OPTIMAL CONTROL APPROACH IN DYNAMICS OF PROTEIN FOLDING

By

Uğur Güner

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
The Degree of

Master of Science

in

Computational Sciences and Engineering

Koç University

June 2005

Koç University

Graduate School of Sciences and Engineering

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

Uğur Güner

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. Atilla Gürsoy

____

Assist. Prof. İsmail Lazoglu

____

Assist. Prof. Özlem Keskin

Date: ______________________________

To my niece and nephew

ABSTRACT

This thesis explores the pathways of protein folding using a 3-D off-lattice model. A dynamic model is proposed and used in an optimization framework. Amino acids are represented by monomer beads centered at C^{α} atoms. The interactions between monomers are represented through spring forces .In addition, a force field is introduced as a control input to the dynamic model. Next, protein folding 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. This approach allows us to generate possible pathways of protein folding from an initial configuration to the given native state. Our model is applied to a fast folding, 36-residue protein, villin headpiece subdomain. The simulated structures resemble the real native state of chicken villin headpiece with the C^{α} based root mean square deviation of 3.97 Å on the average. Starting from several initial conditions that cover a wide range of compactness, the sequence of dynamic events along trajectories are studied. Furthermore, the characteristics of forces on each bead and their relation to the constraints and interactions are analyzed.

ÖZETÇE

Bu tezde üç boyutlu basitleştirilmiş model kullanılarak, proteinlerin katlanırken izlediği yollar incelenmiştir. Bir dinamik model önerilmis ve bu model optimizasyon çatısı altında çalışılmıştır. Amino asitler monomer tanecikler seklinde ifade edilmiştir. Monomerler arası etkileşim çizgisel yay kuvvetleri olarak belirlenmiştir. Ayrıca, dinamik modele kontrol girdisi olarak bir kuvvet alanı tanımlanmıştır. Protein katlanması, dinamik model kestirimi ve fiziksel sınırlayıcılara tabi olan belirli bir enerji formunu minimize eden bir sistemle, optimum kontrol problemi olarak formüle edilmiştir. Bu yaklaşım bize herhangi bir başlangıç yapısından başlayıp doğal yapıya giden proteinin katlandığı olası yolları oluşturmamızı sağlamıştır. Modelimiz, otuz altı tanecikli, hızlı katlanan bir protein olan VILLIN e uygulanmıştır. Modelimiz uygulanması sonucu elde edilen yapılar doğal yapıya ortalama olarak 3.97 Å *rmsd* değerinde yakınlık göstermiştir. Geniş bir yoğunluk yelpazesi taşıyan başlangıç yapılarının doğal yapıya gitmesindeki izlenen dinamik olaylar incelenmiştir. Ayrıca, her tanecik üzerindeki kuvvetlerin özellikleri ve bu kuvvetlerin yapısal sınırlayıcılar ve etkileşimlerle olan ilişkisi incelenmiştir.

ACKNOWLEDGMENT

I would like to acknowledge the valuable supervisions of Prof. Yaman Arkun and Prof. Burak Erman. I would like to thank Prof. Attila Gürsoy for his assistance through my thesis. I would also like to thank the committee members for their critical reading and valuable comments. I am grateful to my family for their encouragement and support all the time.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The name *protein*, from the Greek *proteios,* meaning “first place”, suggests the importance of this class of macromolecules. A protein is a polymer constructed from amino acid monomers. All set of proteins are constructed from a common set of just 20 amino acids. Each amino acid consists of a central carbon atom bonded to four covalent partners. Three of these attachments are common to all 20 amino acids: a carboxyl group, an amino group, and a hydrogen atom. The variable component of amino acids, called the side group, is attached to the fourth bond of the central carbon atom. Each type of amino acid has a unique side group, giving that amino acid its special chemical properties (Figure 1.1).

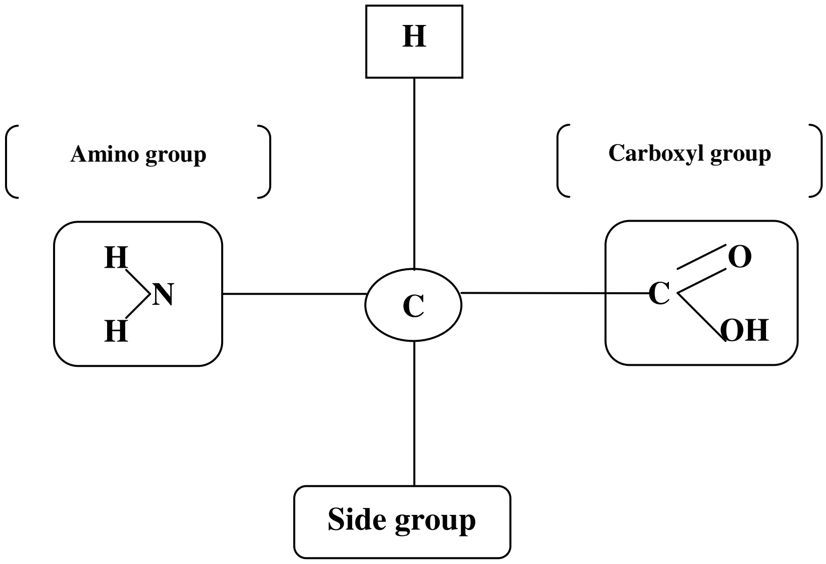


Figure 1.1 General structure of an amino acid.

There are thousands of different kinds of proteins, each with a unique, three-dimensional structure that corresponds to a specific function. There are four classes of proteins: structural proteins, storage proteins, contractile proteins, and transport proteins. Proteins show four kinds of structures. The first one is the primary structure, which is simply the sequence of amino acids. The second type of structure is the secondary structure, which is the regular structure regardless of the type of the amino acids. When long range amino acid interactions stabilize the secondary structure, this is called the tertiary structure. The last level of structure, quaternary, is the way different proteins organize in space into multi-polypeptide chains [1] (figure 1.2).

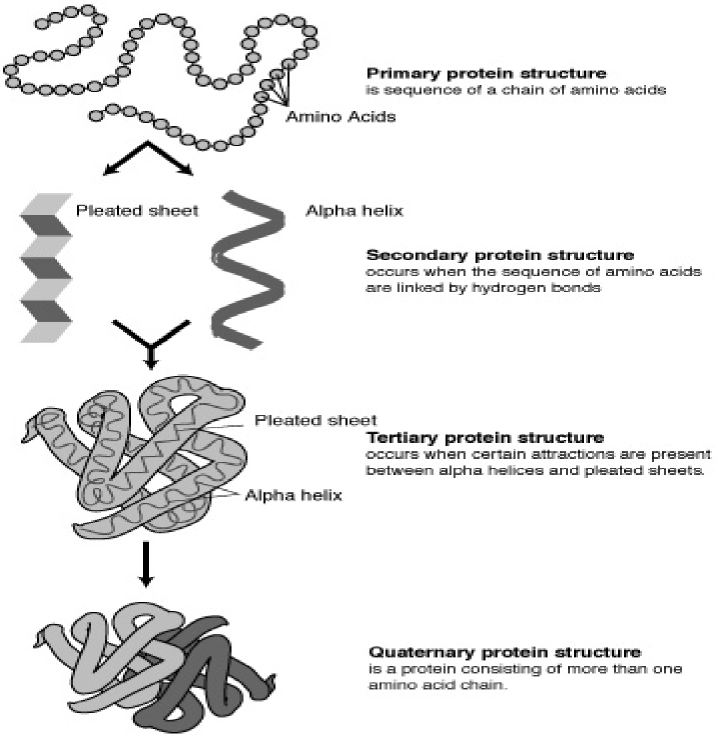

Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one amino acid chain.

Figure 1.2 Structures of proteins [26].

Following its synthesis on the ribosome, a protein must fold successfully to be functional. Although the cellular environment contains many factors that affect the folding process, the code for this process is contained in the primary sequence. Many proteins have been reported to refold from the denatured state in a test tube in the absence of such factors [2].

Understanding the sequence-structure relationships of proteins plays a crucial role in post-genomic era, and it will have deep influence in genetics, biochemistry and pharmaceutical chemistry. Understanding how proteins fold may have great impact on protein design as well as on the rapidly growing field of nanotechnology, in which self-assembling nano-machines may be designed by using synthetic polymers with protein-like properties [3].

Contribution

In this thesis, a new approach for understanding the folding mechanism of proteins is presented. This approach is based on optimal control. In the dynamic model, each amino acid is treated as a spherical bead centered at C^{α} atom. An energy function is constructed based on the known native state of the protein and it is optimized using a deterministic force field. Constraints are formed which are responsible for keeping bonded beads at a fixed distance and avoiding non-bonded beads to come closer than an allowed distance. Optimal control computes the necessary force field to fold the protein. Corresponding to each optimal force field protein folding pathways are generated, analyzed and compared with literature.

Outline

Chapter 2 introduces the recent approaches in the protein folding problem. Our method, its formulation and details are given in Chapter 3. In Chapter 4, we present our results and compare them with those in the literature. In chapter 5, we draw the conclusions. In the Appendix, the matrices that are used in the model section are given.

Chapter 2

PROTEIN FOLDING

Protein folding is the process by which a protein assumes its functional shape or conformation. All protein molecules are simple unbranched chains of amino acids, but it is by coiling into a specific three-dimensional shape that they are able to perform their biological function [20]. The particular amino acid sequence (primary structure) of a protein biases it to fold into its native conformation. Many proteins do so spontaneously during or after their synthesis inside cells. While macromolecules may seem to be “folding themselves”, in fact their folding depends a great deal on the characteristics of their surrounding solution, including the identity of the primary solvent (either water or lipid inside cells), the concentration of salts, the temperature, and molecular chaperons. Chaperons are the proteins whose function is to assist other proteins in achieving proper folding [20]. 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 to folded state is mainly guided by Van der Waals interactions 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 [20].

Incorrectly folded proteins are responsible for illnesses such as Creutzfeldt-Jakob disease, Bovine spongiform encephalopathy, Alzheimer’s disease. These diseases are caused by misfolded proteins aggregating into insoluble plagues.

The entire duration of the folding process varies dramatically depending on the protein of interest. The slowest folding proteins require many minutes or hours to fold. However, small proteins, with lengths of a hundred or so amino acids, typically fold on time scales of milliseconds [20].

The problem of protein folding breaks into three parts ; Genomic question of relating sequence to structure, the operational question of how structure is related to function, and the kinetic question of what pathways lead to folding and how systems find it [5]. To answer the kinetic question of how a protein folds is at the heart of molecular biology. In this chapter, we represent the computational methods that are used to understand the folding mechanism of proteins.

Computational Methods on Protein Folding

There are several computational methods for studying the dynamics of protein folding. We can list them in two major groups as MD simulations, simplified lattice and off-lattice models.

2.1. Molecular Dynamics Simulations

Molecular Dynamics (MD) is the most accurate method in which all atom interactions including the solvent in which protein is suspended (water) are modeled by Hamiltonian or Langevin dynamics. Current single processor computers can only simulate about a nanosecond of real-time folding in full atomic detail per CPU day. It is known that fastest proteins fold in the order of tens of microsecond. There is a great computational gap. To overcome these computational obstacles, great strides in parallel MD have been made. Hundreds of supercomputers in parallel are used for a single MD simulation by Duan and Kollman [8]. This makes significant progress in closing the computational gap. However, this method has some drawbacks. It requires complex, expensive supercomputers due to the need for communication between processors. Moreover, because of the stochastic nature of folding, one must simulate hundreds of microseconds, requiring computing power equal to thousands or tens of thousands of today's processors [7].

All-atom molecular dynamics with explicit representation of water for 1 microsecond on Villin headpiece using parallel computers of increased efficiency was implemented by Dual and Kollman [8]. They showed two distinct phases in folding namely, initial hydrophobic collapse followed by a conformational adjustment phase [8].

One main impediment in computational speed of MD simulation is the presence of solvent in the system. More than 95% of the atoms in the system are those of water molecules. The introduction of implicit solvent model, in which the solvation effect is considered in a mean field representation, can dramatically increase the throughput of simulation and can still provide a reasonable accuracy [8].

An MD simulation of the Villin was conducted by Shen and Freed [9] using an implicit solvent model. They found close correspondence with the all atom simulation of Villin by Duan and Kollman.

Pande et al [10] applied the implicit solvent approximation in an innovative distributed computer approach in the simulation of a small protein like Villin. They observed the initial collapse phase in the molecule that greatly reduces the size of the conformational space to be explored in search of the native state. They obtained an ensemble of folded structures, 1.7 Å away from the native state in C^{α} based rmsd sense [10].

2.2. Simplified Models

Lattice protein models are highly 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 behavior in all-atom detail [20]. Hence, real proteins can not be folded on a computer. 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. The simplification means that they can fold to their energy minima in a time quick enough to be simulated [20].

Lattice proteins are made to resemble real proteins by introducing an energy function, a set of conditions, which specify the energy of interaction between neighboring beads, usually taken to be those occupying adjacent lattice sites. The energy function mimics the interactions between the amino acids in real proteins, which include steric, hydrophobic, and hydrogen bonding effects [20]. The beads are divided into types, and the energy function specifies the interactions depending on the bead type, just as different types of amino acids interact differently. One of the most popular lattice models, *HP*, features just two bead types; hydrophobic (H) and polar (P), and mimics the hydrophobic effect by specifying a negative (favorable) interaction between *H* beads. *HP* model is a free energy model. It is based on the assumption that a major contribution to the free energy of the natural conformation of a protein is due to interactions between hydrophobic amino acids that tend to form a core in the spatial structure shielded from the surrounded solvent by polar amino acids [20]. Figure 2.1.1 displays a two dimensional picture of *HP* lattice model of a model protein that is composed of 20 beads. Filled beads represent the hydrophobic beads, and empty beads stand for the polar residues.

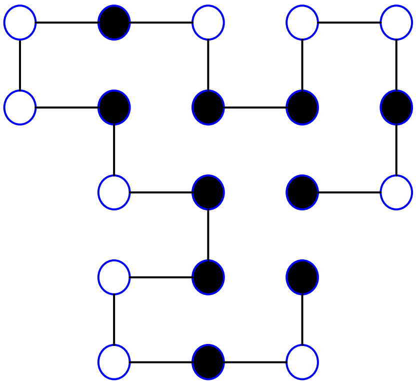


Figure 2.2.1 A minimum energy conformation in the two dimensional *HP* model with 9 non-local H-H contacts.

Due to the energy function, every lattice protein has an 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 [20].

By varying the energy function and the bead sequence of the chain, effects on the native state structure and the kinetics 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 [20].

One of the most popular methods used in lattice models is Monte Carlo simulations. In this method, small random changes are made repeatedly and accepted or rejected according to the rule, which is based on the change of energy [2]. The algorithm is such that the probability is greater for the system to move to conformations of lower energy rather than higher energy. This mimics the situation in a real protein, where native-like interactions are on average more stabilizing than non-native ones. Many moves carried out in succession lead to a folding trajectory that is directed by the potential energy function [2].

Many studies have been carried out on lattice model proteins. The lattice model, even being the simplest possible protein model, could still capture many essential characteristics of the folding problem and the prediction of tertiary structure [21, 22, 23].

In recent studies, Bagchi et al [10] studied the folding dynamics of Villin by using a force-field which incorporates hydropathy scale and the role of helical propensity of amino acids through a non-local harmonic potential. In their simplified model, each amino acid is represented by one side chain atom which is attached to the backbone C^{α} atom. In the simulations, they found out that the protein follows an initial burst phase that is followed by a slow stage [10].

Dinner et al. studied energy surfaces for folding using lattice models with a Monte Carlo Algorithm [2]. Jiang et al. introduced a novel algorithm of *tabu* search with genetic algorithms for the protein folding simulations of the HP model [1].

Recently, there have been studies that involve the coarse-grained models with MD simulations. Micheletti et al. [11] introduced a novel approach combining both methods. The coarse-grained model simplifies the evolution of the protein toward viable starting conformations for MD rapidly. They obtained rmsd of 3.7 Å from the native state in their simulations. Furthermore, important aspects and folding trajectories are obtained by MC (Monte Carlo) -MD method [11].

Another popular coarse-grained model of protein folding is the Go-type model. The basic feature of the Go-type models is that the native configuration of the protein is assumed known. In this model, the beads on a protein chain are subject to a Go-type potential. In this potential, the interactions between the pairs of residues that are in known positions in the native state are assumed known in advance. 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]. The pathway of folding is obtained through the solution of the equation of motion. 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 [24]. Erman developed Langevin dynamics of protein molecule with Go-type potentials. Long time-scale events in the folding of *cytochrome c* were analyzed [19]. Pande and Rokshar studied a protein-like heteropolymer by using direct simulation of a lattice model using Go-model [25]. In the model, the energy of each polymer conformation is taken to be proportional to the number of nearest neighbor native contacts it possesses [25].

Chapter 3

MODEL

3.1 Force Equations

In our dynamic model, we represent the amino acids with spherical beads centered at the C^{α} atoms. The position of i^{th} bead is denoted by the vector r_i with respect to a fixed frame coordinate. Distances between bonded pairs are adopted as 3.8 Å. Forces acting on the chain are divided into three groups: Forces between bonded pairs, non-bonded pairs and friction forces acting on each bead. Forces between bonded pairs can be further categorized as repulsive and attractive forces:

i. Attractive Forces: these forces are treated as linear spring forces.

ii. Repulsive Forces: these forces are active when two beads try to come closer than their respective bond length, 3.8 Å.

Forces between non-bonded pairs can also be divided into two groups:

i. Attractive Forces: these are the forces between i^{th} and j^{th} beads when C_i^{α} and C_j^{α} are in contact in the native state of the protein (two non-bonded beads are defined to be in native contact if the distance between them is within 7 Å.).

ii. Repulsive Forces: these forces act between all non-bonded bead pairs i and j. They act when two beads try to come closer than the hydrogen bond length which is 5.1 Å.

We assume that a friction force is acting on each bead. One well-known way to describe the dynamics of the chain is through the Langevin equation [12] given in its most general nonlinear form:

$$mr'' = -\gamma r' + f(r) + w \tag{3.1.1}$$

Here m is the mass of the residue; γ is the friction coefficient with dimension of (force) (time) / (distance); $f(r)$ stands for the forces between bonded and non-bonded pairs, and w is a random force.

We express the force term $f(r)$ in terms of its components:

$$f(r) = \Gamma_A^B r + f_R^B(r) + f_A^{NB}(r) + f_R^{NB}(r) \tag{3.1.2}$$

In this equation, $\Gamma_A^B r$ represents the attractive forces between bonded pairs. Γ_A^B is not a function of position since it is the linear spring force; $f_R^B(r)$ represents the repulsive forces between bonded pairs; $f_A^{NB}(r)$ includes the attractive forces between non-bonded pairs and $f_R^{NB}(r)$ represents the repulsive forces between non-bonded pairs.

We lump the last three terms and name it u:

$$u = f_R^B(r) + f_A^{NB}(r) + f_R^{NB}(r) \tag{3.1.3}$$

We assume that masses are small so that $mr'' = 0$. Without any loss of generality, we let the friction coefficient to be equal to one and we consider only the deterministic case i.e. $w = 0$. With these assumptions and definitions, equation (3.1.1) simplifies to:

$$\dot{r} = \Gamma_A^B r + u \tag{3.1.4}$$

In order to construct the forces $\Gamma_A^B r$, one can write the individual attractive spring forces between the bonded pairs. For simplicity, we first consider the n-bead system in one-dimension x.

The attractive spring force between two bonded beads is given by:

$$f_{i,i+1} = k_{i,i+1}(x_{i+1} - x_i) = -f_{i+1,i} \qquad i = 1,....,n-1 \qquad \textbf{(3.1.5)}$$

Here, *k* is the spring constant.

Let us denote the total spring forces on the i^{th} bead by F_i:

$$F_i = f_{i,i-1} + f_{i,i+1} = k_{i,i-1}(x_{i-1} - x_i) + k_{i,i+1}(x_{i+1} - x_i) \qquad i = 2...n-1 \qquad \textbf{(3.1.6)}$$

$$F_1 = f_{1,2} = k_{1,2}(x_2 - x_1)$$

$$F_n = f_{n,n-1} = k_{n,n-1}(x_{n-1} - x_n)$$

Equation **(3.1.6)** can be alternatively written as:

$$F_i = -(k_{i,i-1} + k_{i,i+1})x_i + k_{i,i-1}x_{i-1} + k_{i,i+1}x_{i+1} \qquad \textbf{(3.1.7)}$$

The individual forces, F_i, can be collected in a vector:

$$F^T = [F_1 \quad . \quad . \quad F_i \quad . \quad . \quad F_n] \qquad \textbf{(3.1.8)}$$

Thus, the attractive forces between bonded pairs may be expressed in matrix form:

$$F=\begin{bmatrix}F_1\\ \cdot\\ \cdot\\ F_i\\ \cdot\\ \cdot\\ F_n\end{bmatrix}=\begin{bmatrix}-\left(k_{1,2}\right) & k_{1,2} & 0 & 0 & .... & 0\\ k_{2,1} & -\left(k_{2,1}+k_{2,3}\right) & k_{2,3} & 0 & .... & 0\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & 0 & 0 & .... & -\left(k_{n,n-1}\right)\end{bmatrix}*\begin{bmatrix}x_1\\ \cdot\\ \cdot\\ x_i\\ \cdot\\ \cdot\\ x_n\end{bmatrix} \qquad \textbf{(3.1.9)}$$

Or simply:

$$F=\Gamma_A^B x \qquad \textbf{(3.1.10)}$$

If we take the spring constants to be unity, the spring constant matrix is given in tri-diagonal form by:

$$\Gamma_A^B=\begin{bmatrix}-1 & 1 & 0 & 0 & 0 & ... & 0\\ 1 & -2 & 1 & 0 & 0 & ... & 0\\ 0 & 1 & -2 & 1 & 0 & ... & 0\\ \cdot & \cdot & \cdot & \cdot & \cdot & & 0..\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0\\ 0 & ... & 0 & 0 & 1 & -2 & 1\\ 0 & ... & 0 & 0 & 0 & 1 & -1\end{bmatrix} \qquad \textbf{(3.1.11)}$$

For the general three-dimensional case the position vector can be recast in terms of its x, y, z coordinates as follows:

$$r^T=\left[x_1 \quad . \quad . \quad x_n \quad y_1 \quad . \quad . \quad y_n \quad z_1 \quad . \quad . \quad z_n\right]=\left[x^T\, y^T\, z^T\ \right] \qquad \textbf{(3.1.12)}$$

x_1 to x_n represent the x-coordinate values of the n beads and y_1 to y_n represent the y-coordinate values. The rest stands for the z- coordinate values.

We can write the attractive energy between bonded pairs as follows [13];

$$E_A^B = -\frac{1}{2} k r^T \Gamma_A^B r \qquad \textbf{(3.1.13)}$$

One can shift to force equation in all directions from energy equations since force is the negative gradient of the energy;

$$f_x = -\nabla_x E_A^B \qquad f_y = -\nabla_y E_A^B \qquad f_z = -\nabla_z E_A^B \qquad \textbf{(3.1.14)}$$

The position vector r can be written in terms of its components in x, y and z directions with basis vectors e_1 ,e_2 and e_3;

$$r = xe_1 + ye_2 + ze_3 \qquad \frac{\partial r}{\partial x} = e_1 \qquad \frac{\partial r}{\partial y} = e_2 \qquad \frac{\partial r}{\partial z} = e_3 \qquad \textbf{(3.1.15)}$$

Using all these equations, we can obtain force equations in x, y and z directions as follows;

$$f_x = -\frac{\partial E_A^B}{\partial r}\frac{\partial r}{\partial x} = k\Gamma_A^B r \cdot e_1 = k\Gamma_A^B x \qquad \textbf{(3.1.16)}$$

$$f_y = -\frac{\partial E_A^B}{\partial r}\frac{\partial r}{\partial y} = k\Gamma_A^B r \cdot e_2 = k\Gamma_A^B y$$

$$f_z = -\frac{\partial E_A^B}{\partial r}\frac{\partial r}{\partial z} = k\Gamma_A^B r \cdot e_3 = k\Gamma_A^B z$$

Thus, equation (3.1.4) becomes, in three dimensions:

$$\dot{r} = \overline{\Gamma}_A^B r + u \qquad \textbf{(3.1.17)}$$

Or more explicitly:

$$\begin{bmatrix} \dot{x}_1 \\ \cdot \\ \cdot \\ \dot{x}_n \\ \dot{y}_1 \\ \cdot \\ \cdot \\ \dot{y}_n \\ \dot{z}_1 \\ \cdot \\ \cdot \\ \dot{z}_n \end{bmatrix} = \left[\begin{array}{cccccccccccc} & & & & 0 & \cdot & \cdot & 0 & 0 & \cdot & \cdot & 0 \\ & \Gamma_A^B & & & \cdot & \cdot & & \cdot & \cdot & & & \\ & & & & \cdot & & & \cdot & \cdot & & & \cdot \\ & & & & 0 & \cdot & \cdot & 0 & 0 & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & 0 & & & & & 0 & \cdot & \cdot & 0 \\ \cdot & & & \cdot & & \Gamma_A^B & & & \cdot & & & \cdot \\ \cdot & & & \cdot & & & & & \cdot & & & \cdot \\ 0 & \cdot & \cdot & 0 & & & & & 0 & & & 0 \\ 0 & \cdot & \cdot & 0 & 0 & \cdot & \cdot & 0 & & & & \\ \cdot & & & \cdot & \cdot & & & \cdot & & \Gamma_A^B & & \\ \cdot & & & \cdot & \cdot & & & \cdot & & & & \\ 0 & \cdot & \cdot & 0 & 0 & \cdot & \cdot & 0 & & & & \end{array} \right] \bullet \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ x_n \\ y_1 \\ \cdot \\ \cdot \\ y_n \\ z_1 \\ \cdot \\ \cdot \\ z_n \end{bmatrix} + \begin{bmatrix} u_{x_1} \\ \cdot \\ \cdot \\ u_{x_n} \\ u_{y_1} \\ \cdot \\ \cdot \\ u_{y_n} \\ u_{z_1} \\ \cdot \\ \cdot \\ u_{z_n} \end{bmatrix} \qquad \textbf{(3.1.18)}$$

Thus, equation (3.1.18) is a modified Langevin equation with a deterministic force field. In this model, it is important to clarify and stress the role of the force vector $u(t)$.These forces are nonlinear forces that are not known a priori. They define the force field that helps to fold a protein from an initial condition to a final native state in a feasible (i.e. without violating excluded volume and bond length constraints) and an optimal way (i.e. with respect to the defined objective function). In our approach, these forces are automatically computed to deliver the desired folding. In this context, it is appropriate to interpret them as control inputs acting on each bead. In our optimal control approach, which will be discussed in the next section, these control inputs are optimally computed.

3.2 Optimization Problem Statement

In its most general setting, the protein folding problem can be studied as an optimal control problem. We assume that the protein chain starts from an initial configuration and folds to a native state. Our optimization problem focuses on native contact pairs and tries to bring them together during the folding process. This is accomplished by minimizing the attractive energies between non-bonded native contact pairs. Native contact pairs are the pairs that are two or more residues apart with distance less than 7 Å.

We can represent the minimized energy E_A^{NB} by:

$$E_A^{NB} = \int_0^{t_f} \sum_{i,j} \left\| r_i(t) - r_j(t) \right\|^2 dt = \int_0^{t_f} r^T Qr \ dt \qquad \textbf{(3.2.1)}$$

Where i and j represent residue index of the native contact pair. We can now state the optimization problem as follows;

$$\underset{u(t)}{Min} \left[E_A^{NB} = \int_0^{t_f} r^T Qr \ dt \right] \qquad \textbf{(3.2.2)}$$

Subject to:

$\dot{r} = \overline{\Gamma}_A^B r + u(t)$ *Dynamic Model*

$r(t=0) = r_0$ *Initial condition*

$$l-\varepsilon \le r^T H_i r \le l+\varepsilon \qquad \textit{Bond length constraints}$$

$$r^T L_i r \ge d_{ij} \qquad \textit{Excluded volume constraints}$$

$$\begin{array}{lll} \overline{u}_{x_i} \ge u_{x\,i} \ge -\overline{u}_{x_i} & & \\ \overline{u}_{y_i} \ge u_{y_i} \ge -\overline{u}_{y_i} & i=1,2,..,N & \textit{Force magnitude constraints} \\ \overline{u}_{z_i} \ge u_{y_i} \ge -\overline{u}_{z_i} & & \end{array}$$

E_A^{NB} : Objective function for the optimization problem

Q : Matrix that projects position vector to the native contact distances

r : States vector or position vector of the beads

u : Control input vector (the force field)

$\overline{\Gamma}_A^B$: System matrix

H_i : Matrix that relates states to the bond lengths

L_i : Matrix that relates states to the excluded volumes

l : Bond length distance

d_{ij} : Minimum excluded volume distance the i^{th} and j^{th} bead

$\overline{u}_{x_i, y_i, z_i}$: Limits on the x, y and z components of the force acting on the i^{th} bead

t_f : Final time

In this optimization formulation forces are optimally adjusted to bring the native contact pairs together as close as possible. In Equation (3.2.2), the dynamic model equation is included in the form of a state-space model. States are the positions of the beads of the chain. Inputs are the forces that act on each bead in three dimensions. These forces are responsible for satisfying the path constraints and bringing the native contact pairs together and consequently folding the protein. Bond length constraints, which are equality constraints, are relaxed by introducing a small number epsilon**;** thus, they are converted to inequality constraints that are easier to be handled by the optimization.

Optimization computes the optimal force field $u(t)$ and the trajectory $x(t)$ over the time interval $(0, t_f)$. Final time is chosen as long enough to let the chain settle to a final position and short enough to decrease the optimization running time.

The bond length distance l is taken as 3.8 Å. For the native contact pairs, minimum excluded volume distance, d_{ij}, is set equal to the known values at their native states. The choice of excluded volume that gives the closest approach for any two residues is expected to depend on the type of the residue pairs. In this present work, it is chosen as 5.1 Å which is the approximate hydrogen bond length distance.

The maximum and minimum limits on the forces are set equal to 2 and -2 respectively. If these limits are chosen smaller, optimization can not find feasible solution (i.e. folding to native state is not possible). For the higher limits, the changes in the states are very abrupt due to overly aggressive forces. When limit is taken as 2, we obtain smooth state trajectory and observe more realistic folding patterns with good resolution.

3.3. Optimization Technique

PENNON code is used to solve the optimization problem. It is based on an augmented langrangian with a penalty barrier function for matrix inequalities. It is designed for convex semi definite programming problems. It has also been generalized and tested on non-convex, nonlinear problems with success [14].

The optimization problem is written in the AMPL environment to be able to use PENNON as a solver. In PENNON, generalization of PBM (penalty barrier method) is used. PBM method was introduced by Ben-Tal, Zibulevsky [15]. It is a class of iterative methods for convex nonlinear programming. Generalization of PBM approach to convex semi definite problems is coded in PENNON [14]. Pennon solves optimization problems with nonlinear objectives subject to nonlinear inequality and equalities as constraints:

$$\underset{x \in R^N}{Min} \quad f(x) \tag{3.3.1}$$

$$\text{s.t.} \quad \begin{array}{ll} g_i(x) \leq 0, & i = 1,....,m_g \\ h_i(x) = 0, & i = 1,.....,m_h \end{array}$$

Here, f , g_i and h_i are the functions from R^N to R

Our model equations are discretized using orthogonal collocation on finite elements. This method is robust and efficient in handling path constraints. The discretization method as in the study of Biegler et al. is adopted [16]. Monomial basis representation is used [16]. The differential profiles can be written as:

$$r_{i,k} = r_{(i-1),k} + h_{(i-1)} \sum_{l=1}^{l=NCOL} \dot{r}_{(i-1),l,k} \Omega_{l,j} \tag{3.3.2}$$

$i = 1...M \qquad k = 1...N \qquad k = 1...NCOL$

Here, i represents the index of time step. k is the state index of position vector. *NCOL* stands for the number of collocation points in each element. *M* is the number of time steps. *N* is the number of states. $r_{i,k}$ is the value of k^{th} state in the element i. h_i is the length of element i. $\dot{r}_{(i-1),l,k}$ is the value of k^{th} state's derivative in the element $i-1$ at the collocation point *l*, and $\Omega_{l,j}$ is the polynomial of order *NCOL*. t_f is the final time.

Our model equation can be written as;

$$\dot{r}_{i,l,k} = t_f \sum_{k=1}^{k=N} \left(\overline{\Gamma}_{A\ m,k}^{B} r_{i,k} + u_{i,k} \right) \tag{3.3.3}$$

The objective function was defined in equation (3.2.2) as

$$E_A^{NB} = \int_0^{t_f} r^T Qr \; dt \qquad \textbf{(3.3.4)}$$

This is discretized as follows

$$E_A^{NB} = t_f \sum_{i=1}^{i=M} \sum_{j=1}^{j=ncol} E_{A\ i}^{NB} \cdot h_i \cdot \Omega_{j,ncol} \qquad \textbf{(3.3.5)}$$

$E_{A\ i}^{NB}$ is the value of energy (objective function) in the i^{th} element. Bond length and excluded volume constraints are also discretized

$$l - \varepsilon \le r_i^T \cdot H_m \cdot r_i \le l + \varepsilon \qquad m = 1...NB \qquad i = 1...M \qquad \textbf{(3.3.6)}$$

Here, NB represents the number of bond lengths in a chain. r_i is the position vector in the i^{th} element. Similarly, excluded volume constraints are rewritten in discretized form

$$r_i^T L_n r_i \ge d_{kl} \qquad n = 1...NEXC \qquad i = 1...M \qquad \textbf{(3.3.7)}$$

$NEXC$ represents the number of excluded volume constraints.

Chapter 4

RESULT AND DISCUSSIONS

4.1 Behavior of Short and Long Range Contacts

In our study, we applied our method to 1vii, chicken villin headpiece which is the smallest protein that can fold autonomously. It has three short helices. We refer to them as helix 1, 2 and 3. These helices contain the residues 4-8, 15-18, and 23-30, respectively. They are held together by a loop between residues 9-14, and a turn between residues 19-22. Figure 4.1.1 shows the 3-D structure of 1vii in tube representation.

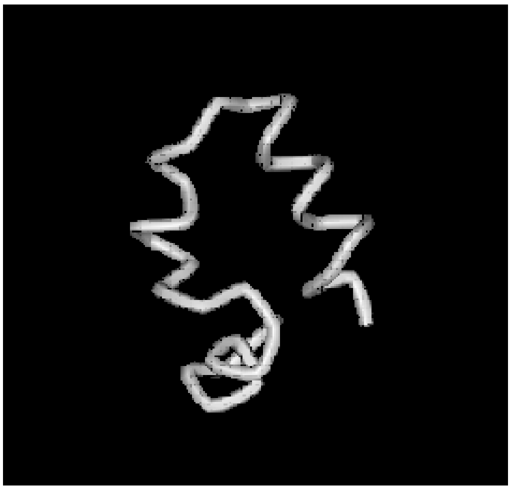


Figure 4.1.1 3-D structure of 1vii with tubular representation

The secondary structures are shown in Table 4.1.1.

Table 4.1.1 Secondary structure units and their corresponding residue numbers.

Residue Number	Secondary Structure
4-8	Helix1
9-14	Loop
15-18	Helix 2
19-22	Turn
23-30	Helix3

Starting from random 19 initial configurations folding trajectories were obtained and analyzed. The results are discussed next.

We can separate the native contact pairs into two groups: short range and long range contact pairs. We define the long range pairs as the pairs which are 5 or more residues apart. Short range pairs are less than 5 residues apart.

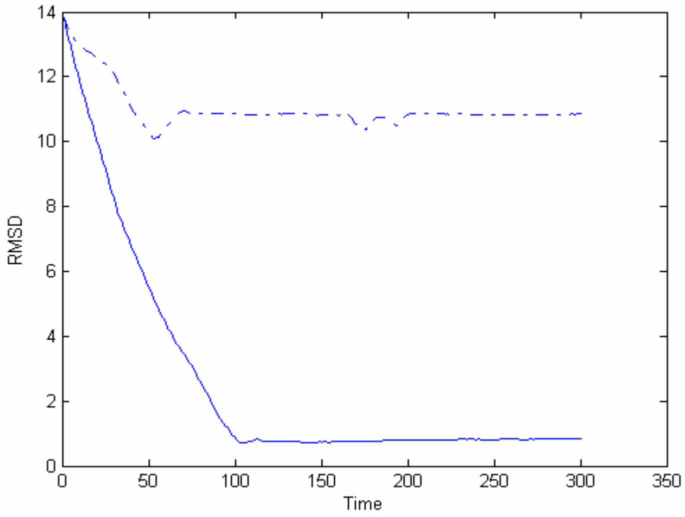


Figure 4.1.2 *rmsd* changes for the chain with all native contacts (solid line); and with only short range native contacts (-.-.)

Figure 4.1.2 shows that when all native contacts (short and long range) are included in the objective function, the *rmsd* (root mean square deviation) value is decreased to very small values, below 1Å. If we omit the long term contacts in the objective function, the time to form the chain becomes shorter as shown by how fast the *rmsd* values settle to their final values. This is expected because long range contacts are reported to slow down the folding process [8]. However the *rsmd* value for the chain increases significantly when long range contacts are omitted. This is due to the fact that although the secondary structures successfully form (as indicated by their small individual *rmsd* values), the overall configuration of the chain can deviate from the native configuration. This is illustrated in Figure 4.1.3.

Table 4.1.2 Final *rmsd* values of substructures and whole chain for the optimization cases, with all native contacts and with only short range native contacts.

.	Final Rmsd Values for the optimization with only short range native contacts	Final Rmsd Values for the optimization with all native contacts
Helix1	0.090	0.0638
Loop	1.140	0.0999
Helix2	0.100	0.0656
Turn	0.741	0.1190
Helix3	0.060	0.0595
All chain	10.86	0.8258

In Table 4.1.2, it can be seen that final *rmsd* of 0.83 Å is reached for the whole chain is in the case of optimization with all native contacts. However, it is 10.86 Å when long range contacts are omitted in the optimization.

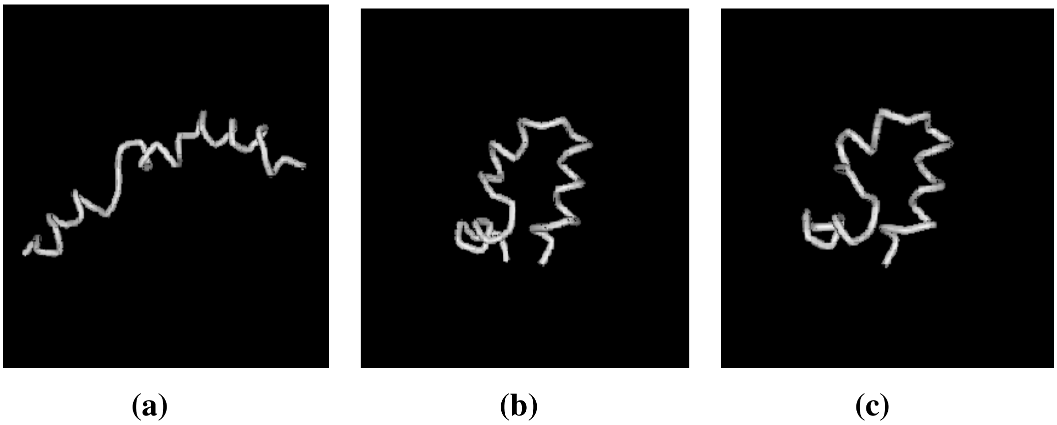


Figure 4.1.3

(a) The final configuration reached by our model when long range contacts are omitted

(b) Final configuration reached by our model when all native contacts are included

(c) Native state of protein

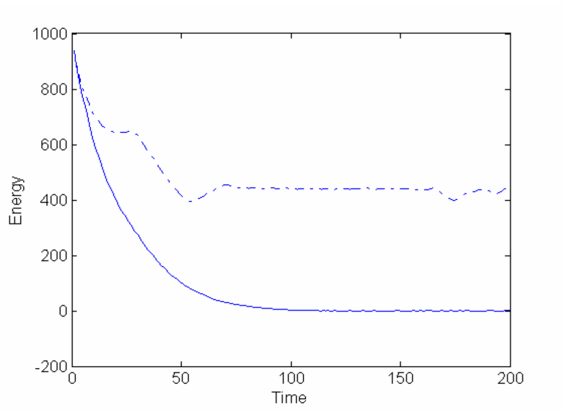
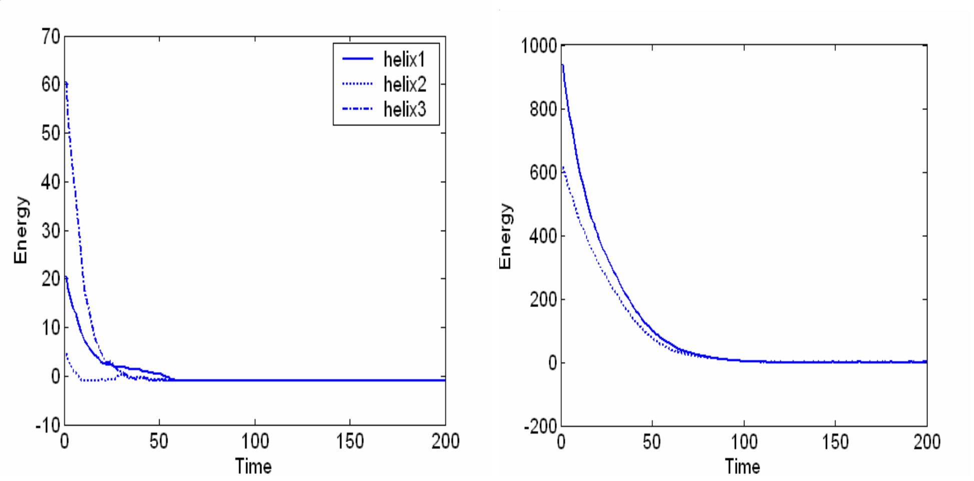


Figure 4.1.4 Energy change curves for the folding process with only short range contacts (-.), and with all contacts (-).

In our optimization there are 89 native contacts; 8 of them are the long range contacts and 81 of them are the short range contacts. In Fig 4.1.4, we can see that if we omit to minimize the long range contacts, the final energy is relatively high. Thus, long range contacts are dominant in terms of energy in our energy function formulation.

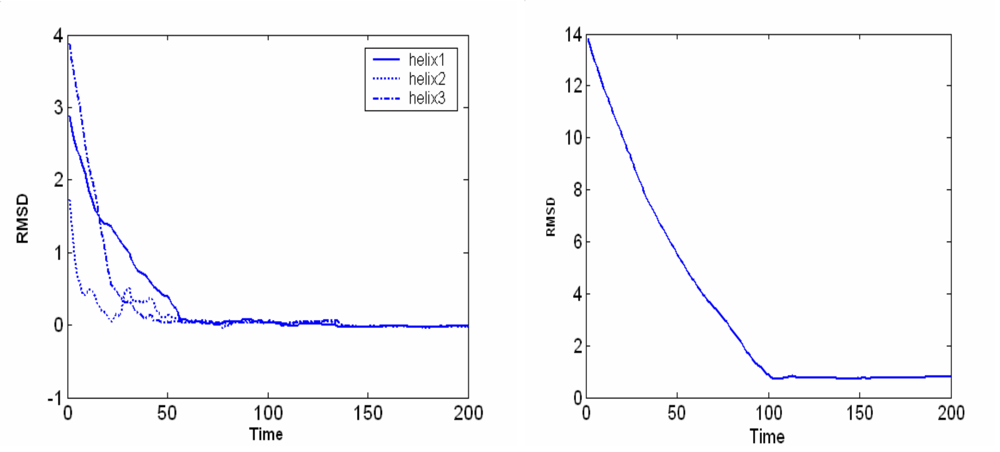


(a) **(b)**

Figure 4.1.5

(a) Energy changes for the individual helices.

(b) Energy changes for long range contacts and energy change for the whole system.



(a) **(b)**

Figure 4.1.6

(a) *rmsd* changes for the individual helices.

(b) *rmsd* change for whole chain.

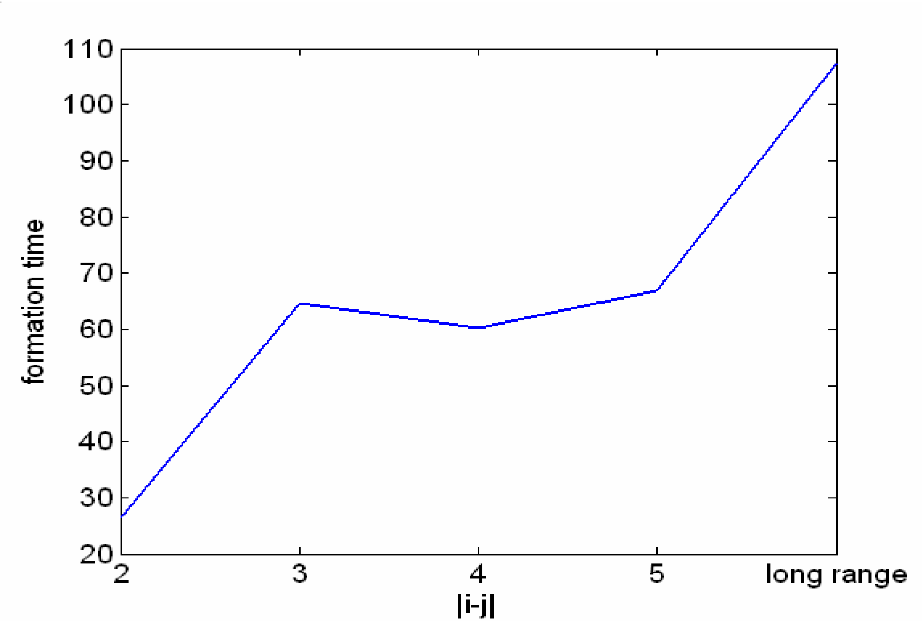


Figure 4.1.7 Average formation time for 19 initial conditions with respect to residue index difference 2, 3, 4, 5, and long range contact pairs.

In figure 4.1.5(a), the trend of individual helix energies is shown. In figure 4.1.5(b), the dotted line represents the energy of the contacts that are five and more residues apart. It can be seen that the energy of long range contacts settle down at the same rate with the energy of the whole chain. This result is in accordance with the result that the long range contacts are less likely to form in the early stages of folding [8]. Tertiary contacts are responsible for tertiary structure formation. It is clearly seen that the local contacts, which account for helix energies, form prior to tertiary contacts.

In figure 4.1.6, *rmsd* change for helices and whole chain is displayed. It is obvious that, individual helices form earlier than the whole chain. It shows that local contacts form first. In figure 4.1.7, average formation time with respect to residue index difference is shown. It can be seen that the fastest formation occurs when the residues are closest in position (2 residues apart). When the residue index difference is 3, 4, 5, formation time is close to each other. However, formation time increases fast, when the pairs are more 5 residues apart. This plot also shows us that, local structures form first, and long range contacts form in the last stages of folding. Here, we define the formation time as the first arrival time of the *rmsd* value of a secondary structure to its final value.

4.2 Analysis of Sequence of Events during the Folding Process

In this part, we try to clarify the sequence of events during folding from an arbitrary initial condition. We studied sequence of events in three groups, namely, first formation time, deformation period and settling time. First formation time is the first arrival time of the *rmsd* value of a secondary structure to its final value. Deformation period is the time intervals in which *rmsd* values exceed the limits defined around their final values. To illustrate these concepts, we show the *rmsd* change of helix1 for the first initial condition in figure (4.2.1). The dashed line in the middle shows the final value of *rmsd* of helix 1. The other two dashed line represent the interval in which the settling can occur. It can be seen that the formation time of helix 1, is 58, which is the first arrival of *rmsd* value to its final value. Between times 62-87, the *rmsd* values fall outside the settling interval. This period is called the deformation period as the structure has deformed. After time 87, the *rmsd* values remain in the settling interval again, so settling time for helix 1 is 87.

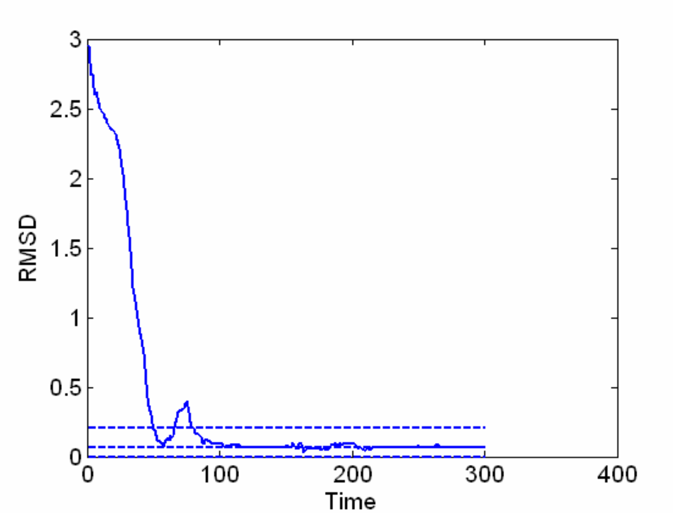


Figure 4.2.1 *rmsd* changes of helix 1 for first initial condition (solid line), settling interval of *rmsd* changes (--).

Table 4.2.1 Times for first formation of three helices for 19 different initial conditions. Those calculations are based on the rmsd values of helices.

Initial Condition Num.	Time of First Formation Helix1	Time of First Formation Helix2	Time of First Formation Helix3	Time of First Formation Loop	Time of First Formation Turn
1	58	18	61	101	62
2	82	26	59	53	110
3	84	24	79	23	98
4	92	35	70	216	83

5	85	56	62	226	77
6	73	18	48	32	20
7	86	15	60	24	25
8	47	45	51	40	60
9	41	43	76	50	85
10	55	58	67	70	81
11	75	22	44	102	76
12	52	48	76	84	55
13	88	34	70	17	21
14	72	69	69	74	52
15	90	87	71	165	54
16	63	43	51	31	26
17	30	19	38	38	13
18	40	58	37	19	28
19	48	23	27	13	65
Averages	66.36	39	58.73	72.52	57.42

In table 4.2.1, we can see the first forming times of three helices, loop and the turn in villin. These are the recorded times in which individual *rmsd* values reach their final values for the first time. Standard deviation for each substructure for the first forming times data is calculated. For helices 1 and 2, 3, loop and turn, standard deviation of 19.07, 19.46, 14.44, 61.63, and 27.95 were obtained respectively. It can be concluded that, helix 3 has the lowest standard deviation, so its first formation time is less dependent on initial condition when compared to other structures. However, the loop has the highest standard deviation value, which shows that first formation time for it highly depends on initial condition. It can be observed that, helix 2 forms first for 15 out of 19 initial conditions. As we look at the average forming times, we see that the helix2 is the fastest forming structure. The same conclusion appears in Duan and Kollman's study on Villin [8]. They found out that second helix and the N-terminal of helix3 are the initiation sites of folding [8]. This result can be explained as follows. Helix 2 with three contacts is the smallest structure in the protein. Our optimization tries to bring the contact distances to their native values by minimizing the energy. Since the number of contacts that define formation of helix 2 is small, optimization finds it easiest to form the smallest structure first. Optimization, having the predictive capability based on the available dynamic model, also knows that it can temporarily deform the same helix, if this is going to help the formation of the more energy demanding helices 1 and 3 as further discussed below.

Table 4.2.2 Deformation time intervals for three helices for 19 initial conditions.

Initial Conditior Num.	Deformation Time intervals for Helix1	Deformation Time intervals for Helix2	Deformation Time intervals for Helix3
1	(62-87)	(52-67), (72-101),(152-183).	-
2	-	27-51	-
3	-	25-70 /99-111	-
4	280-298	-	-
5	86-133	(107-118),(126-131),(188-225)	-
6	-	30-68	-
7	-	240-256	-
8	-	-	-
9	-	59-95	-
10	-	(100-136), (167-216)	-
11	-	23-54	-
12	-	255-266	-
13	144-169	176-188	-

14	-	-	-
15	-	145-159	-
16	-	62-71	-
17	-	-	-
18	-	-	-
19	-	-	-

In our simulations, it is observed that protein folding goes through some distinct phases. For example, the first phase is the initial formation phase explained above. The second phase is the restructuring phase in which some of the structures go through a transition period and change their configurations. Finally the last phase is the final formation phase. In table 4.2.2, time intervals for deformation phase are tabulated for different structures. Deformation time intervals correspond to the intervals in which *rmsd* values increase and stay above the forming threshold that is defined for Table 4.1.1 (i.e. within 5% of the difference of initial and final *rmsd* values). Table 4.2.2 shows that helix 2 goes through a number of deformations in the specified time intervals for most of the initial conditions. On the other hand, helix 1 and 3 are more stable once they form. Helix 3 is stable for all initial conditions and helix 1 goes through reconstruction for four initial conditions. In our model, there are 3, 6 and 15 native contacts defined for helices 2, 1, and 3 respectively. Thus, the energy contribution of helix 1 and 3 into the objective function is greater than that of helix 2. Consequently, optimization tries to keep helix 1 and especially helix 3 more stable by not allowing them to go through unnecessary deformation. Helix 2 is between the loops and other helices. Therefore, it plays a damping role between these structures. It moves freely to achieve the tight regulation of the formation of helix 1 and 3. As a result it has to undergo more reconstructions.

Table 4.2.3 Settling times for three helices and loop and the turn for 19 initial conditions. The numbers higher than 250 are indicated with stars.

Initial Condition Num.	Settling Time Helix1	Settling time Helix2	Settling time Helix3	Settling time Loop	Settling time Turn
1	87	183	61	209	194
2	82	51	59	95	110
3	84	111	79	278*	117
4	298*	268*	70	272*	240
5	133	226	62	226	225
6	73	68	48	290*	293*
7	86	256*	60	299*	187
8	47	45	51	207	195
9	41	95	76	184	85
10	55	215	67	251	237
11	75	54	44	102	78
12	52	266*	76	265*	278
13	169	188	70	200	200
14	72	69	68	93	87
15	90	159	71	165	202
16	63	71	51	283*	288*
17	30	19	38	120	101
18	40	284*	37	285*	292*
19	48	23	27	142	32
Average	85.52	139.52	58.68	208.73	181.10

Table 4.2.3 shows the final formation phase dynamics by giving the settling times (i.e. times at which the structures reach and stay within 5% of difference of their initial and final *rmsd* values). The recorded times, which are more than 250, are indicated with stars.

These are the structures which can not settle down. Based on the average values for formation time, it can be noticed that, loop and the turn are the last forming structures. Even in many initial conditions, loop and turn are not settled to their final values. (These numbers indicated with star). Helix 3 is the first settling structure as far as average values are concerned. Since helix 3 is costly in energy function, it doesn't undergo deformation phases, so it has the shorter settling time.

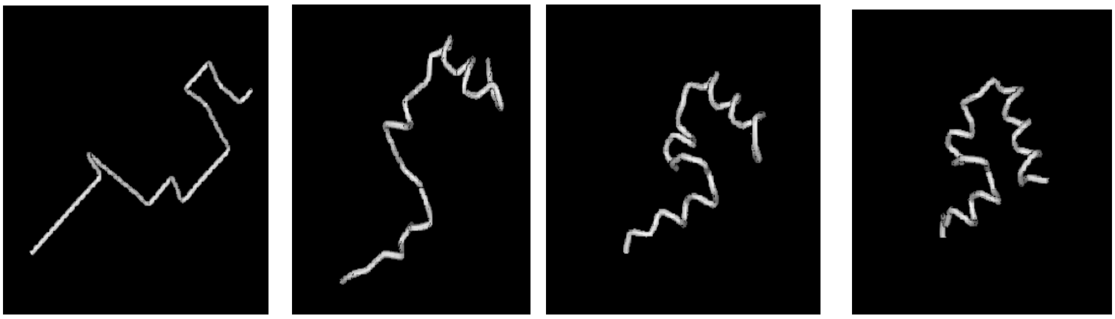


Figure 4.2.2 Snapshots from the folding process starting from an arbitrary initial condition. They show the configurations in the times, 0, 25, 50, and 75 from left to right.

In figure 4.2.2, the progress of conformations throughout time can be observed. The first figure shows the initial configuration of the chain. In the second figure, formation of helices is seen. In third part, almost all helices formed. In the last picture, the orientations of helices in three dimensions are reached with the long distance contacts formation.

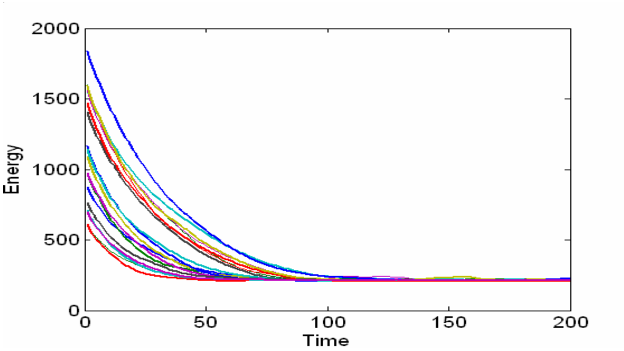


Figure 4.2.3 Superimposed energy changes for 19 different initial conditions

In figure 4.2.3, we plotted the energy changes with respect to time for the 19 different initial conditions. It can be seen that, starting initial energies cover a wide range in energy, 500-2000. The behavior of the energy is smoothly decreasing in all initial conditions. And the initial conditions starting from lower energy values, reach final value faster than the initial conditions starting from higher energy values.

As we discussed before, the force field acting on each bead in three dimensions. These forces are responsible for: (i) bringing the native contact pairs together; (ii) provide repulsive effect on bonded pairs to keep them in certain distance (bond length); and (iii) provide repulsive effects on the non-bonded pairs when they tend to become closer than their allowed distance (excluded volume effect). Thus, for the 36 residue protein 1vii, there are (36x3=108) forces components acting at each time step during the folding process.

A particular type of principal component analysis known as Karhunen-Loeve expansion (KLE) is used. The application and details of this expansion on folding trajectories of the protein CI2 can be found Palazoglu et al [17]. Using KLE, input (i.e. force) trajectories are decomposed into modes. We can summarize the KL expansion equations shortly as follows;

First, NxM input matrix U is defined as:

$$U = \left[u(t_1) ....u(t_M) \right] \qquad \textbf{(4.2.1)}$$

Where M is the number of time steps, and N is the number of residues.

Next, the covariance matrix Φ of the input matrix is constructed:

$$\Phi = \frac{1}{M} UU^T \qquad \textbf{(4.2.2)}$$

Eigenvalue decompositions of Φ results in:

$$\Phi\phi_j = \lambda_j \phi_j \qquad \textbf{(4.2.2)}$$

Where, λ_j is the j^{th} eigenvalue and ϕ_j is the j^{th} eigenvector of Φ. The original matrix can be reconstructed in terms of KL expansion.

$$u(t_m) = \sum_{j=1}^{N} c_j(t_m)\phi_j \qquad \textbf{(4.2.3)}$$

In this equation, c_j stands for the time varying amplitude of the j^{th} eigen vector. It can be calculated from:

$$c_j(t_m) = \phi_j^{\;T} u(t_m) \qquad \textbf{(4.2.4)}$$

Original U matrix obtained from optimal folding is reconstructed using first mode, first two, first three, and first ten modes. All these reconstructed inputs are applied to our dynamic model. The outputs from these simulations give us the states. From these states, the behavior of constraints and objective function are computed. In figure 4.2.3, we can see the behavior of the objective function that is evaluated using the output of the simulations from the reconstructed input matrix data. The solid line stands for the simulation results of the original optimal input matrix. The dash-dotted line, dashed and dotted lines represent the outcome of simulations from reconstructed input matrix from 1, 2, and 3 modes respectively.

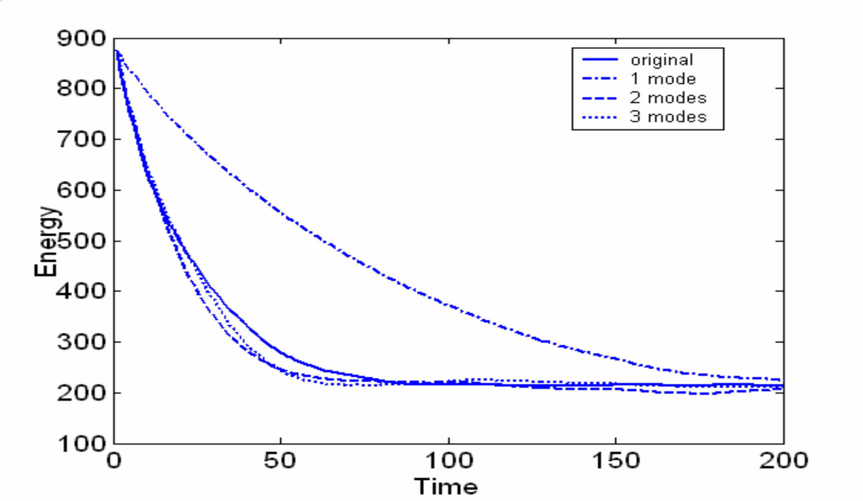


Figure 4.2.3 Energy change with respect to time for the simulations results of input matrix reconstructed with first 1, 2, 3 modes.

Figure 4.2.3 shows that when reconstruction of only first three modes of input matrix is used, we obtain very close behavior of energy change to the case when original input matrix is used. Energy is defined as the sum of the squares of native contacts distances. Thus, we can conclude that, first three modes of the input matrix can explain the attractive forces. It was told before that the forces acting in our model include attractive forces between native contacts, and the repulsive forces that are acting on bonded pairs to satisfy bond length constraints and the repulsive forces for excluded volume effect. We can check the behavior of constraints for the simulations with reconstructed forces in order to see the effects of different modes. Figure 4.2.4 shows the trend of first bond length for the output from original input matrix and reconstructed input matrices for several modes.

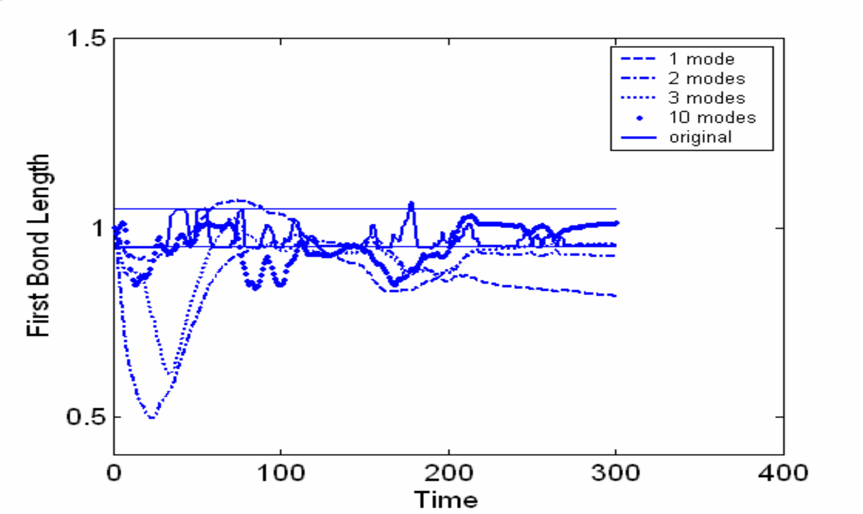


Figure 4.2.4 Change of the first bond length with respect to time for the simulation outputs of reconstructed input matrix from the first, 1, 2, 3 and 10 modes. The dashed line indicates bond length change corresponding to the result of the simulation with reconstructed input matrix with only one mode. The dash-dotted, dotted and bold dotted line explains the results for reconstructed input matrix of 2, 3 and 10 modes respectively. Solid line shows the original trend of bond length with the limits on 0.95 and 1.05 indicated with dashed straight lines.

Figure 4.2.4 shows the bond length changes for the simulations outputs of the reconstructed input matrix. We can see the behaviors for the data which is reconstructed for 1, 2, 3 and 10 modes. Even with input data that is reconstructed with 10 modes, bond length constraint can not stay within its limits. It means that bond length constraints can be explained with the inclusion of higher modes. On the other hand, we observed that attractive energies can be explained just two or three modes significantly. To see the change of force in modes, we pick the force acting on the first bead in x-direction, and compare it with its trend in the reconstructed input matrix with 3 modes.

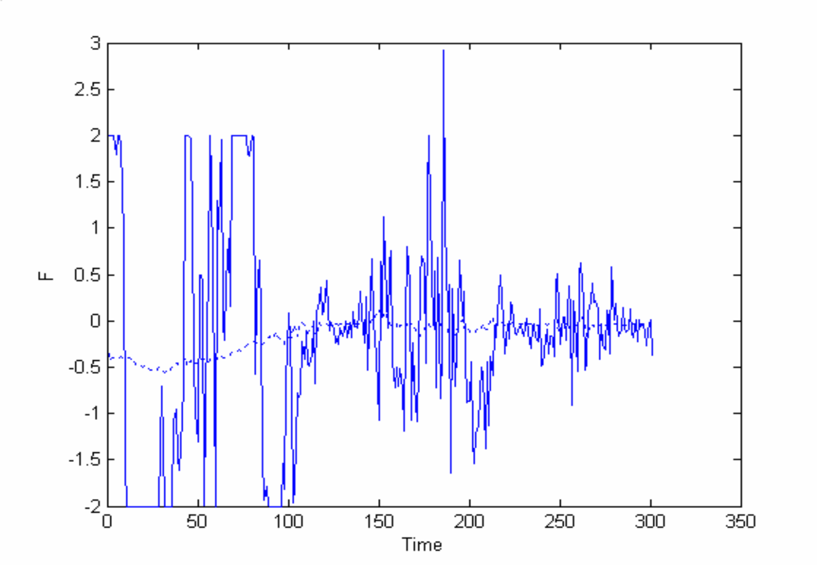


Figure 4.2.5 Change of force (F) on first bead in x-direction (solid line), change of F from the results for the reconstructed input matrix with three modes (dotted line).

Figure 4.2.5 displays the change of force on the first bead in x-direction in the original input matrix and in the input matrix that is reconstructed with three modes. It can be seen that, the force that is obtained from the first three modes shows a smooth behavior. However, the original data is very oscillatory. We showed that the energy of the folding is represented very well with the result of reconstructed data from only three modes. Therefore, we can conclude that the high frequency behavior of the force field is due to repulsive forces originating from the bond length and excluded volume constraints.

Chapter 5
CONCLUSION

The protein folding problem is regarded as a problem of great importance in natural sciences. The folding of an extended protein to its unique three-dimensional folded native state is a complex problem which has attracted a great deal of interest in recent years [18]. Despite numerous theoretical and experimental studies, a comprehensive understanding of many aspects of the protein folding is still lacking [18].

Various pathways are available to a folding protein that starts from a random initial configuration and ends in the native state. Each path is determined by the succession of self-interactions of the elements that make up the protein molecule [19]. Pathways can be described in terms of sequence of events. Formations of secondary structures, deformations of a helix, or entrance of a loop between two structures are few examples of such events [19]. Computer simulations can provide rich information about pathways. In this regard, we introduced a new optimal control approach based on a simplified model in which residues are taken as linked beads. We assume that native structure of the target protein is known and we obtain the native contact data from the native structure. We first model the interaction between covalently bonded beads as linear springs. The other interactions are defined as a force field that helps to fold the protein from an initial condition to a final native state in a feasible (i.e. without violating excluded volume and bond length constraints) and an optimal way (i.e. with respect to the defined objective function). In our approach, these forces are automatically computed to deliver the desired folding. We studied with a small, fast folding protein, Chicken Villin Headpiece. It has 36 residues that form 3 short helices, a loop and a turn.

We obtained 19 pathways starting from random initial conditions. We obtained the sequence of events for each pathway. This sequence of events includes formation of helices, loops and turns, deformation periods of these substructures and settling times for each structure. We observed that there is no unique pathway for proteins starting from random initial conditions. Each pathway has its own succession of events. However, we also observed some common characteristics for most of the pathways. It has been observed that the helix 2 is the fastest forming structure in folding of villin headpiece for different pathways. The result is in accordance with the MD simulations of Villin by Duan and Kollman [8]. Long range and short range contacts and their formation are also studied. We found that, short range contacts which are responsible for the formation of helices and orientation of successive substructures are formed prior to the long range contacts. Long range contacts are the tertiary contacts and they provide the orientation of secondary structures in 3-D space. This result also conforms to the finding that tertiary contacts are less likely to form in the early stages of folding [8].

Lastly, we tried to analyze the force field we obtained from the optimization program using a particular type of principal component analysis known a KLE expansion. Force field trajectories are decomposed into modes. It is observed that high frequency behavior of the force field is due to the repulsive forces that try to satisfy the bond length constraints. On the other hand, energy of the protein (minimum attractive energies between non-bonded native contact pairs) can be represented quite well with the result of simulations with the few modes of force field. Therefore, we deduced that, first few modes of the force field could explain the attractive forces between non-bonded pairs, and higher modes capture more of the higher frequency repulsive forces.

Our method is a new approach and it is easy to construct and implement. Since it is a simplified model, it lacks the molecular details that MD simulations include. However, we can generate feasible and optimal pathways by the machinery of the proposed optimal control formulation and use these pathways to get further insight into folding. In particular, we can extract valuable information about general characteristics of optimal folding pathways and sequence of events and properties of force fields acting on molecules. Besides, we have the flexibility to change and improve an objective function, i.e., energy definition of the system.

As future work, our model can be implemented for other proteins. Modifications in the dynamic model and objective function and improvements in computation time may lead to better results.

APPENDIX

Appendix A

A1. Bond length constraints.

In section 3.2, we refer to the matrices that define bond length and excluded volume constraints. We can state bond length constraints in most open for as follows:

$$l - \varepsilon \leq \left\| r_i - r_{i+1} \right\| \leq l + \varepsilon \qquad i = 1..n - 1 \qquad \textbf{(A1.1)}$$

Here ε is a very small number**,** l is the bond length distance, r_i represents the position of i^{th} bead. n is the number of beads in the sequence.

Let us define,

$$h_i = \begin{bmatrix} 0 & . & . & 0 & 1 & -1 & 0 & . & . & 0 \end{bmatrix} \qquad r = \begin{bmatrix} r_1 \\ . \\ . \\ . \\ r_i \\ r_{i+1} \\ . \\ . \\ . \\ r_n \end{bmatrix} \qquad \textbf{(A1.2)}$$

So that;

$$(r_i - r_{i+1}) = h_i r \qquad \textbf{(A1.3)}$$

Then;

$$\|r_i - r_{i+1}\| = (r_i - r_{i+1})^2 = (h_i r)^T h_i r = r^T h_i^T h_i r \qquad \textbf{(A1.4)}$$

Let

$$h_i^T h_i = H_i \qquad \textbf{(A1.5)}$$

We obtain;

$$\|r_i - r_{i+1}\| = (r_i - r_{i+1})^2 = r^T H_i r \qquad \textbf{(A1.6)}$$

A2 Excluded volume Constraints.

Excluded volume constraints can be written in open form as follows:

$$\|r_i - r_j\| \geq d_{ij} \quad \text{where} \quad |i - j| \geq 2 \qquad \textbf{(A2.1)}$$

Here d_{ij} is the minimum allowable distance between i^{th} and j^{th} bead, r_i and r_j represents the position of i^{th} and j^{th} bead respectively.

Let us define,

$$l_i = [0 \quad .. \quad 0 \quad 1 \quad ... \quad -1 \quad 0 \quad . \quad . \quad 0] \qquad \textbf{(A2.2)}$$

In l_{il} vector i^{th} element taken as unity and j^{th} element is taken as minus unity.

Using (A2.2), we can get;

$$\left\| r_i - r_j \right\| = \left(r_i - r_j \right)^2 = \left(l_i r \right)^T l_i r = r^T l_i^T l_i r \qquad \textbf{(A2.3)}$$

Let us define;

$$L_i = l_i^T l_i \qquad \textbf{(A2.4)}$$

We may restate (A2.3) as;

$$\left\| r_i - r_j \right\| = \left(r_i - r_j \right)^2 = r^T L_i r \qquad \textbf{(A2.5)}$$

BIBLIOGRAPHY

[1] Jiang, T., Cui, Q., Shi, G., Ma, S. (2003). Protein folding simulations of the hydrophobic-hydrophilic model by combining tabu search with genetic algorithms. *J. Chem. Phys*., **119 (8),** 4592-4596.

[2] Dinner, A.R., Sali, A., Smith, L.J., Dobson, C.M., Karplus, M. (2000).Understanding protein folding via free-energy surfaces from theory and experiment. *TIBS*, **25,** 331-339.

[3] Zagrovic, B., Snow, C. D., Shirts, M.R., Pande, V.S. (2002).Simulation of Folding of a Small Alpha-helical Protein in Atomistic Detail using Worldwide-distributed Computing. *J. Mol. Biol.* **323**, 927-937.

[4] Fernandez, A., Shen, M., Colubri, A., Sosnick, T.R., Berry, R.S., Freed, K. F. (2003). Large-Scale Context in Protein Folding: Villin Headpiece. *Biochemistry*, **42**, 664-671.

[5] De Mori, G.M.S., Colombo, G., Micheletti, C. (2005). Study of the Villin Headpiece Folding Dynamics by Combining Coarse-Grained Monte Carlo Evolution and All-Atom Molecular Dynamics. *PROTEINS: Structure, Function, and Bioinformatics*. **58**, 459-471.

[6] Kubelka, J., Eaton, W.A., Hofrichter, J. (2003). Experimental Tests of villin Subdomain folding Simulations. *J. Mol. Biol.* **329**, 625-630.

[7] 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.(2003).Atomistic Protein folding Simulations on the Sub millisecond time Scale Using Worldwide Distributed Computing. *Biopolymers*, **68,** 91-109.

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

[9] Shen, M., Freed, K.F. (2002). All-Atom Fast Protein Folding Simulations: The villin Headpiece. *PROTEINS: Structure, Function, and Bioinformatics*, **49**, 439-445.

[10] Mukherjee, A., Bagchi, B. (2003). Correlation between rate of folding, energy landscape, and topology in the folding of a model protein HP-36. *J. Chem. Phys.,* **118,** 4733-4747.

[11] De Mori, G.M.S., Micheletti, C., Colombo, G. (2004). All-Atom Folding Simulations of the villin Headpiece from Stochastically Selected Coarse-Grained Structures. *J. Phys. Chem. B*., **108**, 12267-2270.

[12] Cieplak M., Hoang, T.X, (2003). Universality Classes in Folding times of Proteins. *Biophysical Journal*, **84,** 475-488.

[13] Erman, B., Dill, K. (2000).Gaussian Model of Protein Folding. *Journal of Chemical Physics*, **112**, 1050-1056.

[14] Kocvara, M., Stingl, M. (2003). A Generalized Augmented Lagrangian Method for Semi definite Programming. *G. Di. Pillo and A. Murli (Eds*), 297-315.

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

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

[17] Palazoglu, A., Gursoy, A., Arkun, Y., Erman, B. (2004).Folding Dynamics of Protein from Denatured to Native State: Principal Component Analysis. *Journal of Computational Biology*, **11**, 1149-1168.

[18] Srinivas, G., Bagchi, B. (2003). Study of Pair Contact Formation among Hydrophobic Residues in a Model HP-36 Protein: Relationship between Contact Order Parameter and Rate of Folding and Collapse. *J. Phys. Chem. B, **107**, 11768-11773*.

[19] Erman, B. (2001).Analysis of Multiple Folding Routes of Proteins by a Coarse-Grained Dynamics Model. *Biophysical Journal,* **81**, 3534-3544.

[20]WikipediaEncyclopedia.http://en.wikipedia.org/wiki/Protein_folding#The_relationship_between_folding_and_amino_acid_sequence.

[21] Yue, K., Dill, K.A.(1993). *Phys. Rev. E.*, **48**, 2267.

[22] Scolnick, J., Kolinski, A. (1990). *Science,* **250,** 1121.

[23] Kolinski, A., Galazaka, W., Skolnick, J. (1996). *Proteins,* **26**, 271.

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

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

[26] http://folding.stanford.edu/