure A.4. 6 Addition of the corresponding hydrogen bonding terms to the statistical weight matrices of AY and YA dipeptides.
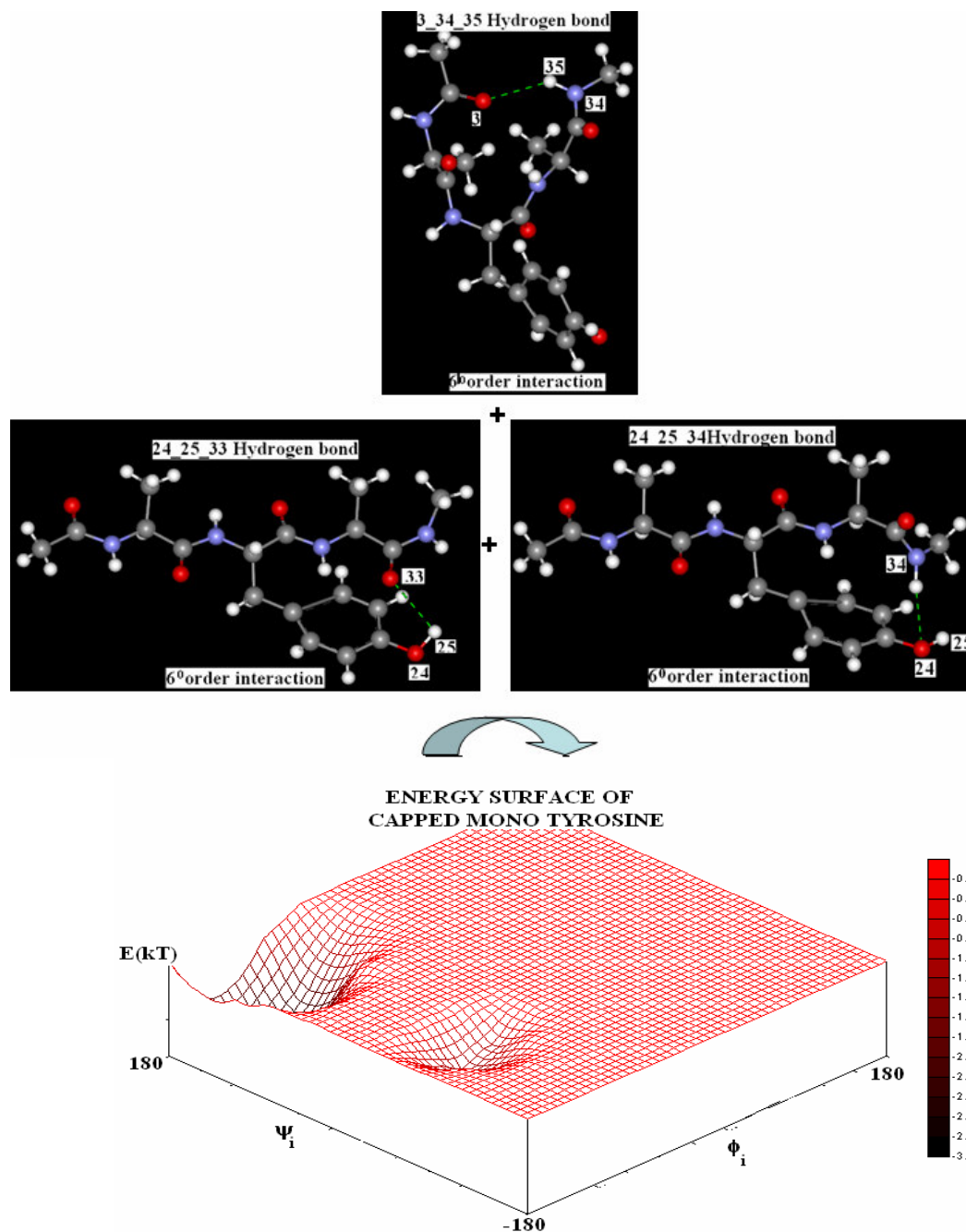
Figure A.4. 7 Addition of the corresponding hydrogen bonding terms to the statistical weight matrix of monotyrosine (Hydrogen bonds were shown as dotted green lines).
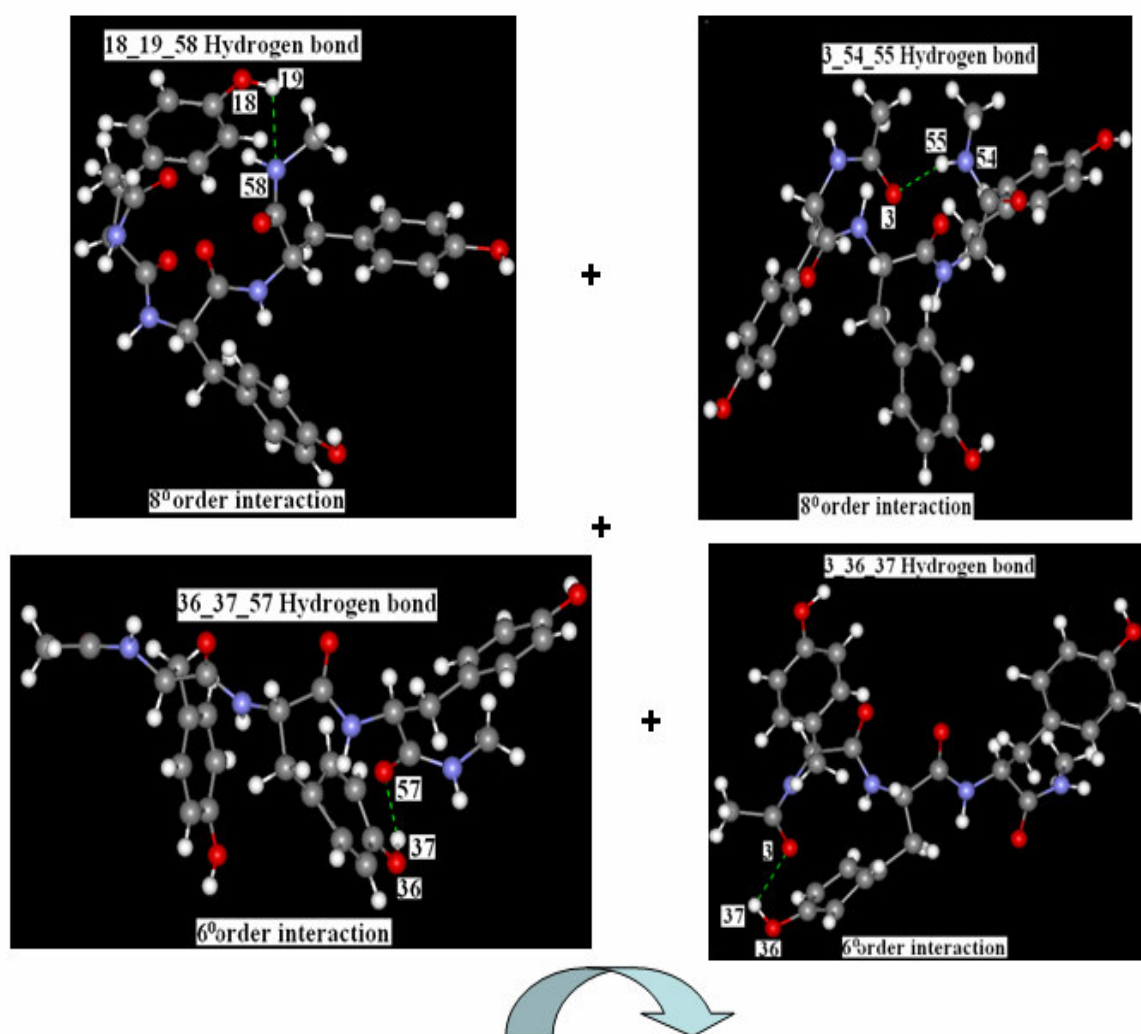
Figure A.4. 8 Addition of the corresponding hydrogen bonding terms to the statistical weight matrix of monotyrosine (Hydrogen bonds were shown as dotted green lines).
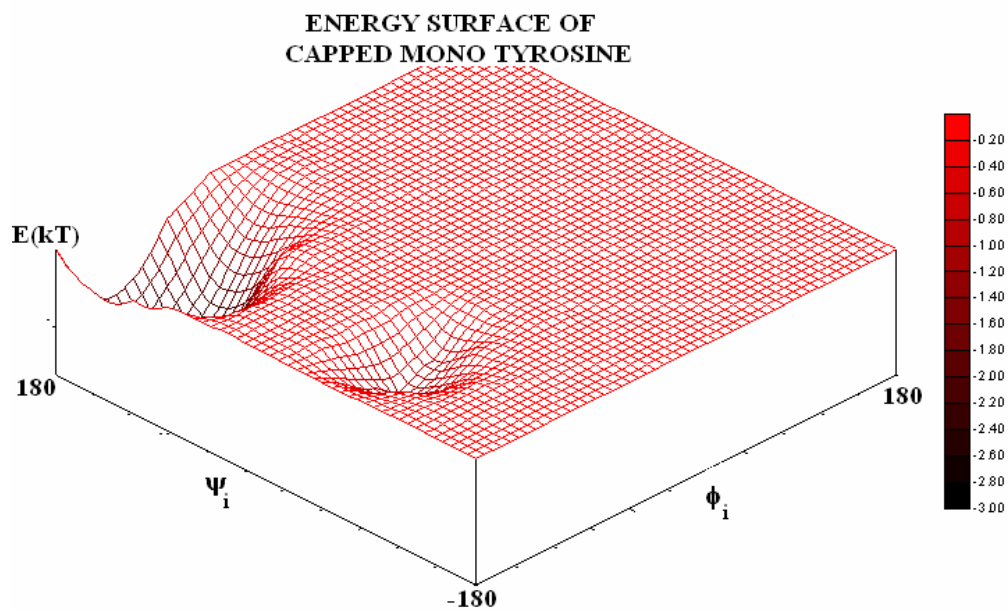
Figure A.4. 8 (cont'd) Addition of the corresponding hydrogen bonding terms to the statistical weight matrices of monotyrosine. (Hydrogen bonds were shown as dotted green lines).
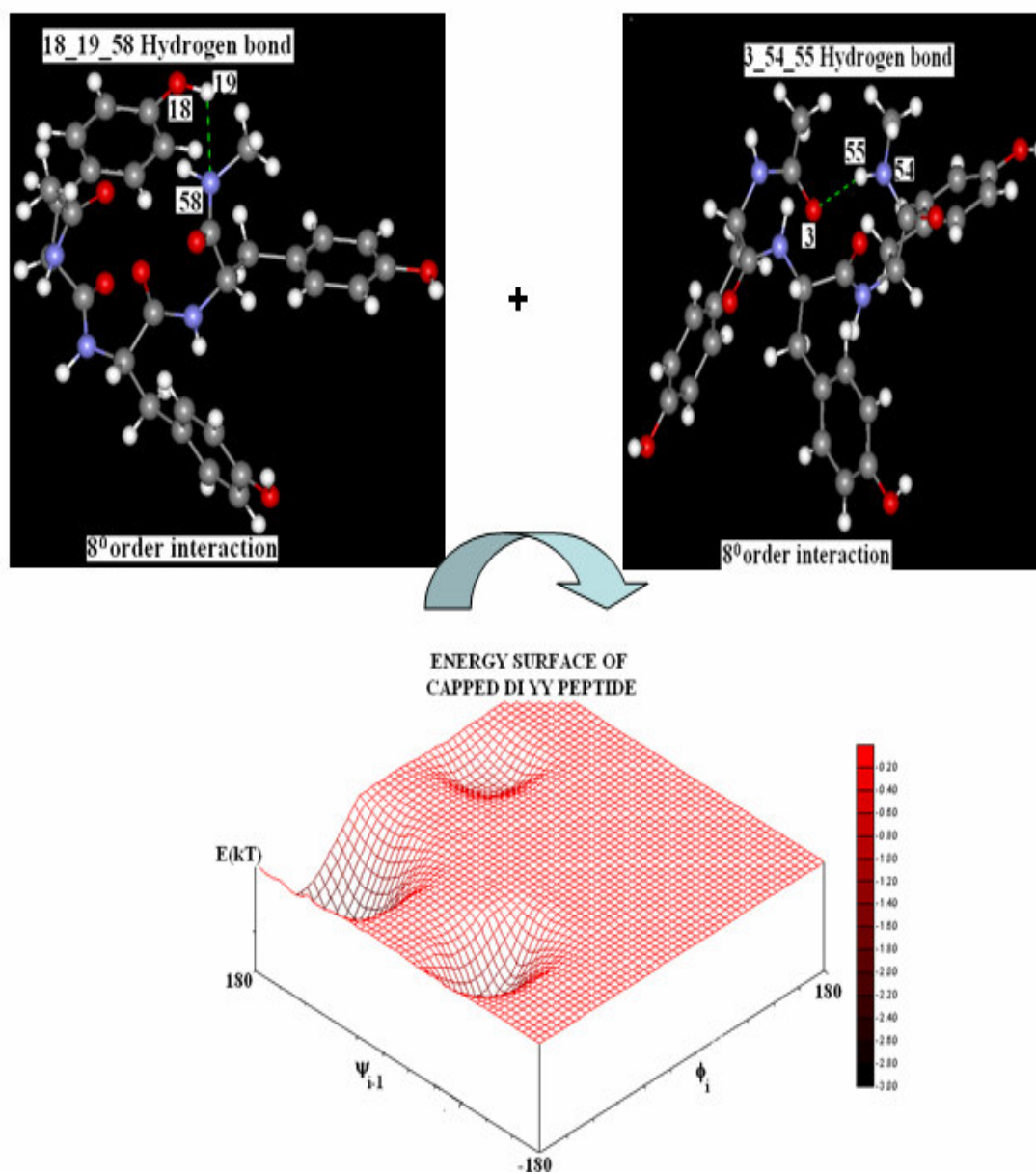
Figure A.4. 9 Addition of the corresponding hydrogen bonding terms to the statistical weight matrix of tyrosine dipeptide (Hydrogen bonds were shown as dotted green lines).
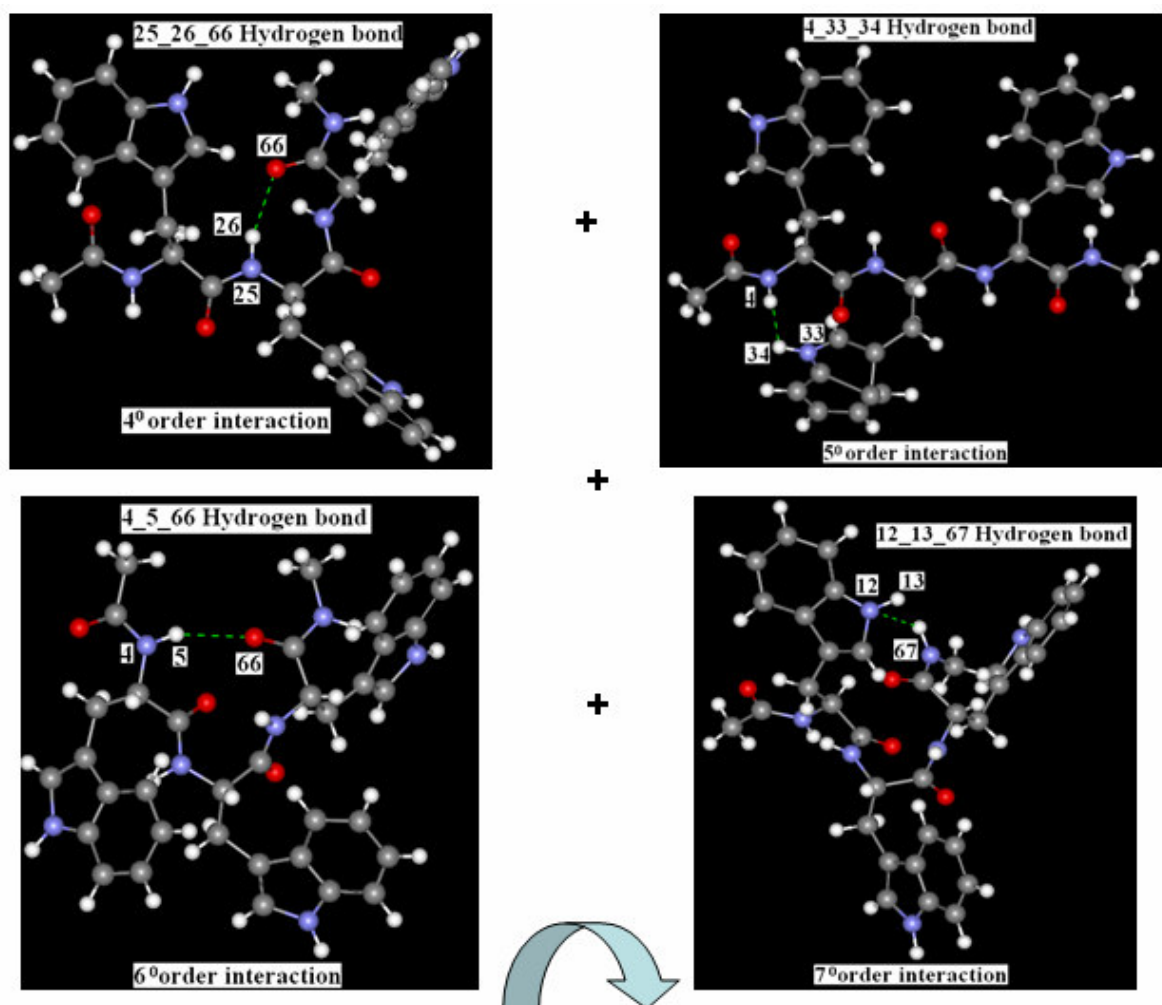
Figure A.4. 10 Addition of the corresponding hydrogen bonding terms to the statistical weight matrix of monotryptophan (Hydrogen bonds were shown as dotted green lines).

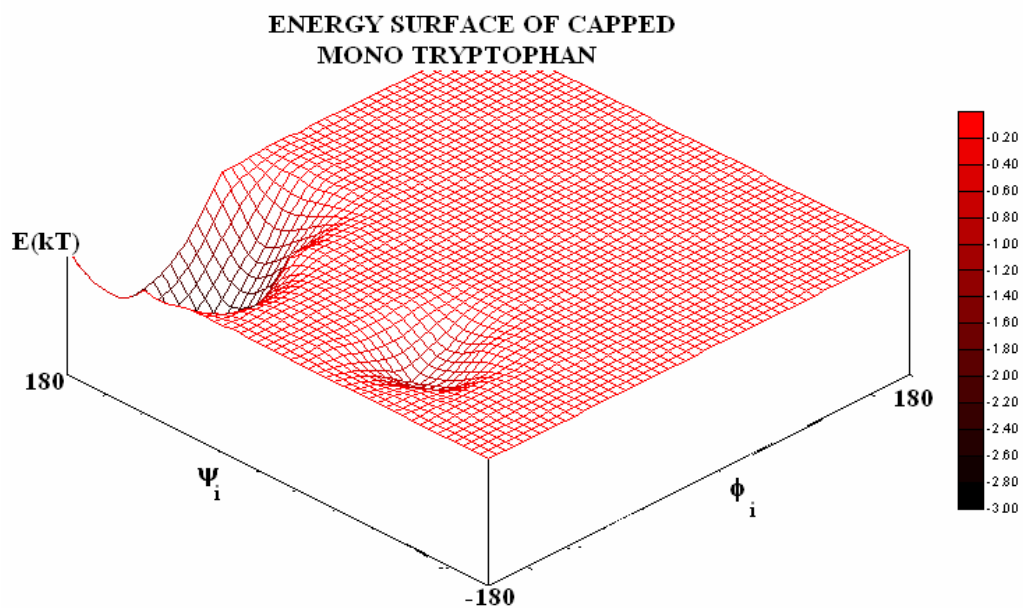**ENERGY SURFACE OF CAPPED MONO TRYPTOPHAN**

Figure A.4. 10 (cont'd) Addition of the corresponding hydrogen bonding terms to the statistical weight matrix of monotryptophan

Figure A.4. 11 Addition of the corresponding hydrogen bonding terms to the statistical weight matrix of WW dipeptide (Hydrogen bonds were shown as dotted green lines).

Figure A.4. 12 Addition of the corresponding hydrogen bonding terms to the statistical weight matrix of monoalanine (Hydrogen bonds were shown as dotted green lines).

Figure A.4. 12   (cont'd) Addition of the corresponding hydrogen bonding terms to the statistical weight matrix of monoalanine

Figure A.4. 13 Addition of the corresponding hydrogen bonding terms to the statistical weight matrix of AW and WA dipeptides (Hydrogen bonds were shown as dotted green lines).

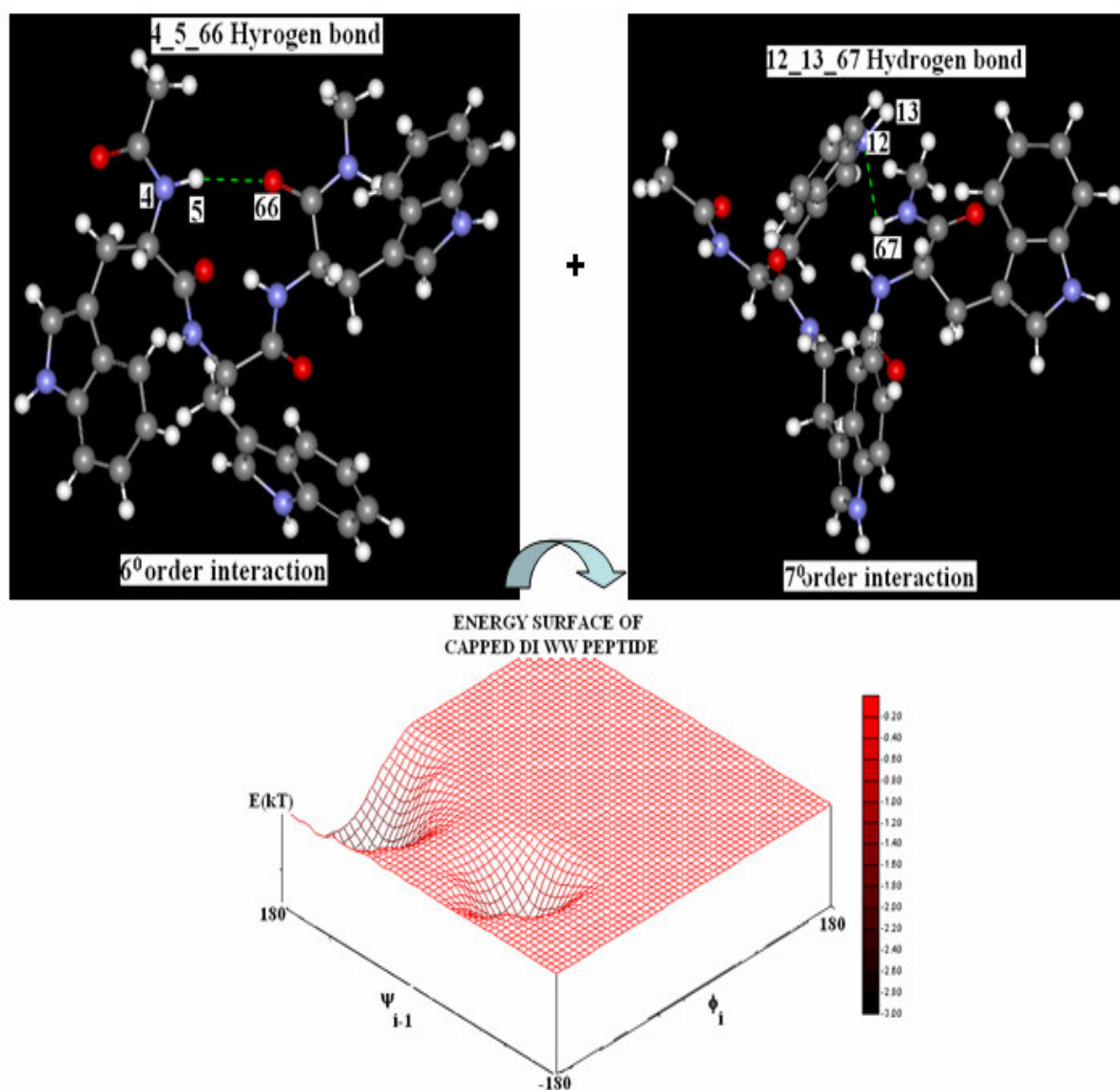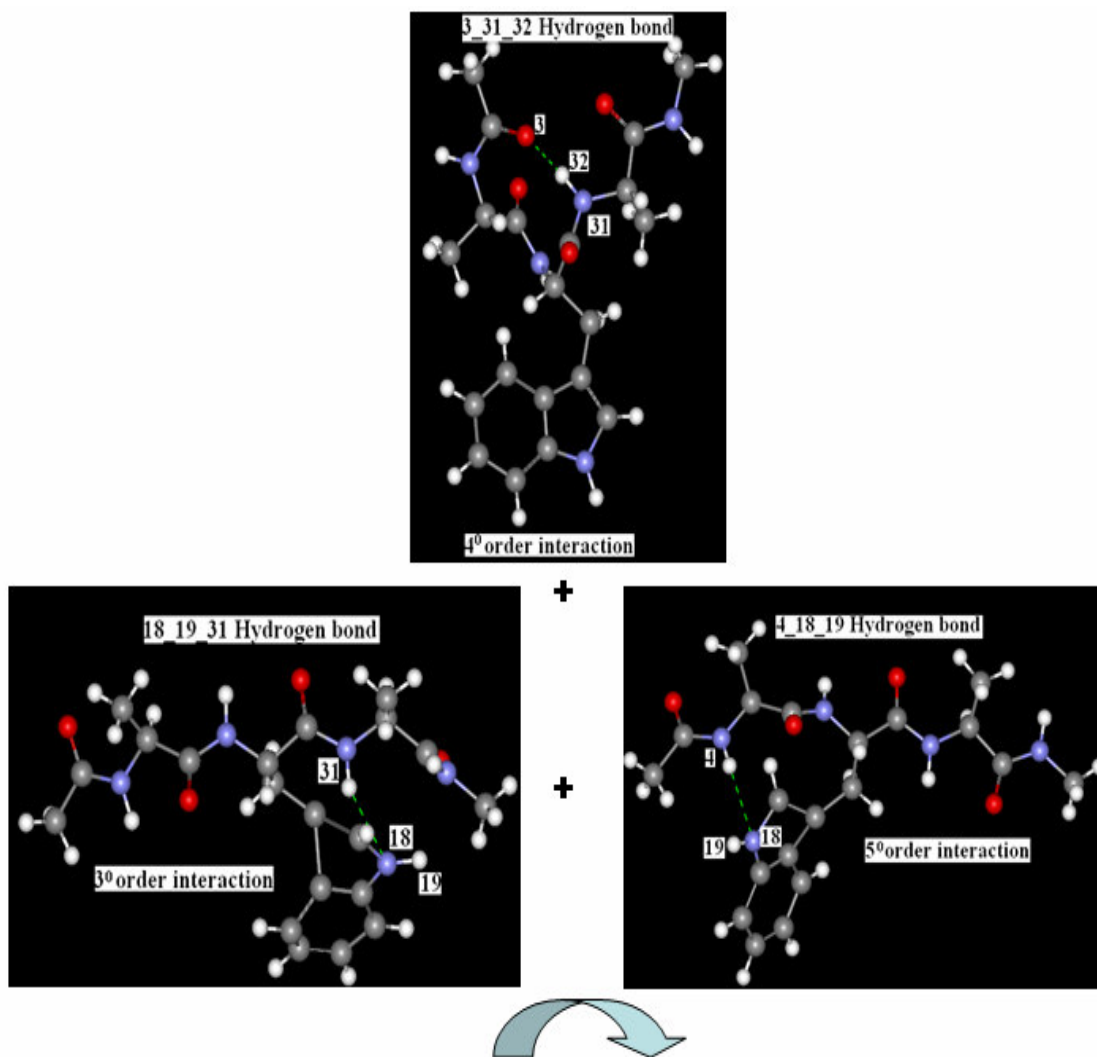Figure A.4. 14 Addition of the corresponding hydrogen bonding terms to the statistical weight matrix of mono tryptophan (Hydrogen bonds were shown as dotted green lines)
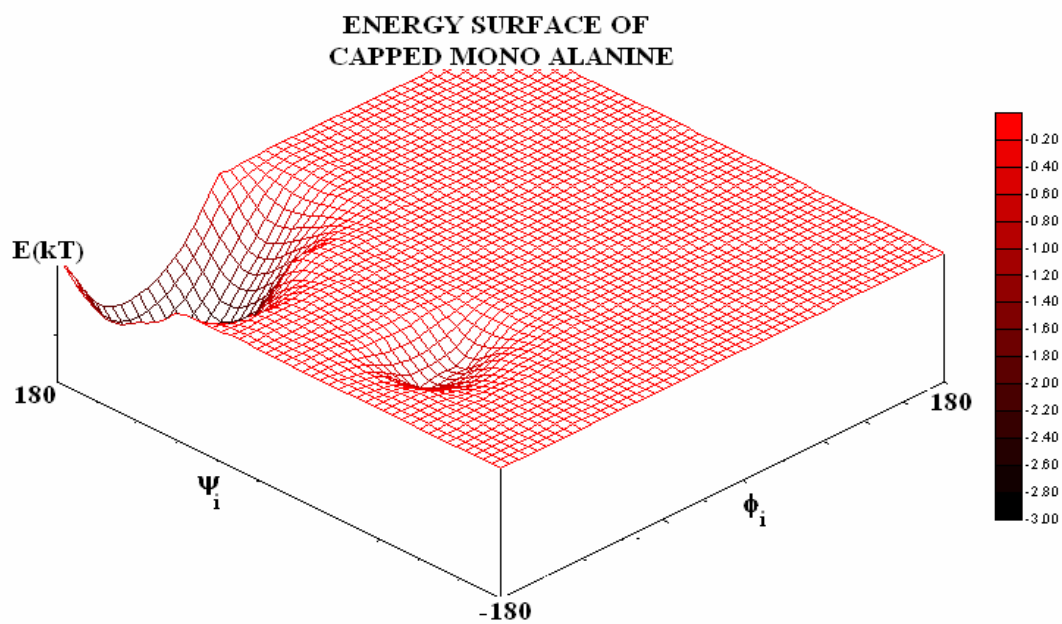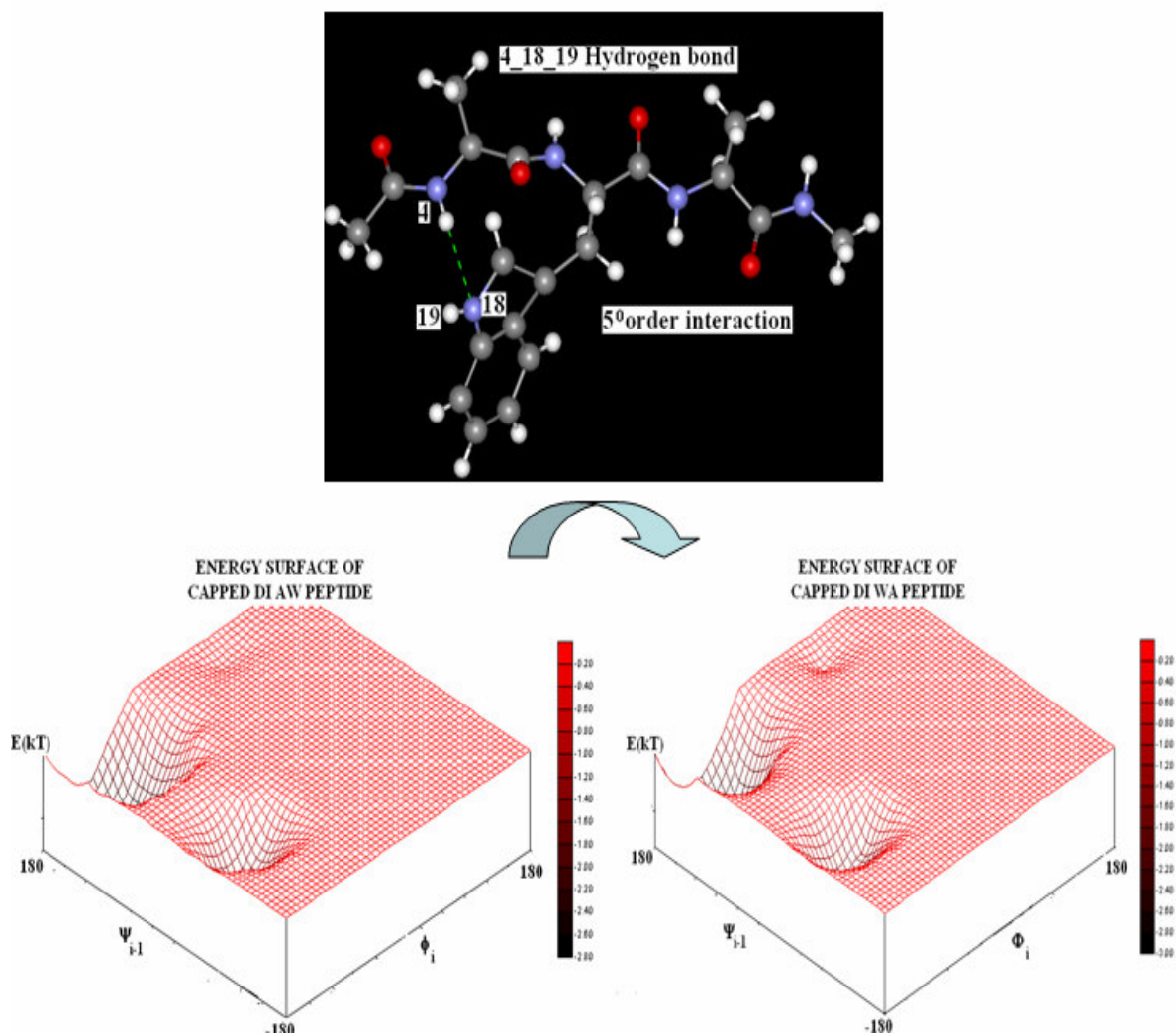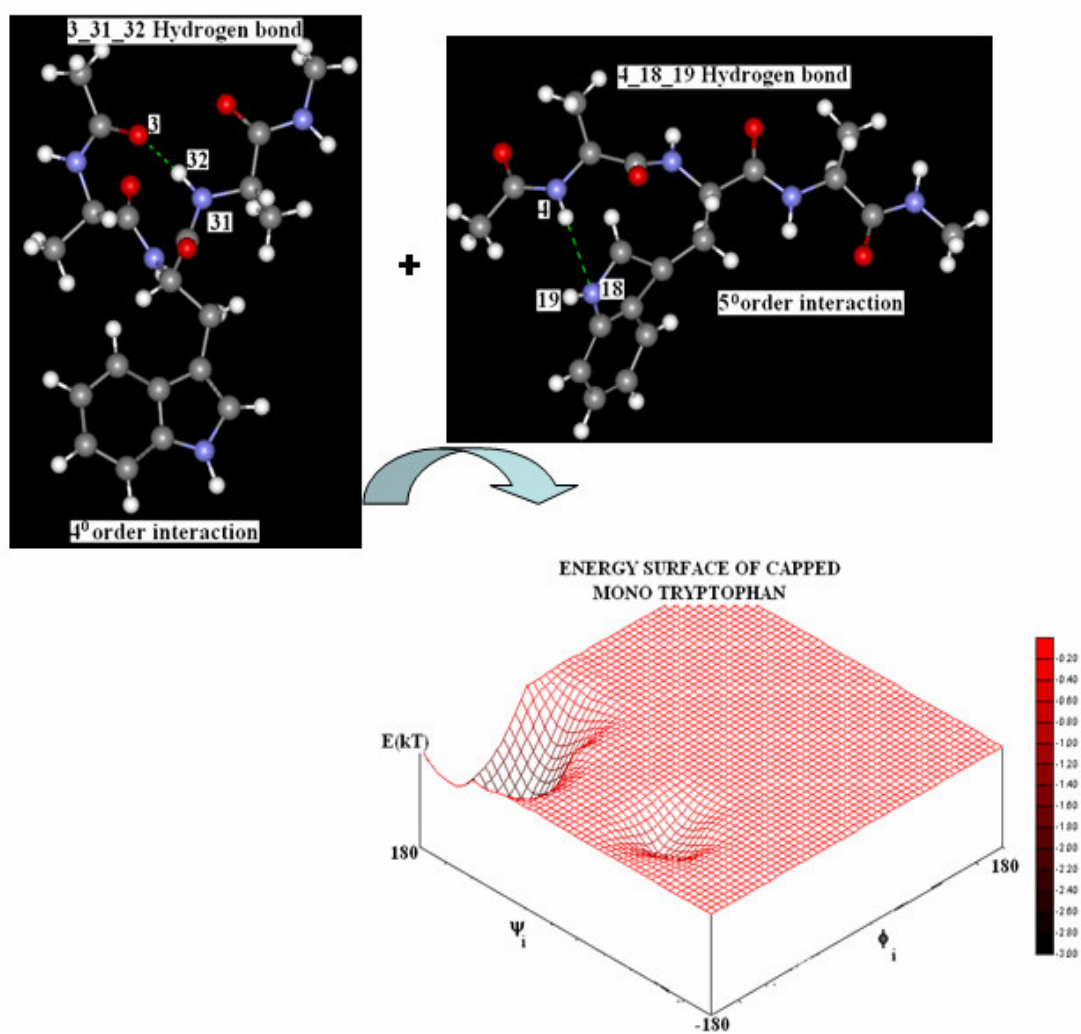
BIBLIOGRAPHY

1.   http://www.science.siu.edu/microbiology/micr302/figure%207.07.jpg.
2.   Pauling L. The nature of the chemical bond, 3rd edition., Cornell University Press, 1960.
3.   http://www.langara.bc.ca/biology/mario/Assets/peptidebond3.jpg.
4.   Ho BK, Thomas A, Brasseur R: Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. *Protein Sci* 2003, 12(11):2508-2522.
5.   F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M.D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi: The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol* 1977, 112:535-542.
6.   Horton, H.R., L.A. Moran, R.S. Ochs, J.D. Rawn, Principles of Biochemistry, Prentice-Hall, Inc. , 2000.
7.   Avbelj F, Fele L: Role of main-chain electrostatics, hydrophobic effect and side-chain conformational entropy in determining the secondary structure of proteins. *J. Mol. Biol* 1998, 279(3):665-684.
8.   C.Tanford: Protein Denaturation. *Adv.Prot.Chem.* 1970,24: 1-95.
9.   Shortle D: The denatured state and its role in protein stability. *FASEB J* 1996, 10: 27- 34.
10.   Smith LJ, Bolin KA, Schwalbe H, MacArthur MW, Thornton JM, Dobson CM: Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations. *J.Mol.Biol* 1996, 255(3):494-506.
11.   Sreerama N, Woody RW: Molecular dynamics simulations of polypeptide conformations in water: A comparison of alpha, beta, and poly(pro)II conformations. *Proteins* 1999, 36(4):400-406.
12.   Fitzkee NC, Rose GD: Sterics and solvation winnow accessible conformational space for unfolded proteins. *J. Mol. Biol* 2005, 353(4):873-887.
13.   Ohkubo YZ, Brooks CL, 3rd: Exploring Flory's isolated-pair hypothesis: statistical mechanics of helix-coil transitions in polyalanine and the C-peptide from RNase A. *Proc Natl Acad Sci U S A* 2003, 100(24):13916-13921.
14.   Zaman MH, Shen MY, Berry RS, Freed KF, and Sosnick TR: Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the Flory isolated-pair hypothesis for peptides. *J.Mol.Biol* 2003, 331(3):693-711.

15.   Cowan, P. M and McGavin, S: The polypeptide chain configuration of collagen. *Nature* 1955, 501-503.

16.   Rapaport, D.C., The Art of Molecular Dynamics Simulation. 1997.

17.   Colubri A, Jha AK, Shen MY, Sali A, Berry RS, Sosnick TR, Freed KF: Minimalist representations and the importance of nearest neighbor effects in protein folding simulations. *J.Mol.Biol* 2006, 363(4):835-857.

18.   Oostenbrink C., Mark A. E., Gunsteren W.F: A Biomolecular force field based on the free enthalpy of hydration and solvation:The Gromos Force-field parameter sets 53A5 and 53A6. *J.Comput. Chem* 2004, 25:1656-1676.

19.   Pappu RV, Srinivasan R, Rose GD: The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc Natl Acad Sci U S A* 2000, 97(23):12565-12570.

20.   Jha AK, Colubri A, Zaman MH, Koide S, Sosnick TR, Freed KF: Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library. *Biochemistry* 2005, 44(28):9691-9702.

21.   Fowler, R. H., and Guggenheim, E.A: Statistical Thermodynamics. Cambridge Univ. Press, London, 6.

22.   Eker F, Cao X, Nafie L, Schweitzer-Stenner R: Tripeptides adopt stable structures in water. A combined polarized visible Raman, FTIR, and VCD spectroscopy study. *J. Am.Chem..Soc* 2002, 124(48):14330-14341.

23.   L. J. Smith, K. M. Fiebig, H. Schwalbe, and C. M. Dobson: The concept of the random coil - Residual structure in peptides and denatured proteins. *Fold. Des* 1996, 1, R95-R106.

24.   Pappu RV, Rose GD: A simple model for polyproline II structure in unfolded states of alanine-based peptides. *Protein Sci* 2002, 11(10):2437-2455.

25.   Serrano L: Comparison between the phi distribution of the amino acids in the protein database and NMR data indicates that amino acids have variuos phi propensities in the random coil conformation. *J.Mol.Biol* 1995, 254:322-333.

26.   Solov'yov IA, Yakubovich AV, Solov'yov AV, Greiner W: Ab initio study of alanine polypeptide chain twisting. *Phys Rev E Stat Nonlin Soft Matter Phys* 2006, 73(2 Pt 1):021916.

27.   Yurtsever E, Yuret D, Erman B: Quantum mechanical calculations of tryptophan and comparison with conformations in native proteins. *J.Phys.Chem A* 2006, 110(51):13933-13938.

28.   Fitzkee NC, Fleming PJ, Rose GD: The Protein Coil Library: a structural database of nonhelix, nonstrand fragments derived from the PDB. *Proteins* 2005, 58(4):852-854.

29. Kelly A. M., Rucker L. A., Troutman J. M., Fried M. G., Miller F., and Creamer T.: Host-Guest Study of Left-Handed Polyproline-II Helix Formation. *Biochemistry* 2001, 40:14376-14383.

30. Munoz V, Serrano L: Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. *Proteins* 1994, 20(4):301-311.

31. Swindells MB, MacArthur MW, and Thornton JM: Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. *Nat Struct Biol* 1995, 2(7):596-603.

32. Fleming PJ, Fitzkee NC, Mezei M, Srinivasan R, Rose GD: A novel method reveals that solvent water favors polyproline II over beta-strand conformation in peptides and unfolded proteins: conditional hydrophobic accessible surface area (CHASA). *Protein Sci* 2005, 14(1):111-118.

33. Kentsis A, Mezei M, Gindin T, Osman R: Unfolded state of polyalanine is a segmented polyproline II helix. *Proteins* 2004, 55(3):493-501.

34. Shi, Z., Woody, R. W. , and Kallenbach, N.R: Is polyproline II a major backbone conformation in unfolded protein? Adv. Protein Chem 2002, 62: 163-240.

35. Eker F, Griebenow K, Cao X, Nafie LA, Schweitzer-Stenner R: Preferred peptide backbone conformations in the unfolded state revealed by the structure analysis of alanine-based (AXA) tripeptides in aqueous solution. *Proc Natl Acad Sci U S A* 2004, 101(27):10054-10059.

36. Avbelj F, Baldwin RL: Origin of the neighboring residue effect on peptide backbone conformation. *Proc Natl Acad Sci U S A* 2004, 101(30):10967-10972.

37. Hu H, Elstner M, Hermans J: Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine "dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. *Proteins* 2003, 50(3):451-463.

38. Avbelj F, Baldwin RL: Role of backbone solvation and electrostatics in generating preferred peptide backbone conformations: distributions of phi. *Proc Natl Acad Sci U S A* 2003, 100(10):5742-5747.

39. Soto P, Baumketner A, Shea JE: Aggregation of polyalanine in a hydrophobic environment. *J.Chem.Phys* 2006, 124(13):134904.

40. Roccatano D, Fioroni M, Zacharias M, Colombo G: Effect of hexafluoroisopropanol alcohol on the structure of melittin: a molecular dynamics simulation study. *Protein Sci* 2005, 14 (10):2582-2589.

41. Eswar N, Ramakrishnan C: Deterministic features of side-chain main-chain hydrogen bonds in globular protein structures. *Protein Eng* 2000, 13(4):227-238.

42. Spoel D., Hess B., Buuren A.R.,Apol E., Meulenhoff P.J., Tieleman D.P. , Sijbers A.L.T.M., Feenstra K.A., Drunen R., and Berendsen H.J.C: Gromacs User Manual Version 3.2. 2004.

43. van Gunsteren WF, Berendsen HJ, Hermans J, Hol WG, Postma JP: Computer simulation of the dynamics of hydrated protein crystals and its comparison with x-ray data. *Proc Natl Acad Sci U S A* 1983, 80(14):4315-4319.

44. Schuettelkopf A. W., and Aalten D. M. F: PRODRG - a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallographica* 2004, D60: 1355-1363.

45. O'Connell M.T.,Tropsha A., and Hermans J: The 'Random Coil' State of Proteins: Comparison of Database Statistics and Molecular Simulations. *Proteins: Structure, Function, and Genetics* 1999, 36:407-418.

46. Cornell WD, Cieplak, P., Bayly C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Coldwell, J.W., Kollman, P.A: Application of the Multimolecule and Multiconformation RESP Methodology to Biopolymers. *J.Am.Chem.Soc* 1995, 117:5179-5197.

47. Jorgensen WL, Maxwell, D.S., Tirado-Rives, J: Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J.Am.Chem.Soc* 1996, 118:11225-11236.

48. MacKerell A.D. BD, Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., JosephMc Carthy, Kuchnir, L., Kuazua, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Pradhon, B., Reiher, W.E, Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Yin, D., Karplus, M: All atom emprical potential for molecular modeling and dynamics studies of proteins. *J.Phys.Chem B* 1998, 102:3586-3616.

49. Scott WRP, Hünenberger, P.H., Tironi, I.G., Mark, A.E., Billeter, S.R., Fennen, J.Torda, A.E., Huber, T., Krüger, P. and van Gunsteren, W.F.: The GROMOS biomolecular simulation program package. *J.Phys.Chem A* 1999, 103:3596-3607.

50. Flory P: Statistical Mechanics of Chain Molecules. 1969.

51. Mattice WL, and U.W., Suter: Conformational theory of large molecules. The rotational isomeric state model in macromolecular systems. 1994.

52. Keskin O, Yuret D, Gursoy A, Turkay M, and Erman B: Relationships between amino acid sequence and backbone torsion angle preferences. *Proteins* 2004, 55(4):992-998.

53. http://www.ch.embnet.org/MD_tutorial/pages/MD.Part1.html

54. Panasik N, Jr., Fleming PJ, Rose GD: Hydrogen-bonded turns in proteins: the case for a recount. *Protein Sci* 2005, 14(11):2910-2914.

55.   Avbelj F, Grdadolnik SG, Grdadolnik J, and Baldwin RL: Intrinsic backbone preferences are fully present in blocked amino acids. *Proc Natl Acad Sci U S A* 2006, 103(5):1272-1277.

56.   Nemethy, G., and Printz, M: The gama turn, a possible folded conformation of the polypeptide chain. Comparison with the beta turns. *Macromolecules* 1972, 5(6):755-758.

VITA

Özge Engin was born in Istanbul, on 21 August, 1982. She received her first B. Sc. Degree in Molecular Biology and Genetics, and second B.Sc. Degree in Chemistry from Istanbul Technical University, Istanbul, 2005. From 2005 to present she has been worked as teaching and research assistant at Koç University, Istanbul. She studied to develop a coarse-grained model for the representation of unfolded peptides.