An Optimization Approach To Study the Dynamics of Co-translational Folding

by

Serife Senturk

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, 2006

Koç University

Graduate School of Sciences and Engineering

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

Serife Senturk

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:

Prof. Yaman Arkun (Advisor)

Prof. Burak Erman (Co-advisor)

Assist. Prof. Attila Gursoy

Assist. Prof. Ozlem Keskin

Assist. Prof. Alper Erdogan

Date: _____

To my father-in memory

ABSTRACT

An optimization model is implemented to obtain the possible pathways for the folding of the protein chain as the protein folds in the exit tunnel of the ribosome during and following its synthesis. In this model, the folding problem is formulated as an optimal control problem in which a particular form of energy is minimized subject to the dynamic model predictions and physical constraints. It is assumed that the chain grows and folds at the same time while it is still being synthesized. The optimization models change according to the length of the partial chain. The co-translational folding dynamics of a fast-folding protein, chicken villin headpiece protein, is simulated. The model is implemented using different growth rates for the protein to fold and the results for the different growth rates are compared. The folding dynamics is analyzed as a process composed of early stage, intermediate stage and late stage according to the secondary structure property of chicken villin protein. Also, the important role of long-range contact pairs is presented and the input variables are analyzed.

ÖZETÇE

Bu tezde, protein zincirlerinin ribozom tarafından üretimi sırasında nasıl katlandığının anlaşılması için bir optimizasyon modeli kullanılmıştır. Proteinler ribozomda üretilirken, zincirin oluşan kısmı ribozom tünelinde katlanmaya başlamaktadır ve ayrıca ribozomdan çıkan amino asitler kendi aralarında özelliklerine uygun şekillere bürünmektedir ve aynı zamanda zincirin geri kalan kısmının üretimi devam etmektedir. Bu optimizasyon modelinde, katlanma problemi fiziksel belirleyicilere ve dinamik model hesaplamalarına uygun olarak, bir tür enerji şeklini minimize eden bir kontrol problemi olarak formule edilmiştir. Kullanılan optimizasyon modelinin büyüklüğü zincirin o anda kaç tane amino asitten oluştuğuna göre değişmektedir. Bu çalışmada, otuz altı amino asitten oluşan ve hızlı katlandığı bilinen bir protein, chicken villin headpiece, üzerine çalışılmıştır. Model, proteinin değişik üretim hızlarında nasıl katlandığının anlaşılması için her amino asitin değişik hızlarda daha önceden var olan zincire eklenmesi şeklinde kurgulandırılmıştır. Proteinin katlanma süreci, proteinin ikincil yapılarına bakılarak ilk aralık, orta aralık ve son aralık olarak sınıflandırılmış ve bu aralıklar için zincirlerin katlanma mekanizması analiz edilmiştir. Ayrıca, lokal olmayan etkileşmelerin proteinin katlanması üzerine olan önemli etkisi vurgulanmıştır.

ACKNOWLEDGMENTS

Firstly, I would like to thank my advisors, *Prof. Dr. Yaman Arkun* and *Prof. Dr. Burak Erman* for their great guidance throughout the entire study and for always being patient with me. During the past two years, they gave me the best research-education I have ever had in my life. I would also like to thank the members of the thesis committee for their critical reading and valuable comments.

I would like to thank all of my friends for their support. Special thanks go to my friends from ITU, my friends from Koc University, my roommates in Acarlar, my friends who opened their house to me during the last days of this study and my current roommate.

Finally, I am grateful to my family for their great support all my life. I thank my brother *Ismail* for always being there with me, my sister *Sefika* for always encouraging me, my mother *E. Dudu* for being such a good mother, my brother-in-law *Taner* for being my second brother. I am very happy to inform that the birth of my nephew was a joyful experience during my thesis study. I am also indebted to my grandmother *Serife* and my aunt *Lutfiye* for their support.

I dedicate this study to my father.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

One of the most essential units of cells of all organisms is the protein. Water makes up about 55% of the mass of an average person and proteins make up about 15% of the mass, so it can be understood why proteins play a vital role in our cells functioning in an enormous variety of different ways.

Proteins are polymers which are composed of different combinations of 20 amino acids, in other words the monomers of proteins are amino acids. These 20 amino acids have different chemical properties. The sequence of amino acids in a protein chain is determined by the gene that encodes the protein. The chemical properties of amino acids determine the biological function of the protein. Thus, the sequence is the main key to the three dimensional structure and the function of the protein.

In Figure 1.1, the structure of an amino acid can be seen. An amino acid is composed of an alpha carbon, a carboxyl group, an amino group, and a side chain. The amino group and the carboxyl group and the side group are linked to alpha carbon atom. All amino acids have the same structure except for the side chain atoms. The side chain varies in different amino acids, so the chemical variety comes from the side chain atoms.

Amino acids can be classified as hydrophobic and hydrophilic amino acids according to the property of side chain atoms. The hydrophobic amino acids tend to form the interior of proteins as they repel the aqueous environment. The hydrophilic amino acids generally take place on the exterior surfaces of proteins as they have a tendency to interact with the aqueous environment.

Amino acids are covalently bonded together by peptide bonds. Peptide bond

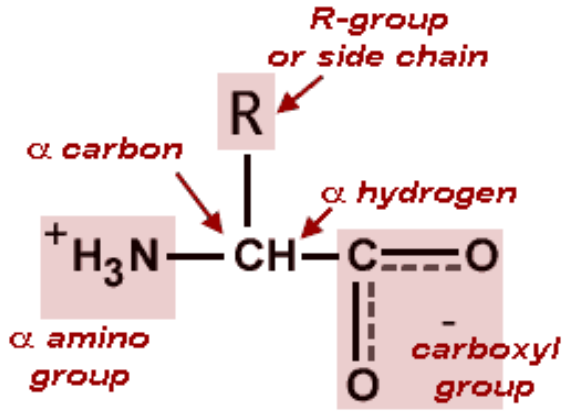


Figure 1.1: Structure of an amino acid

formation, a chemical reaction, is polymerization of amino acids into peptides and proteins. The peptide bond is formed by the condensation of the carboxyl group of one amino acid with the amino group of the second amino acid involving loss of a molecule of water. A dipeptide is the simplest peptide which contains two amino acids linked by a peptide bond. In Figure 1.2, two amino acids can be seen. Peptides are small groups of amino acids and contain around 30 amino acids. Longer chains are called polypeptides and proteins.

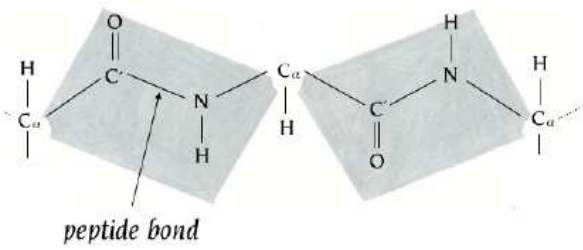


Figure 1.2: Structure of two amino acids linked by a peptide bond.

The individual amino acids, which are linked in the polypeptides, are called residues. The polymer chain has directionality due to the chemical structure of amino acid atoms. The end of the protein with a carboxyl group is known as the C-terminus,

the end with a free amino group is known as the N-terminus.

The proteins are synthesized in the cellular machinery called ribosome. The newly synthesized protein folds into a 3-dimensional structure. The conformation into which a protein naturally folds is called the native state of the protein. Structural features of proteins can be described at four levels of complexity: primary structure, secondary structure, tertiary structure and quaternary structure. These structures can be seen in Figure 1.3. The primary structure is the linear arrangement of amino acids in a protein, in other words, it is just the sequence of the protein. In the secondary structure, there are areas of folding and coiling within the protein. The main secondary structures are alpha helix, pleated sheets, and random coil regions. These structures are stabilized by hydrogen bonding. Secondary structures are usually local, so a protein can be composed of many different individual secondary structures. Alpha helices are one-dimensional structures; the hydrogen bonds are aligned with the axis of the helix and there are 3.6 amino acids per turn. Beta (pleated) sheets are quasi two-dimensional structures and the hydrogen bonds are perpendicular to the strands. The tertiary structure is the final 3-dimensional structure of the protein where many secondary structures are present. A large number of non-covalent interactions take place between amino acids in the tertiary structure. Most commonly, a hydrophobic core is formed. The tertiary structure is often the overall shape of the folded protein. The quaternary structure is the shape or structure which is formed by more than one protein molecule. These molecules are held together by non-covalent interactions.

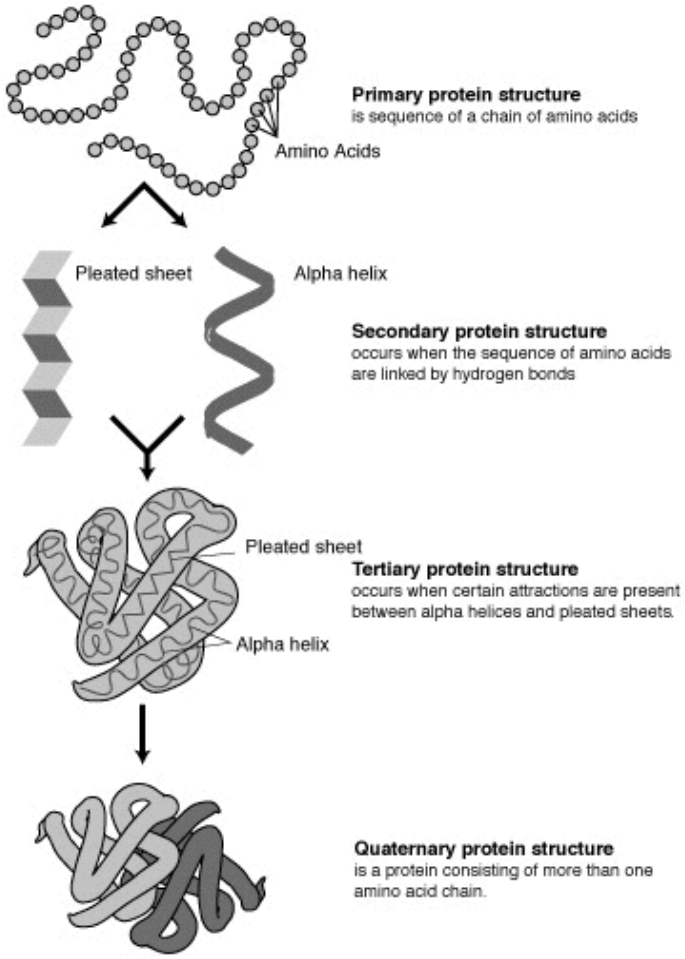


Figure 1.3: Structures of proteins. Figure is borrowed from [5].

Contribution

In this thesis, a new approach to simulate the birth of a protein is presented. This approach is based on the energy minimization of the protein while it is being synthesized in the exit tunnel of the ribosome. The model is implemented step by step to analyze the folding mechanism of the nascent protein. The numerical analysis is performed using a coarse-grained topology-based model. The method is implemented using different folding rates. The results for these different folding rates are analyzed and compared with the literature. The results are grouped according to the three stages of folding event: early stage, intermediate stage and late stage. In addition to

these three stages, a final rearrangement stage was performed to form a more compact structure.

Outline

In Chapter 2, the related approaches for the protein folding problem are introduced. Chapter 3 illustrates the model to simulate the folding dynamics of the nascent protein chain. In Chapter 4, the method is implemented using different growth rates and the results are analyzed and the importance of long-range contact pairs is presented. Chapter 5 concludes the thesis study.

Chapter 2

THE PROTEIN FOLDING PROBLEM

Proteins are the basic units of life for all living cells. Understanding their structure and function is necessary to understand how life works.

Protein folding is the process by which a protein assumes its functional shape or native conformation. By folding into a specific three-dimensional shape they are able to perform their biological function. The amino acid sequence of the protein determines the structure of the protein and the structure of the protein determines the function. Also, it can be noted that the function of the protein depends on the ability of the protein to fold rapidly and reliably to its native structure.

Many proteins fold into their three dimensional structures during their synthesis inside cells. Folding depends on the characteristics of their surrounding solution. These characteristics are mainly the type of the solvent which the protein is synthesized, the concentration of salts, the temperature and molecular chaperons. Chaperons are proteins whose function is to assist other proteins in achieving proper folding. Many proteins can fold in the absence of chaperons, but some proteins strictly require them [3]. The essential fact of folding, however, remains that the amino acid sequence of each protein contains the information that specifies both the native structure and the pathway to attain that state: Folding is a spontaneous process. The passage of the folded state is mainly guided by Van der Waals forces and entropic contributions to the Gibbs free energy: an increase in entropy is achieved by moving the hydrophobic parts of the protein inwards, and the hydrophilic ones outwards. This endows surrounding water molecules with more degrees of freedom [3].

In certain solutions and under some conditions proteins may not fold. Temperatures above or below the range that cells tend to live in will cause proteins to unfold or

denature (this is why boiling makes the white of an egg opaque). High concentrations of solutes and extremes of pH can do the same. A fully denatured protein lacks both tertiary and secondary structure, and exists as a so-called random coil [3].

Problems may occur in the process of protein folding. Wrongly folded proteins may cause many diseases. For instance, prion is related to the illnesses such as Creutzfeldt-Jakob disease and Bovine spongiform encephalopathy (mad cow disease), and amyloid is related to illnesses such as Alzheimer's Disease. So, in recent years, protein folding has become a focus of attention in pharmaceutical research: it is probable that new approaches to the treatment of diseases such as cancer and Alzheimer's disease are to be found within its convoluted pathways [3].

The type of the protein determines the duration of the folding process. Small proteins, which have one hundred or so amino acids, can typically fold on time scales of milliseconds. The very fastest known proteins can fold within a few microseconds. The Levinthal paradox, proposed by Cyrus Levinthal in 1969, says that, if a protein had folded by searching all possible conformations randomly, it would take an astronomical amount of time to reach the native state, even if the conformations were generated rapidly. For instance, for a protein with 100 amino acids, when it is assumed that each amino acid can adopt only 3 possible conformations, the total number of conformations would be $3^{100} = 5 * 10^{47}$, so it would take $10^{13}s$ to change each conformation, the time required to test all possible conformations would be $5 * 10^{34}s$ or 10^{27} years, which is longer than the age of the universe ($14 * 10^{9}yr$). But, the protein can fold within seconds. Based upon this calculation, it can be stated that proteins fold much faster in real life. Levinthal then proposed that folding process is not composed of random conformation sampling, and the protein follows a pre-determined path while folding [4].

Folding and unfolding rates also depend on environment conditions like temperature, solvent viscosity, pH and more. The folding process can also be slowed down (and the unfolding sped up) by applying mechanical forces, as revealed by single-molecule experiments [3].

2.1 Energy Landscape Theory of Protein Folding

The reversible folding of a single protein means that the protein in the native state is thermodynamically stable, and therefore that the native state has the global minimum free energy of all kinetically accessible structures. Furthermore since the folded state is a small ensemble of conformational structures compared to the conformational entropy present in the unfolded ensemble, the folded structure must then have the lowest internal energy of all kinetically accessible conformational structures [4].

2.2 Computational Methods for Protein Folding

De novo or *ab initio* techniques for computational protein structure prediction employ simulations of protein folding to determine the protein's final folded shape. The methods for studying protein folding can be listed in two groups: Molecular Dynamics simulations and coarse-grained models.

2.2.1 Molecular Dynamics

The method of molecular dynamics simulations (MD) is one of the most important tools in the theoretical study of biological molecules. This computational method calculates the time dependent behavior of a molecular system. MD simulations provide detailed information on the fluctuations and conformational changes of proteins and nucleic acids. These methods are routinely used to investigate the structure, dynamics and thermodynamics of biological molecules and their complexes. They are also used in the determination of structures from x-ray crystallography and from NMR experiments [7].

Molecular dynamics simulations generate information at the microscopic level, including atomic positions and velocities. The conversion of this microscopic information to macroscopic variables such as pressure, energy, heat capacities, etc., requires statistical mechanics knowledge. Statistical mechanics is fundamental to the study of biological systems by molecular dynamics simulation [8].

In a molecular dynamics simulation, the macroscopic characteristics of a system can be analyzed through microscopic simulations, for instance, the changes in the binding free energy of a particular drug candidate can be calculated, or the energetics and mechanisms of conformational change can be observed. The connection between microscopic simulations and macroscopic properties is made via statistical mechanics which provides the rigorous mathematical expressions that relate macroscopic properties to the distribution and motion of the atoms and molecules of the N-body system; molecular dynamics simulations provide the means to solve the equation of motion of the particles and evaluate these mathematical formulas. With molecular dynamics simulations, both thermodynamic properties and/or time dependent phenomenon can be studied [30].

The molecular dynamics simulation method is based on Newton's second law or the equation of motion, $F = ma$, where F is the force applied on the particle, m is its mass and a is its acceleration. From the force on each atom, it is possible to determine the acceleration of each atom in the system. Integration of the equations of motion then yields a trajectory that describes the positions, velocities and accelerations of the particles as they change with time. From this trajectory, the average values of properties can be calculated. The method is deterministic; once the positions and velocities of each atom are known, the state of the system can be predicted at any time in the future or the past. Molecular dynamics simulations can be time consuming and computationally expensive. However, nowadays computers are getting faster and cheaper. Simulations of solvated proteins can be calculated up to the nanosecond time scale [31].

2.2.2 Monte Carlo Simulations

MC is distinguished from other simulation methods (such as molecular dynamics) by being stochastic, that is nondeterministic in some manner, usually by using random numbers (or more often pseudo-random numbers), as opposed to deterministic algorithms [6].

The use of MC methods to model physical problems allows us to examine more complex systems than we otherwise can. Solving equations which describe the interactions between two atoms is fairly simple; solving the same equations for hundreds or thousands of atoms is impossible. With MC methods, a large system can be sampled in a number of random configurations, and that data can be used to describe the system as a whole [6].

2.3 Coarse-Grained Models

Lattice proteins are simplified computer models of proteins which are used to investigate protein folding. Because proteins are such large molecules, containing hundreds or thousands of atoms, it is not possible with current technology to simulate more than a few microseconds of their behaviour in all-atom detail. Lattice proteins, however, are simplified in two ways: the amino acids are modeled as single beads rather than modeling every atom, and the beads are restricted to a rigid (usually cubic) lattice. This simplification means that they can fold to their energy minima in a time quick enough to be simulated [9].

Lattice proteins are made to resemble real proteins by introducing an energy function, a set of conditions which specify the energy of interaction between neighbouring beads, usually taken to be those occupying adjacent lattice sites. The energy function mimics the interactions between amino acids in real proteins, which include steric, hydrophobic and hydrogen bonding effects. The beads are divided into types, and the energy function specifies the interactions depending on the bead type, just as different types of amino acid interact differently. One of the most popular lattice models, the HP model, features just two bead types - hydrophobic (H) and polar (P) - and mimics the hydrophobic effect by specifying a negative interaction between H beads [9].

For any sequence in any particular structure, an energy can be easily calculated from the energy function. For the simple HP model, this is simply the sum of all the contacts between H residues that are in contact in the structure, but not adjacent atoms in the chain. Most researchers consider a lattice protein sequence protein-like

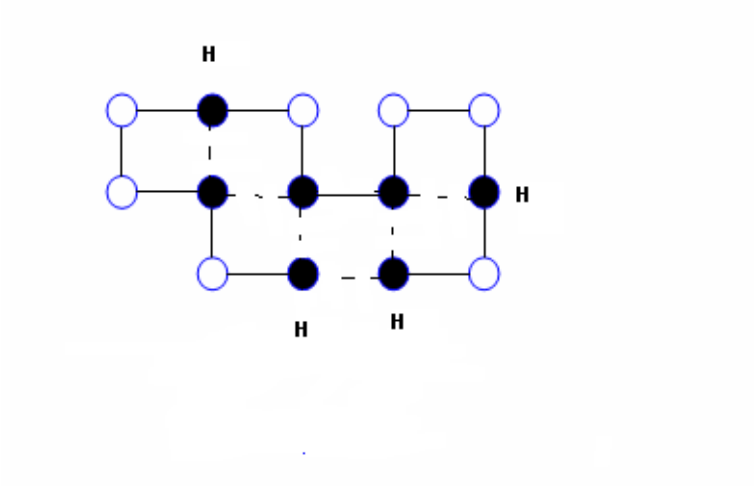


Figure 2.1: A minimum energy conformation in the 2D HP model with 6 nonlocal H-H contacts. Hydrophobic ones are in black color.

only if it can be found in a single structure with an energetic state lower than in any other structure. In Figure 2.1, a lattice protein which has six non-local interactions can be seen. This is called as energetic ground state, or native state. The relative positions of the beads in the native state constitute the lattice protein's tertiary structure. Lattice proteins do not have genuine secondary structure, although some researchers have claimed that they can be extrapolated to real protein structures which do include secondary structure, by appealing to the same law by which the phase diagrams of different substances can be scaled onto one another [9].

By varying the energy function and the bead sequence of the chain (the primary structure), effects on the native state structure and the kinetics (rate) of folding can be explored, and this may provide insights into the folding of real proteins. In particular, lattice models have been used to investigate the energy landscapes of proteins, i.e. the variation of their internal free energy as a function of conformation [9].

Another popular coarse-grained model of protein folding is the Go-type model. In the Go-type model, the native configuration of the protein is assumed known. It is assumed that the beads of the chain are subject to a Go-type potential energy. In this potential function, the interactions between pairs of residues that are in known

positions in the native state are assumed known in advance, thus the use of a go-type model essentially tells the beads of a protein where to go at the end of trajectory, but not how to go [19]. Erman developed Langevin dynamics of protein molecule with Go-type potentials. Long time-scale events in the folding of cytochrome c were analyzed [25].

Extensive studies have been performed using Go-type models. A coarse-grained model was introduced by Hoang and Cieplak, where the Langevin equation is solved for a protein chain whose beads are subject to a Go-type potential [26]. Pande and Rokshar studied a protein-like heteropolymer by using direct simulation of a lattice model using Go Model [27]. In this model, the energy of each polymer conformation is taken to be proportional to the number of nearest neighbor native contacts it possesses [27].

2.4 Folding Dynamics of Nascent Protein

A ribosome is an organelle composed of ribosomal proteins and ribosomal RNA. It translates messenger RNA into a protein. As discussed in the previous chapter, the nascent protein chain must fold into its native conformation so that it can perform its function.

The geometry of the polypeptide exit tunnel is widely analyzed using the crystal structure of the ribosome. The tunnel is a component of a much larger, interconnected system of channels accessible to solvent that permeates the subunit and is connected to the exterior at many points [34]. The structure referred to as the tunnel is the only passage in the solvent channel system that is both large enough to accommodate nascent peptides, and that traverses the particle. At no point is the tunnel big enough to accommodate folded polypeptides larger than alpha-helices [34]. Recent studies have investigated the mechanism of peptide bond formation catalyzed by the large ribosomal subunit, the interaction of nascent polypeptides with the ribosomal exit tunnel, and the role of ribosomal proteins in the recruitment of accessory factors that assist protein folding and targeting [33].

Recent studies focus on the determinants of compact structure formation inside the tunnel. Using an extended nascent peptide as a molecular tape measure of the ribosomal tunnel, helix formation inside the tunnel is presented by Lu et. al. [12, 35]. They monitored the formation of compact structure in the nascent peptide. They showed that there are zones of secondary structure formation inside the ribosomal exit tunnel. It can be suggested that these zones have an active role in nascent-chain compaction [35].

Studies by Woodland et al. and Gilbert et al. provide experimental evidence for the folding problem in the exit tunnel of ribosome [21] [22]. They used stalled ribosome to monitor how a newly translated polypeptide chain travels through the tunnel of the ribosome. Their studies show that the compaction of chain segments and burial of some hydrophobic amino acids take place before the chain leaves the ribosome, provided that this compaction is not reversed as the chain moves down the ribosomal tunnel. In Figure 2.2, these observations can be seen in simple terms [23].

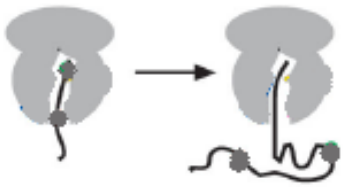


Figure 2.2: The nascent protein begins to fold as it moves out of the ribosomal tunnel [23].

The problem of how proteins fold into their native conformations is an appealing study field in computational biology. Most previous research has focused on how proteins fold from denatured conformations *in vitro*. However, Fedorov et. al. presented that the fully unfolded form of a complete polypeptide does not exist within the living cell [37]. As discussed in the previous section, it is essential to understand the protein folding as it occurs *in vivo*. Recently, many experimental studies present that proteins begin to fold while being synthesized. Thus, more realistic methods are performed to simulate the protein folding problem [36, 11]. Elcock et. al. analyzed

the folding problem as the chains synthesized by the ribosome. They performed their method on three proteins: chymotrypsin inhibitor 2 (CI2), barnase, and semliki forest virus protein (SFVP), and compared the folding during ribosome-mediated synthesis with their refolding from random, denatured conformations as performed in many computational methods [36]. This kind of folding is named as co-translational folding [2, 10]. Elcock et. al. proposed that multi-domain proteins fold co-translationally [36]. Fedorov et. al. also suggested that that co-translational folding contributes to the rapid formation of the native structure in the cell [37].

Chapter 3

THE OPTIMIZATION MODEL

In this chapter, the mathematical formulation of the optimization problem of protein folding is introduced. The optimization model developed by Guner et. al. is implemented [13].

3.1 Theory of the Model

In this model, a coarse grained model, C^{α} representation for proteins is used. The position of the i^{th} residue is denoted by the vector r_i according to the position of the alpha carbon atom of the residue.

The total energy of the protein is composed of energy which results from the bonded and non-bonded interactions within the protein, these energies are denoted by E^B and E^{NB}, respectively. Both of the energies have an attractive and a repulsive component. Thus, the energy formula for a system with N beads can be written as:

$$E = \sum_{i=1}^{N-1} \left(E^B_{i,i+1;A} + E^B_{i,i+1;R} \right) + \sum_{i>j+1}^{N} \left(E^{NB}_{i,j;A} + E^{NB}_{i,j;R} \right) \tag{3.1}$$

Here, the residue indices are represented by the subscripts i and j, A and R stands for the attractive and repulsive parts of the energy, and the superscripts B and NB represent the bonded and non-bonded parts, respectively. The force vector f_i operating on the i^{th} bead can be calculated from the total energy which is denoted by E.

$$f_i = -\nabla_{r_i} E \quad \text{, for } i = 1, 2, \ldots N \tag{3.2}$$

In addition to this force it is proposed that each bead a friction force along the

direction opposite to the velocity. According to the Newton´s second law, the equation of motion of a residue becomes

$$m\frac{\partial^2(r_i)}{\partial^2 t} = -\gamma\frac{\partial(r_i)}{\partial t} + f_i \quad \text{for } i = 1, 2, \ldots N \tag{3.3}$$

Where m denotes the mass of the residue; γ denotes the friction coefficient with dimension of $(force)(time)/(distance)$; f_i is the total force operating on the i^{th} bead.

Using Equation (3.2), the force can be calculated in terms of repulsive and attractive parts:

$$f_i = f_{i,A}^B + f_{i,R}^B + f_{i,A}^{NB} + f_{i,R}^{NB} \tag{3.4}$$

Attractive forces between bonded beads $f_{i,A}^B$ are considered as linear spring forces. The attractive forces can be obtained from the potential energy function [28]:

$$E_A^B = -\frac{1}{2}\sum_{i,j}\frac{a_{ij}}{r_{ij}^m} \tag{3.5}$$

where a_{ij}'s and m are constants and $r_{ij} = ||r_i - r_j||$ is the distance between the i^{th} and j^{th} bead. Guner et. al. assumed the energy as a Hookean spring with $m = -2$ and for all i and j, Equation(3.5) can be written as:

$$E_A^B = -\frac{1}{2}\sum_{i=1}^{N-1} a\,||r_{i+1} - r_i||^2 \tag{3.6}$$

Here, a is a constant which changes according to the adopted empirical energy function. So, assuming $a = 1$, Equation(3.6) can be rewritten as:

$$E_A^B = \frac{1}{2}r^T\Gamma_A^B r \tag{3.7}$$

Where $r = [r_1\, r_2\, \ldots\, r_N]^T$ is the position vector set of the beads. Γ_A^B is the linear connectivity matrix [14]. It is a symmetric Toeplitz matrix whose first off-diagonal elements are equal to -1 and the diagonal elements are equal to the negative sum of the corresponding row without its diagonal element.

The attractive bonded forces for all beads can be shown by f_A^B:

$$f_A^B = \begin{bmatrix} f_{1,A}^B \\ f_{2,A}^B \\ . \\ . \\ f_{N,A}^B \end{bmatrix} \tag{3.8}$$

From Equation(3.2) and Equation(3.7), the attractive force between bonded pairs can be derived:

$$f_A^B = \begin{bmatrix} f_{1,A}^B \\ f_{2,A}^B \\ . \\ . \\ f_{N,A}^B \end{bmatrix} = \begin{bmatrix} -\nabla_{r_1} E_A^B \\ -\nabla_{r_2} E_A^B \\ . \\ . \\ -\nabla_{r_N} E_A^B \end{bmatrix} = -\nabla_r E_A^B = \Gamma_A^B r \tag{3.9}$$

The sum of all remaining forces in Equation (3.2) (i.e. bonded repulsive, non-bonded attractive and non-bonded repulsive forces) is denoted by u_i :

$$u_i = f_{i,R}^B + f_{i,A}^{NB} + f_{i,R}^{NB} \tag{3.10}$$

These forces for all the beads can be written in one term:

$$u = \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ u_N \end{bmatrix} \tag{3.11}$$

The left hand side of Equation(3.3) can be equated to zero as it is a very small term compared to the other terms and also the friction coefficient can be taken equal to unity as an assumption.

Taking into account these definitions, the equation of motion for the protein model can be described by:

$$\frac{dr}{dt} = \Gamma_A^B r + u \tag{3.12}$$

3.2 The Optimization Formulation

The optimization model focuses on the native contact pairs. Using the energy between these pairs as the main force which drives the protein into its native state, the problem is stated as a minimization problem. The native contact pairs are the pairs which are two or more residues apart and separated by less than $7\mathring{A}$ in the native state.

The attractive energy of non-bonded native contact pairs is represented as:

$$E_A^{NB} = \frac{1}{2} \sum_{i>j+1} b_{ij} \, ||r_i - r_j||^2 \tag{3.13}$$

where b_{ij} are constants and $r_{ij} = ||r_i - r_j||$ is the distance between the native contact pair which results from i^{th} and j^{th} beads. Assuming b_{ij}'s are equal to unity without any loss of generality, Equation(3.13) can be written in quadratic form in (3.14), where Q is the matrix which relates the state vectors to the sum of the distances between native contact pairs.

$$E_A^{NB} = \frac{1}{2} r^T Q r \tag{3.14}$$

The optimization problem can be expressed as a constrained optimal control problem to solve the equations of the state-space model stated above. In this state-space model, state variables are the positions of the beads of the chain, input variables are the forces that act on each bead in x, y, z dimensions. These forces drive the protein into its native state starting from an initial condition. Throughout this pathway, there are path constraints to be satisfied. Optimization tries to find the most appropriate structure using the information about the distances of native contact pairs, thus the protein folds. The minimization problem is solved for the time period between initial time $t = 0$ and final time t_{final}. t_{final} is chosen as long enough to let the chain settle

to the most native-like structure. For an N bead chain, the minimization problem and the necessary constraints are written below.

$$\min_{u(t)} \left[\int_0^{t_f} E_A^{NB}(t)\, dt = \frac{1}{2} \int_0^{t_f} r^T(t) Q r(t)\, dt \right] \tag{3.15}$$

Subject to:

$$\dot{r} = -\Gamma_A^B r + u(t) \tag{3.16}$$

$$r(t=0) = r_0 \quad (Initial\, condition) \tag{3.17}$$

Constraints:

$$l - \varepsilon \leq r^T H_i r \leq l + \varepsilon \qquad Bond\, length\, constraints\, (i = 2, 3, \cdots N)$$

$$r^T L_i r \geq d_{ij} \qquad Excluded\, volume\, constraints\, (i = 1, 2, ..., N)$$

$$u_* \leq u(t) \leq u^* \qquad Force\, magnitude\, constraints$$

Where:

H_i : $Matrix\, that\, relates\, the\, state\, r\, to\, the\, bond\, lengths$

L_i : $Matrix\, that\, relates\, the\, state\, r\, to\, the\, excluded\, volumes$

l : $Bond\, length\, distance$

d_{ij} : $Minimum\, excluded\, volume\, distance\, between\, i^{th}\, and\, j^{th}\, bead$

u_* : $Lower\, and\, upper\, limits\, on\, the\, forces\, acting\, on\, the\, beads$

p : $Number\, of\, bond\, length\, constraints$

q : $Number\, of\, excluded\, volume\, constraints$

t_f : $Final\, time$

Here, $r^T H_i r$ means the square of the distances between two adjacent atoms, so they indicate the bond lengths between adjacent atoms. The bond constraints are inserted with a 10% tolerance, thus the constraints are inserted as:

$$0.9\, l_b^2 \leq r^T H_i r \leq 1.1\, l_b^2 \tag{3.18}$$

Here, l_b is the virtual bond length which is 3.8 $\mathring{A}$.

$r^T L_i r$ means the distances between non-adjacent beads. The excluded volume constraints are inserted for the pairs which are 2 or more beads apart from each other on the chain. d_{ij} is the excluded volume limit value for that pair, which means that these 2 beads can not come closer than this value during folding process. The limit value for the native contact pairs is the distance values calculated from the native state of the protein, the limit values for the other pairs are all 5.1$\mathring{A}$, which is the approximate hydrogen bond length [15].

The input variables are limited to be between 2 and -2, because for the smaller input limit values, the optimization can not find a feasible solution. For the bigger limit values on the input variables, the changes in the states are very unexpected due to the bigger input variables in the state equations. Smooth state trajectories are obtained when the limit is taken as 2 and more realistic folding patterns with good resolution are obtained. Thus, the input constraints become $-2 \leq u(t) \leq 2$ [13].

The optimization problem is solved using the AMPL environment and PENNON solver is used via AMPL environment [18]. PENNON solver is designed to solve optimization problems with non-linear objectives subject to non-linear inequalities and equalities as constraints such as:

$$\min_{x \epsilon R_N} F(x) \tag{3.19}$$

$$Subject\ to: g_i(x) \leq 0,\ i = 1, ..., m_g \tag{3.20}$$

$$h_i(x) = 0,\ i = 1, ..., m_k \tag{3.21}$$

Where f, g_i and h_i are the functions from R^N to R. $F(x)$ is the energy function of the optimization problem.

The minimization problem is represented as a non-linear program (NLP) [17]. The state variables $x(t)$ and input variables $u(t)$ are expressed in terms of finite elements by using Lagrange polynomials. The discretization method presented by Biegler et al. is used [19]. The problem is solved in the time interval $(0, t_f)$, the time interval is

divided into *ne* intervals, such that $(t_0 < t_1 < t_2 ... < t_{ne} = t_f)$.

The state-space model equations are stated using orthogonal collocation on finite elements as written in Equation 3.22. This collocation method is efficient and robust in order to solve optimization problems while handling non-linear path constraints.

$$r_{i,k} = r_{(i-1),k} + h_{(i-1)} \sum_{l=1}^{l=NCOL} \dot{r}_{(i-1),l,k}\, \Omega_{l,j} \tag{3.22}$$

where, $i = 1, ..., NE$, $k = 1, ..., N$, $l = 1, ..., NCOL$

Here, the index of the time step is denoted by i. k is the index of the position vector. $NCOL$ represents the number of collocation points used on the finite element i. NE is the total number of time steps. N is the number of state variables. $r_{i,k}$ is the k^{th} state variable representing the state of k^{th} bead in the i^{th} time step. h_i is the length of the finite element i. $\dot{r}_{(i-1),l,k}$ is the derivative of k^{th} state variable in the i^{th} time step at the collocation point l. $\Omega_{l,j}$ is the order of the polynomial of order $NCOL$. Inside each finite element, differential state equations are satisfied at the collocation points. t_f is the final time.

The model equation in Equation 3.17 is written as:

$$\dot{r}_{i,l,k} = t_f \sum_{k=1}^{k=N} \left(\Gamma^B_{A,m,k}\, r_{i,k} \; + \; u^k_i\right) \tag{3.23}$$

The objective function in Equation 3.15 can be discretized as:

$$\min_{u(t)} \int_0^{t_f} E^{NB}_A(t)\, dt = \frac{1}{2} \int_0^{t_f} r^T(t) Q r(t)\, dt \;\; = t_f \sum_{i=1}^{NE} \sum_{j=1}^{j=NCOL} E^B_{A,i}\, h_i\, \Omega_{j,NCOL}$$

Here, $E^B_{A,i}$ is the energy value of the i^{th} time step. The sum of the energy values forms the objective function for the minimization problem. The excluded volume constraints and bond constraints are also inserted into the minimization.

In this optimization problem, the optimization uses the input variables $u(t)$'s to drive the protein into its native state. The dynamic model given by 3.17 governs the

motion of the beads under the optimal force field. The optimal force values and the trajectories are computed over the time interval $(0, t_f)$. The bond length constraints and excluded volume constraints are satisfied during folding, thus optimal folding trajectories are obtained for the protein starting from different initial conditions.

3.3 The Optimization Model for Nascent Protein

As discussed in the previous chapter, the nascent protein chain begins to fold as it leaves the exit tunnel of the ribosome. The protein tries to reach its native state as the following residues leave the ribosome one by one. Based on this experimental fact, the optimization model is implemented to monitor the birth of the protein as the chain grows step by step. We define the folding time of a partial chain as the length of time horizon in the optimization formulation. The growth rate is the number of beads exiting the ribosome per unit time.

First, we assume that as the first residues leave the ribosome, they fold into a compact structure according to their folding dynamics and determined by the possible native contact pairs among them. These native pairs are determined as explained in the previous section. Then, the following bead leaves the ribosome, and the previous partial chain and this lastly-added bead endeavor to reach a state which is close to the native state. This process continues until the whole protein chain leaves the ribosome.

The optimization model described in the previous section is implemented for our time-varying system. The folding of a partial chain of a particular length can be thought as a subsystem. So, our system consists of successive subsystems which simulate the whole folding altogether. Equation 3.12 is the equation of motion for the discrete events of the folding process for a particular subsystem i.e. partial chain. The optimization statements and the dimensions of the system for these events are defined according to the length of the structure at that folding period. r is the state vector, its dimension is $3 * N$ for an N-bead chain. Also, u, which implies the input vector, has $3 * N$ elements. As the length of the chain changes, the dimensions of the state vectors change. These dimensions vary in the range starting from $6 * 3 = 18$ to

$36 * 3 = 108$ as our system starts from the 6-bead chain and finishes at the 36-bead chain of chicken villin headpiece protein. Also, the dimensions of the connectivity matrix in Equation 3.12 vary in this range.

The energy term described in Equation 3.13 consists of native contact pairs. As the chain grows, the number of native contact pairs increases. Figure 3.1 shows the number of native contact pairs as a function of the number of beads of the growing chain. There are 89 native contact pairs for the 36-bead chicken villin protein, 8 of these are long range contact pairs.

At any time instance (discrete sample time) optimization computes and implements M control actions. This defines M folding steps that the partially grown chain goes through before the next bead is added and optimization is repeated to compute the next M folding steps.

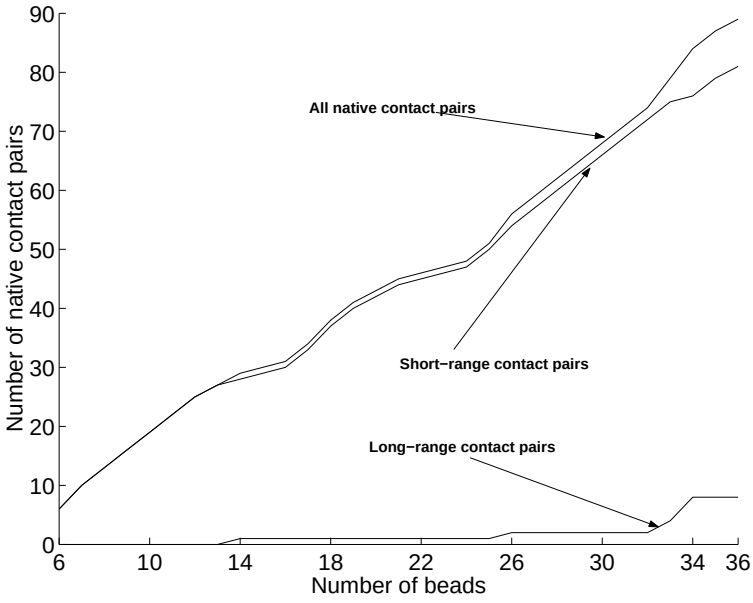


Figure 3.1: Numbers of native contact pairs for discrete subsystems, also numbers of long-range pairs and short-range pairs.

Chapter 4

RESULTS AND DISCUSSION

In this study, the optimization technique is implemented to analyze the folding dynamics of the 36-residue protein, chicken villin headpiece whose protein data bank keyword is 1VII. Chicken villin headpiece protein is the smallest protein that can fold autonomously. This protein is extensively studied in the literature because it is the smallest polypeptide that has all of the properties of a single domain protein [32]. Our technique relies on the fact that one bead is added per M time steps. In other words, growth rate of the protein is one bead per M time steps. In this chapter, the technique is implemented for M = 1, 10 and 15 folding steps and the results are presented.

The secondary structure elements of the chicken villin headpiece protein are listed in Table 4. The protein has three helices which are held together by a loop and a turn. These helices contain residues 4-8, 15-18, 23-30 respectively. The loop is formed by residues 9-14 and the turn is formed by residues 19-22.

Table 4.1: Secondary structures and their corresponding residue numbers.

Residue Number	Secondary Structures
4-8	Helix1
9-14	Loop
15-18	Helix2
19-22	Turn
23-30	Helix3

Throughout this study, for simplicity of analysis, the folding process is divided

into 3 stages according to the folding of the secondary structures of the chicken villin headpiece protein: early stage, intermediate stage and late stage. The early stage is composed of the folding state of first 14 beads. These beads form the first helix and the loop. The folding of the first 30 beads starting from the folding of first 15 beads is called intermediate stage. In this stage, the helix 2 and turn and helix 3 form. From the 30th bead optimization to 36th bead optimization, it is called late stage. The results are analyzed for all these stages. In addition to these three stages, there is rearrangement stage where the protein folds to form a more compact structure.

The native configuration of the chicken villin headpiece protein can be seen in Figure 4.1.

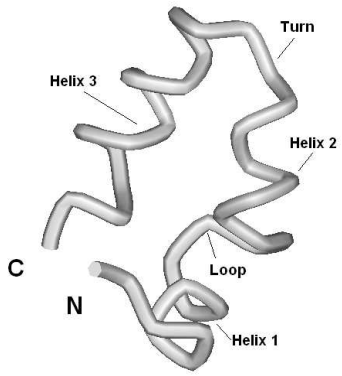


Figure 4.1: 3-D Structure of chicken villin headpiece protein in tube representation (N denotes the N-terminus of the protein, C denotes the C-terminus of the protein).

The native contact pairs defined in Chapter 3 can be classified into 2 main groups: long range contact pairs and short range contact pairs. Long range contact pairs are the pairs which are 5 or more residues apart on the chain of the protein. According to this definition, the whole chain of chicken villin headpiece protein has 89 native contact pairs, 8 of these pairs are long-range contact pairs.

Throughout the study, RMSD value between vectors x and y is calculated according to the Equation 4.1. Here, x and y are vectors of the same size N.

$$RMSD(x, y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)(x_i - y_i)^T} \tag{4.1}$$

Here y and x are sequences of position vector coordinates representing the pair of structures. Each column in y (or x) contains the (x, y, z) coordinates of an atom or point in the structure.

4.1 The Optimization Technique

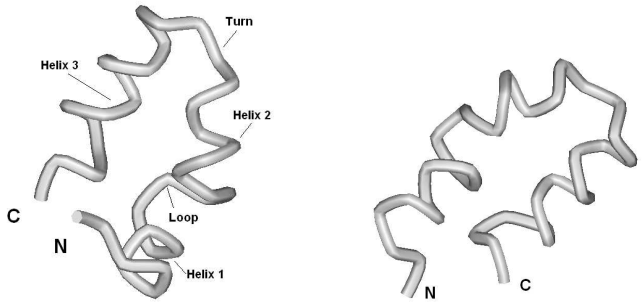
In order to analyze the effect of different growth rates, the optimization technique defined in Chapter 3 is implemented for M = 1, 10 and 15 folding steps.

4.1.1 The Case with 1 Bead per 10 Time Steps

This case considers that after the new bead leaves the ribosome the partial chain folds for 10 time units (takes 10 folding steps) and the system forms a compact structure. A new residue joins the present chain at the 11th time step. The process continues until the 36th bead of the protein is added to the newly-folded 35-bead chain.

Starting from the first five residues which has formed a compact structure, beads begin to join the chain one by one. State variables and input variables are calculated for the whole 10-time-step horizon. The calculated input variables are implemented to fold the protein. At the 10-th time step, the next bead joins the folding process.

The configuration obtained for the 36-bead chain after folding of all beads can be seen in Figure 4.2. The RMSD value between the last configuration obtained in this technique and the native state of the chicken villin protein is 3.39 $\mathring{A}$.



(a) Native structure (b) Configuration obtained

Figure 4.2: Native configuration of villin protein and the obtained configuration in the case with the growth rate one bead per 10 time units.

Figure 4.3, 4.4 and 4.5 show the RMSD change in the early stage, intermediate stage and late stage respectively. The configuration obtained in the early stage and the native state of first 14 beads can be seen in Figure 4.6. The configuration obtained in the intermediate stage and the native state of first 30 beads can be seen in Figure 4.7.

As seen in the RMSD figure of the early stage, each partial chain reduces its RMSD to a base value less than 0.2 $\mathring{A}$ until the 15th bead is added. When the 15th bead is added, the RMSD value jumps to around 1.2 $\mathring{A}$. The first 14-bead part of the protein is composed of one helix and one loop. The 15-th bead is the first bead of helix 2, so when this 15-th bead is out of the ribosome, a native-like structure can not be reached due to the inconsistency of the loop and the first bead of the following helix. When the 17-th bead and the 18-th beads join the chain, RMSD values significantly decrease to lower values than the values of the beginning of that optimization run. The results are not appreciable until the optimization program is run for the 33rd bead. This fact points out the importance of the long range contacts which was stated for the folding of the chain at once [13]. There are 8 long range native contact pairs for the whole protein chain. These pairs are 2-34, 7-14, 7-34, 10-33, 10-34, 11-33, 11-34, 19-26. 2 of these pairs take place in the 33rd bead optimization run, while 4 of them play in the 34th bead optimization run. So, in the last 4 optimization runs,

RMSD value decreases significantly to desirable values.

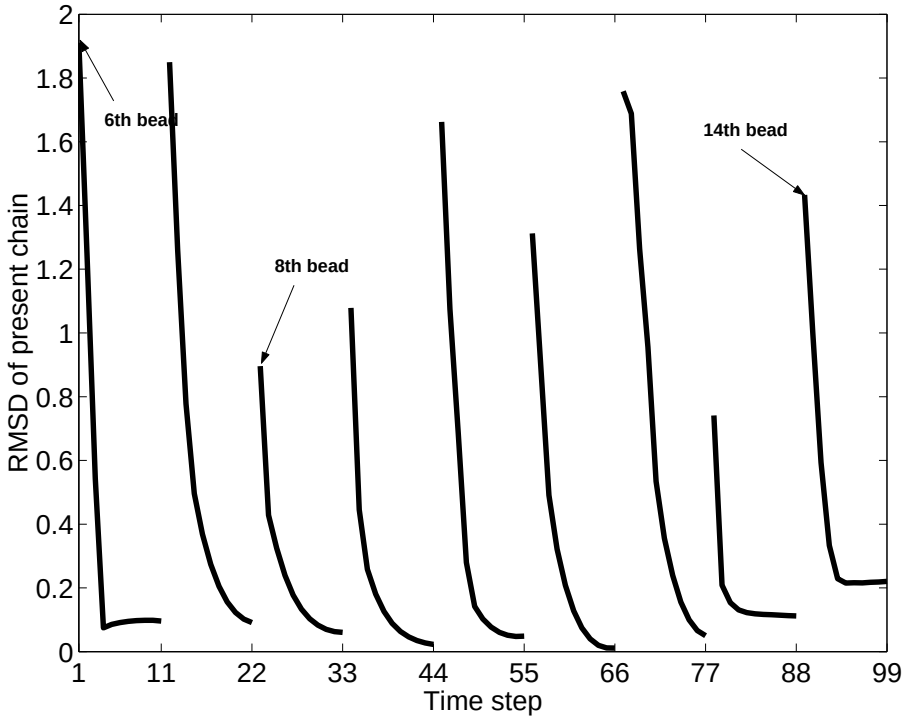


Figure 4.3: RMSD in the early stage of the case with 10 time units per bead. This stage starts with the 6-th bead and ends when the 14-th bead leaves the ribosome.

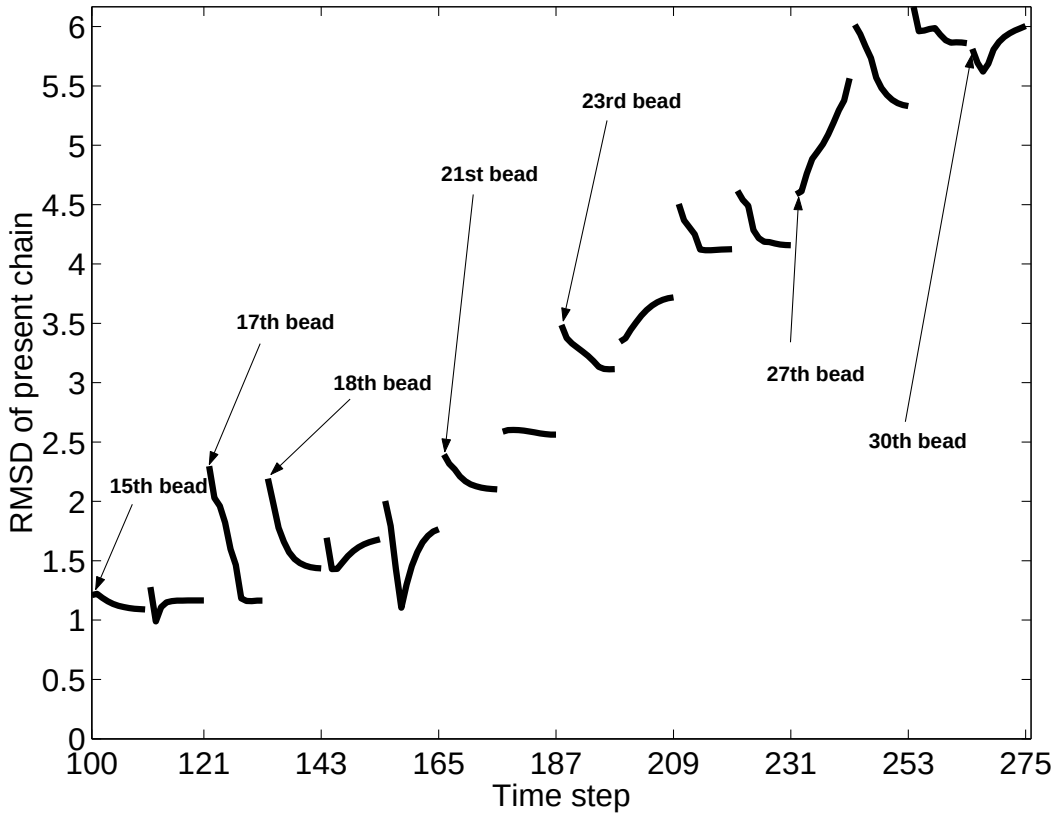


Figure 4.4: RMSD in the intermediate stage of the case with 10 time units per bead.

However, the RMSD value 3.39 $\mathring{A}$ is a large value. This large value indicates the lack of folding time for the chain to fold into a more compact form. So, in the final rearrangement stage, it is assumed that the protein folds for 20 time steps. This stage is called as final rearrangement stage. Starting from the obtained configuration with the first 35-bead structure, the optimization is run for 20 time units. In this case, RMSD value is lower than the previous one, it is 2.92 $\mathring{A}$. The RMSD figure regarding this extra optimization run can be seen in Figure 4.8. As can be seen from the figure, the chain visits the local minimum in the 6-th time step. The reason for this decrease is that the chain endeavors to reach a more native-like state changing the conformation of the first four residues and the last residues.

In order to see how the protein will behave in case it is given some more time to form a more compact structure, the whole chain (36 beads of villin protein) is

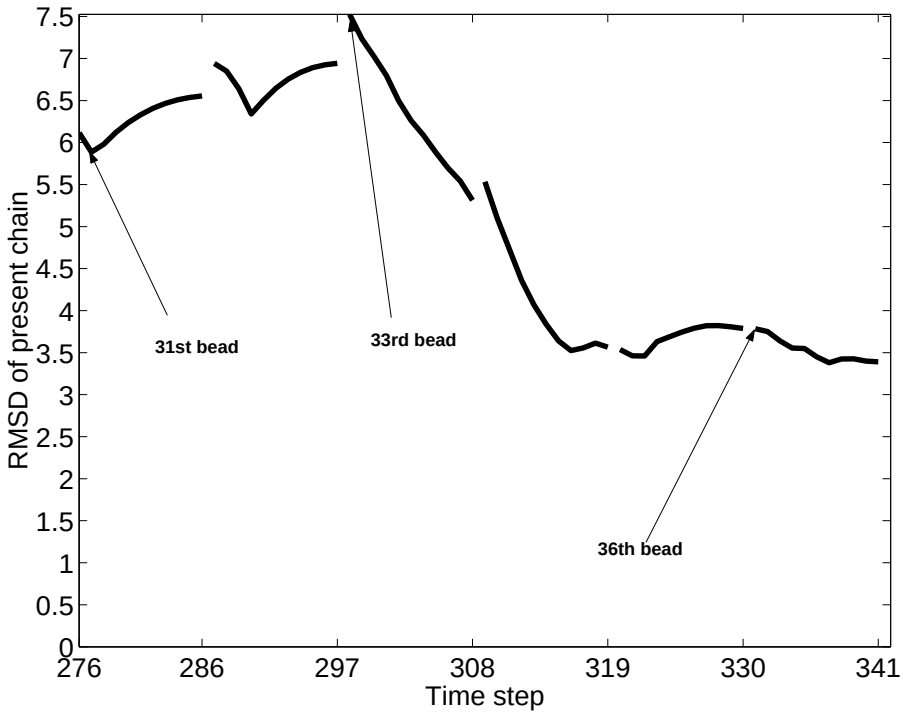
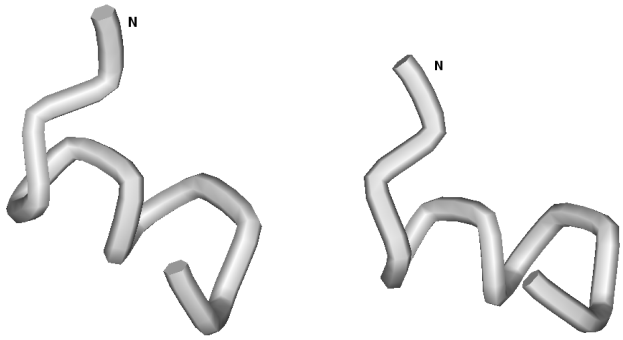


Figure 4.5: RMSD in the late stage of the case with 10 time units per bead.

assumed to fold for 40 time steps. In this case, the obtained RMSD value is 0.26 $\mathring{A}$ which is a much better result. The RMSD change for this 41-time step optimization can be seen in Figure 4.9. As can be seen from the figure, the RMSD value increases between the 10-th time step and 30-th time step after a decrease in the first 10 time steps, and then after the chain collapses into a more compact structure, it decreases. The obtained configuration can be seen in Figure 4.10.

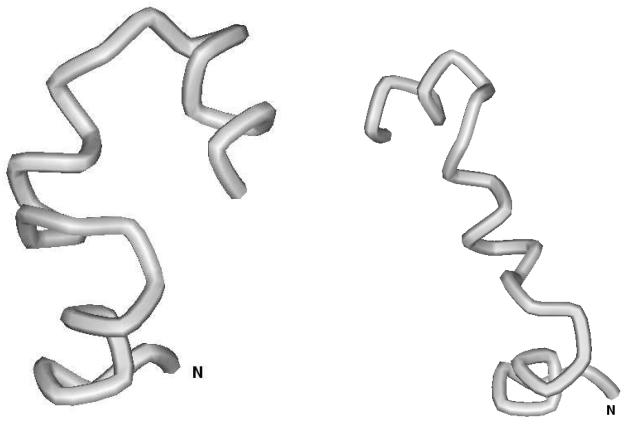
The RMSD values are calculated for the present chain and they are plotted for the 3 stages. The first stage is the one where RMSD can be maintained after the new bead is added. In the second stage, which is the intermediate stage of the folding process, the RMSD can not be decreased to a lower value because of the the absence of the long range-contacts. In the late stage, RMSD values begin to decrease because of the presence of long range contacts. Figures 4.3, 4.4, 4.5 show the RMSD changes in the early stage, intermediate stage and late stage respectively.

As can be seen from the RMSD change figure of the early stage, the chain can



(a) Native structure (b) Configuration obtained

Figure 4.6: Native configuration of villin protein and the obtained configuration in the early stage of the case with the growth rate one bead per 10 time units.



(a) Native structure (b) Configuration obtained

Figure 4.7: Native configuration of villin protein and the obtained configuration in the intermediate stage of the case with the growth rate one bead per 10 time units.

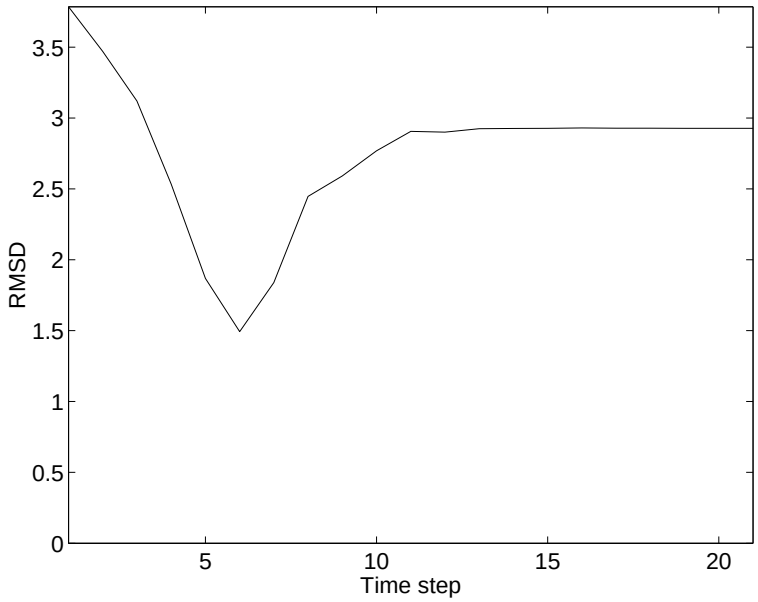


Figure 4.8: Rmsd change when the whole protein is assumed to fold for 20 time units during the rearrangement stage.

conserve the appropriate RMSD value until 15-th bead is added to the chain. The first 14-bead part of the protein is composed of one helix and one loop. The 15-th bead is the first bead of the helix 2, so when this 15-th bead is out of the ribosome, a native-like structure can not be reached due to the inconsistency of the first secondary structures and the first bead of the following bead. When the 17-th bead and the 18-th beads join the chain, RMSD values significantly decrease to lower values than the values of the beginning of the optimization. The results are not appreciable until the optimization program is run for the 33rd bead. This fact points out the importance of the long range contacts which was stated for the folding of the chain at once [13]. There are 8 long range native contact pairs for the whole protein chain as stated above. When 33-rd bead joins the chain contributing to the folding process with an amount of 25% of the total long range contacts through its 2 long range contact pairs, RMSD values decrease considerably. 4 of them take place in the 34th bead optimization run. So, in the last 4 optimization runs, RMSD value decreases significantly to a desirable value.

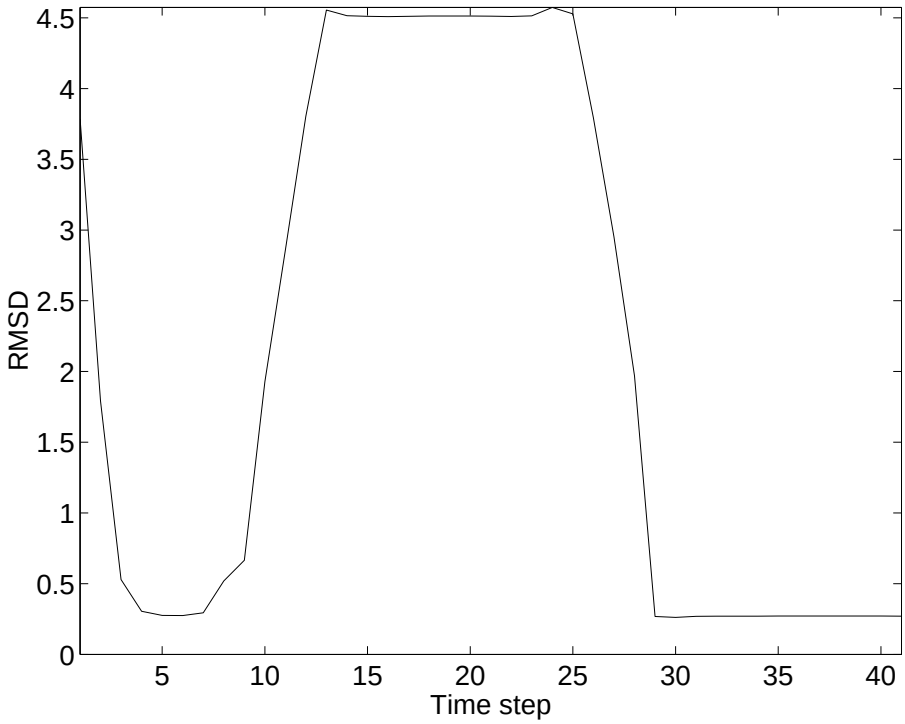
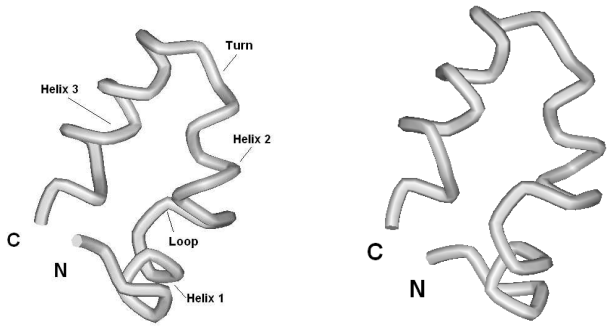


Figure 4.9: Rmsd change when the 36-bead chain is assumed to fold for 40 time units during the rearrangement stage.

Figure 4.11 shows the RMSD change for helix 1 which is the chain starting from the 4-th residue to the 6-th residue. As can be seen from the figure, the conformation of helix 1 does not change much during time after settling to a desirable value starting from a high value in the beginning of the folding process. However, when the interactions of the 34-th bead come into effect, RMSD value increases a bit. The reason for this RMSD change is that it is difficult for the protein to maintain the present RMSD value as helix 3 ends at the 30-th residue and the residues between 30-th bead and the 36-th bead have a turn-like secondary structure. So, RMSD value increases a bit then decreases again in the last stages of the folding.

The RMSD change for loop 1 which is composed of the residues between 9-th and 14-th can be seen in Figure 4.12. After the helix 3 is out of the ribosome completely, the protein can not maintain the RMSD for loop 1, so RMSD increases significantly up to 2.5 A because of the complication of the secondary structures of the last residues.



(a) Native structure (b) Configuration obtained

Figure 4.10: Native configuration of villin protein and the obtained configuration in the case with the growth rate 10 time units per bead followed by a 40 time step rearrangement period.

The RMSD of loop increases in the last 4 optimization steps of folding process, the residues following helix 3 do not have a proper secondary structure, so it is difficult to form compact structures with these residues.

The RMSD change for helix 2 can be seen in Figure 4.13. The conformation of helix 2 does not change much according to the native state of this part, because it is a small alpha helix with 4 beads, thus it is a helix with one helical turn on its own. In addition to being a small helix, for this helix, there are 3 possible contact pairs between its beads and they are all native contact pairs. These pairs are 15-17, 15-18 and 16-18. So, the RMSD value does not change much until the last 4 beads are added to the chain. In the last step, it decreases to 0.15 A.

Figure 4.14 shows the RMSD change for the turn with the residues between the 19-th and 22-th residues. At most stages of the folding process, the RMSD value for turn is larger than the values of the helices, because the protein can not find enough time to reach a native-like turn before the next bead is out of the ribosome. Also, the RMSD value changes rapidly from the beginning of the folding process. Its final value is 0.4 $\mathring{A}$.

Figure 4.15 demonstrates the RMSD change for helix 3, which is the longest helix of the protein. RMSD value does not change a lot once it decreases after the 30-th

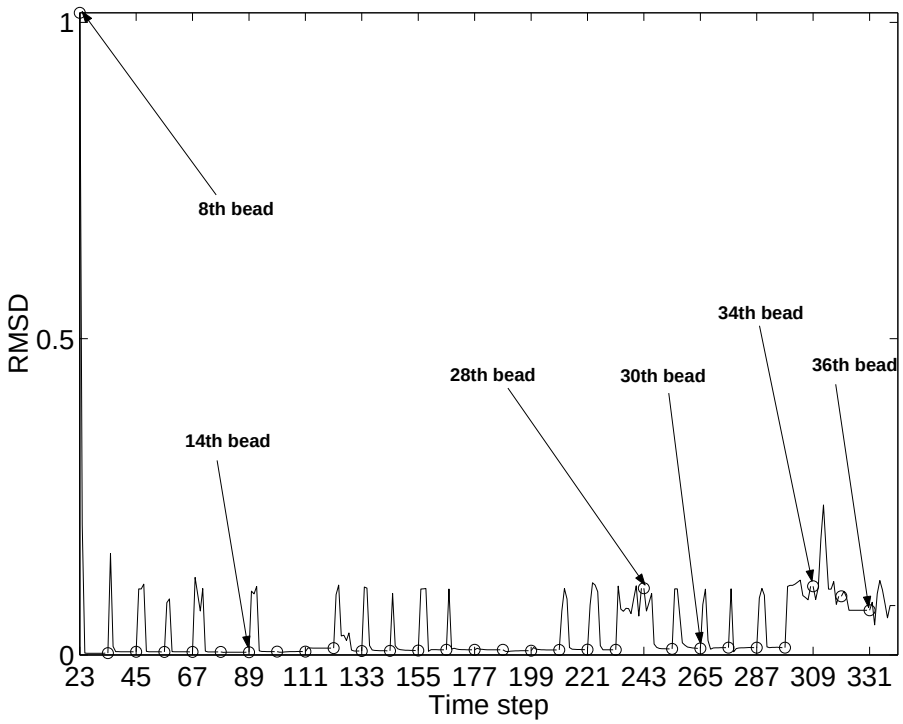


Figure 4.11: Rmsd change of helix 1 (beads 4-8) during folding in the case with 10 time units per bead.

bead is added to the chain.

The change of the energy value for early stages, intermediate stages and late stages can be seen in Figures 4.16, 4.17, 4.18. For the case with 1 bead per 10 time steps, energy values per number of native contact pairs does not change much during time.

4.1.2 The Case with 1 Bead per 15 Time Steps

When the method is conducted with 15 time steps, the RMSD change figures for the 3 stages of folding event are shown in Figures 4.19, 4.20, 4.21 respectively. The early stages are composed of the runs till the 14th bead optimization. The intermediate stages are the ones until the 32th bead leaves the ribosome. The late stages are the ones where 33th bead leaves the ribosome till the whole chain folding.

As can be seen from Figure 4.19, in the early stages of folding, RMSD values are lower than the values for the case with 10 time units, this reveals the fact that

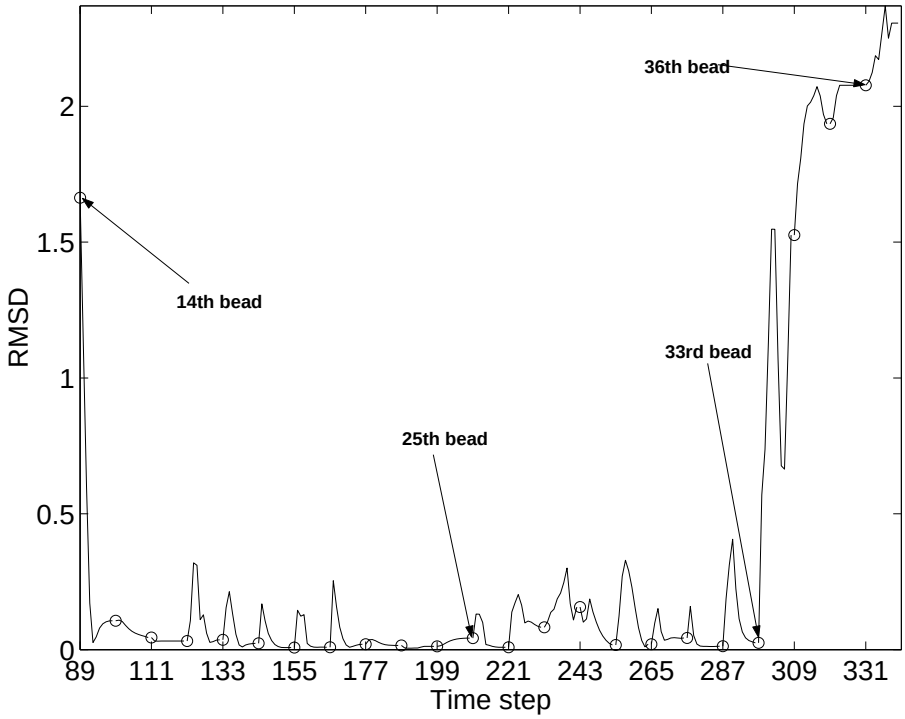


Figure 4.12: Rmsd change of loop (beads 9-14) during folding in the case with 10 time units per bead.

the chain can fold into a native-like structure whenever a new bead joins the present chain.

In the intermediate stages of folding, RMSD increases significantly, because the chain can not maintain the RMSD value because of the lack of proper native contact pairs.

Figure 4.22 shows the native structure of the protein and the conformation obtained in the case with growth rate 15 time units per bead.

The RMSD values of the individual secondary structures from helix 1 to the last helix can be seen in Figures 4.23, 4.24, 4.25, 4.26, 4.27, respectively. As can be seen from Figure 4.23, the RMSD value for the first helix decreases to a desirable value starting from a high value, then it does not change much. However, the RMSD change of loop fluctuates during all folding process. Unlike helix 1, helix 2 has an unstable RMSD property during folding, its RMSD value fluctuates. The second helix is a

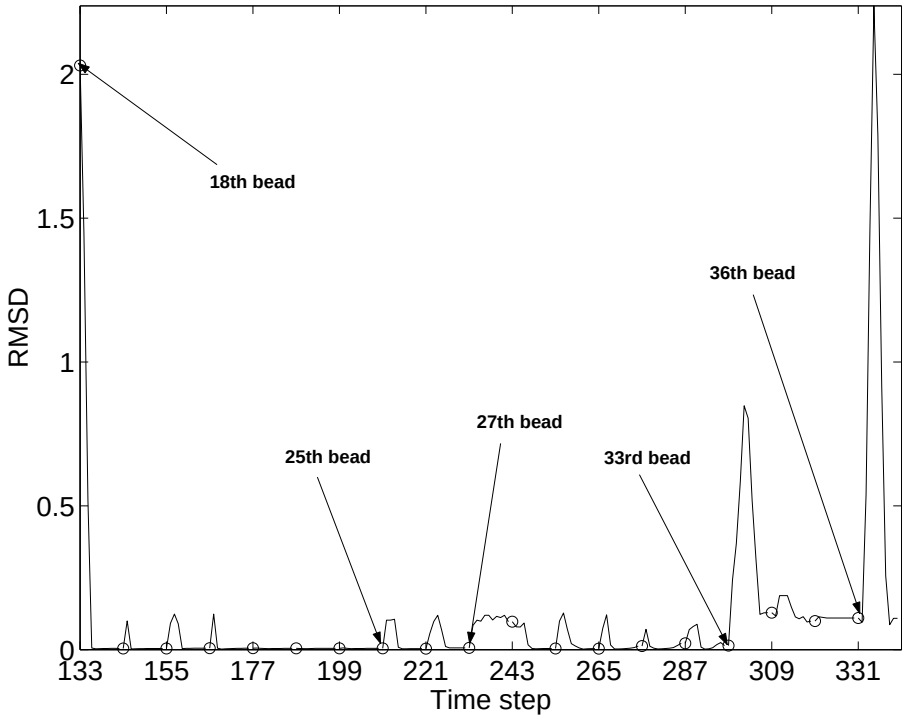


Figure 4.13: Rmsd change of helix 2 (beads 15-18) during folding in the case with 10 time units per bead.

short helix, so it is easy for the optimization model to change the helix 2 whenever necessary [13]. The turn has a large RMSD value overall, because the structure turn is a link between helix 2 and helix 3, so it fluctuates as new native contact pairs are added to the model.

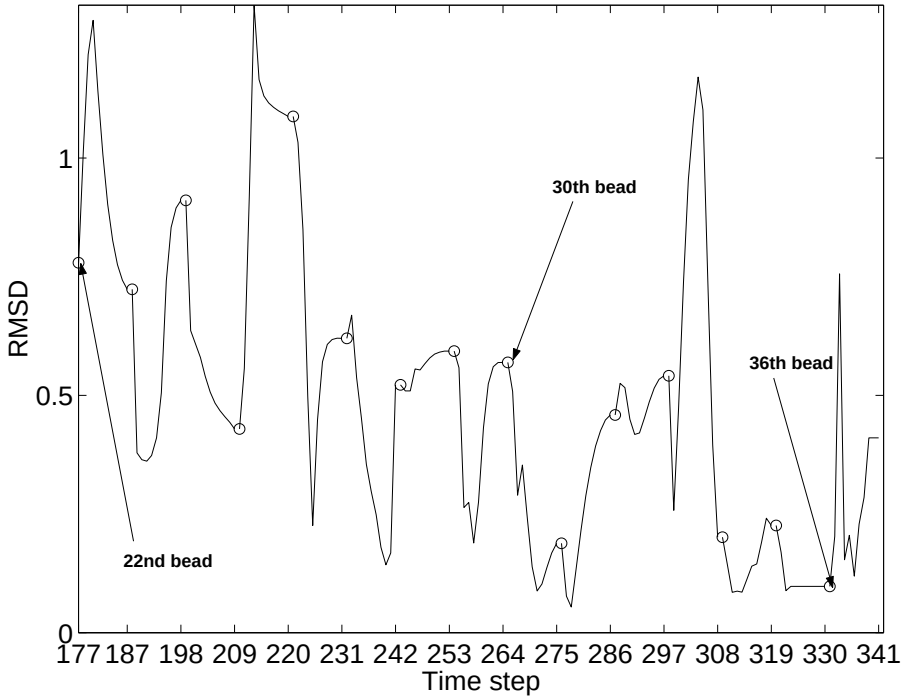


Figure 4.14: Rmsd change of turn (beads 19-22) during folding in the case with 10 time units per bead.

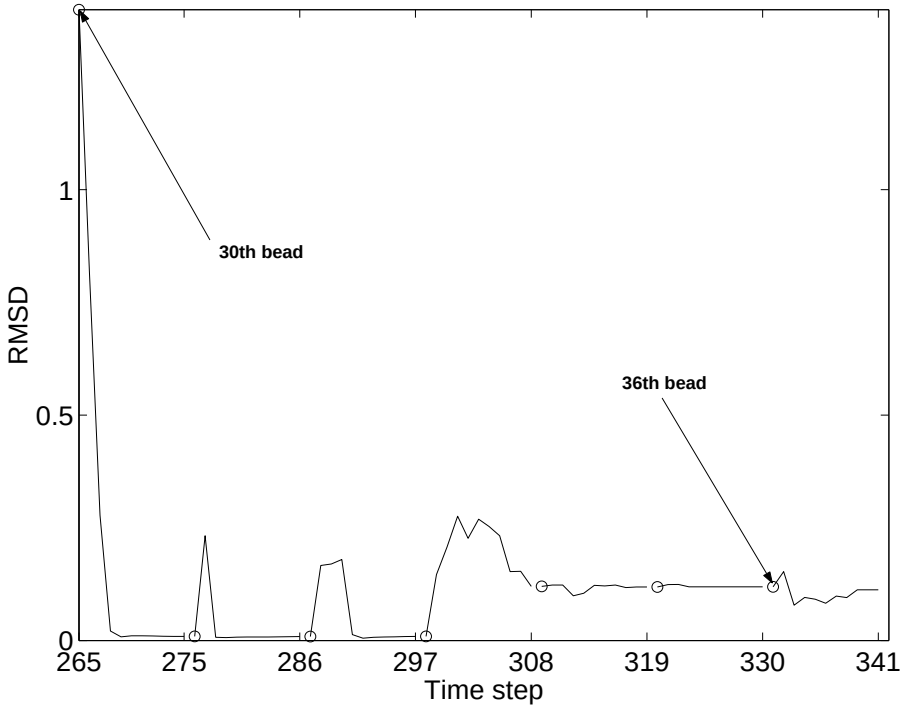


Figure 4.15: Rmsd change of helix 3 (beads 23-30) during folding in the case with 10 time units per bead.

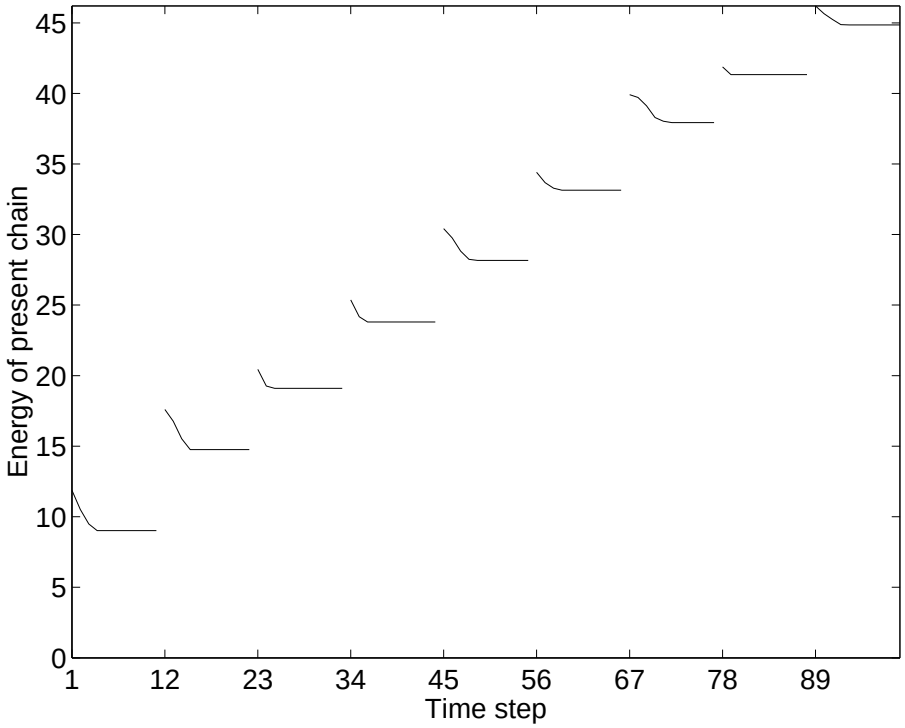


Figure 4.16: Change in minimized energy during folding in the early stage of the case with 10 sample time per bead.

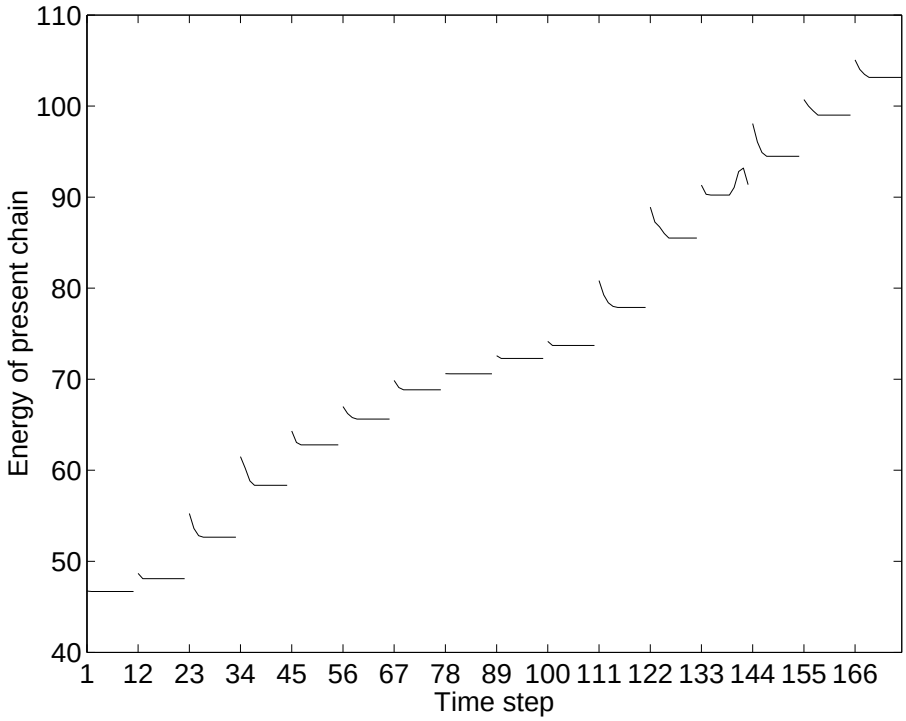


Figure 4.17: Change in minimized energy during folding in the intermediate stage of the case with 10 sample time per bead.

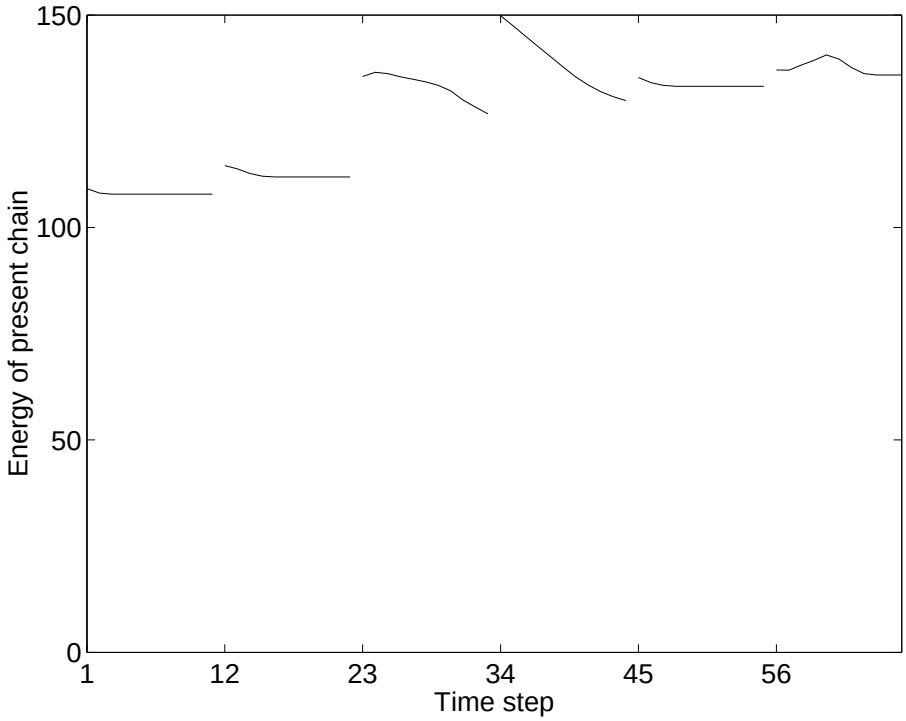


Figure 4.18: Change in minimized energy during folding in the late stage of the case with 10 sample time per bead.

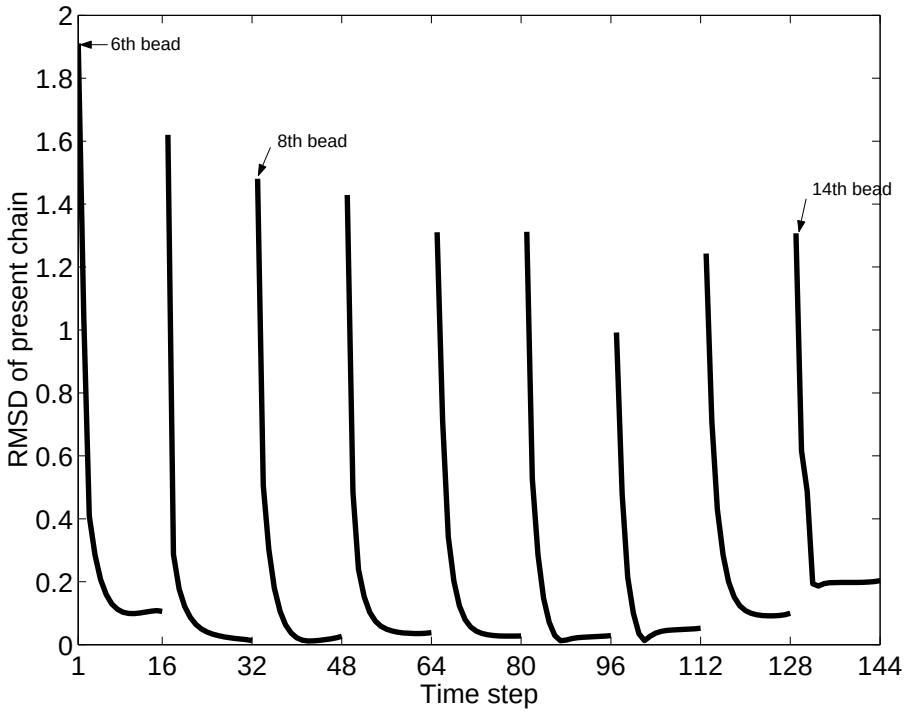


Figure 4.19: Change in rmsd value during folding in the early stage of the case with 15 sample time per bead.

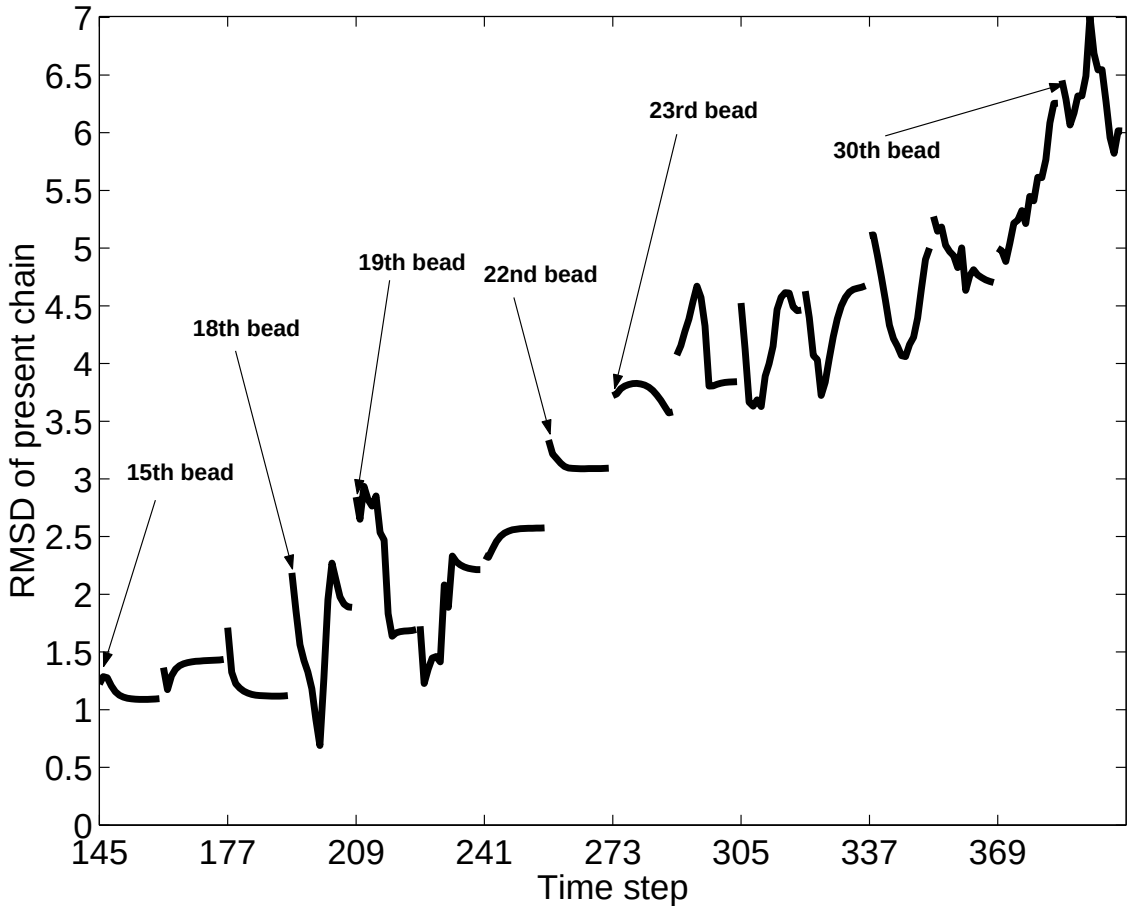


Figure 4.20: Change in rmsd value during folding in the intermediate stage of the case with 15 sample time per bead.

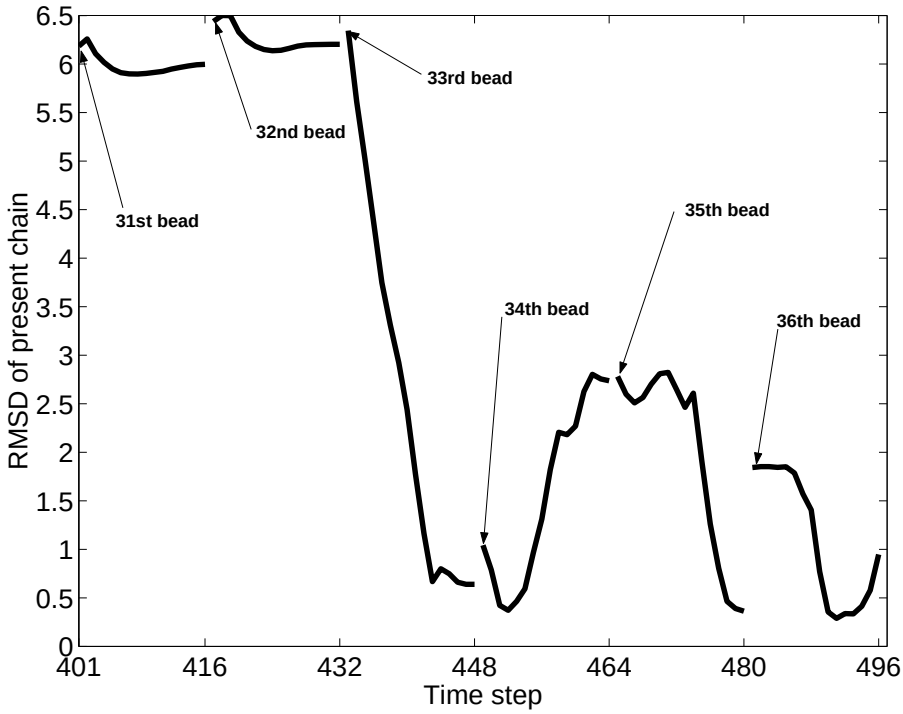
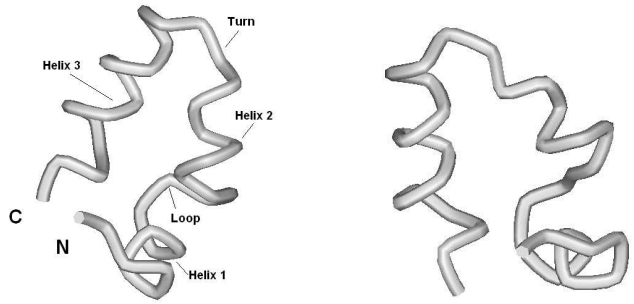


Figure 4.21: Change in rmsd value during folding in the late stage of the case with 15 sample time per bead.



(a) Native structure (b) Configuration obtained

Figure 4.22: Native configuration of villin protein and obtained configuration in the case with the growth rate 15 time units per bead.

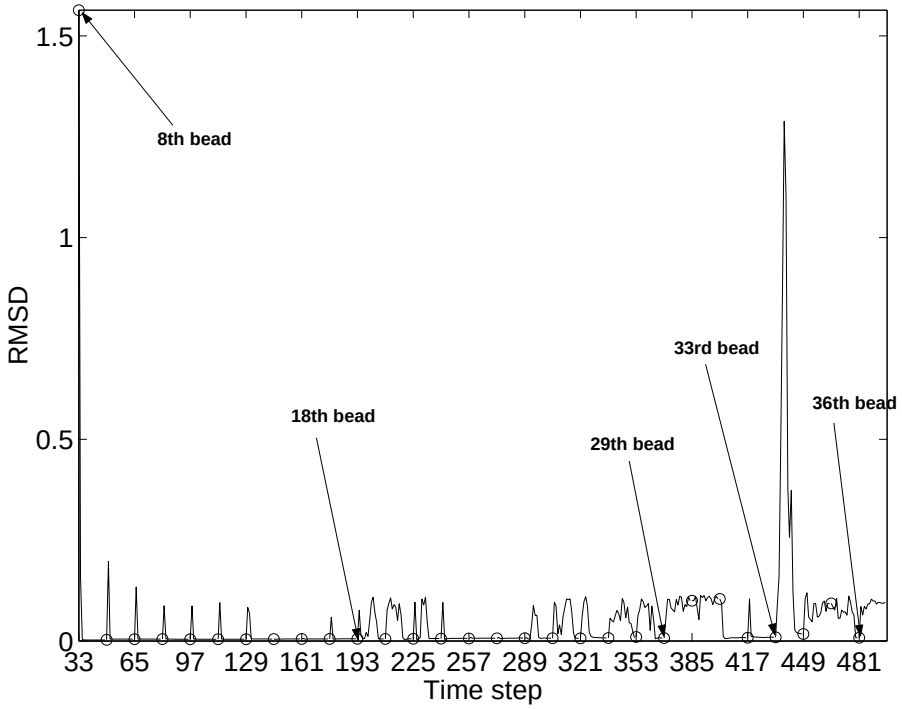


Figure 4.23: Rmsd change of helix 1 during folding in the case with 15 time units per bead.

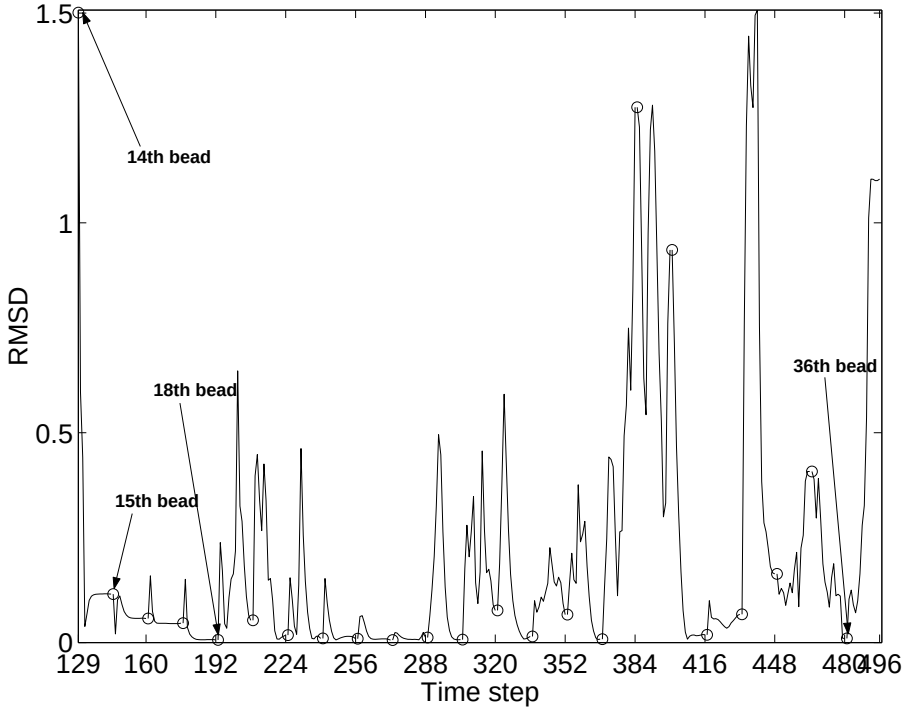


Figure 4.24: Rmsd change of loop during folding in the case with 15 time units per bead.

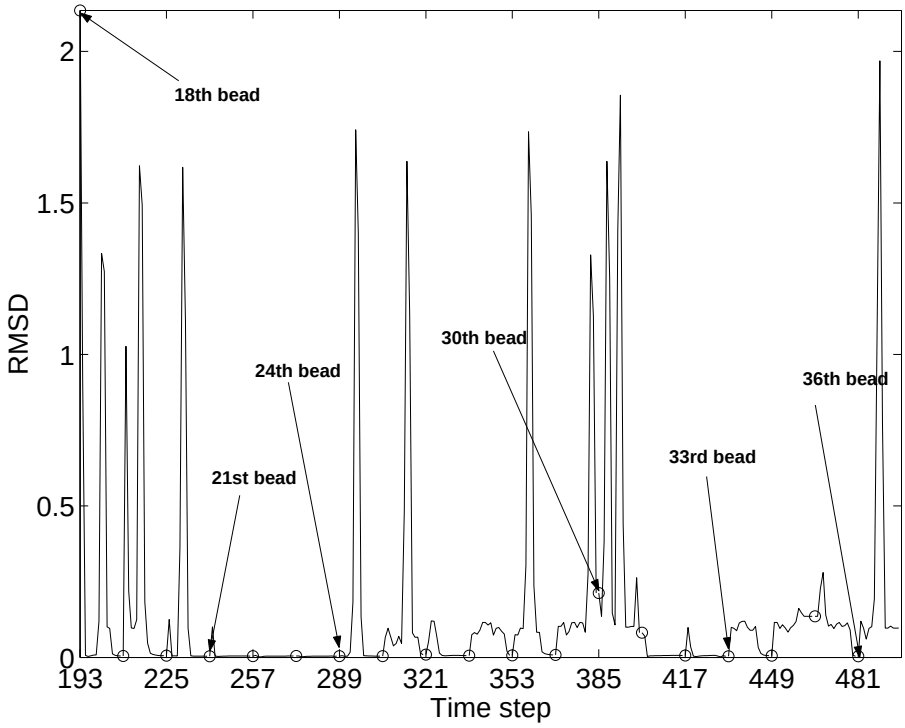


Figure 4.25: Rmsd change of helix 2 during folding in the case with 15 time units per bead.

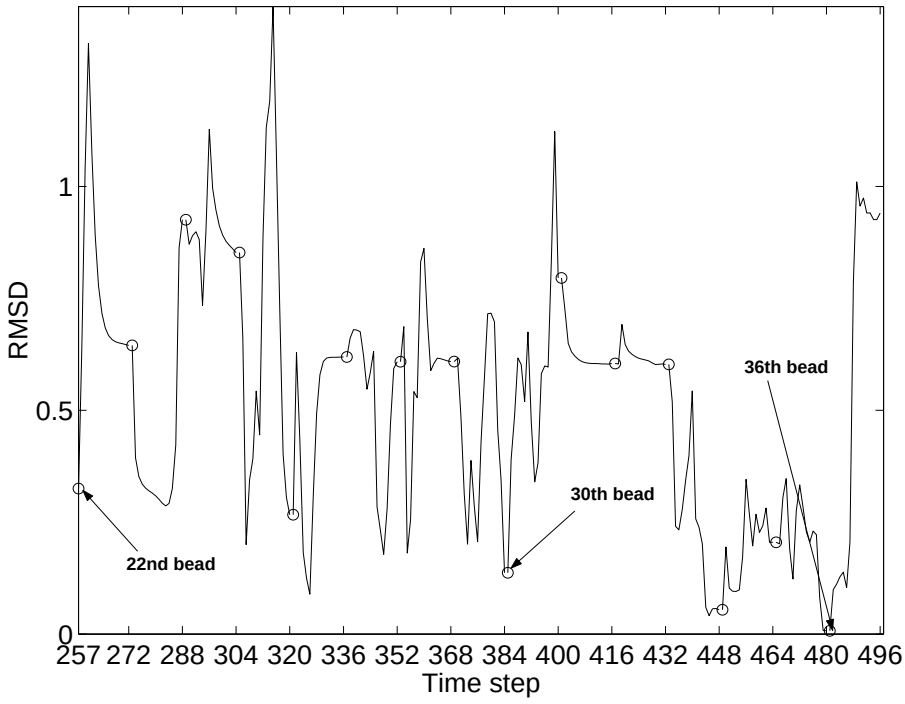


Figure 4.26: Rmsd change of turn during folding in the case with 15 time units per bead.

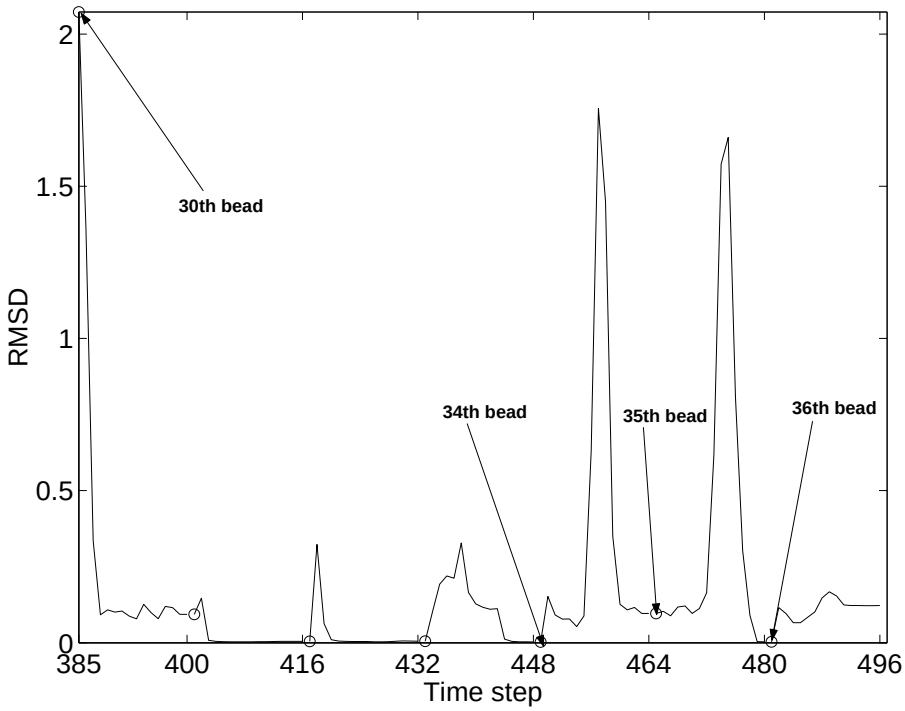


Figure 4.27: Rmsd change of helix 3 during folding in the case with 15 time units per bead.

4.1.3 The Case with 1 Bead per 1 Time Step

In order to see how the protein will fold given less time step per bead, the growth rate of the chain is taken as 1 bead per one time step. A bead joins the chain, then the present chain folds for 1 time step, then another bead is out of the ribosome.

The RMSD change figure for the folding process can be seen in Figure 4.28.

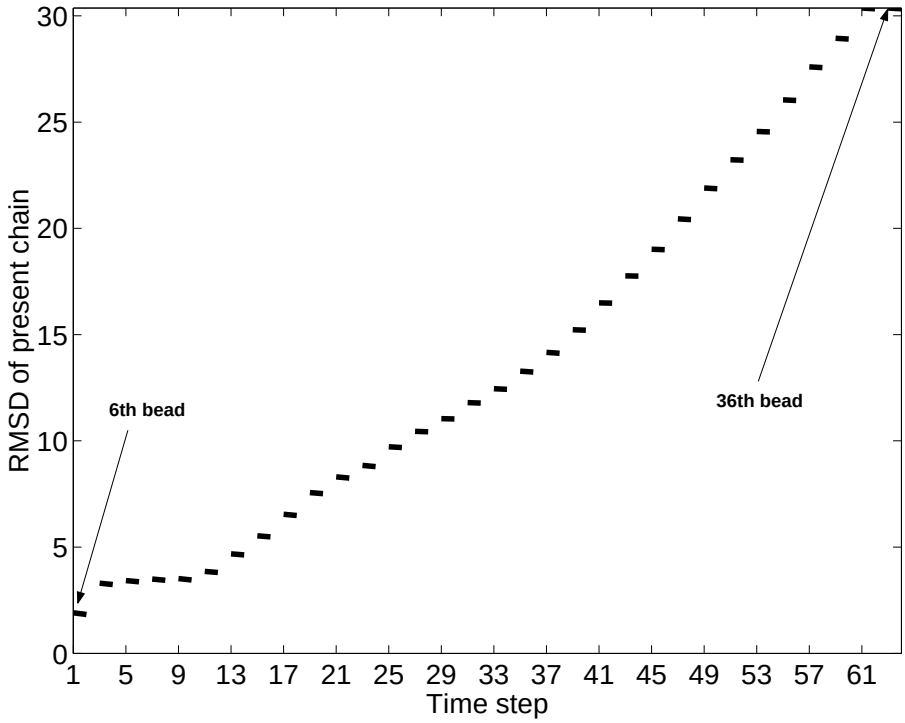
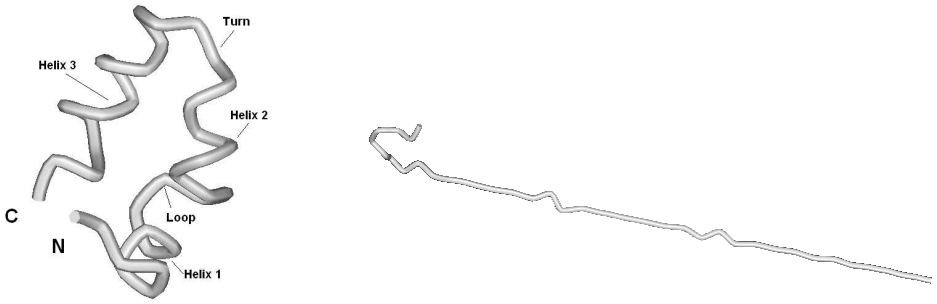


Figure 4.28: Rmsd change in the whole folding process with the growth rate 1 sample units per bead.

This figure reveals the fact that the protein can not fold into a compact structure because of the lack of folding time for the partial chain between the addition of the beads.

In Figure 4.29, the native configuration and the obtained conformation in the case with 1 time unit can be seen.

As can be seen from Figure 4.29, the chain has a 30.33$\mathring{A}$, which is a big value. In the rearrangement stage, the obtained configuration is assumed to fold for another 300 time steps. The RMSD value for the rearrangement stage can be seen in Figure



(a) Native structure (b) Obtained configuration

Figure 4.29: Obtained configuration of villin protein and obtained configuration in the case with the growth rate 1 time unit per bead.

4.30. The chain can reach a compact structure in this stage.

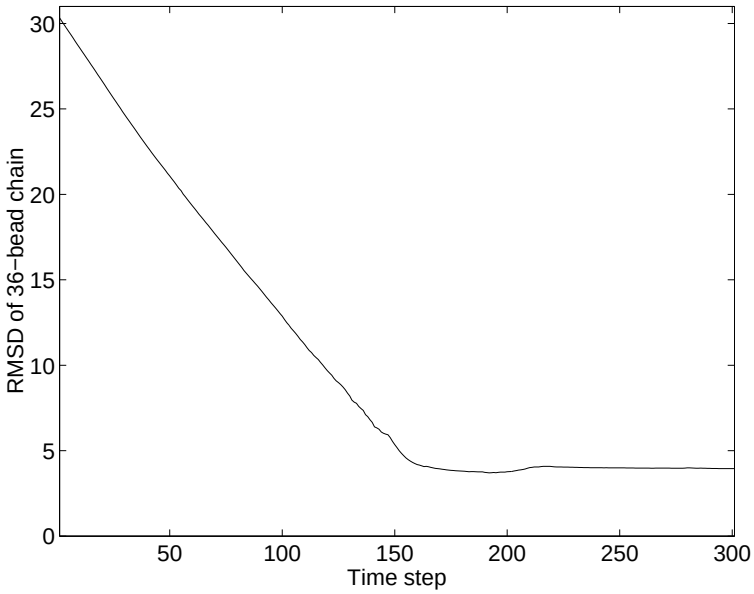
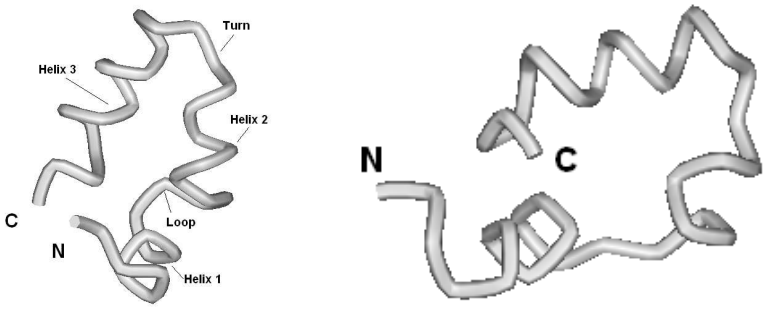


Figure 4.30: Rmsd change in the rearrangement stage of the case with the growth rate 1 sample units per bead.



(a) Native structure (b) Obtained configuration

Figure 4.31: Obtained configuration of villin protein and obtained configuration in the rearrangement stage of the case with the growth rate 1 time unit per bead.

4.2 Effects of Long Range Contact Pairs

In order to see the importance of long range native contacts in folding dynamics, the optimization method with 10 time steps is conducted by omitting the long range contact pairs.

According to the secondary structures, the RMSD figures are again grouped as early stages, intermediate stages and late stages again. The RMSD changes for these stages can be seen in Figures 4.32, 4.33, 4.34.

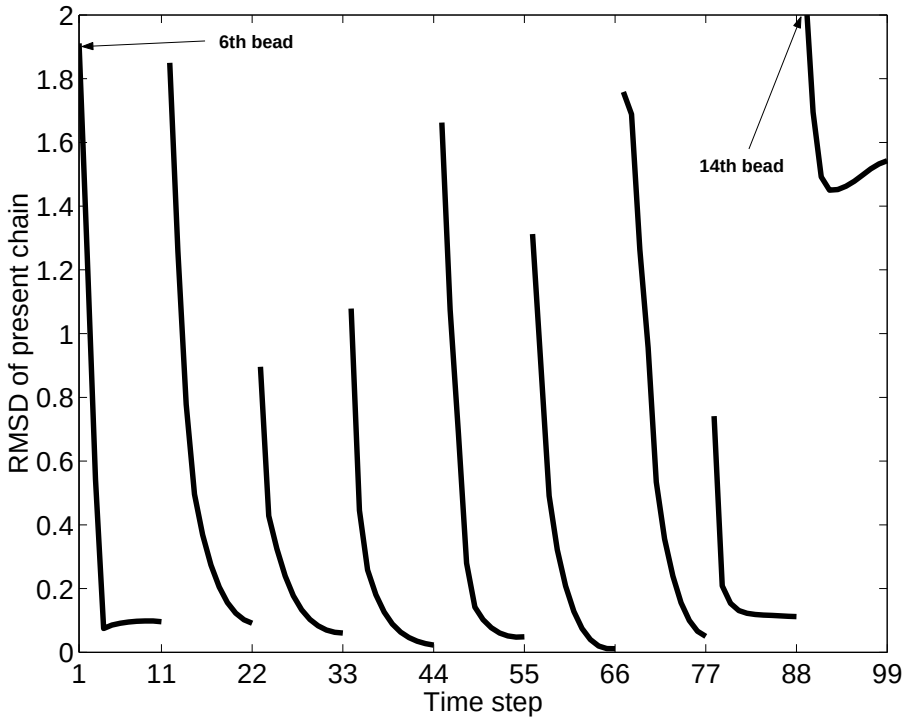


Figure 4.32: Rmsd change in the early stage of the case with 10 time units per bead in the absence of long range contact pairs.

These results show that the long range contact pairs play important roles in protein folding. The RMSD changes in the presence of all long-range contact pairs were analyzed in Figures 4.3, 4.4 and 4.5. As can be seen in the early phase plots, the chain can attain desirable RMSD values until the 14-th bead is added. The first long range contact pair beads are the 7-th bead and the 14-th bead. So, the first difference can be seen when 14-th bead leaves the ribosome. In the intermediate stage, RMSD

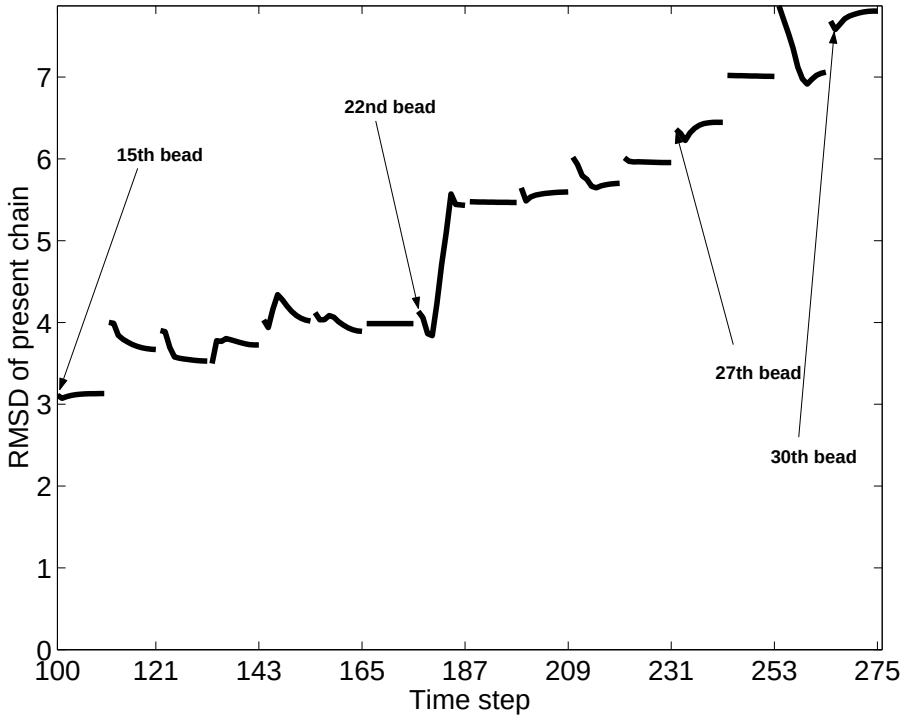


Figure 4.33: Rmsd change in the intermediate stage of the case with 10 time units per bead in the absence of long range contact pairs.

values increase more than the values in the presence of all contact pairs. In the late stage, RMSD again increases. Similar analysis about the long range contact pairs were presented in the Section 4.1.1. 6 of 8 long-range contact pairs are taken into account in this late stage. The final RMSD value is 11.7 $\mathring{A}$.

In Figure 4.35, the native configuration and the obtained conformation can be seen.

4.3 Comparison of Different Growth Rates

In summary, the RMSD values for these 4 cases are plotted in Figure 4.36. The RMSD values of the last time steps of that particular subsystems are plotted. It can be seen that the chain can not reach a compact structure in case it leaves the ribosome too fast. Also, the figure reveals that long range contact pairs are the main determinants of forming a compact structure.

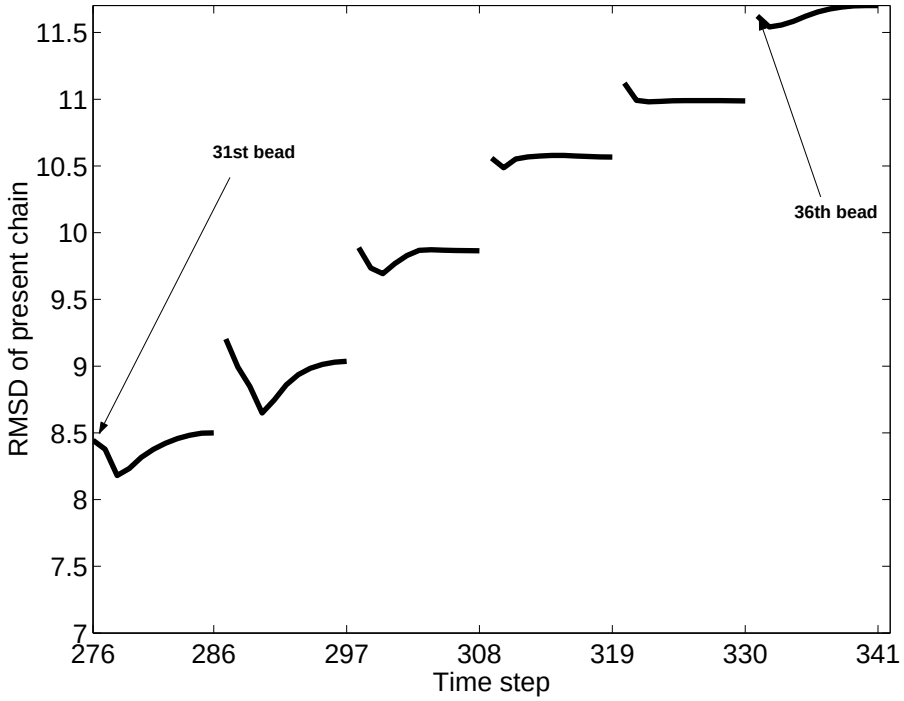
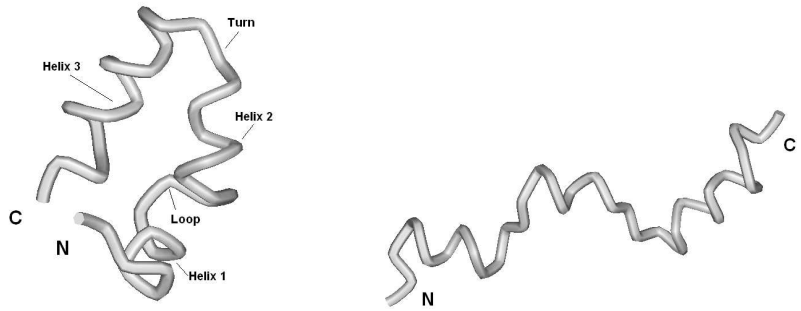


Figure 4.34: Rmsd change in the late stage of the case with 10 time units per bead in the absence of long range contact pairs.



(a) Native structure (b) Configuration obtained

Figure 4.35: Native configuration of villin protein and obtained configuration in the absence of long range contact pairs.

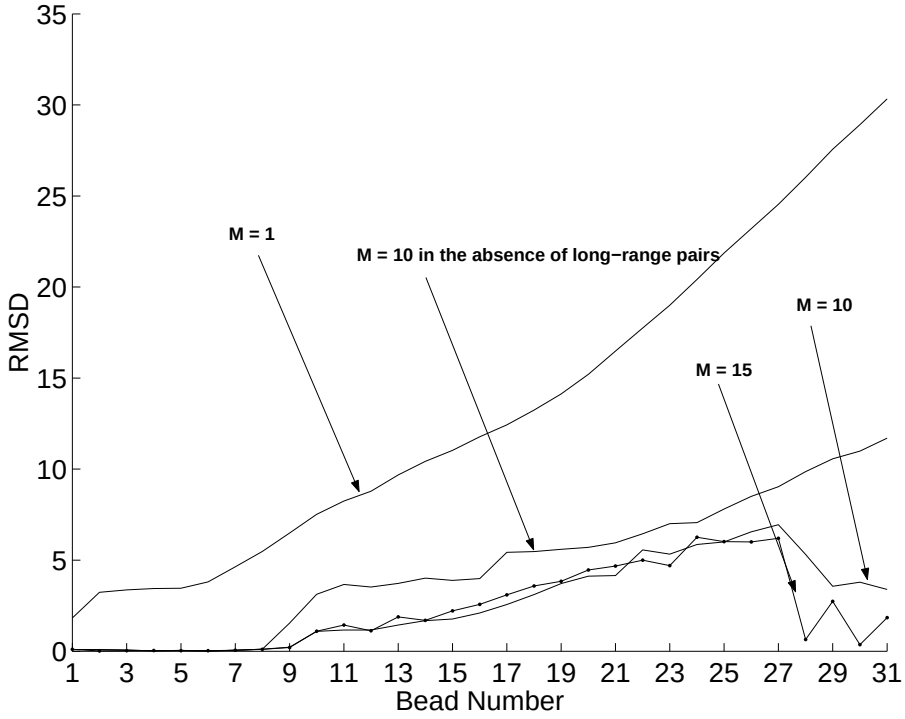


Figure 4.36: Rmsd values for 4 cases analyzed. M is the growth rate of the chain.

Table 4.2: Final RMSD values of substructures and the whole chain for 4 cases analyzed: case with 1 bead per 1 time step, case with 1 bead per 10 time steps, case with 1 bead per 15 time steps, case with 1 bead per 10 time steps in the absence of long-range contact pairs.

Structure	Final RMSD value ($\mathring{A}$)			
	1 step	10 steps	15 steps	No long-range contacts
Helix1	2.6264	0.0780	0.0078	0.0069
Loop	3.2956	2.3061	0.0098	2.0084
Helix2	2.5123	0.1088	0.0038	0.0119
Turn	2.1460	0.4105	0.0070	1.0244
Helix3	5.2489	0.1125	0.0031	0.0069
The chain	30.3316	3.3916	0.9477	11.7023
Folding time	62	341	496	341

4.4 Analysis of Input Variables

As discussed in Chapter 3, in the optimization model, the input variables determine the state variables.

In Figure 4.37, the norms of the input variables for beads 5, 13, 17, 22 and 30 are plotted respectively from left to right. These beads can be thought as the representative beads for the five secondary structures of chicken villin headpiece protein respectively, for instance bead 17 is in helix 2, bead 30 is placed in helix 3. To compare these force values with the case in the absence of long-range contact pairs, the norms of the input variables for the case in the absence of long-range contact pairs can be seen in Figure 4.38. These figures confirm the fact that the optimization optimally adjusts state and input variables adjusting input variables whenever a new bead is added to the chain. The last sub-figures of Figures 4.37 and 4.38 represent the change of input variables for bead 30. The other sub-figures are not much different from each other. However, the force values for bead 30 has a different behaviour. In the absence of long-range contact pairs, the chain can not form a compact structure, so the input variables do not change much because the number of the contact pairs are not sufficient to adjust the states properly.

The force change figures for the same beads for the cases with 1 bead per 1 time step and with 1 bead per 15 time step can be seen in Figures 4.39 and 4.40. For the case with 1 bead per 1 time step, similar results about the change in the RMSD value can be presented for input variables. The chain can not fold due to the lack of necessary time step, so the input variables simply fluctuate from one value to the next value when the lastly-added bead joins the folding process.

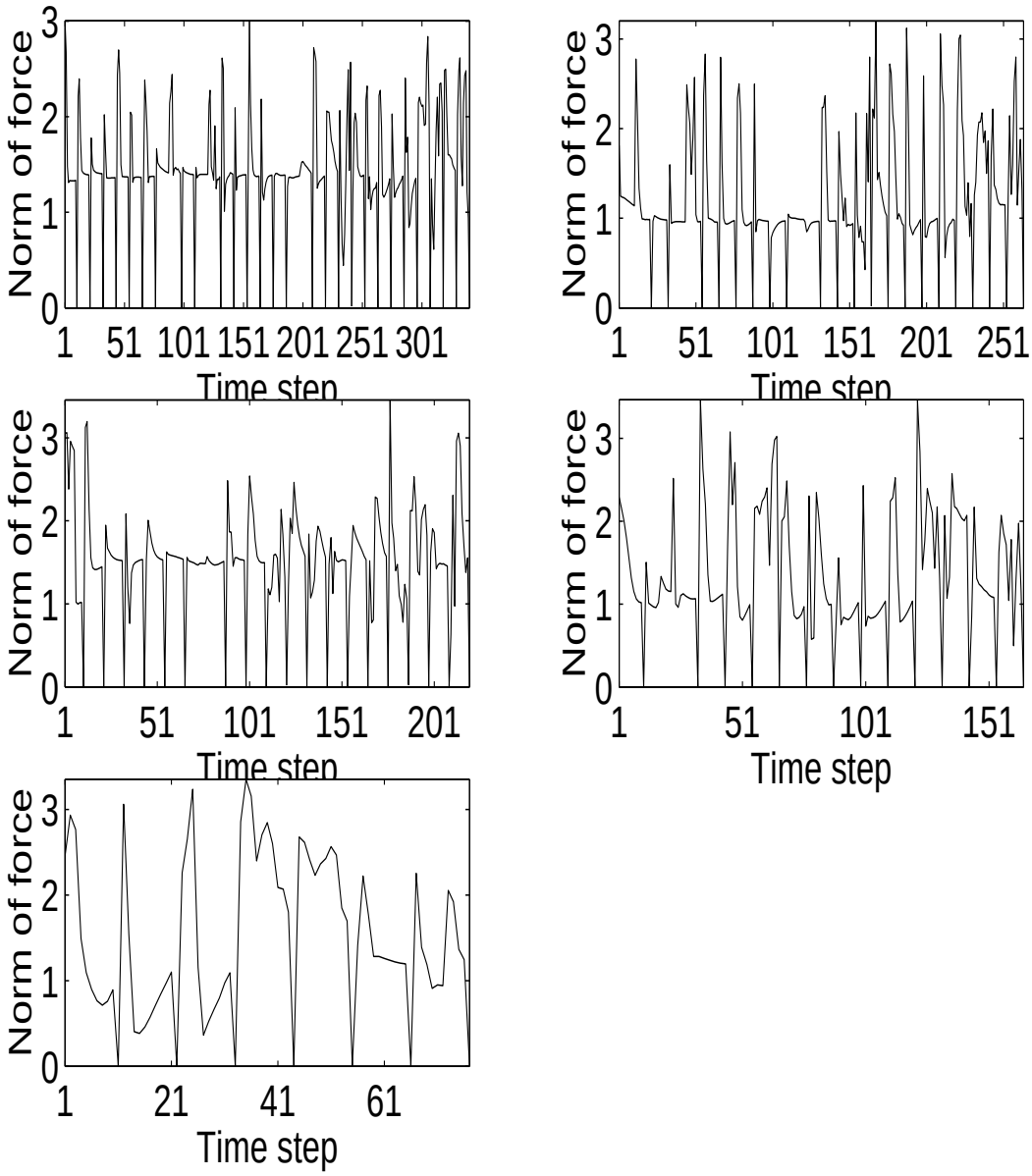


Figure 4.37: Norms of the input variables for beads 5, 13, 17, 22 and 30 for the case with 1 bead per 10 time units are plotted respectively.

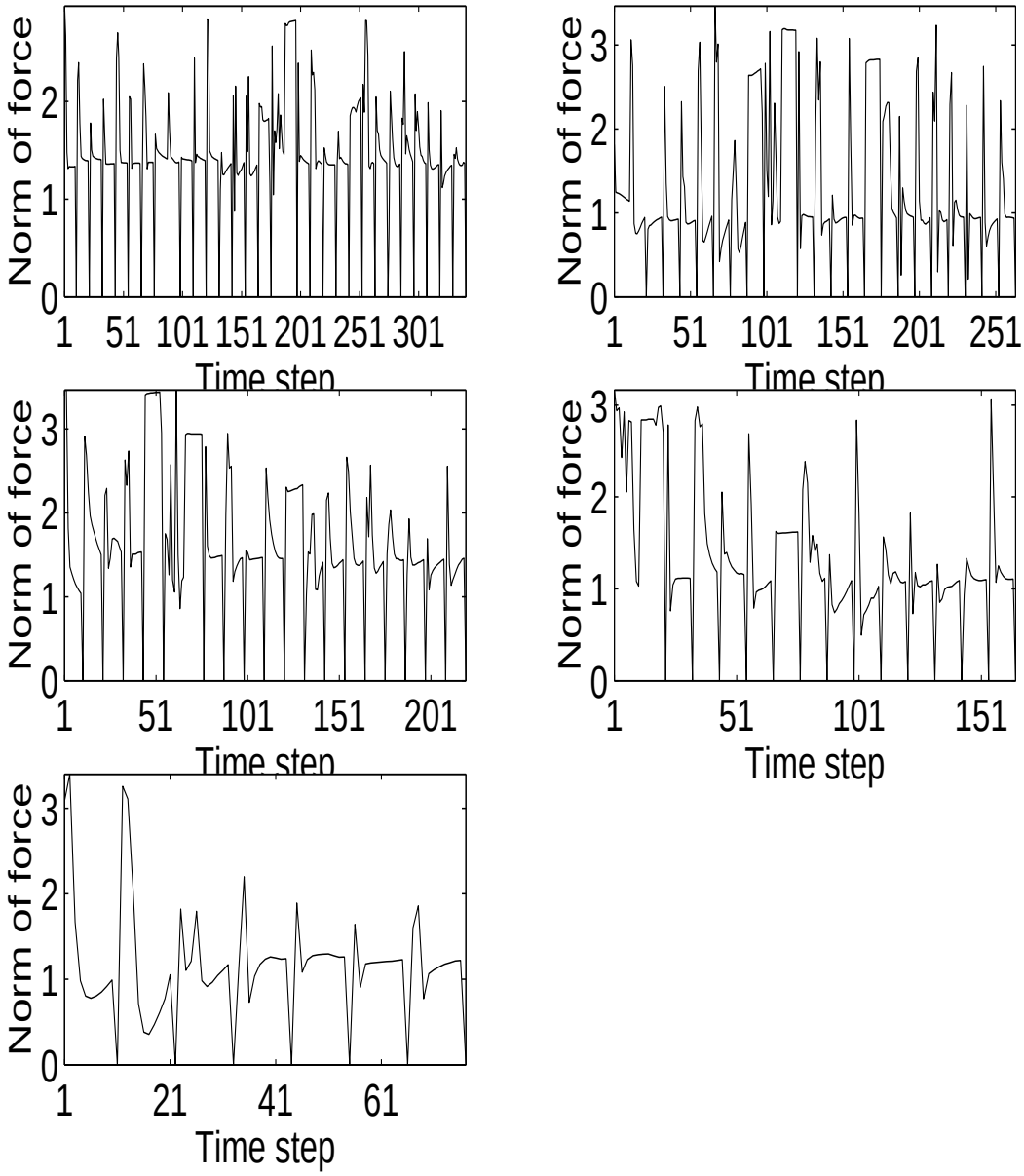


Figure 4.38: Norm of the input variables for beads 5, 13, 17, 22 and 30 for the case with 1 bead per 10 time units in the absence of long-range contact pairs are plotted respectively.

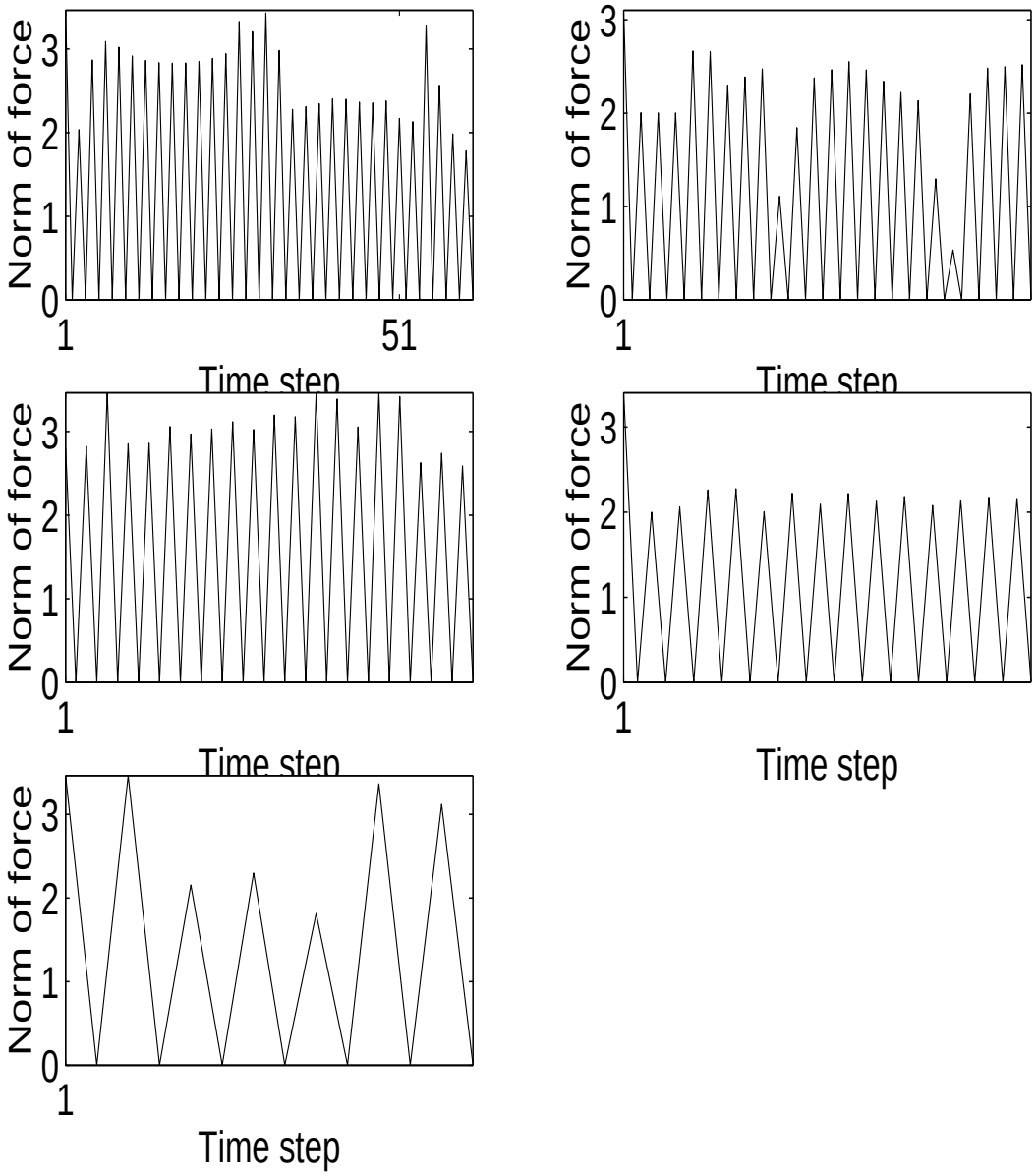


Figure 4.39: Norms of the input variables for beads 5, 13, 17, 22 and 30 for the case with 1 bead per 1 time units are plotted respectively.

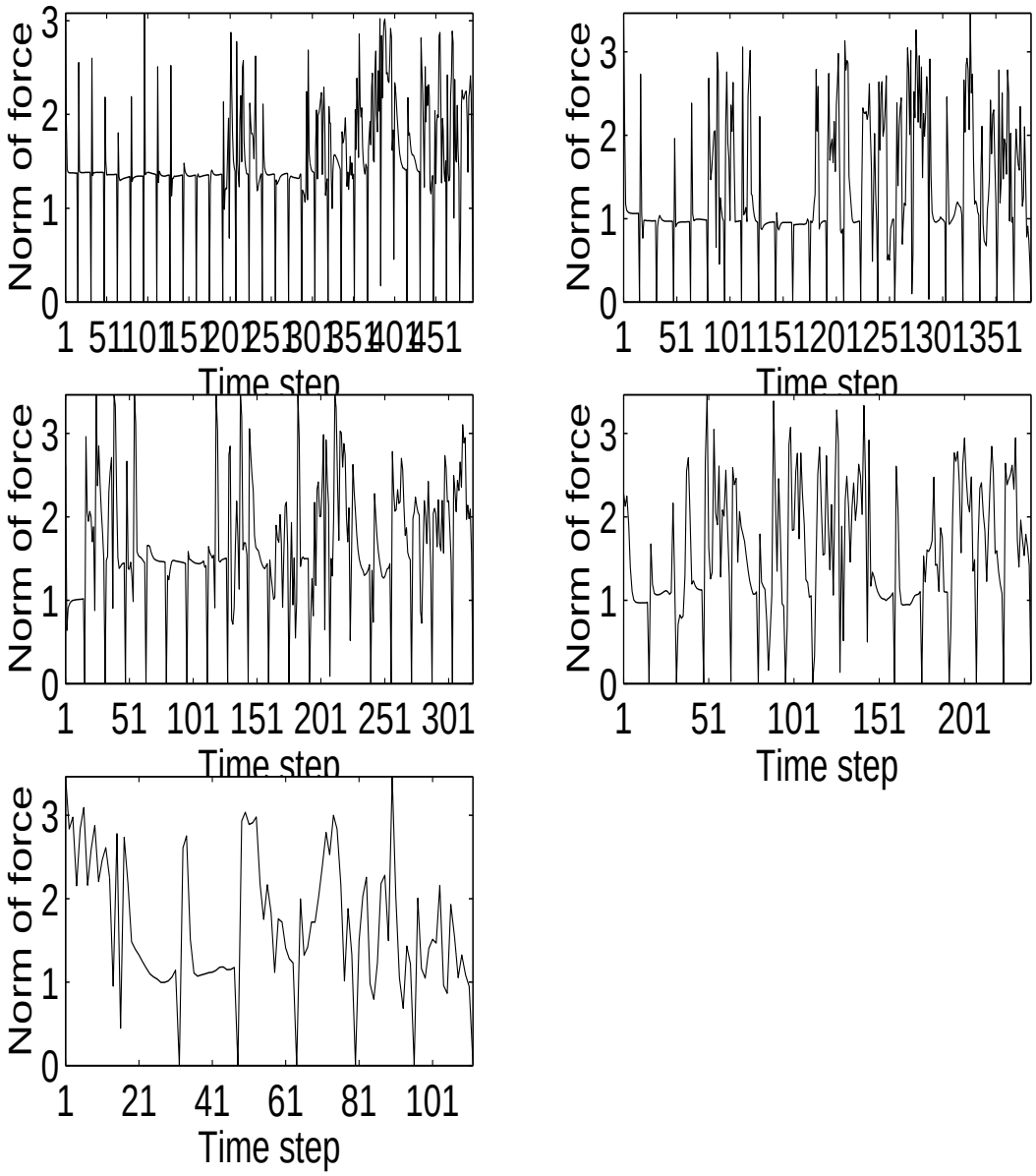


Figure 4.40: Norms of the input variables for beads 5, 13, 17, 22 and 30 for the case with 1 bead per 15 time units are plotted respectively.

Chapter 5

CONCLUSION

Protein folding is an important issue in computational biology, so there are numerous experimental and theoretical studies about understanding the properties of protein folding problem.

Recently, experimental studies have been performed to understand the real phenomena which underlie the folding mechanism. These studies propose that the nascent protein chain begins to fold in the ribosomal exit tunnel [37]. Thus, more realistic methods are performed to simulate the folding of the protein chain [36, 11].

In our study, we used an optimization model to obtain the possible pathway for the folding of the protein chain as if the protein folds in the exit tunnel of the ribosome while it is still being synthesized. In this optimization model, the native configuration of the protein is assumed known and the covalently bonded beads are modeled as linear springs. The other interactions are defined as a force component that drives the protein into its most native-like structure without violating the excluded volume and bond constraints. In the model, the energy function of the present chain is taken as the objective function of the minimization problem. The energy function is composed of the non-local interactions between native contact pairs. Starting from the initial configuration, the possible pathway for that particular chain is obtained [13]. The optimization model calculates the force variables which drive the chain into the obtained configuration which is a feasible and optimal way according the constraints and the objective function.

In our approach, we simulated the nascent folding dynamics of a fast-folding protein, chicken villin headpiece protein which is composed of 3 helices, one loop and one turn. We set the conformation of the first five beads to the native conformation

and assumed that the protein begins to fold while being synthesized after first five beads are out of the ribosome and that the folding process of the protein is composed of sub-systems which change according to the length of that particular chain. We also assumed that the growth rate of the protein depends on the present properties of the cell and the characteristics of the protein chain. So, we defined different growth rates for the protein to fold and compared the results for these different growth rates. We concluded that the chain can not form a compact structure in case it leaves the ribosome too fast. We defined three stages of the folding process: early stage, intermediate stage and late stage. In addition to these three stages, a final rearrangement stage was performed in case the chain is not compact. We also analyzed the important effect of long-range contact pairs during folding process. We observed that the chain can not form a compact structure in the presence of only short-range contact pairs as it can reach a more compact form in the presence of all long-range and short-range contact pairs. This important characteristic of long-range contact pairs were also emphasized by Duan et. al. [24]. We also analyzed the force field obtained in the optimization model.

The optimization model implemented in this study uses the simplified model for protein representation, so the molecular details which can be easily obtained in Molecular Dynamics simulations can not be obtained in this model.

As future work, the model can be changed so that it can be implemented using all atom representation of proteins, thus more realistic results can be obtained. Furthermore, the model can be implemented for larger proteins.

BIBLIOGRAPHY

[1] Baker, D., A surprising simplicity to protein folding. *Nature*, 2000. 405: 39-42.

[2] Kolb, V.A., Co-translational protein folding. *Mol Biol*, 2001. 35: 584590.

[3] Darby, N.J., and Creighton, T.E. Protein structure, Imprint Oxford, IRL Press at Oxford University Press. New York, 1993.

[4] Zwanzig, R., Szabo, A. and Bagchi, B., *Proc Natl Acad Sci U S A*. 1992 January 1; 89(1): 2022.

[5] Wikipedia Encyclopedia, http://en.wikipedia.org/wiki/Protein Folding.

[6] Metropolis, N. and Ulam, S. The Monte Carlo Method, *J American Statis Assoc*, 1949. 44: 335.

[7] Stillinger, F. H. and Rahman, A. *Journal of Chemical Physics*. 1974. 60: 1545.

[8] McQuarrie D., Statistical Mechanics , Harper & Row. New York, 1976.

[9] Kit Fun Lau, Ken A. Dill. A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins. *Macromolecules*, 1989. 22: 3986-3997.

[10] Kramer, G., Ramachandiran V., Hardesty B., Cotranslational folding: Omnia mea mecum porto? *Int J Biochem Cell Biol*, 2001. 33: 541553.

[11] Morrissey M.P., Ahmed Z., Shakhnovich E.I., The role of cotranslation in protein folding: A lattice model study. *Polymer*, 2004. 45: 557571.

[12] Prat Gay G., Ruiz-Sanz J., Neira J.L., Itzhaki L.S., Fersht A.R., Folding of a nascent polypeptide chain in vitro: Cooperative formation of structure in a protein module. *Proc Natl Acad Sci USA* , 1995. 92: 36833686.

[13] Guner, U., Arkun Y., Erman B., Optimum Folding Pathways of Proteins. Their Determination and Properties. *J Chem Phys*, 2006. 124/13: 297-309.

[14] Haliloglu T., Bahar I., Proteins: Structure, Function, and Genetics. it J Comp Chem, 1998. 31: 271.

[15] Branden C., Tooze J. Introduction to protein structure. Garland Pub, New York. 1991.

[16] Bryson and Ho, Applied Optimal Control: Optimization, Estimation and Control. Taylor and Francis, New York. 1975.

[17] Ben-Tal A., Zibulevsky, M., Penalty/Barriers Multiplier Methods for Convex Programming Problems. *Siam J. Optim.*, 1997. 7: 347-366.

[18] Kocvara, M., Stingl, M., A Generalized Augmented Lagrangian Method for Semidefinite Programming. G. Di . Pillo and A. Murli (Eds), 2003. 9: 297-315.

[19] Biegler, L.T., Cervantes, A.M., Waetcher, A., Advances in simultaneous strategies for dynamics process optimization. *Chemical Engineering Science*, 2002. 57: 575-593.

[20] Dill, K.A. and H.S. Chan, From Levinthal to pathways to funnels. *Nat Struct Biol*, 1997. 4(1): 10-19.

[21] Woolhead , C.A., McCormick P.J., Johnson, A.E., Nascent Membrane and Secretory Proteins Differ in FRET-Detected Folding inside the Ribosome and in Their Exposure to Ribosomal Proteins. *Cell*, 2002. 116: 725-736

[22] Gilbert R.J.C., Fucini P., Connell S., Three-Dimensional Structures of Translating Ribosomes by Cryo-EM. *Mol Cell*, 2004. 14: 57-66.

[23] Etchells, S.A., Hartl, F.U., The Dynamic Tunnel. *Nat Str and Mol Biol*, 2004. 11(5): 392-393.

[24] Duan, Y., Kollmani P.A., Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science*, 1998. 282: 740-744.

[25] Erman, B., Analysis of Multiple Folding Routes of Proteins by a Coarse-Grained Dynamics Model. *Biophys J*, 2001. 81: 3534-3544.

[26] Hoang, T. X., Cieplak, M., Two-state expansion and collapse of a polypeptide . *J Chem Phys*, 2000. 112: 6851-6862.

[27] Pande, V.S., Rokhsar, D.S., Folding Pathway of a lattice model for proteins. *Proc. Natl. Acad. Sci. USA.*, 1999. 96: 1273-1278.

[28] Erman, B., Dill, K. , Gaussian Model of Protein Folding. *Jour Chem Phys*, 2000. 112: 1050-1056.

[29] Cieplak M., Hoang, T.X, Universality Classes in Folding times of Proteins. *Biophy Jour*, 2003. 84: 475-488.

[30] Wilde R. E. and Singh S., Statistical Mechanics, Fundamentals and Modern Applications, John Wiley & Sons, Inc. New York, 1998.

[31] Pande, V.S., Baker, I., Chapman, J., Elmer, S.P., Khaliq, S., Larson, S.M., Rhee, Y. M., Shirts, M.R., Snow, C.D., Sorin, E.J., Zagrovic, B. Atomistic Protein foding Simulations on the Submillisecond time Scale Using Worldwide Distibuted Computing. *Biopolymers*, 2003. 68: 91-109.

[32] Kubelka J., Eaton W.A. and Hofrichter J. *J Mol Biol*, 2003. 329: 625.

[33] Jenni S., Ban N., The chemistry of protein synthesis and voyage through the ribosomal tunnel, *Current Opinion in Structural Biology*, 2003. 13,4,: 533.

[34] Voss N.R., Gerstein M., Steitz T.A., Moore P.B, The Geometry of the Ribosomal Polypeptide Exit Tunnel. *J Mol Biol*, 2006.

[35] Lu J., Carol D., Folding zones inside the ribosomal exit tunnel. *Nat Struc Mol Biol*, 2005. Nov 20.

[36] Elcock, H.A., Molecular Simulations of Cotranslational Protein Folding: Fragment Stabilities, Folding Cooperativity, and Trapping in the Ribosome. *PLoS Computational Biol*, 2006. 2: 98.

[37] Fedorov, A.N., Baldwin, T.O. , Contribution of cotranslational folding to the rate of formation of native protein structure. *Biochemistry*, 1995. 92: 1227-1231.

VITA

Serife Senturk was born in Izmir, Turkey on November 23, 1981. She received her B.Sc. degree in Mathematical Engineering from Istanbul Technical University in 2004. From October 2004 to June 2006, she worked as a teaching and research assistant in Koç University, Istanbul, Turkey.