Derivation of Pair Potentials for Optimal Folding of Secondary Structures

by

Sefer Baday

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
November, 2008

Koc University

Graduate School of Sciences and Engineering

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

Sefer Baday

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)

Assoc. Prof. Özlem Keskin

Assoc. Prof. Attila Gürsoy

Assoc. Prof. Alper Tunga Erdoğan

Date: 14.11.2008

ABSTRACT

We present a method to calculate the pair potentials for folding of protein secondary structures. In the first part of the method necessary training data are generated to compute the potentials. For this purpose a Go-type model and dynamic optimization is used to compute the optimal folding trajectories. A coarse-grained model for a helix and a beta sheet, each consisting of 12 residues has been constructed by representing each amino acid as a bead. The dynamic optimization gives the total optimal force acting on each residue (bead) to fold the protein from an initial configuration to its native state. Next, forces between pairs of residues are derived from this data. This is done by first projecting the optimal residue forces onto the pair-wise directions between residues and expressing these mean forces as (nonlinear) functions of pair-wise distances. We show how to compute the forces between pairs from the mean forces. We next incorporate the derived pair forces into the dynamic model. Thus, for new initial conditions folding is achieved in a predictive way by simulating this model without any need for optimization. We further show that the folding pathways obtained by such "simple" simulation are similar to folding pathways which can be obtained by the rigorous dynamic optimization. To measure similarity between folds we use MPCA (Multi-way Principal Component Analysis). In addition, mean forces between pairs are presented and analyzed.

ÖZETÇE

Tezimizde, protein ikincil yapılarını optimal olarak katlayan amino asit çiftleri arasındaki potensiyelleri hesaplama metodu sunmaktayız. Bu metodun ilk bölümünde çiftler arası potensiyelleri hesaplamak için gerekli olan veriler üretilmektedir. Bu nedenle optimal katlanma yollarını hesaplamak icin Go-modeli ve dinamik optimizasyon kullanılmıştır. Herbiri 12 amino asit içeren bir heliks ve bir β- tabakalı yapı oluşturuldu. Bu şekilde, her amino asidi bir boncuk ile gösterilen kaba ölçekli yapılar elde edildi. Dinamik optimizasyon, proteinin bir başlangıc konfigurasyonundan onun doğal haline katlanmasını sağlayan her boncuk uzerindeki toplam kuvveti bulmaktadır. Sonraki adımda bu veriden amino asit çiftleri arasındaki kuvvetler çıkartılmaktadır. Bu da şöyle yapılmaktadır: ilk olarak optimal kuvvetlerin amino asit çiftleri yönünde izdüşümleri alınr ve sonra bu ortalama kuvvetler amino asit çiftleri arasındaki uzaklığın doğrusal olmayan bir fonksiyonuyla ifade edilir. Bu ortalama kuvvetlerden amino asit çiftleri arasındaki kuvvetlerin nasıl hesaplandığını göstermekteyiz. Sonra çıkardığımız çiftler arası potensiyelleri dinamik modelimize eklemekteyiz. Böylece, verilen herhangi bir başlangıç konfigurasyonu icin, önceden tahmin edici bir yol olarak benzetim yolu ile katlama başarılmıştır. Bu modelde optimizasyona ihtiyaç kalmamaktadır. Ayrıca, bu basit modelle oluşturulan katlanma yollarının dinamik optimizasyon sonucu elde edilen katlanma yollarıyla benzer olduğunu gostermekteyiz. Bu katlanmalar arasındaki benzerliği göstermek icin çok boyutlu temel bileşen analizi kullanmaktayız. Ek olarak, amino asit çiftleri arasındaki ortalama kuvvetler gösterilip analiz edilmektedir.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my advisors, Prof. Dr. Yaman Arkun and Prof. Dr. Burak Erman for their guidance and support throught my graduate study and during the completion of this thesis. I am proud of being one of their students. I thank also Associate Prof. Özlem Keskin, Associate Prof. Attila Gürsoy and Associate Prof. Alper Tunga Erdoğan for their participation in my thesis committee and for critical reading of my thesis.

I would like to thank all my friends at Koç University who made the two years an enjoyable period of time; Cengiz Ulubas, Osman Yogurtcu, Besray Unal, Nurcan Tuncbag, Ozge Engin, Bora Karasulu, Osman Sengul, Ramazan Erduran, Bahar Ondul, Gulay Ergul, Fatih Toy, Ozkan Egri, Ekin Tuzun, Gozde Kar, Emre Guney, Aslıhan Aslan, Ayse Kucukyılmaz, Abdullah Turan, Huseyin Seren, Baybora Baran and Yasemin Demir.

Finally, I am grateful to my parents Salatın and Latif; my brother Murat and my sister-in-law Rachel; my sister Meryem and my little brother Fatih for their love, affection and support throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Proteins are one of the most abundant macromolecules in living organisms and they have crucial functions in biological processes. They can serve as catalyst for biochemical reactions, store and transport other molecules and carry signals from cell to cell. Some proteins have important role in immune responses, cell adhesion and cell cycle. Proteins also provide mechanical support and generate movements to cells and tissues, such as actin and myosin in muscle[1] .

The diversity of functions performed by proteins arises from the enormous number of three dimensional shapes they have. After being synthesized, the majority of proteins must be converted to folded compact structures in order to function. This process is known as Protein Folding. Only correctly folded proteins have long term stability in biological environment and failure in this process may cause severe diseases. Some of these diseases such as cystic fibrosis result from the fact that incorrectly folded proteins can't do their functions properly. Sometimes, misfolded proteins form aggregates in the cell and deposition of these aggregates may cause several diseases such as Alzheimer's, and Parkinson's diseases[2, 3].

Understanding the protein folding is important to discover cures for diseases that stem from misfolding or aggregation. Protein folding has been studied in many aspects by computational and experimental methods. More information for protein folding is given in section 2.2.

In this thesis dynamics of folding is investigated. Specifically, folding dynamics is determined by the pair interaction potential between residues. Several types of pair potentials were developed and those potentials are used to fold the protein. One type of pair potentials is the physics-based potentials which are used in simulations to acquire detailed information about the folding dynamics [4]. However, calculation of these interaction potentials is computationally expensive. Another type of potential is the knowledge based potential [4]. These are statistical potentials which are derived from known protein structures using distribution of atom pairs. Usually they are used for protein structure prediction; however, the application of these potentials in protein folding dynamics studies is not common. In section 2.3 detailed information about pair potentials is introduced.

One aim of this thesis is to develop pair potentials for optimal folding of secondary structures. These potentials are derived for a coarse grained system in which an amino acid is represented by a bead. In order to obtain the data for the computation of the potentials an optimization model which was developed by Guner et al.[5] is used. In this model, the dynamic model is the Newton's equation of motion. This model is explained in section 3.1. Optimization gives the total force acting on each bead to fold the protein from an initial configuration to its native state.

Pair interaction potential is defined as mean force acting in a pair direction and these mean forces are modeled as a function of pair distance using results of optimizations starting from random different initial conformations. It is called mean force because the average effect of other beads on the interaction of a certain pair is included implicitly. The pair potential derivation method is given in section 3.2.

Another objective of this thesis is to use these learned pair potentials in folding simulations. By doing so, optimal folding path is achieved without carrying out an optimization again. In section 3.3 a method for using these mean forces for folding simulations is presented. It should be checked whether the trajectories obtained using derived pair potentials are similar to those of optimization simulations since it is claimed that these derived pair potentials result from optimal paths. In order to compare trajectories MPCA (Multi-way Principal Component Analysis) method is implemented. This method is explained in section 3.3.

Optimizations were carried out for secondary structures; helix and beta sheet (see section 2.1 for the information about structure of a helix and a beta sheet). Pair potentials are obtained using results of these optimizations. Then, these derived potentials are implemented into folding simulations. Trajectories of these simulations are analyzed whether they represent optimal path or not, using MPCA method. In chapter 4 results are presented and analyzed. In chapter 5 thesis is concluded with a short summary and suggestions for future work.

Chapter 2

LITERATURE REVIEW

2. 1 Protein Structure

Proteins are linear polymers composed of monomers which are called amino acids. The function of a protein is directly related to its three dimensional shape which in turn is determined by its amino acid sequence. There are 20 different amino acids and these amino acids are linked each other by peptide bond when forming proteins. Amino acids are composed of four parts; amino group, carboxylic acid group, a hydrogen atom and a distinctive R group (Figure 1). The difference in chemical properties of amino acids stem from R group which is often referred as side chain. According to these side chains amino acids may be hydrophobic, polar or charged properties.

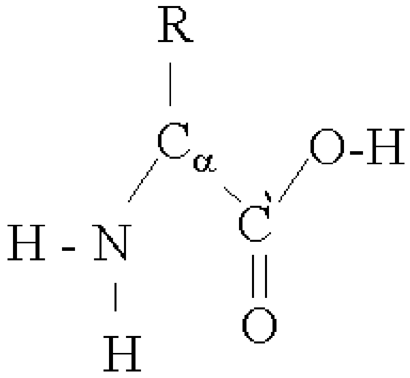


Figure 1 The general structure of an amino acid[6]

The three dimensional structure of a protein is also known as protein conformation. Conformation of protein is determined by phi (Φ) and psi (Ψ) angles. Phi and psi are the angles of rotation around N-C^{α} and C^{α} -C' (C^{α} is the backbone carbon atom bounded side chain and C' is the carbon atom in the carboxyl group, Figure 1) bonds respectively. In addition, only certain combinations of phi and psi angles are allowed because of steric hindrance between backbone and side chain atoms.

Protein structure can be interpreted in four levels. Amino acid sequence is the primary structure of proteins. Some structures regularly repeat in proteins such as α helices, β sheets and loops; these structures are the secondary structures. Tertiary structure is formed by the combination of secondary structure elements into compact globular units (domains). Final structure is quaternary structure and it refers to the spatial arrangements of polypeptide chains (Figure 2).

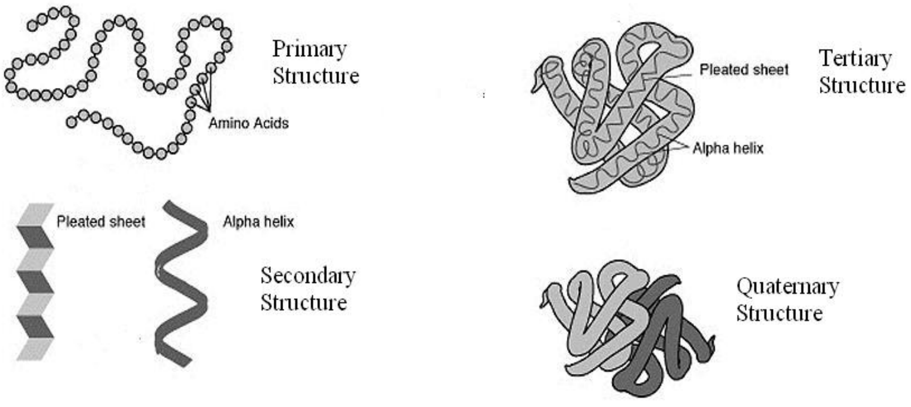


Figure 2 Protein structures[7]

Since, this study focuses on folding of secondary structures; some more information should be given about helices and beta sheets. Helix is a rod like structure which tightly coiled backbone forms inner part of the rod and the side chains extend outward in a helical form (Figure 3). Helices can be right-handed or left-handed. If you hold helix, if it moves away in clockwise direction then it is right handed otherwise it is left handed. The most common form of helices found in nature is the α helix. Most of the α helices are right handed and Ψ and Φ angles in consecutive residues are approximately -60^0 and -50^0. The α helix has 3.6 residues per turn and this structure has hydrogen bonds between C’=O of residue n and NH of residue n+4. These hydrogen bonds have very significant role in stabilization of helices[1, 2, 8]. In this study, α helix is studied due to the fact that generally this type of helix is used most in the literature.

Beta (β) sheets are major structural elements of globular proteins. β sheets are almost fully extended and built by polypeptide chains which are called β strands. β strands are usually constituted by 5 to 10 residues and are in almost fully extended conformations. A β sheet is built by linking two or more β strands such that hydrogen bonds form between C’=O group of one β strand and NH group of adjacent β strand. β strands may form β sheets in two way; antiparallel and parallel. If adjacent beta strands are opposite in direction, NH group to C’=O group, then it is an antiparallel β sheet. Otherwise it is parallel β sheet. These two kinds of β sheets have different pattern of hydrogen bonds[1, 2, 8] .

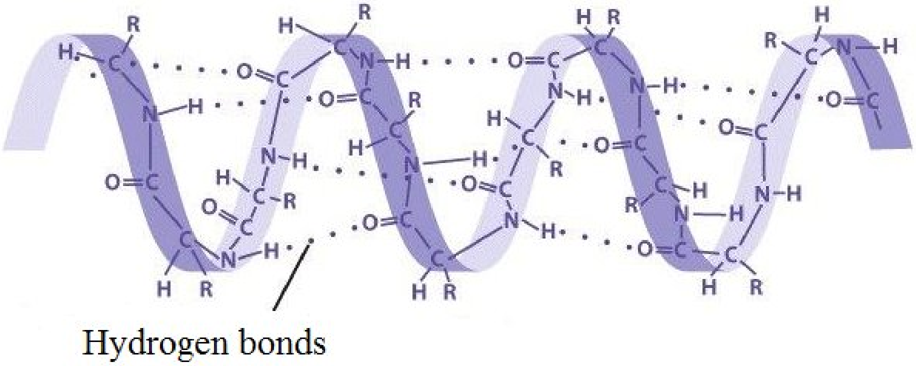


Figure 3 A right handed α helix structure[9]

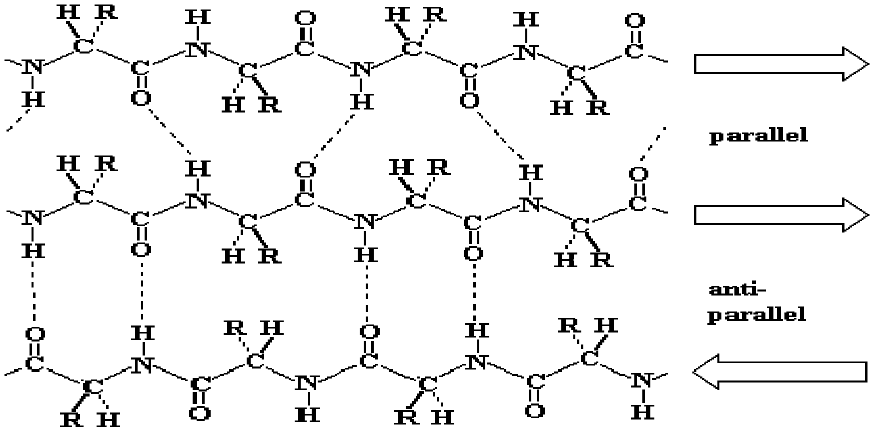


Figure 4 Antiparallel and parallel beta sheets [10]

2. 2 The Protein Folding Problem

2.2.1 Introduction

Protein folding is the process by which a polypeptide chain folds into its characteristic structure, known as native state. The question of how an amino acid sequence dictates its native state is the most fundamental problem of folding. There have been many studies in the field of protein folding since 1960s (Figure 5). The main objectives of these studies can be summarized as understanding the mechanism of protein folding and fundamental principles of folding transition, determination of principles for sequence structure, identification of major driving forces for folding[11, 12]. In the rest of this section, several aspects of folding dynamics will be mentioned since it is the main objective of this thesis.

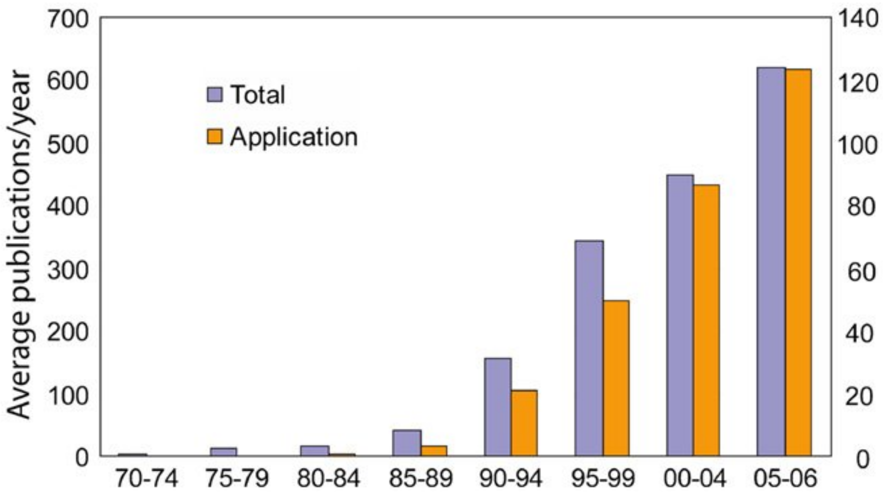


Figure 5 Growth of The Protein folding field The average number of publications per year in protein folding field (left y axis) and the average number of publications per year that are dedicated to application (right y axis) were plotted every five years between 1970 and 2004, and 2005–2006. The first dataset was generated by searching articles in PubMed that contain the keyword 'protein folding' or 'protein unfolding' in either title or abstract. The second dataset was extracted from the previous dataset by searching with the following additional keywords: 'engineering', 'design', 'misfolding', 'aggregation', 'amyloid' and 'amyloid disease'[11].

2.2.2 Levinthal Paradox

There has been enormous number of studies in order to understand the protein folding kinetics. The question of how proteins find their native structure so quickly is one of the main goals of these studies. Cyrus Levinthal one of the pioneer scientists who worked on this problem, made a comment about complexity of folding in late 1960s which was later called "Levinthal paradox". It can be expressed as follows. The number of possible configurations of a protein consisting of 101 amino acids is 3^{100} (each bond can have one of three states) and if the protein were able to search 10^{13} configuration per second then it would take 10^{27} years to search all possible configurations. However, in reality, proteins fold on a time scale of minutes at most. Levinthal concluded that proteins must follow specific folding pathways rather than randomly searching conformational space[13].

2.2.3 Energy Landscape Theory

The fastness of folding (compared to random search) can be explained by the energy landscape theory. An energy landscape describes the free energy of protein as a function of conformational properties, such as dihedral angles. Each conformation is represented as a point in a multi-dimensional surface. To achieve efficient folding, the energy landspace should resemble a funnel, because the conformational space is reduced as the native state is approached (Figure 6).

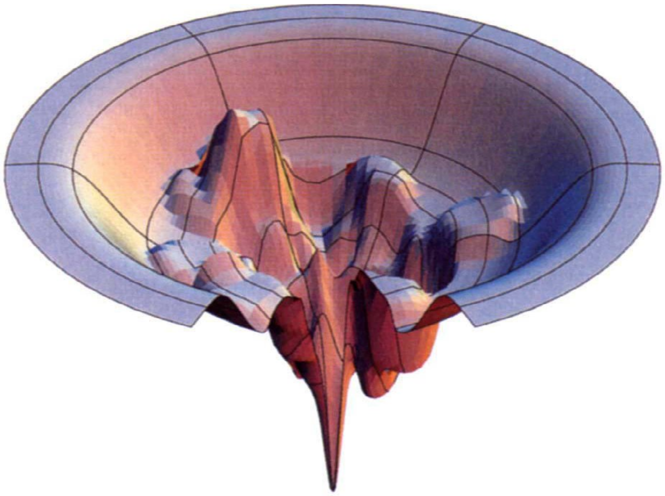


Figure 6 A rugged energy landscape [14]

In addition the shape of this funnel shows the kinetics of folding. For example in bumpy landscapes local minima are the transition states and they determine the rate of folding (i.e. the funnel of the fast two state kinetics has smoother shape).Furthermore, according to this theory transition states may have different conformations (ensemble of states) rather than specific conformations [14].

2.2.4 Folding Mechanisms

Different models have been proposed to explain mechanism of protein folding. First, nucleation-growth mechanism, which proposes the rapid formation of tertiary structure from an initial nucleus of secondary structure, was suggested. However, this model lost its favor after the studies of folding intermediates because it predicts the absence of folding intermediates. Then, the framework model has been suggested. According to this model, secondary structures form first followed by tertiary structure forms. Another model is the hydrophobic collapse which claims that the hydrophobic collapse is the main driving force to make protein take its compact shape. After the formation of the compact shape, secondary structures fold more easily due to searching narrower conformational space[15, 16].

In addition to models described above, several models have been proposed lately. The first one is the nucleation-condensation model which unites the hydrophobic collapse and framework mechanisms. This model remarks that a transition state consists of combination of long range tertiary interactions and secondary structure. Furthermore, it proposes that secondary structures are significant driving force of folding, but they need to be stabilized by tertiary interactions[15-17].

Moreover, it has been proposed that some proteins fold by stepwise assembly of foldon units rather than one amino acid at a time. In this model, folding proceeds by sequential stabilization; previously foldon units guide and stabilize subsequent foldon units to built native state[18].

Another model is zipping and assembly mechanism (ZA) which hypothesizes that local structures form at independent sites along the chain and then these structures grow (zip) and assemble with other structures. According to this model, protein can fold so quickly because different small peptide pieces of chain can form local structure on the fastest timescales. Growing of these structures happen on slower time scales. The ZA method is much faster than straightforward Monte Carlo and molecular dynamics simulations because it doesn’t search for all conformational space[19-21].

2.2.5 Experimental and Computational Methods

Many experimental techniques have been developed to study protein folding problem. Some of these are fluorescence resonance energy transfer (FRET) methods that can observe the formation of particular contacts, hydrogen exchange methods that can give detailed information about structure and properties of folding intermediates, laser temperature jump methods that can provide information about folding kinetics. Techniques which give structural information are Nuclear Magnetic Resonance (NMR), mass spectrometry, atomic force microscopy (AFM) methods[11].

Experimental methods are not sufficient to explore protein folding. Therefore computational methods have been developed to have better interpretation of experimental data by making simulations to obtain detailed information about microstates during folding. Molecular Dynamics (MD) simulation, which uses physics based potentials to find the interactions between atoms, is one of the commonly used simulation methods. Duan and Kollman achieved folding of 36-residue villin headpiece up to 4.5 Å rmsd by performing microsecond all atom explicit solvent MD simulation[22]. Recently, IBM Blue Gene group of Pitera and Swope folded 20-residue Trp-cage peptide to nearly 1 Å rmsd using implicit solvent replica-exchange molecular dynamics simulation (REMD)[23]. Pande et al accomplished folding of villin to a rmsd of 3 Å by folding@home project (a distributed grid computing system)[11, 20]. Another simulation method is Monte Carlo which is a stochastic simulation method. Vila et al. folded a 46 residue Protein A to 3.5 by performing an implicit solvent Monte Carlo dynamics[24].

Although all atom MD and MC simulations give detailed information about the folding dynamics, they require excessive computational time and resource. In order to make these simulations faster some modifications such as Coarse-graining have been done. In Coarse-grained models, not all atoms are represented explicitly; pseudo atoms are used to represent a group of atoms. Discontinuous Molecular Dynamics (CG-DMD) and Go models are examples of Coarse-grained models [25-27]. Dokholyan et al. obtained ensemble of transition states of Src Homology 3 (SH3) by using simplified models and rapid sampling DMD[11].

In addition, lattice models have been developed to make fast search on simplified conformational space. In lattice models amino acids are represented by beads and these beads have restricted conformations as defined by their moves on the lattice. One of the most popular lattice models is HP model. In this model, each bead can be either H (non polar) or P (polar). For any sequence and structure, the interaction energy can be easily calculated since the system is very simplified. Excluded volume constraints are included by avoiding occupation of one position by two beads. Two dimensional HP models resemble the general properties of globular proteins. Lattice studies have shown that secondary structures in proteins are stabilized by chain compactness[28].

2. 3 Pair Potentials for Protein Folding

In folding studies, when pair potential concept is considered, two types of potentials are usually used. First type is physics based potentials which account for fundamental interactions between atoms. Second type is knowledge-based potentials derived from experimentally determined structures. The advantage of physics based potential is to give detailed information about folding kinetics and pathways. However, computation of these potentials in a simulation takes too much time. On the other hand, using knowledge-based potential doesn't require much computational time and they are very useful in discrimination of native folds and misfolded structure and protein structure prediction studies[4].

The concept of knowledge based potentials was first suggested by Tanaka and Scheraga[29]. Miyazawa and Jernigan[30] determined the effective inter-residue contact potentials which have been used extensively in protein native structure prediction studies. Miyazawa and Jernigan[31] updated these contact potentials using 1661 protein subunits. There have been many studies about these pair potentials and generally potential of mean force (PMF) method is used to calculate these pair interaction potentials. The general definition of PMF is:

$$E(r) = -kT \ln\left[f(r)\right] \tag{2.1}$$

where r is the distance (or some other parameter such as dihedral angle); $E(r)$ is the energy at r; $f(r)$ is the probability density; k is Boltzmann's constant and T is the absolute temperature [32]. The use of PMF can have many variations due to choice of these parameters such as selection of interaction sites to characterize residue-residue interactions and the reference state used for calculating the probability. The most common way of selecting interaction site is to represent each residue by a single interaction site and it is usually the C_α and C_β atoms or centroids of side chains. Kocher et al. showed that selecting side chains as the interaction site gives better threading results than selecting the C_α and C_β atoms [33]. Bahar and Jernigan selected different atoms for each type of amino acid regarding them to be near the side-chain terminus. Advantages of taking multiple interaction sites are that specific interactions are explicitly taken at their original locations and are not smoothed out. Using multiple sites can expand the sample size and increase smoothness of the data[34].

Knowledge based potentials can be used in many ways in simulations. They can be used in on lattice simulations which give the overall tertiary fold, rather than details of secondary structures[35]. In addition, these potentials can be used in off-lattice simulations such as the work of Gunn et al.[36] in which a hierarchical method is used to determine tertiary structure of protein. In this method, helices and β strands are represented as cylinders and spheres are used to represent loops. It is assumed that secondary structures form before the formation of tertiary structures. MC simulated annealing and genetic algorithm is used to make simulation. Hydrophobic potentials of Casari-Sippl[37] are used in these simulations. The method is applied to myoglobin and folded structure and a rmsd of 6.2 Å was obtained.

There have been several studies regarding the distance dependency of interaction potentials. Bahar and Jernigan [34] developed a method to derive residue-specific potentials for the interaction of side-chain pairs and side-chain backbone pairs. Multiple interaction sites are chosen to improve the specificity and smoothness of the distance dependent potentials. Results of this study suggest that the most favorable attractive potentials between hydrophobic pairs are in the distance between 4 and 6 Å, whereas polar and charged pairs have stronger interactions in close interval between 2 and 4 Å. Tobi and Elber[38] obtained distance-dependent pair potentials by results optimization. Energy function is the sum of pairwise interactions. The distance between two interaction sites, centers of side chains, is divided into 13 intervals and the energy of each interval is optimized independently. The result of this study is well aligned with the result of Bahar and Jernigan's[34] work in the aspect of hydrophobic component of the potential. In addition, Mukherejee et al. acquired distance and orientation dependent pair potentials for residues. Interaction sites are side chains which are represented by a single ellipsoidal site. The potential is four dimensional; distance and angles between principal axis of and axis linking centers of ellipsoids and torsion angles[39].

Besides residue level potentials, atomic level pair potentials have been developed as well. Lu and Skolnick [4] developed a heavy atom distance-dependent knowledge-based pair potential. Atoms are selected based on a residue specific and on intra-residue position specific properties; that is, C_α of ALA is different from C_β of ALA and is also different from C_α of ILE. Total number of different atoms is 167 and hydrogen atoms are neglected. The distance between any two atoms is divided 14 intervals and PMF method is used to find interaction potential. This atomistic pairwise potential has better selectivity for near-native structures. Moreover, another all atom level potential derivation method was suggested by Melo and Feytmans. 40 different heavy atom types are selected as the interaction site depending on bond connectivity, chemical properties and location (side-chain or backbone)[40].

Generally, in the knowledge-based pair potentials the effect of chain connectivity is neglected. Skolnick at al. developed a method to find pair potential of effective inter-residue interactions that explicitly contain chain connectivity. Gaussian chain is taken as the reference for the constraint of chain connectivity. Keskin et al. extracted intermolecular inter-residue potentials for interactions taking place protein-protein interfaces including the effect of chain connectivity [41, 42].

In addition to knowledge-based potentials, Erkip et al. developed a method for finding optimal parameters for Gaussian model of protein folding[43]. In Gaussian model, each amino acid is assumed as a bead and all covalent and non-covalent interactions are represented by Hooke's law springs. Spring constants are the parameters to be optimized and these parameters depend on the type of amino acid pairs. The advantage of Gaussian model is that forces linearly depend on displacements so the global minimum (native conformation) can be easily found by matrix algebra. Minimum energy conformation, starting from initial conformation, is reached by iteratively computing the parameters. By this method minimum energy conformation comes closer to the native state. The method was applied to several small proteins, such as BPTI (predicted with a rmsd of 1.7 Å) [43, 44].

Chapter 3

METHODS AND MODELS

3. 1 Optimization Model and Formulation

The optimization model developed by Guner et al.[5] is taken as a base for the optimizations performed in this thesis. In this model, each amino acid is represented by a C^{α} atom. The distance between bonded pairs is assumed to be 3.8 Å. The dynamic model is based on Newton's equation of motion:

$$m\,\frac{d^{2}r_{i}}{dt^{2}} = -\gamma\,\frac{dr_{i}}{dt} + f_{i} \quad \text{for i=1,2, …., N} \qquad (3.1)$$

Where r_i denotes the position vector of C^{α} atom of i^{th} residue with respect to a fixed frame coordinate; m denotes the mass of the residue; γ is the friction coefficient and f_i stands for the non-friction force acting on i^{th} bead (Figure 7). This equation is a deterministic equation since no random forces act on the residues.

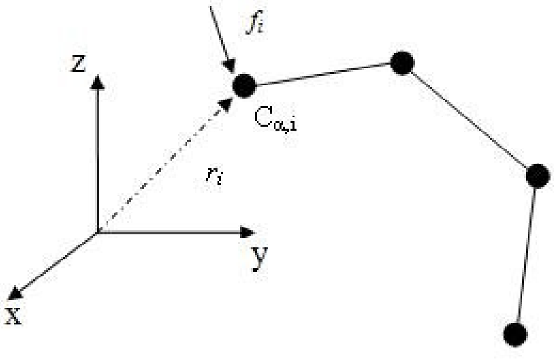


Figure 7 Example representation of 4-beads system

It is assumed that left hand side of Equation (3.1) can be taken as zero since this term is much smaller than intermolecular forces. Since left hand side is equated to zero the friction coefficient can be taken as unity. In addition it is assumed that f_i is composed of forces between bonded and nonbonded residues. f_i can be expressed in terms of its components as:

$$f_i = f_{i,A}^{B} + f_{i,R}^{B} + f_{i,A}^{NB} + f_{i,R}^{NB} \tag{3.2}$$

Here, subscript *i* denotes the residue index, *A* and *R* denote the attractive and repulsive components and superscripts *B* and *NB* denote bonded and nonbonded components.

Attractive forces between bonded residues ($f_{i,A}^{B}$) are represented by linear springs. Other components of f_i represented by a single force u ;

$$u_i = f_{i,R}^{B} + f_{i,A}^{NB} + f_{i,R}^{NB} \tag{3.3}$$

With these assumptions the equation of motion (Equation 3.1) can be written as:

$$\frac{dr}{dt} = -\Gamma r + u \qquad (3.4)$$

Where $r = [r_1 \ r_2 \ r_3 ..... \ r_N]^T$ and $u = [u_1 \ u_2 \ u_3 ..... \ u_N]^T$ are the sets of vectors for position of the beads and forces acting on the beads respectively. Connectivity matrix Γ [45] is a symmetric Toeplitz matrix (first off-diagonal elements are -1 and the diagonal elements are the sum of the corresponding row without its diagonal element).

Optimization in this thesis aims to fold a protein to its native state by bringing pair distances to their native state values. Therefore modeling of pair potentials as a function of pair distances is the major objective of this study. By the minimization of pair distances, forces acting on beads are related to pair distances.

The optimization problem can be stated as a constrained optimal control problem that makes following minimization for the time interval between time t = 0 and final time t_f:

$$\underset{u(t)}{Min}\left[\frac{1}{2}\int_0^{t_f} \tilde{r}^T(t)\,\tilde{r}(t)dt\right] \qquad (3.5)$$

Where $\tilde{r}$ is an error vector representing the difference between the actual separation at time t and target separation at the native state for all non-bonded pairs:

$$\tilde{r}(t) = \begin{bmatrix} r_{13}(t) - r_{13}^{n} \\ r_{14}(t) - r_{14}^{n} \\ \cdots\cdots\cdots\cdots \\ r_{(N-2)N}(t) - r_{(N-2)N}^{n} \end{bmatrix} \qquad (3.6)$$

Here, $r_{ij}(t) = r_j(t) - r_i(t)$ is the distance vector for *ij* pair at time t and r_{ij}^n is the known distance difference vector of pair *ij* at native state. $\tilde{\boldsymbol{r}}$ is constructed for nonbonded pairs and the dimension of $\tilde{\boldsymbol{r}}$ is $\frac{N \times (N-1)}{2} - N$ for N-beads system. $\tilde{\boldsymbol{r}}$ can be written more clearly as:

$$\tilde{\boldsymbol{r}}(t) = \begin{bmatrix} [r_3(t) - r_1(t)] - [r_3^n - r_1^n] \\ [r_4(t) - r_1(t)] - [r_4^n - r_1^n] \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ [r_N(t) - r_{N-2}(t)] - [r_N^n - r_{(N-2)}^n] \end{bmatrix} \tag{3.7}$$

The states *r* (t) have to satisfy the equation of motion or the state space model:

$$\frac{dr}{dt} = -\Gamma r + u(t) \tag{3.8}$$

$$r(t = 0) = r_0 \quad \text{(Initial condition)}$$

The model contains bond length and excluded volume constraints in order to make the folding simulation more realistic. The bond length constraint is set with a ±10% tolerance and it is formulated as in [5]:

$$0.9 l_b^2 \le r^T H_i r \le 1.1 l_b^2 \qquad \text{for all i=1.2…N-1} \tag{3.9}$$

Where H_i is the matrix that relates states to bond length, $r^T H_i r$ is the square of the bond length for adjacent atoms and l_b is the bond length.

Excluded volume constraints are included in a similar way and they are expressed in optimization as in [5]:

$$ev_{ij} \ge d_{ij} \tag{3.10}$$

Where ev_{ij} is the square of the excluded volume distance between ith and jth bead and d_{ij} is the square of the minimum excluded volume distance. The square of the excluded volume distance is calculated as:

$$ev_{ij} = \left\| r_i - r_j \right\| = r^T L_{i,j} r \qquad (3.11)$$

Where $L_{i,j}$ is the matrix that relates position to excluded volumes.

Excluded volume limit (d_{ij}) means that two beads can not come closer than this value and it depends on the native state structure. The limit values for native pairs (having distance less than 7 Å in native state) are the distance between these pairs in native structure and 5.1 Å (approximate hydrogen bond length) is taken for other non-native pairs.

In addition, there are constraints on input (u (t)) values:

$$-2 \le u_i(t) \le 2 \qquad (3.12)$$

-2 and 2 are set as the limit values for inputs because for the smaller values optimization can not find feasible solution and for the higher values there can be abrupt changes due to excessive force. Using these limits, smooth trajectories can be obtained [5].

The optimization problem is solved using PENNON solver in AMPL environment using the method proposed in [5].

The optimization adjusts the forces u(t) to drive the protein to desired structure. The dynamic model (Equation 3.8) provides the motion under the optimal force field. In addition, optimization tries to satisfy excluded volume and bond length constraints during the folding.

For many different initial protein configurations optimization provides us with data (i.e. optimal forces and pair distances) that will be used to derive the pair potentials as described in the next section.

3. 2 Pair Potential Derivation Method

Here, the method for the derivation of pair potentials is presented. The dynamic model used for optimal folding is given in equation (3.4) as:

$$\frac{dr}{dt} = -\Gamma r + u$$

For the demonstration purposes let's write this equation for three beads without loss of generality.

$$\frac{dr_1}{dt} = k(r_1 - r_2) + u_1 \tag{3.13}$$

$$\frac{dr_2}{dt} = k(r_2 - r_1) + k(r_2 - r_3) + u_2 \tag{3.14}$$

$$\frac{dr_3}{dt} = k(r_3 - r_2) + k(r_3 - r_4) + u_3 \tag{3.15}$$

where k is the spring coefficient between bonded pairs.

If we subtract Equation (3.15) from Equation (3.13), one gets:

$$\frac{dr_{13}}{dt} = -kr_{12} - kr_{23} + kr_{34} + (u_3 - u_1) \tag{3.16}$$

Here, $r_{ij} = r_j - r_i$ is the distance vector between ith and jth beads. The general form of Equation (3.16) can be represented as:

$$\frac{dr_{ij}}{dt} = A_{ij}{}^{T}\,\mathbf{r}_{ij} + (u_j - u_i) \tag{3.17}$$

Where $\mathbf{r}_{ij} = [r_{12} \quad r_{13} \; \dots . r_{N-1,N}]^{T}$ is the set of vectors for distance vectors of all pairs. In this set there are $\frac{N\times(N-1)}{2}$ vectors, so the dimension of $\mathbf{r}_{ij}$ is $3\times\frac{N\times(N-1)}{2}$ since each vector has 3 elements. A_{ij} is the vector with a dimension $3\times\frac{N\times(N-1)}{2}$ that gives relation for bonded terms. Example representation of A_{ij} for 4 beads system for Equation (3.16) is given below.

$$A_{13} = k \times \left[\underbrace{\begin{matrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{matrix}}_{r_{12}} \underbrace{\begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix}}_{r_{13}} \underbrace{\begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix}}_{r_{14}} \underbrace{\begin{matrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{matrix}}_{r_{23}} \underbrace{\begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix}}_{r_{24}} \underbrace{\begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}}_{r_{34}} \right] \tag{3.18}$$

It is clear that $r_{ij}{}^{T}\, r_{ij} = \left\| r_{ij} \right\|^{2}$. Now take the derivative of both sides with respect to time and to get the following equation:

$$2\, r_{ij}{}^{T} \frac{dr_{ij}}{dt} = 2\left\| r_{ij} \right\| \frac{d\left\| r_{ij} \right\|}{dt} \tag{3.19}$$

Rearranging Equation (3.19) leads to :

$$\frac{d\left\| r_{ij} \right\|}{dt} = \frac{r_{ij}^{T}}{\left\| r_{ij} \right\|} \frac{dr_{ij}}{dt} = v_{ij}^{T} \frac{dr_{ij}}{dt} \tag{3.20}$$

Where, $v_{ij}^{T} = \frac{r_{ij}^{T}}{\left\| r_{ij} \right\|}$ is the unit vector in *ij*-direction. Multiplying Equation (3.17) by v_{ij}^{T} leads to:

$$v_{ij}^{T} \frac{dr_{ij}}{dt} = v_{ij}^{T} A_{ij}{}^{T} \boldsymbol{r}_{\boldsymbol{ij}} + v_{ij}^{T} (u_j - u_i) \tag{3.21}$$

Using Equation (3.20) Equation (3.21) can be written as:

$$\frac{d\left\| r_{ij} \right\|}{dt} = v_{ij}^{T} A_{ij}{}^{T} \boldsymbol{r}_{\boldsymbol{ij}} + v_{ij}^{T} (u_j - u_i) \tag{3.22}$$

From Equation (3.22) it is seen that change of pair distance is related to $v_{ij}^{T} A_{ij}{}^{T} \boldsymbol{r}_{\boldsymbol{ij}}$ and $v_{ij}^{T} (u_j - u_i)$ terms. Pairwise mean force can be defined as net force acting on *ij*-direction:

$$v_{ij}{}^{T} \left(u_j - u_i \right) = g_{ij} \left(\left\| r_{ij} \right\| \right) \tag{3.23}$$

Optimization finds u (t) and r (t) for all time steps. The term $v_{ij}^{T} (u_j - u_i)$ in Equation (3.23) can be easily calculated. The pair distance data ($\left\| r_{ij} \right\|$) is also calculated using position data ($r(t)$). Next modeling is done to compute an appropriate function g for the Equation (3.23). Pair distance data and mean force data are calculated using results of optimizations for many different initial conditions. Example representation is given below for pair 1-3 for many optimization runs (results of many optimizations are merged into a vector).

$$X=\begin{bmatrix}\|r_{13}\|_1\\ \|r_{13}\|_2\\ \|r_{13}\|_3\\ \cdots\cdots\\ \cdots\cdots\\ \|r_{13}\|_n\end{bmatrix} \qquad Y=\begin{bmatrix}v_{ij}^{T}(u_j-u_i)_1\\ v_{ij}^{T}(u_j-u_i)_2\\ v_{ij}^{T}(u_j-u_i)_3\\ \cdots\cdots\\ \cdots\cdots\\ v_{ij}^{T}(u_j-u_i)_n\end{bmatrix} \tag{3.24}$$

Where $\|r_{13}\|_i$ is the distance between pair 1-3 at i$^{\text{th}}$ time step.

X vector is ordered from minimum to maximum value and Y vector is ordered according to new order of X vector (time step indices should be same in both X and Y vectors. Then X and Y vectors are divided to 50 intervals and mean forces, which lie in the same interval, are averaged. As a result, for each distance interval corresponding mean force is obtained. Finally, curve fitting is made to this data to obtain the form of $g_{ij}(\|r_{ij}\|)$. This procedure is done for all pairs. At the end, mean forces have been modeled as a function of pair distance data.

3. 3 Using Derived Mean Potentials for Folding Purposes

In this section a method for using mean forces in a folding simulation is presented. The mean forces are defined as a function of pair distance using the method explained in section 3.2.

Let's set $u_i = \sum_j c_{ij} v_{ij}$ where v_{ij} is the unit vector in *ij*-direction (Figure 8). $c_{ij} v_{ij}$ is component of u_i in *ij*-direction. Substitution of this equation into Equation (3.23) gives:

$$v_{ij}^{T}\left(\sum_{k\neq j} c_{jk} v_{jk} - \sum_{k\neq i} c_{ik} v_{ik}\right) = g_{ij}\left(\|r_{ij}\|\right) \tag{3.25}$$

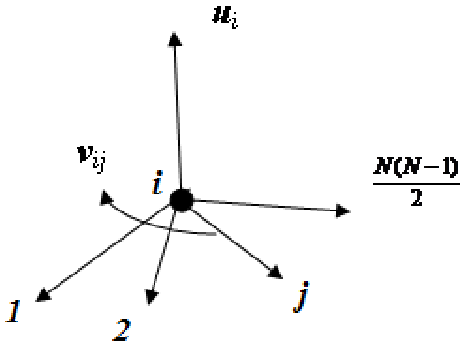


Figure 8 Resolution of the individual forces, u_i

At each time step during the simulation, mean forces are calculated using pair distance data ($\left\|r_{ij}\right\|$), since mean force pair distance relation $g_{ij}\left(\left\|r_{ij}\right\|\right)$ is known. By knowing the right hand side of Equation (3.25), c_{ij} 's can be calculated by solving Equation (3.25) simultaneously for all pairs. Example representation for the Equation (3.25) for 4 beads system is given below. Since forces between bonded pairs are represented by a spring they are not included in this calculation.

$$\begin{bmatrix} v_{13}^T * v_{13} & v_{13}^T * v_{14} & 0 & -v_{13}^T * v_{31} & 0 & 0 \\ v_{14}^T * v_{13} & v_{14}^T * v_{14} & 0 & 0 & -v_{14}^T * v_{41} & -v_{14}^T * v_{42} \\ 0 & 0 & v_{24}^T * v_{24} & 0 & -v_{24}^T * v_{41} & -v_{24}^T * v_{42} \end{bmatrix} \begin{bmatrix} c_{13} \\ c_{14} \\ c_{24} \\ c_{31} \\ c_{41} \\ c_{42} \end{bmatrix} = \begin{bmatrix} g_{13} \\ g_{14} \\ g_{24} \end{bmatrix} \qquad (3.26)$$

where g_{ij} is the mean force in *ij*-direction.

Equation (3.26) can be written in general form as:

$$\underline{A}\,\underline{\mathbf{c}} = \mathbf{g}, \tag{3.27}$$

where $\mathbf{g} = \left[g_{13} \ldots . g_{N,N-2}\right]$ $\mathbf{c} = \left[c_{13} \ldots . c_{N,N-2}\right]^{T,}$ and A is the matrix composed of projections.

For 12-bead system the dimension of A is 55 by 110. c_{ij}'s are obtained by the solution of Equation (3.27). Since A is rank-deficient in general, $\mathbf{c}$ is solved by:

$$\mathbf{c} = A^{+}\,\mathbf{g} \tag{3.28}$$

Where A^{+} is the pseudo-inverse computed by using singular value decomposition method.

After obtaining c_{ij} 's, they are used in the dynamic model:

$$\frac{dr_i}{dt} = k\,(r_i - r_{i-1}) + k\,(r_i - r_{i+1}) \;\; + \sum_j c_{ij} v_{ij} \tag{3.29}$$

Equation (3.29) can be written for the whole system as:

$$\frac{dr}{dt} = -\,\Gamma r + V\,\mathbf{c} \tag{3.30}$$

Where, V is the matrix composed of unit vectors between pairs.

$$V = \begin{bmatrix} v_{1,3} \;\ldots . v_{1,N} \;.0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . 0 \\ \ldots\ldots\ldots\ldots\ldots v_{2,4} \;\ldots . v_{2,N} \;.0 \ldots\ldots\ldots\ldots\ldots\ldots . 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . v_{N,1} \ldots\ldots . v_{N,(N-2)} \end{bmatrix} \tag{3.31}$$

3. 4 Folding Trajectory Comparison Method

In order to assess the usefulness of derived pairwise forces, one needs to compare optimal folding paths obtained from (3.4) with those of obtained from dynamic simulation using (3.30).

Multi-way principal component analysis (MPCA) method is used to make this comparison. MPCA method is generally used to monitor batch processes in chemical industry. This method is a powerful tool for analyzing historical and diagnosing problems in plant operations [46-48].

MPCA is equivalent to performing ordinary PCA on a large two-dimensional matrix formed by unfolding a three-dimensional array in certain ways. For monitoring of batch processes, the most meaningful way of unfolding is that different batches are aligned in vertical side and in horizontal side measurement variables are aligned for all times (Figure 9). In normal PCA analysis, original data is rotated and projected into new reduced space defined by first few principal components. First principal component describes the largest amount of variation in the data and second one describes the second largest amount of variation and it goes like this for other principal components. By PCA method, data is represented adequately in a simpler and more meaningful way. Likewise, MPCA method can help to understand the variation of the measured variables about their average trajectories[48].

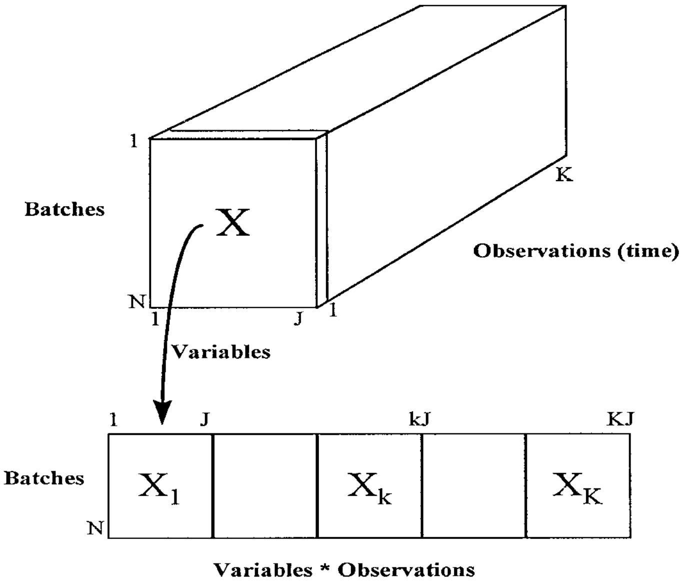


Figure 9 Unfolding three dimensional array [49]

The aim of this version of MPCA is to decompose three-way array into series of principal components of score vectors (t_r) and loading matrices (P_r) plus residual matrix E. t_r vectors are orthogonal and related only to batches whereas P_r are orthonormal and related to variables and their time variation. Each element of t_r vector corresponds to a single batch and gives the overall variability of this batch to other batches. This decomposition can be shown as:

$$X = \sum_{r=1}^{L} t_r \otimes P_r + E \tag{3.32}$$

Where L is the number of principal components used and ⊗ denotes the following multiplication: $X(i, j, k) = t(i)P(j, k)$.

In order to make this decomposition, the nonlinear iterative partial least squares (NIPALS) algorithm [50](see Appendix A) is used. It is a simple and effective algorithm to find principal components in a sequential way.

It is now explained how MPCA method can be used for discriminating trajectories or batches. For the data obtained from simulations or batches, the three dimensional matrix, X (I, J, K), is constructed. Where, I is the number of simulations, J is the number of variables and K is the number of time steps in a simulation. In order to remove the nonlinear behavior of process, data should be mean centered. It is accomplished by subtracting each column of X by its mean. Then, by MPCA decomposition method, score vectors (t_r) and loading vectors (P_r) are acquired for the $\overline{X}$ matrix (mean centered X). Similar or different simulations can be easily differentiated by means of plot of first and second score vectors (t_1 and t_2) . Because, points (t_1 and t_2) for similar simulations are clustered whereas scatter points of different simulations stay separate from this cluster. In addition, plot for second and third score vectors can be used together with the plot of first and second score vectors. Additional to score plots, sum of squares of the residual is also used to discriminate simulations. The sum of squares of residual (SSR) is calculated as:

$$Q_i = \sum_{k=1}^{K} \sum_{j=1}^{J} E(i,k,j)^2 \qquad (3.33)$$

Where Q_i is the sum of squares for i^{th} simulation and E is the residual obtained from MPCA decomposition. In Q plot, similar simulations have close Q values whereas different simulations don't have. The criteria for being close to a certain region or a value are determined by the confidence limits. Confidence interval for the t score plot is determined by the equation:

$$\pm t_{n-1,\alpha/2} s_{ref} \left(1 + 1/n\right)^{1/2} \qquad (3.34)$$

Where, n, s_{ref} are the number of observations and the estimated standard deviation of the t-score sample $t_{n-1,\alpha/2}$ is the critical value of the t-distribution with n-1 degrees of freedom at significance level α/2. In order to use this equation it is assumed that elements of score vectors are normally distributed. The confidence interval is calculated for each of score vector by using Equation (3.34) and elliptical contour, whose center is 0 and distance on the principal axes is the confidence limits for score vectors, is drawn. The simulations whose score points lie in this regions is assumed as similar simulations, otherwise they are accepted as different[48, 51].

Control limit for SSR is given as:

$$SSR_{\alpha} = \left(v / 2m\right)\chi^2_{2m^2/v,\alpha} \tag{3.35}$$

Where, m is the sample mean, v denotes the variance of the sample, α is the significance level, $\chi^2_{2m^2/v,\alpha}$ denotes the critical value of the chi-squared variable with $2m^2/v$ degrees of freedom at significance level α[48, 51].

Up to now, discrimination of simulations within a data set is studied. Now, analysis of a simulation which is not inside a data set is presented. It is again questioned whether this given simulation, X_{new}(J,K) , is similar to reference data set or not. First, column mean of reference data set is subtracted from X_{new} . Then, score vectors of X_{new} are predicted using loading matrices of reference data set (obtained from MPCA decomposition). This is achieved by following equation:

$$t_r = \overline{X}_{New} P_r \tag{3.36}$$

Where t_r is the rth predicted score for the new simulation $\overline{X}_{New}$ is the mean centered data, Pr is the loading matrix of reference data set. After obtainingt_r , it is checked whether the plot of first and second predicted scores is in the acceptable region which is defined already by using score vectors of reference data set.

Data for the new simulation is reconstructed as in the Equation (3.37) using these predicted score vectors and loading vectors of reference data set.

$$X_{new,\Pr edicted} = \sum_{r=1}^{R} t_r \otimes P_r + \overline{X}_{reference} \tag{3.37}$$

Where $\overline{X}_{reference}$ is the mean of reference data.

Square prediction error (SPE) is calculated using predicted data and original data by using Equation (3.38).

$$SPE = \sum_{k=1}^{K}\sum_{j=1}^{J}\left(X_{new,\Pr edicted}(j,k) - X_{new}(j,k)\right)^2 \tag{3.38}$$

The calculated SPE is checked whether it is in the acceptable limit or not. Acceptable limit is the confidence limit generated for SSR using reference data set.

If SPE is high (not under the confidence limit) and predicted score is in the acceptable limit, then model is not correct. In other case, if SPE lies inside the control region but score values are not in the confidence interval, then model is correct however, simulation is different from reference simulations. If both SPE and predicted scores are in the confidence interval then this new simulation is similar to reference data set[48].

Now, implementation of MPCA method to our data is presented. It is important to define the variables for the comparison of trajectories. In our simulations, position of data is obtained during the folding. There are several ways of defining the data for trajectory comparison. One way is using the pair distance data. For the 12 beads system, total number of pairs is 66 and distances between pairs are easily calculated using position data. Matrix having 66 columns (each column represents time variation of a pair distance) is formed for each simulation and then this data is used in MPCA method. Choosing pair distance as a comparison method is meaningful because it is rotational invariant and analyzing variation of all pair distances can give better result for the discrimination of trajectories.

Another method is to construct the variable matrix. The matrix can be formed by the following properties of a simulation; root mean square deviation (rmsd), number of contact pairs (if the distance between pairs is less than certain value, i.e. 7Å, then it is assumed that these pairs are in contact), and energy of the protein (can be defined as sum of squares of pair distances). In this case, matrix having three columns is constructed for each simulation. Then MPCA method is applied to this data.

Chapter 4

RESULTS AND DISCUSSION

The model given in section 3.1 is applied to helix and beta sheet secondary structures: The coarse-grained models are used throughout our calculations. 550 optimization runs were made for a helix with 12 residues and 500 runs for a beta sheet containing 12 residues. Using the results of these optimizations the pair potentials are found as described in section 3.2. Then, these pair potentials are used in simulations which start from different random initial conformations. The method given in section 3.3 is used to perform these simulations. In order to analyze whether the trajectories obtained from these simulations are in optimal folding region, the method presented in section 3.4 is applied. Results are presented and analyzed for the helix and the beta sheet in sections 4.1 and 4.2, respectively.

.

4. 1 Results for the Helix

The backbone representation of the helix is given in Figure 10.

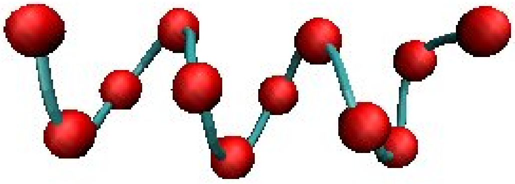


Figure 10 Backbone representation of the helix

4.1.1 Pair Potentials for the Helix

The pair potential is defined as the mean force acting between residues. The pair potentials for the helix are derived using data obtained from optimizations. The mean force vs. pair distance data is obtained for all pairs. In order to simplify the analysis of the potentials easier, these pair potentials are clustered according to the sequence separation of pairs. To enhance clustering, the plots are based on Mean Force vs. $\left\|r_{ij}\right\| - \left\|r_{ij}^{n}\right\|$ (r_{ij}^{n} is the pair distance at native state). The plots for 9 clusters are given in. In these plots the unit of pair distances is Angstrom (Å).

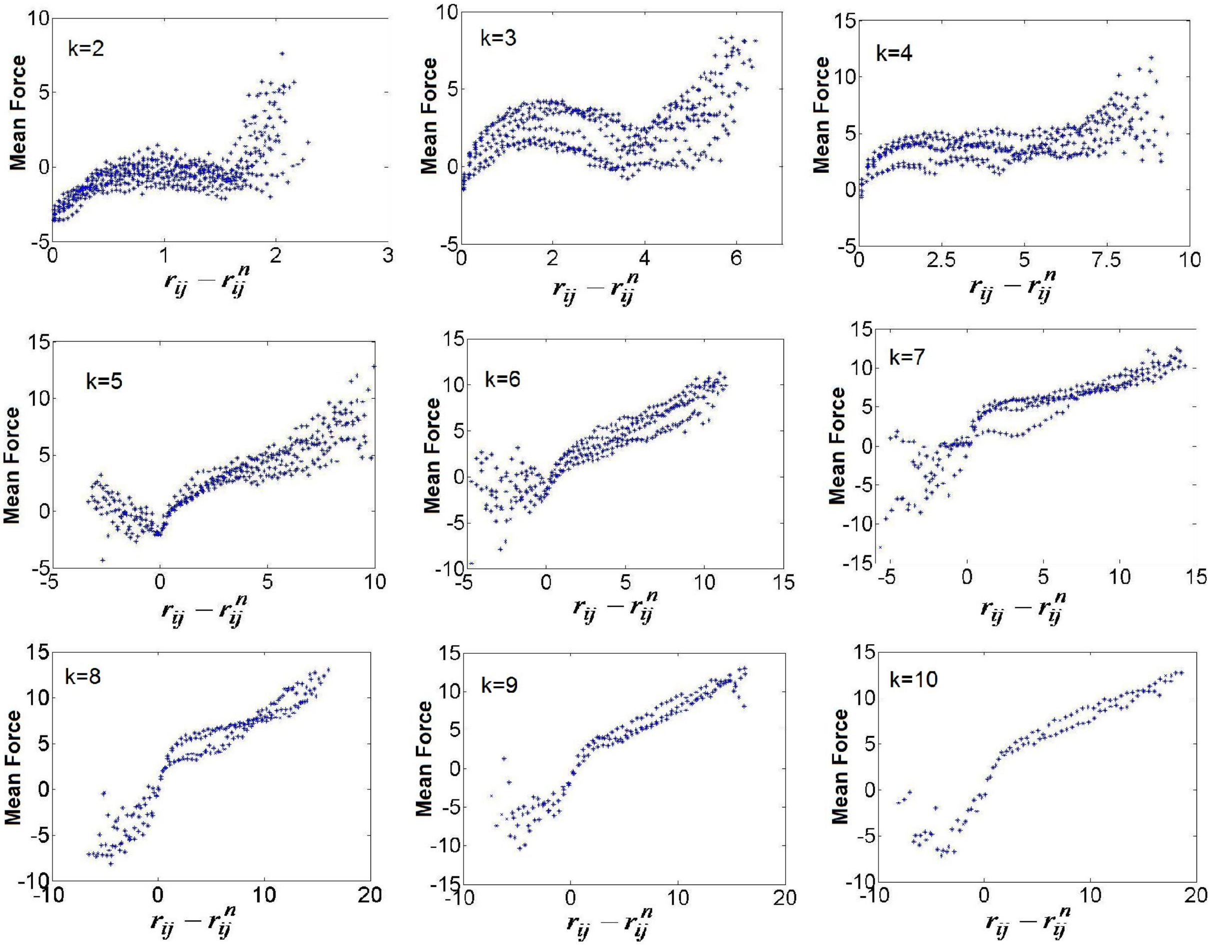


Figure 11 Clusters of mean forces according to sequence separations of pairs. r_{ij} denotes $\left\|r_{ij}\right\|$

From the plots in, it is seen that potentials can be clustered according to sequence separation. That is, pairs whose sequence separations are the same, have similar trends for potentials. In these plots, all potentials are superimposed for pairs which have same sequence separation. Average potential for each cluster is calculated and these average potentials are given in Figure 12 and Figure 13 for pairs having sequence separation *k* = 5-10 and *k* =2-4 respectively. In Figure 12 it is seen that potentials are very similar for pairs having greater sequence separation. On the other hand, the potentials for pairs which are close in sequence don't have similar trends. In fact for helix structure close pairs on sequence are native pairs. These two figures show the important characteristics of pair potentials for folding of helix. It can be said that interactions between native pairs are more specific than long interactions for helix.

In addition, it is observed that potentials have two distinct regions around native state distance for pairs having sequence separation *k* = 5-10 (Figure 12): $\|r_{ij}\| - \|r_{ij}^{n}\| > 0$ and $\|r_{ij}\| - \|r_{ij}^{n}\| < 0$. In all potentials, at large distances potentials have linear regimes. This is the first stage of folding. After coming closer to certain distances, interactions become more specific to residue pairs. In this part of folding, pair potentials don't have linear behavior. This structure of potentials can be better interpreted using movies of these optimizations. In these movies, it is seen that in the first part of folding residues come closer and then ordering of residues occur at short distances between pairs. In the region $\|r_{ij}\| - \|r_{ij}^{n}\| < 0$ data are more scattered. In this region Mean Force can be either attractive or repulsive. Repulsive forces may be due to the fact that optimization should repel pairs to bring them to their native distances when the distance between these pairs smaller than native state value. Moreover, in some cases optimization may have to bring pairs closer temporarily in order to allow some other residues to rearrange themselves. In such cases, mean force would be attractive.

In Figure 13 it is seen that pair distances don't become less than native value. This is explained as follows. In the optimization the excluded volume limits for native pairs are set equal to their native state distances. Since optimization satisfies excluded volume constraints, pair distances don't be less than native values. In general the mean forces for native pairs are attractive except for the pairs having sequence separation *k*=2.

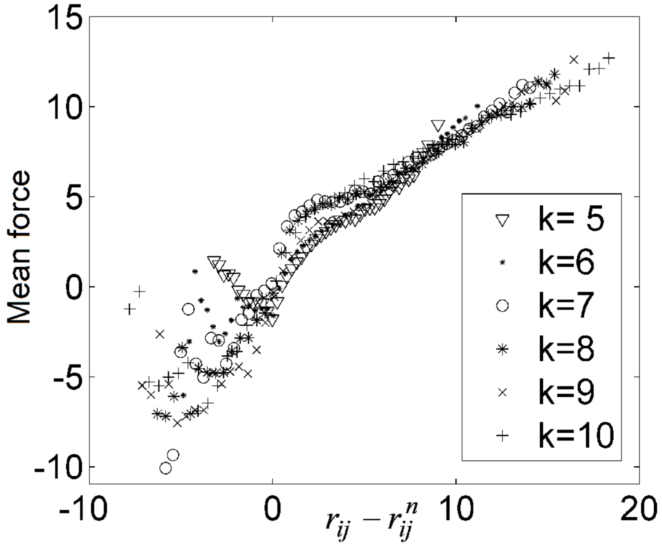


Figure 12 Average mean forces for clusters having sequence separation 5-10, r_{ij} denotes $\left\|r_{ij}\right\|$

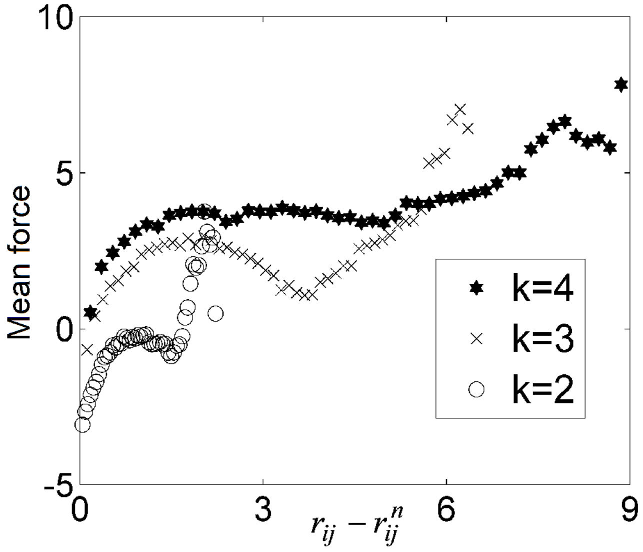


Figure 13 Average mean forces for clusters having sequence separation 2, 3, and 4, r_{ij} denotes $\left\|r_{ij}\right\|$

4.1.2 Folding Test Simulations of the Helix Using Obtained Pair Potentials

The derived pair potentials were used in folding simulations for helix by applying the method described in section 3.3. The functional forms of these pair potentials for helix are given in Appendix A.1. For 100 new initial conditions, which were randomly generated, folding simulations were performed. In Figure 14 RMSD distribution is given for these test simulations. It is observed that some of simulations fold to left-handed helix. RMSD distribution is corrected by taking two reference structures for RMSD calculation: right-handed and left-handed helix. RMSD of a simulation is calculated for both reference structures and the one with the lower value is accepted. The reason why some simulations fold to left-handed helix is due to the fact that distances between pairs in native state are the same for right-handed and left-handed helices. Since our potential tries to bring residue pairs to their native state values, some initial conditions simulations may converge to left-handed helix structure. In addition, Figure 14 shows that these pair potentials successfully fold the helix.

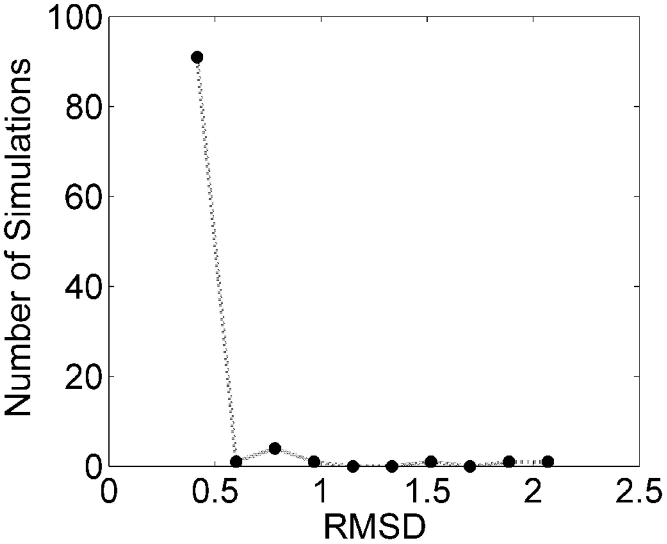


Figure 14 RMSD distribution of test simulations

4.1.3 Trajectory Analysis of Test Folding Simulations

The trajectories obtained from test folding simulations for 100 random initial conditions are analyzed using the method described in section 3.4. Pair distance data of the simulations are used to make this analysis. It is investigated whether the trajectories of these test simulations belong to the family of similar trajectories that are obtained from optimization simulations. Figure 15 shows *t* score plots for optimization simulations and test simulations with 99% confidence limit. In this figure, it is observed that most of *t* scores are in control region. This can show that the trajectories of test simulations are in optimal folding region. However, *t* score plot is not sufficient by itself to make this conclusion. Square prediction error (SPE) plot should support the same result. Figure 16 shows SPE plots for optimization and test simulations. The SPE values for most of the test simulations lie in the confidence interval. As a result, by looking at Figure 15 and Figure 16 together, it can be said that acceptable folding pathways for helix can be achieved with the derived pair potentials.

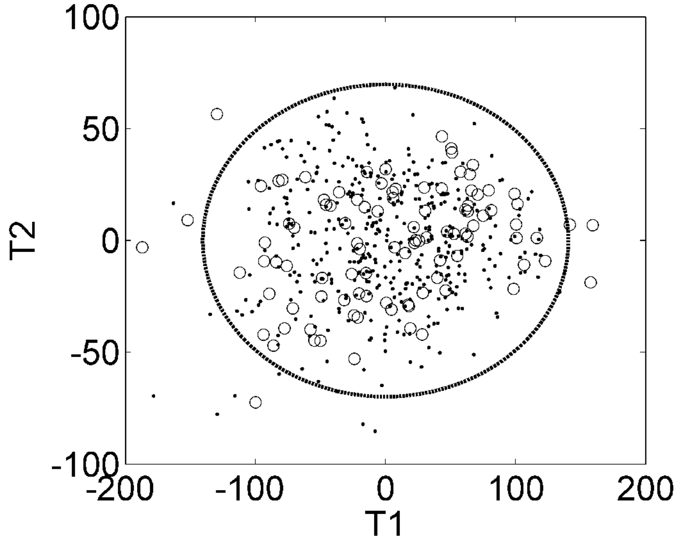


Figure 15 *t* score plots for optimization simulations and test simulations. Dot: Optimization simulations, Circle: Test simulations, elliptical contour: control region with 99% confidence limit.

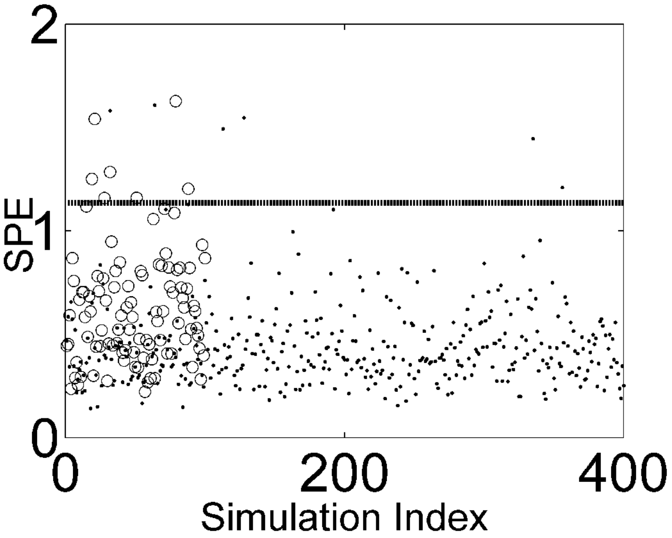


Figure 16 Square prediction error (SPE) plot for optimization simulations and test simulations. Dot: Optimization simulations, Circle: Test simulations, line: control region with 99% confidence limit

In order to check whether this trajectory comparison method is reliable, a negative test is performed. These test simulations were performed as follows. From the mean force for each pair, $g_{ij}(\|r_{ij}\|)$, the component of in *ij*- direction only is extracted. It is achieved by taking the terms *k=i* in the first summation and *k=j* in the second summation in Equation 3.25. As a result of this operation, coefficients are obtained as:

$$c_{ij} = c_{ji} = \frac{1}{2} g_{ij}(\|r_{ij}\|) .$$

These coefficients are next used in the Equation 3.30 in order to make folding simulations. These simulations fold to helix to certain degree (the mean RMSD value of these test simulations is 2.2 Å). Figure 17 and Figure 18 show *t* score plots and SPE values for both these simulations and optimization simulations. Neither *t* scores nor SPE values of these simulations are in control region. So, the trajectories of these simulations are not similar to those of optimization simulations. This test illustrates that this method can discriminate the trajectories.

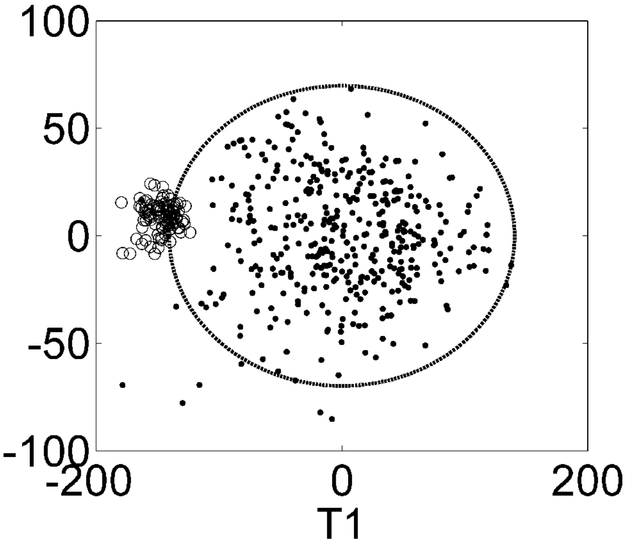


Figure 17 *t* score plots for optimization simulations and negative test simulations. Dots represent optimization simulations, circles are for negative test simulations. The elliptical contour shows the control region with 99% confidence limit

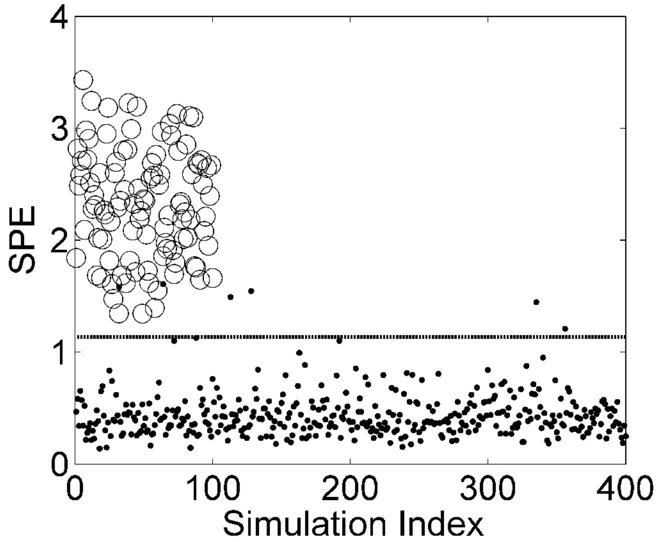


Figure 18 Square prediction error plot optimization simulations and negative test simulations. Dots show the optimization simulations, circles denote the negative test simulations. The line shows the control region with 99% confidence limit

4.1.4 Force analysis at equilibrium

It is analyzed that whether forces acting on beads at equilibrium obtained from test folding simulations converge to theoretical values. The theoretical values for the forces acting on the residues at equilibrium can be obtained by setting $\frac{dr}{dt} = 0$ in the dynamic equation $\frac{dr}{dt} = -\Gamma r + u$. Then theoretical residue forces at equilibrium are obtained as $u^{N} = \Gamma r^{N}$. This can be easily calculated since the native state positions of residues are assumed to be available. After incorporating the pair forces the dynamic model is given by Equation (3.30) $\frac{dr}{dt} = -\Gamma r + V\boldsymbol{c}$ where the residue forces were reconstructed by the Equation:

$$u = V\boldsymbol{c} \tag{4.1}$$

Figure 19 shows the theoretical forces and reconstructed forces for one simulation. It is clearly seen that they are very similar.

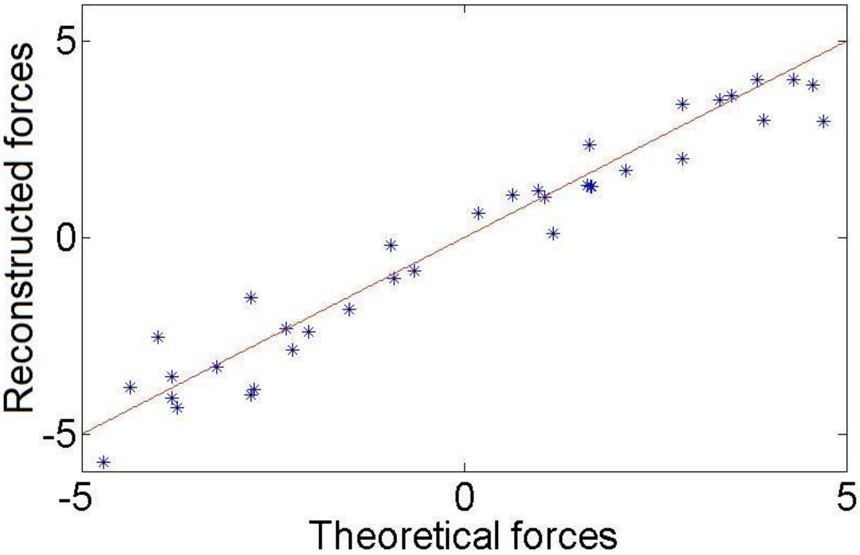


Figure 19 Theoretical forces vs. calculated forces for one simulation. x,y and z components of forces are plotted. red line: y=x

The equilibrium force analysis can be made in another way as following. The dynamic model used in folding test simulations is $\frac{dr}{dt} = -\Gamma r + V\boldsymbol{c}$. Left-hand side of this equation becomes zero in the equilibrium. Theoretical solution of coefficients ($\boldsymbol{c}$) at equilibrium can be obtained by minimum norm solution. In folding simulations which are performed by using obtained pair potentials the coefficients are calculated (see section 3.3). Here it is checked that whether the calculated coefficients are the same or similar to the theoretical coefficients. Figure 20 shows theoretical coefficients and calculated coefficients for one simulation. It is seen that they are highly correlated.

It should be noted that these calculated coefficients and reconstructed forces don’t have to be same with the theoretical ones. In fact the theoretical and calculated values are not similar for all test simulations. They are similar for the simulations which have better RMSD values. As a conclusion, calculated coefficients and reconstructed forces at equilibrium may converge to theoretical values in our algorithm.

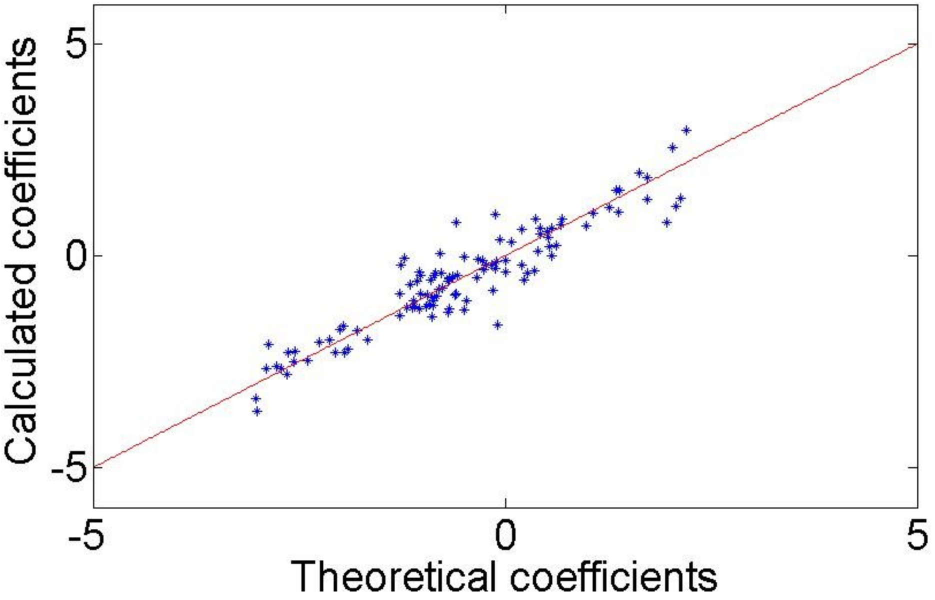


Figure 20 Theoretical coefficients vs.calculated coefficients vs. for one simulation, red line: y=x

4.1.5 Which Pairs Are More Important for Folding of the Helix?

Now we want to identify the critical pairs for folding. Are the interactions of all pairs needed to fold a helix? In section 4.1.1 it is shown that potentials of short pairs are more specific compared to the potentials of long pairs. In order to see the effect of pairs, simulations are performed by using potentials of certain pairs. First only potentials of pairs having sequence separation two are used. Then potential of pairs having sequence separation 3 is added and simulations are performed. For the same initial conditions simulations are repeated by increasing the number of pairs used. Mean RMSD values of these simulations are given in Figure 21 . It is seen that addition of the pairs having sequence separation greater than 6 doesn’t improve the folding much. On the other hand, the effect of pairs having sequence separation up to 6 is seen very clearly in Figure 21. It can be concluded that the interactions for short pairs are more effective and important for folding of a helix than interactions of long pairs.

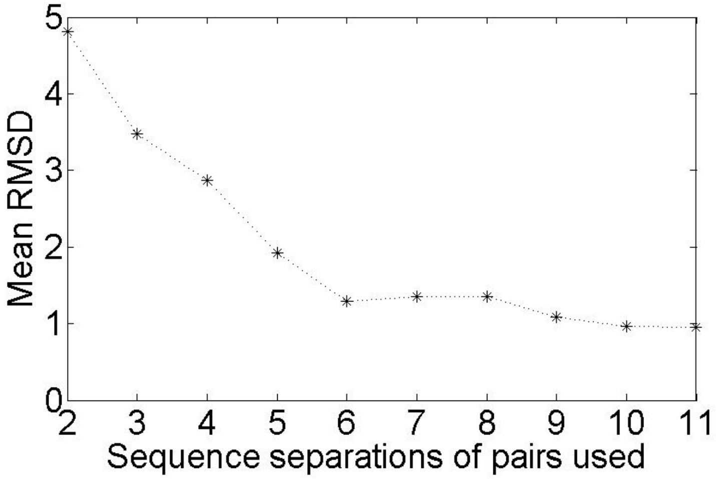


Figure 21 Mean Rmsd plots of folding simulations performed by using potentials of certain pairs

In order to interpret Figure 21 better, final conformations of the simulations using increasing number of pairwise interactions are shown in Figure 22 . These simulations were performed starting from the same initial conformation. When only the native pairs with relatively small separation (*k=2-4*) are used, the conformation is very compact. With the addition of other pairs with separation up to 7, the final conformation is much closer to the native state.

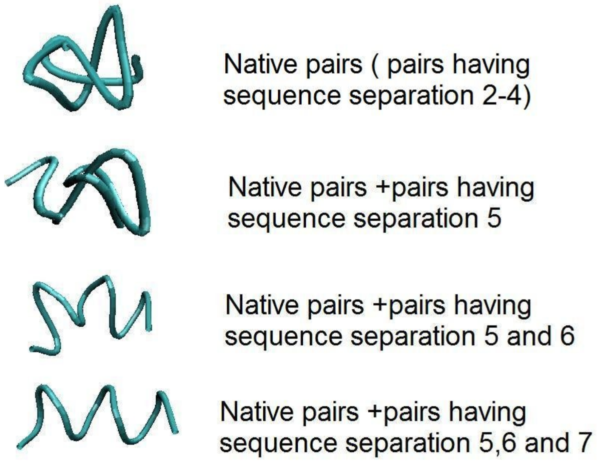

Native pairs (pairs having sequence separation 2-4)
Native pairs +pairs having sequence separation 5
Native pairs +pairs having sequence separation 5 and 6
Native pairs +pairs having sequence separation 5,6 and 7

Figure 22 Evolution of the final fold using different pair interactions

We reconstructed the total force acting each bead by its pairwise components and we used these pairwise forces in dynamic model (3.30). For each residue there are 10 non-bonded pairwise components. The contribution of these components for each residue is analyzed. The magnitudes of these pairwise forces were calculated for each residue for each time step using data of 100 test simulations. Then, for each residue the number of occurrence of its pairwise interactions with the three highest magnitudes was recorded. The data are clustered according to sequence separation of pairwise components. Figure 23 shows the frequency plot of these pairwise forces. The trend is similar to Figure 21. Pairs having sequence separation *k*=2-6 have the most significant contribution to the force field.

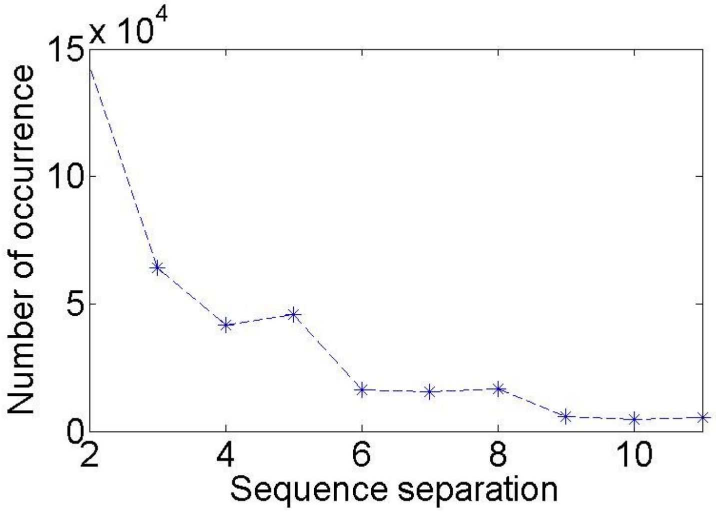
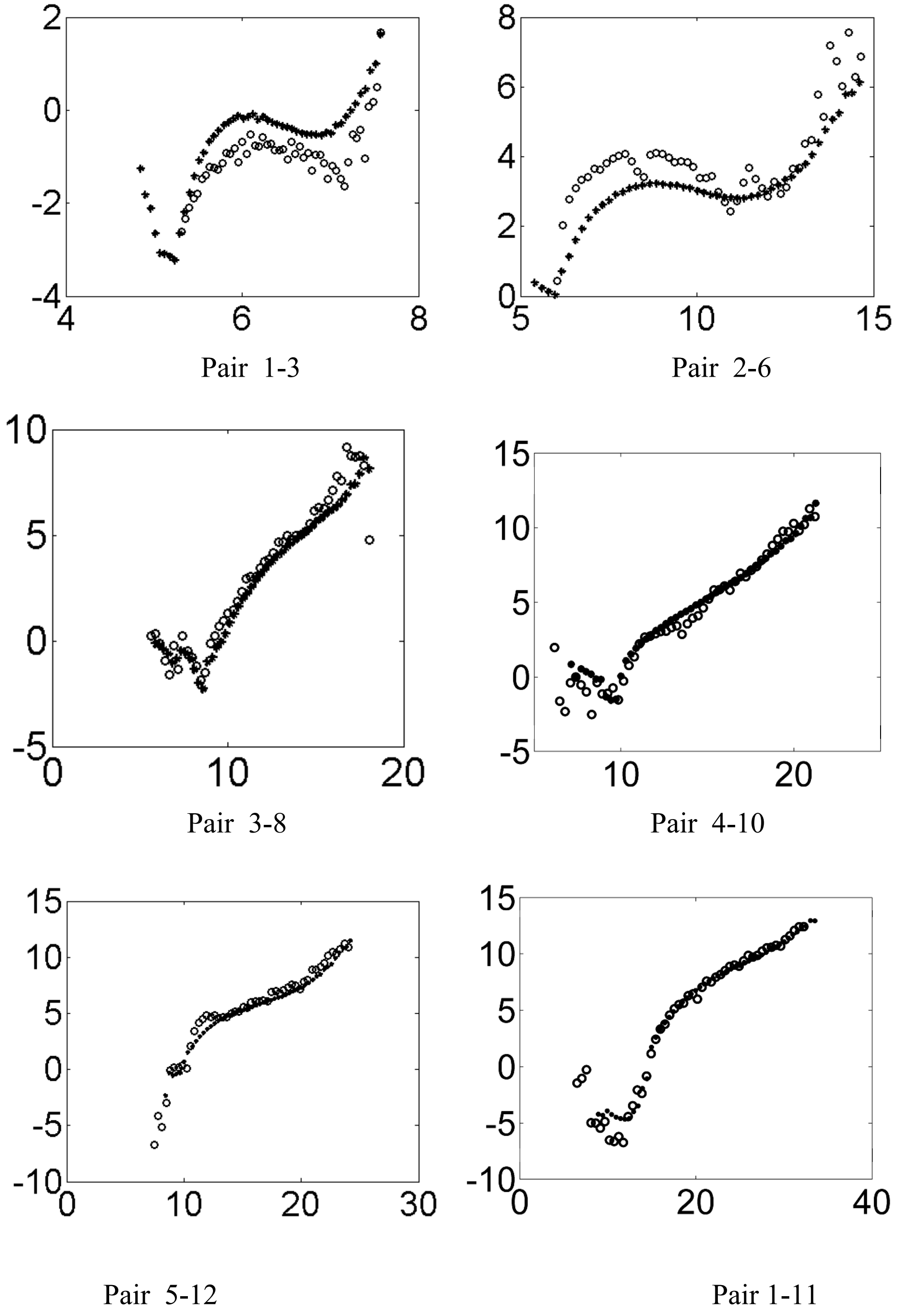


Figure 23 Frequency plot of pairwise components with maximum interactions

4.1.6 Analysis of Mean Potential Recovery in Folding Simulations

Here, we want to find out how much of the mean potentials are recovered in folding simulations. Since, as mentioned in section 3.3, there is a rank deficiency of the system, mean potentials are not recovered exactly. However folding simulations give very good results. One way of making this analysis is to calculate the pair potentials using the data acquired from the simulations. So, folding simulations are performed for 200 random initial conditions by using derived pair potentials. Results (trajectory and force data) of these simulations are used to calculate pair potentials. For several pairs the mean potential derived from optimization data and the recovered mean potential calculated from simulations are given in Figure 24. These plots show that pair potentials are recovered very well for long pairs. For short pairs potentials are not recovered as good as in the case of long pairs. However, for short pairs this result is acceptable since the difference is not much and general trend of potential is conserved.

However in simulations excluded volume constraints may be violated due to errors in force reconstruction. In our test simulations, it was seen that maximum average violation of only 10% in some of the pairs. Excluded volume violation is unavoidable since there are always errors due to averaging.



Pair 1-3

Pair 2-6

Pair 3-8

Pair 4-10

Pair 5-12

Pair 1-11

Figure 24 Pair potential plots for derived from both optimization data and simulation data for several pairs, *: derived from simulations, o : derived from optimizations, x-axis: Pair distance, y-axis : Mean force

4. 2 Results for the Beta Sheet

Beads 1 through 6 and beads 7 through 12 form the two strands of the beta sheet, respectively as shown below (Figure 25).

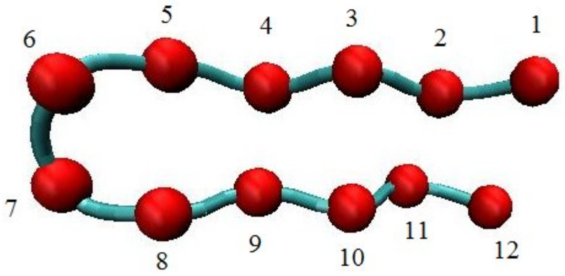


Figure 25 Backbone representation of the beta sheet

4.2.1 Pair Potentials for the Beta Sheet

For a beta sheet, clustering of forces according to sequence separation does not lend itself to any meaningful result as in the case of helix. This is because potentials for residues on the same strand and for residues on the opposite strands may not have the same characteristics. For instance, pair 7-11 (both residues on the same strand) and pair 5-9 (residues on different strands) behave differently despite having the same sequence separation *k=4*. This is illustrated in Figure 26. The mean force for the pair 5-9 is mostly attractive while the opposite is true for the pair 7-11.

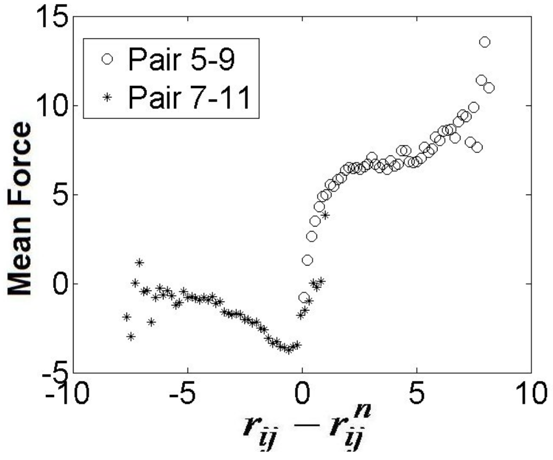


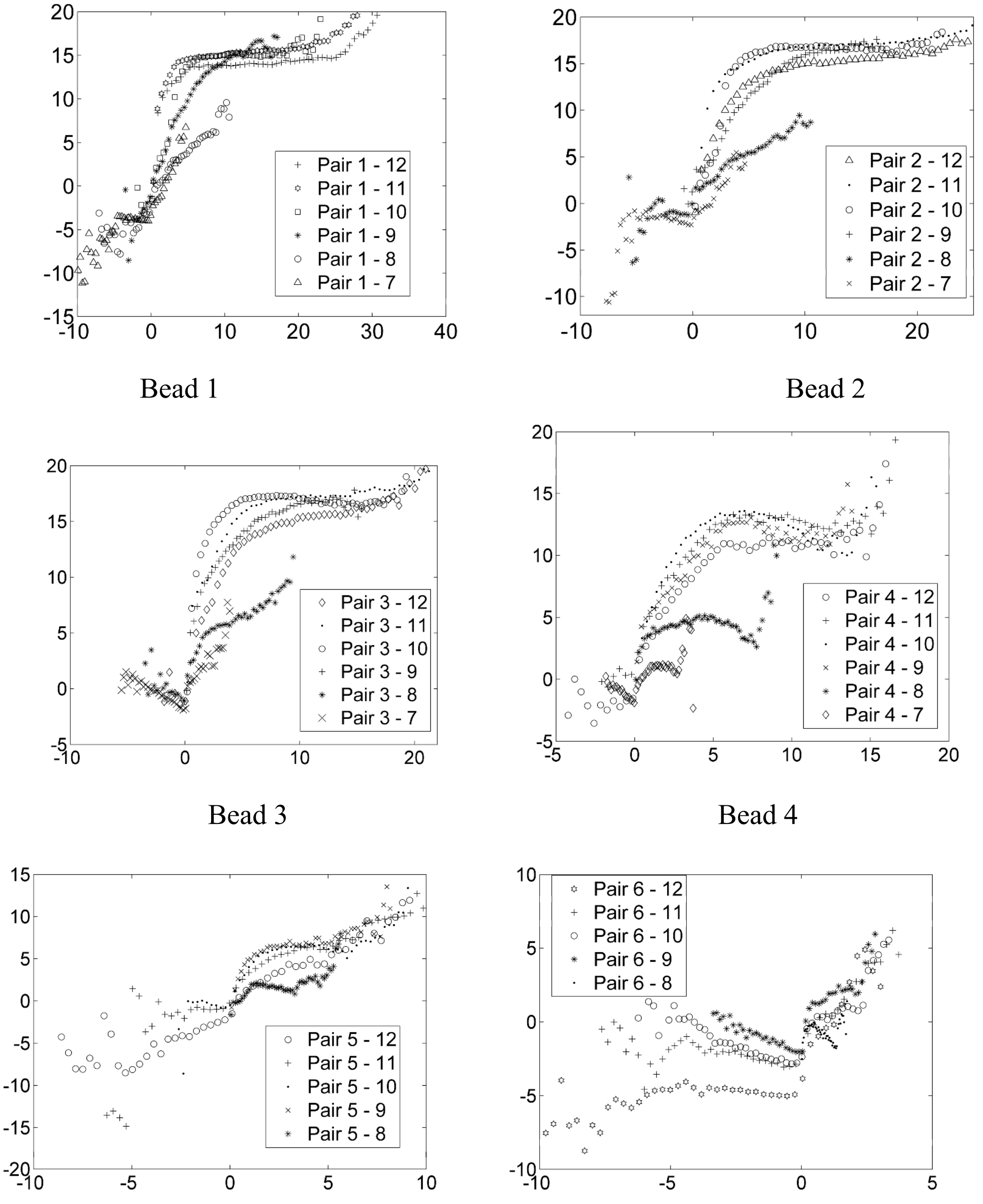
Figure 26 Mean Forces of pairs 7-11 and 5-9. r_{ij} denotes $\|r_{ij}\|$

Due to this difference in behavior, we classify the pairs into two groups: pairs whose residues are on the different strands; and pairs whose residues are on the same strand. First interactions acting across the two strands are analyzed. Figure 27 shows the mean forces for the interactions between residues 1-6 on one strand and residues 7-12 on the opposite strand. In the first three plots, all six mean forces for the interactions between residues (1, 2, and 3) and (7-12) have the same trends. As sequence separation increases, the attractive mean forces have relatively steady forces for a wider range of $\|r_{ij}\| - \|r_{ij}^n\|$. This situation is effective for the interactions among beads with maximum sequence separation, at the two ends of the strands away from the turning point, i.e. between (1,2,3) and (9, 10, 11 and 12). Mean forces acting between these residues are as high as possible to bring them to their native state separation.

Interactions of beads (1, 2, 3) with the other two residues 7, 8 are of the same nature but last for a much smaller range of $\|r_{ij}\| - \|r_{ij}^n\|$(no flat region is observed). They are also more repulsive and smaller in magnitude. This is due to the fact that the sequence separations between (1-3) and (7-8) are smaller than (1-3) and (9-12).

In addition, interactions across the two strands become more specific as one approaches the turning point. We see this change in the characteristics of interactions between the rest of the residues (4, 5, and 6) and the residues on the opposite strand (7-12). As seen in the fourth plot in Figure 27, residue number 4 is in transition, bearing similarities to the residues (1, 2, and 3) but at the same time having its own distinct behavior. We see more of this distinction in the interactions of residues 5 and 6. The magnitudes of interactions of (4, 5, and 6) are smaller and they are more repulsive than those of (1, 2, and 3). In general it can be concluded the further away the residues are from the turning point, the more similar their mean interaction forces across the beta strands are.

Finally the mean forces among the residues that are located on the same strand are analyzed. Figure 28 shows these mean forces. These forces are mostly repulsive since these residues repel each other in order to stretch and form the strand structure.



Bead 1

Bead 2

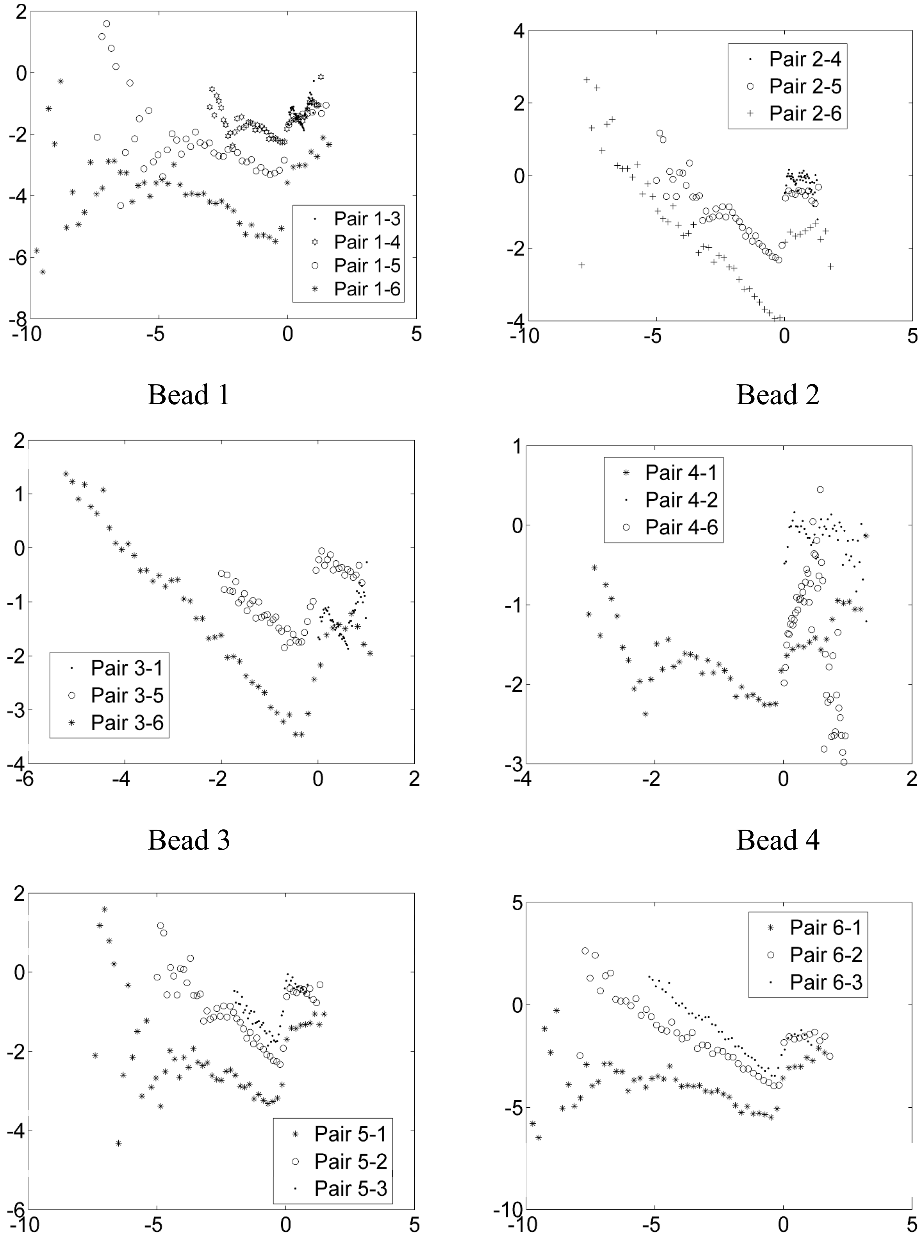
Bead 3

Bead 4

Bead 5

Bead 6

Figure 27 Mean forces for pairs whose residues are on different strands, x-axis: pair distance, y-axis : Mean force



Bead 1

Bead 2

Bead 3

Bead 4

Bead 5

Bead 6

Figure 28 Mean forces for pairs whose residues are on the same strands, x-axis: $\left\|r_{ij}\right\|-\left\|r_{ij}^{n}\right\|$, y-axis : Mean force

4.2.2 Folding Test Simulations of the Beta Sheet

Derived pair potentials for the Beta sheet is used in folding simulations as described in section 3.3. The functional forms of pair potentials are given in Appendix A.2. For 200 random different initial conformations, the folding simulations were performed. Figure 29 shows RMSD distribution of these simulations. The mean RMSD of these simulations is 1.42 Å. Although, this RMSD value is not as good as in the case of helix, it is still a good result for folding.

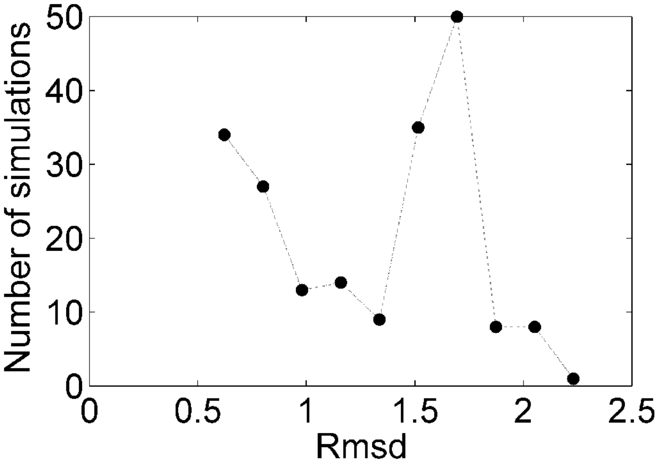


Figure 29 The RMSD distribution of Test Simulations

4.2.3 Trajectory Analysis of the Test Simulations

t score plots and SPE plot for these test simulations are given in Figure 30 and Figure 31 respectively. Both *t* score and SPE points of most of the test simulations lie in the same region with those of optimization simulations. So, it can be concluded that the trajectories of these test simulations are similar to those of optimizations.

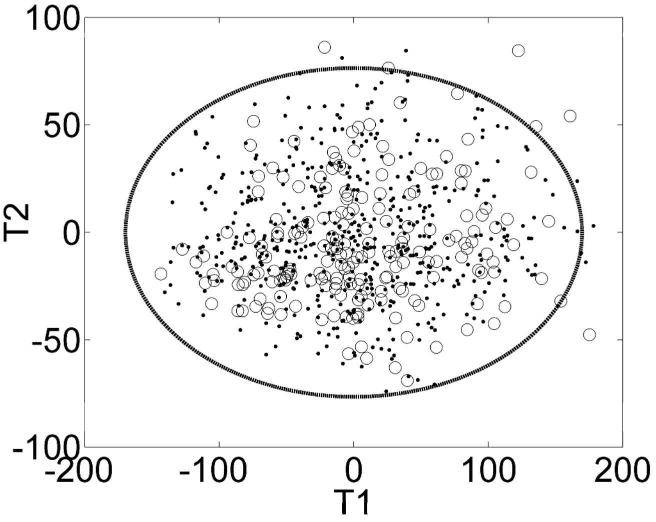


Figure 30 *t* score plots for optimization simulations and test simulations. Dots represent the optimization simulations, circles show test simulations, the elliptical contour shows the 99% confidence limit

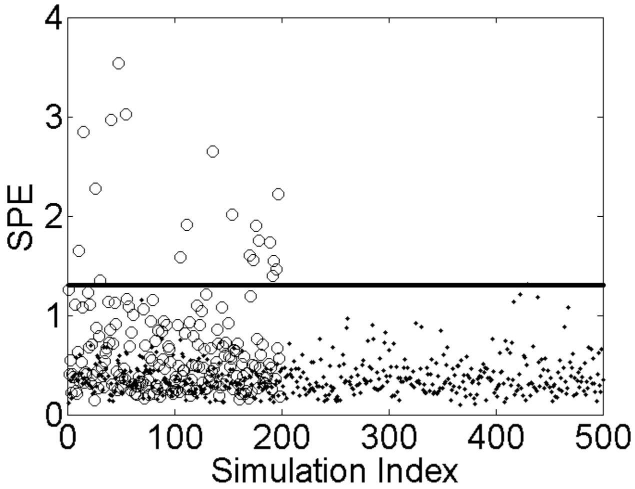


Figure 31 Square prediction error (SPE) plot for optimization simulations and test simulations. See legend for figure 30.

4.2.4 Force Analysis at equilibrium

Theoretical values of forces and coefficients at equilibrium are calculated as in section 4.1.4. Figure 32 and Figure 33 shows theoretical and calculated values of both forces and coefficients for one simulation. It is seen that theoretical and calculated coefficients are close to each other.

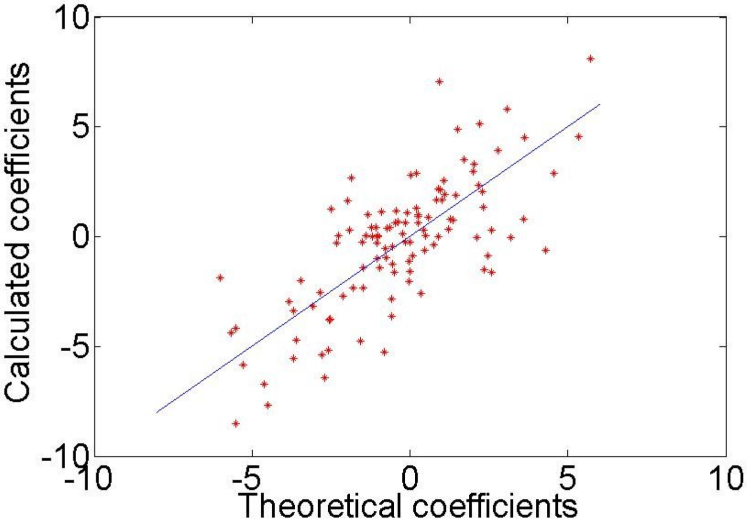


Figure 32 Theoretical coefficients vs. calculated coefficients vs. for one simulation, blue line: y=x

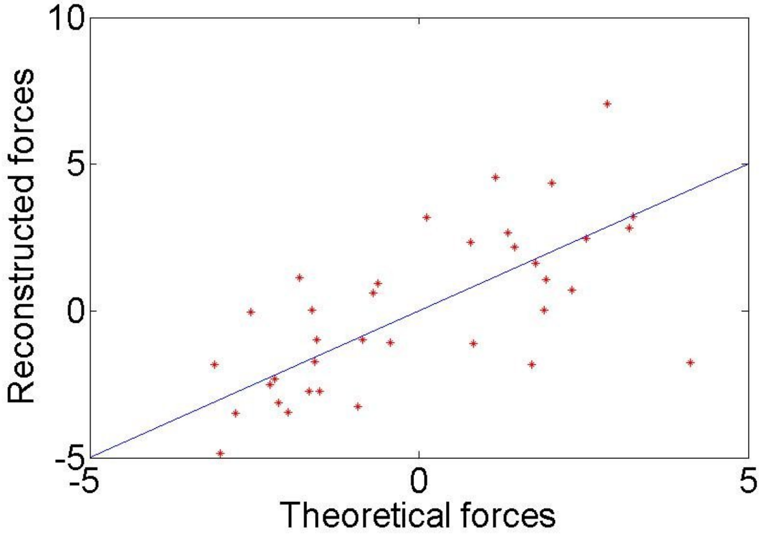


Figure 33 Theoretical forces vs. Reconstructed forces for one simulation, blue line: y=x

4.2.5 Important Pair interactions for folding of Beta Sheet

The analysis of effect of the pairs for folding of beta sheet is different than that of helix. In helix, pairs are clustered according to sequence separation and it is shown that the effects of short pairs are more important. However, for beta sheet it is not useful to cluster pairs according to sequence separation. At first the effect of native pairs is analyzed. In order to do this, a folding simulation was performed by using only the interaction of native pairs. In this simulation 2.06 RMSD was achieved. The structure of turning point couldn't be achieved well by using native pairs only. So some pair interactions are needed in order to acquire stretched strands (Figure 35). Then for the same initial conformation another folding simulation is performed by adding two pairs (1-6 and 7-12) to native pairs. Two pairs (2-6 and 7-11) which can result stretching of strands are added and simulation was performed again. As the pairs on the same strand are added better RMSD value is obtained. When all pairs are used 0.68 RMSD is obtained (Figure 34). As a result it can be concluded that native pairs can bring two strands closer effectively. However structure of turning point and structure of strands are not adequately obtained using only native pairs. Other pairs are needed to make folding better.

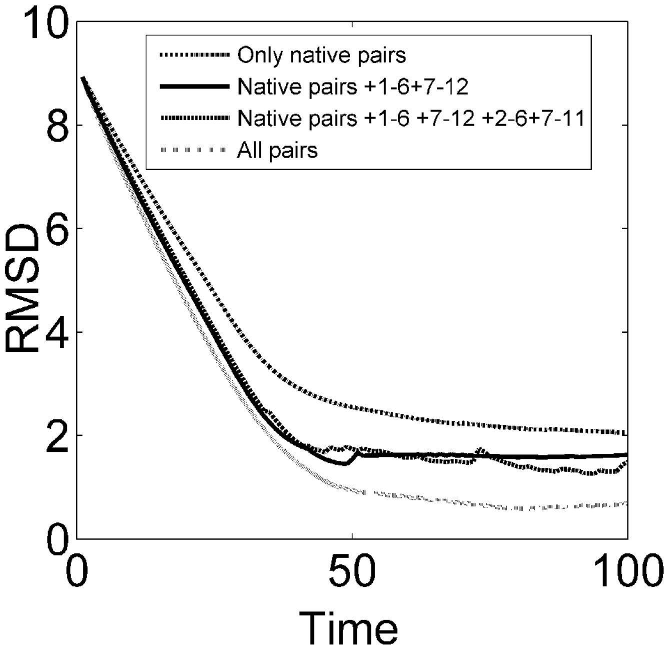


Figure 34 RMSD plots of several simulations

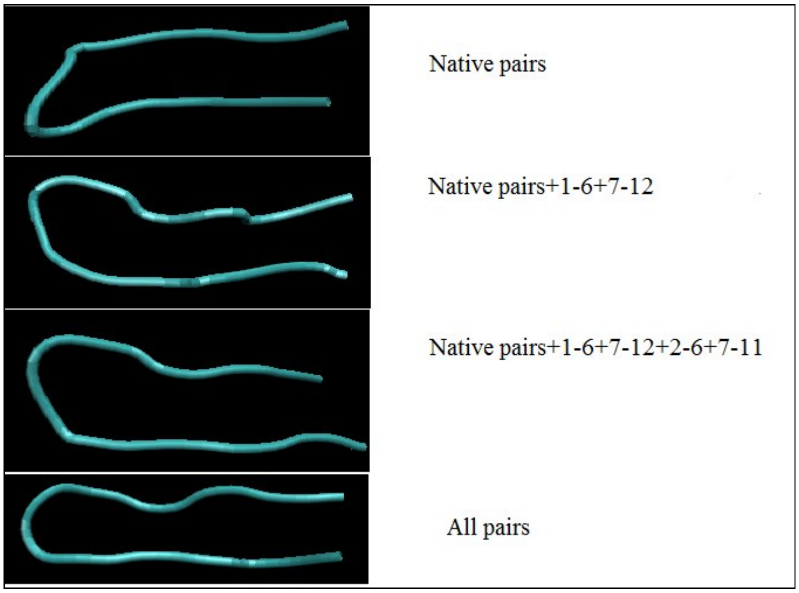
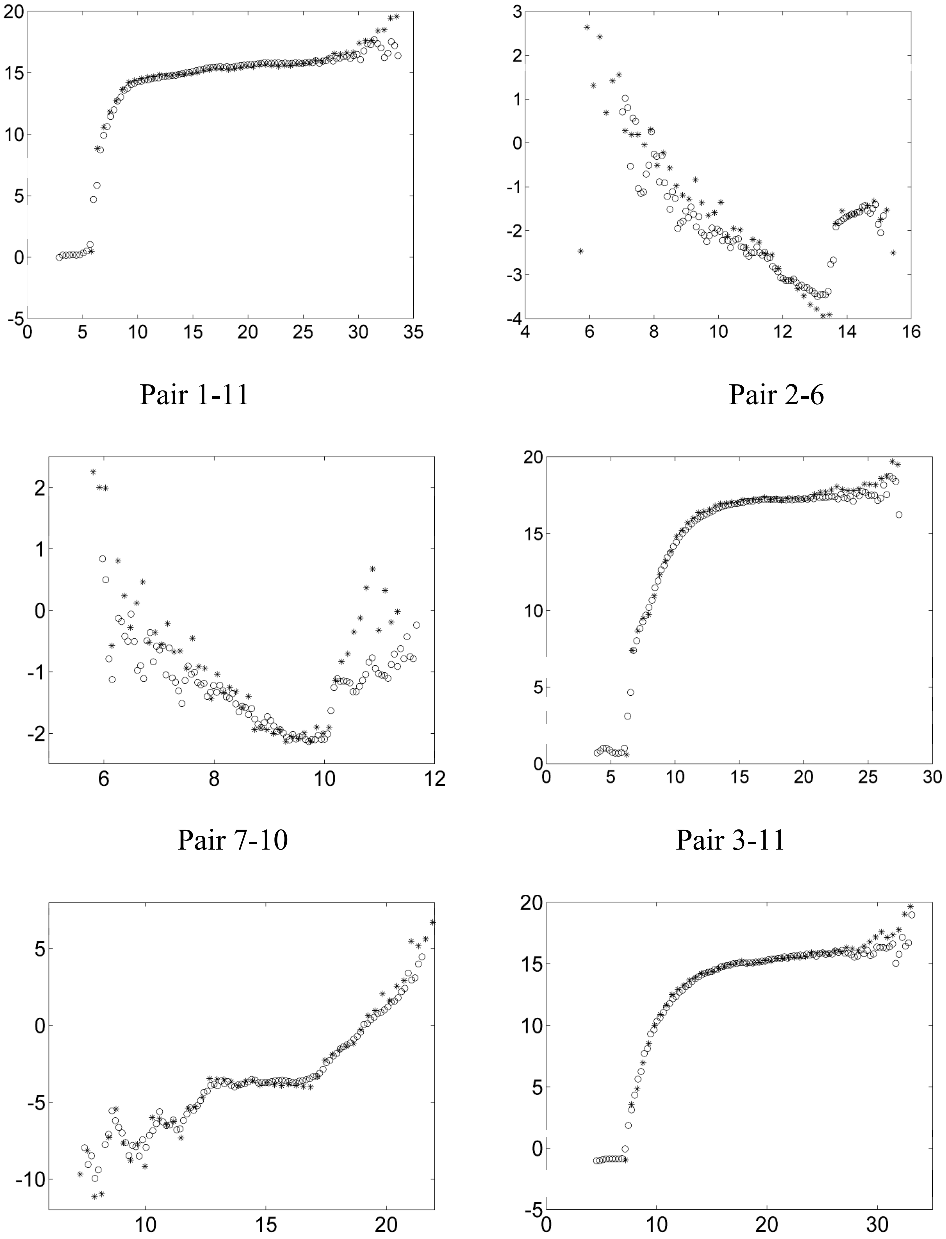


Figure 35 Final structures of several simulations

4.2.6 Analysis of Mean Force Recovery

It is questioned whether mean forces are recovered as well as helix case. It is done as in section 4.1.6. Starting from 500 random different initial conformations folding simulations were performed the pair potentials obtained from optimization data. Results (trajectory and force data) of these simulations are used to calculate pair potentials. For several pairs the mean potential derived from optimization data and the recovered mean potential calculated from simulations are given in Figure 36. These plots show that pair potentials are recovered very well in folding simulations of the beta sheet.



Pair 1-11

Pair 2-6

Pair 7-10

Pair 3-11

Pair 1-7

Pair 2-12

Figure 36 Mean force plots for derived from both optimization data and simulation data for several pairs, o: derived from simulations, * : derived from optimization, x-axis: Pair distance, y-axis : Mean force

Chapter 5

CONCLUSION

Study of pair potentials for folding is critical for understanding the process of protein folding. In this thesis, we obtained pair potentials for optimal folding of secondary structures (helix and beta sheet). The pair potential is defined as mean force for pairwise interaction. A method is developed to derive the mean forces from folding pathways. In order to obtain optimum folding pathways a dynamic optimization method developed by Guner et al. is used. In this method amino acids are represented by beads. The dynamic model is based on the Newton's equations of motion. Folding is achieved by bringing pairs to their native state distances. Deviation of pair distances from their native state values is minimized during folding. Excluded volumes and bond lengths are implemented as constraints. Forces acting on each bead and position of beads are obtained using this optimization method. Using this force and position data, obtained from many optimizations, pairwise mean forces are derived.

In this study, optimal folding paths for a generic helix and beta sheet, both containing 12 residues, are obtained for many different random initial conformations. Pairwise mean forces are derived using the optimization data. A method for using these pair potentials in dynamic folding simulation is presented. Starting from many different random initial conformations, the folding test simulations were performed using derived potentials. In these simulations, helix and beta sheet are successfully folded with mean RMSD values, 0.44 and 1.42 Å respectively. Multi-way Principle Component Analysis (MPCA) method is implemented to analyze whether the trajectories of these test simulations are similar to those of optimizations. Pair distance data of trajectories are used for this analysis. It was shown that the trajectories of these test simulations are similar to those obtained from optimizations.

We have shown the important characteristics of these pair potentials. For example, in helix contribution of the pairs having long sequence separation (k > 7) to the folding is relatively insignificant. For beta sheet, interactions between native contact pairs can bring the two strands closer. However, additional some non-native interactions are needed to shape these strands better.

In addition, the pair potentials we obtained are distance-dependent and sequence-specific but not residue-specific. So, ab initio folding is not possible with these potentials but optimal folding pathways can be obtained by knowing the native state structure.

Our pair potential derivation method is not restricted to secondary structures. It can be applied to any dynamic folding pathway. So, potentials for tertiary interactions can be derived using folding pathways for a complete protein structure. However, many optimization runs for folding of a tertiary structure take too much time. This time problem can be overcome as follows: Secondary structures are assumed to be formed beforehand and afterwards optimization can be performed to fold them to compact native structure. Optimization model can be modified to pack these secondary structures optimally. Since pairwise mean forces for secondary structures were developed, only mean forces for tertiary interactions should be determined.

As a future work, a hierarchical method to combine and coordinate the mean forces for secondary and tertiary interactions can be studied. If that can be achieved, near-optimal pathways for folding of a protein will be acquired without performing an optimization simulation. Since folding simulations with these derived pair potentials are much faster than optimizations simulations, rich data for conformational sampling of optimal pathways can be obtained.

APPENDIX

A.1 Pair Potentials for Helix

For helix structure curve fitting is made for two regions of pair potentials; pair distance greater than native value and pair distance smaller then native value. This is done because potentials have different structure in these regions. Linear fitting is made when pair distance is smaller than that of native value and its functional form is :

$$g_{ij} = mr_{ij} + n \tag{A.1}$$

Table 1 Coefficients for the function of pair potential for the distance smaller than $r_{ij} - r_{ij}^{n}$ (for pairs having sequence separation greater than 5)

BEAD	BEAD	m	n
1	6	-2.644	16.33
1	7	-0.332	-2.046
1	8	-0.228	0.712
1	9	1.216	-15.282
1	10	0.928	-19.488
1	11	-0.116	-6.566
1	12	1.428	-22.598
2	7	-0.52	0.868
2	8	-1.194	10.526
2	9	1.11	-11.546

2	10	1.778	-26.344
2	11	-0.084	-6.912
2	12	1.44	-24.382
3	8	-0.93	5.258
3	9	1.954	-21.552
3	10	2.414	-33.568
3	11	0.446	-12.176
3	12	2.554	-42.38
4	9	-2.554	18.888
4	10	-0.596	3.062
4	11	2.292	-20.664
4	12	2.996	-33.412
5	10	-2.816	20.454
5	11	-3.618	33.034
5	12	6.206	-58.348
6	11	-1.578	10.966
6	12	0.42	-10.524
7	12	-2.178	13.798

For pair distance greater than that of native state value third order polynomial fitting is made and its coefficients are given. The function for this fitting is given as:

$$g_{ij} = kr_{ij}^{3} + lr_{ij}^{2} + mr_{ij} + n \tag{A.2}$$

Table 2 Coefficients for the function of pair potential for the distance greater than $r_{ij} - r_{ij}^{n}$ (for pairs having sequence separation greater than 5)

BEAD	BEAD	k	l	m	n
1	6	0.05	-2.02	27.474	-119.5
1	7	0.034	-1.542	24.872	-129.586
1	8	0.02	-0.992	17.248	-87.994
1	9	0.022	-1.268	24.48	-143.61
1	10	-0.006	0.31	-4.458	18.044
1	11	0.006	-0.434	11.994	-97.332
1	12	0.008	-0.624	16.058	-123.15
2	7	0.036	-1.45	20.322	-92.702
2	8	0.024	-1.154	19.65	-104.068
2	9	0.02	-1	17.054	-85.634
2	10	-0.004	0.292	-5.234	32.274
2	11	0	-0.074	4.022	-41.156
2	12	0.006	-0.42	11.27	-91.168
3	8	-0.008	0.188	1.056	-19.56
3	9	0.026	-1.216	20.06	-108.152
3	10	-0.012	0.81	-14.322	80.328
3	11	-0.006	0.414	-7.156	41.896
3	12	0.006	-0.344	8.714	-70.246
4	9	0.074	-2.878	37.612	-156.756
4	10	0.004	-0.17	4.048	-26.334
4	11	0.012	-0.616	11.174	-56.158
4	12	0.014	-0.786	15.43	-89.41
5	10	0.034	-1.338	19.306	-91.456
5	11	0.014	-0.756	14.556	-82.7

5	12	0.014	-0.694	12.348	-64.266
6	11	0.052	-2.144	30.448	-140.014
6	12	0.028	-1.31	21.18	-112.434
7	12	0.042	-1.7	23.302	-103.386

Table 3 Coefficients for the function of pair potential for the distance greater than $r_{ij} - r_{ij}^{n}$ (for pairs having sequence separation less than 5)

BEAD	BEAD	k	l	m	n
1	3	3.974	-76.898	494.076	-1055.82
1	4	0.322	-7.946	64.008	-166.112
1	5	0.076	-2.354	23.176	-64.76
2	4	2.884	-58.578	396.65	-898.058
2	5	0.428	-10.348	81.516	-203.552
2	6	0.12	-3.586	34.646	-101.62
3	5	9.588	-188.214	1229.57	-2674.06
3	6	0.578	-14.302	116.434	-309.788
3	7	0.094	-2.764	27.576	-88.198
4	6	8.638	-168.708	1095.714	-2367.57
4	7	0.54	-13.044	102.596	-256.606
4	8	0.076	-2.35	23.78	-69.514
5	7	11.328	-222.35	1452.724	-3158.07
5	8	0.362	-8.634	67.814	-169.048
5	9	0.112	-3.334	33.084	-99.568
6	8	13.606	-267.756	1755.034	-3831.96
6	9	0.67	-16.436	131.606	-337.46

6	10	0.116	-3.442	34.208	-108.954
7	9	10.98	-212.144	1363.072	-2912.82
7	10	0.436	-10.364	81.046	-204.772
7	11	0.058	-1.758	17.554	-50.642
8	10	11.104	-216.992	1410.502	-3049.67
8	11	0.334	-8.068	63.536	-156.67
8	12	0.06	-1.884	19.314	-57.004
9	11	5.654	-110.888	723.98	-1574.55
9	12	0.426	-10.576	85.306	-221.788
10	12	7.556	-148.412	968.98	-2105.49

A.1 Pair Potentials for Beta Sheet

For beta sheet curve fitting is made for two regions of pair potentials; pair distance greater than native value and pair distance smaller then native value. Third order polynomial is fitted to data:

$$g_{ij} = kr_{ij}^3 + lr_{ij}^2 + mr_{ij} + n$$

Table 4 Coefficients for the function of pair potential for the distance greater than $r_{ij} - r_{ij}^n$

BEAD	BEAD	k	l	m	n
1	3	7.412	-152.912	1050.326	-2405.07
1	4	1.92	-60.436	634.92	-2229.1
1	5	0.844	-36.672	531.754	-2574.08
1	6	-2.012	108.716	-1956.85	11722.88
1	7	-0.062	4.082	-84.358	553.87
1	8	0.024	-1.438	30.128	-209.89
1	9	0.008	-0.654	16.948	-118.634

1	10	0.016	-1.022	22.204	-125.932
1	11	0.006	-0.382	7.662	-18.138
1	12	0.004	-0.32	6.73	-16.574
2	4	-0.612	11.67	-72.948	148.424
2	5	2.722	-90.462	1001.37	-3692.66
2	6	-1.772	75.722	-1077.84	5105.494
2	7	-0.406	19.668	-313.966	1650.222
2	8	0.004	-0.262	6.38	-42.12
2	9	0.006	-0.402	10.36	-53.432
2	10	0.018	-1.078	20.146	-86.29
2	11	0.008	-0.46	8.652	-17.636
2	12	0.008	-0.588	12.818	-60.176
3	5	6.964	-157.882	1191.5	-2993.98
3	6	1.664	-61.096	741.692	-2984.06
3	7	0.964	-36.954	473.582	-2022.57
3	8	0.07	-2.828	38.718	-166.316
3	9	0.018	-0.926	16.108	-60.08
3	10	0.024	-1.158	18.294	-57.056
3	11	0.014	-0.818	15.15	-56.424
3	12	0.016	-1.044	22.348	-127.496
4	6	13.724	-322.826	2521.006	-6539.8
4	7	0.18	-5.224	51.484	-170.292
4	8	0.15	-4.58	45.434	-137.086
4	9	0.034	-1.326	16.936	-47.466
4	10	0.05	-2.27	32.916	-127.56
4	11	0.032	-1.734	30.618	-151.66
4	12	0.03	-1.918	40.448	-261.168
5	7	6.5	-126.506	817.294	-1750.97
5	8	0.466	-11.368	91.808	-241.718
5	9	0.162	-5.228	56.444	-188.938

5	10	0.17	-6.906	92.392	-396.036
5	11	0.034	-1.742	30.682	-173.114
5	12	0.062	-3.646	73.226	-490.13
6	8	9.972	-199.026	1319.896	-2910.22
6	9	2.93	-86.146	843.708	-2749.43
6	10	1.066	-41.384	536.702	-2325.11
6	11	-0.11	5.404	-83.572	409.418
6	12	-1.18	68.422	-1314.39	8360.75
7	9	300.404	-6846.69	52001.71	-131622
7	10	1.586	-57.262	683.492	-2699.59
7	11	66.782	-2884.51	41527.71	-199280
7	12	-3.654	130.512	-1165.2	0
8	10	12.794	-288.274	2161.074	-5392.58
8	11	7.682	-257.662	2878.398	-10710.9
8	12	-30.442	1315.262	-18933.3	90803.09
9	11	-5.164	114.526	-846.5	2084.942
9	12	98.974	-3258.04	35749.2	-130754
10	12	397.826	-8849.39	65613.25	-162156

When pair distance is less than $r_{ij} - r_{ij}^{n}$ 3rd order polynomial and linear fitting is made depending of the behavior of potential. In addition in Table 5 some coefficients are zero because in optimization distance between native pairs can not be less than equilibrium value.

Table 5 Coefficients for the function of pair potential for the distance less than $r_{ij} - r_{ij}^{n}$

BEAD	BEAD	k	l	m	n
1	3	38.698	-807.532	5612.858	-12997.2
1	4	-0.376	9.902	-87.152	252.306
1	5	-0.072	2.254	-23.64	77.452

1	6	0.022	-0.856	10.846	-50.522
1	7	0	0	1.338	-27.716
1	8	0	0	0.57	-15.742
1	9	0	0	1.93	-24.362
1	10	0	0	-2.494	14.736
1	11	0	0	0	0
1	12	0	0	0	0
2	4	0	0	0	0.96
2	5	0	0	-1.096	6.626
2	6	0	0	-1.384	10.652
2	7	0.196	-6.66	74.22	-273.572
2	8	-0.234	5.38	-39.144	86.596
2	9	0	0	-0.926	8.232
2	10	0	0	0	-0.588
2	11	0	0	0	0
2	12	0	0	0	-1.88
3	5	0	0	-1.496	6.35
3	6	0	0	-1.868	12.576
3	7	0	0	-1.022	7.882
3	8	0	0	-1.568	11.242
3	9	0	0	0	-0.342
3	10	0	0	0	-0.56
3	11	0	0	0	0
3	12	0	0	-1.098	7.896
4	6	0	0	0	-3.966
4	7	0	0	-1.932	11.226
4	8	0	0	0	-0.034
4	9	0	0	0	0
4	10	0	0	0	-0.092
4	11	0	0	0.948	-6.602

4	12	0	0	-0.084	-3.032
5	7	0	0	0	-4.19
5	8	0	0	0	-3.25
5	9	0	0	0	-1.548
5	10	1.458	-37.514	318.748	-895.666
5	11	0.444	-13.116	128.056	-416.264
5	12	0	0	0.99	-21.408
6	8	0	0	0	-5.01
6	9	0	0	-1.546	8.672
6	10	0	0	-1.594	11.616
6	11	0	0	-0.398	0.152
6	12	-0.022	0.728	-7.024	5.768
7	9	0	0	-1.91	8.318
7	10	-0.04	1.458	-17.188	61.726
7	11	0.056	-1.876	19.26	-63.984
7	12	-0.004	-0.052	3.244	-34.28
8	10	0	0	0	-1.492
8	11	0	0	-1.728	12.586
8	12	0	0	0.196	-8.936
9	11	0	0	-0.494	0.66
9	12	-0.288	7.116	-57.28	144.084
10	12	0	0	-0.3	-2.722

BIBLIOGRAPHY

1. Bray A, L.J., Walter RR, *Essential Cell Biology*. International student ed. 1998, New York: Garland Publishing.
2. Berg MJ, T.J., Stryer L, *Biochemistry*. Fifth ed. 2002, New York: W.H.Freeman and Company.
3. Dobson, C.M., *Principles of protein folding, misfolding and aggregation.* Semin Cell Dev Biol, 2004. **15**(1): p. 3-16.
4. Lu, H. and J. Skolnick, *A distance-dependent atomic knowledge-based potential for improved protein structure selection.* Proteins, 2001. **44**(3): p. 223-32.
5. Guner, U., Y. Arkun, and B. Erman, *Optimum folding pathways of proteins: their determination and properties.* J Chem Phys, 2006. **124**(13): p. 134911.
6. Encyclopedia, W.,
7. Institute, N.H.G.R.
8. Tooze J, B.C., *Introduction to Protein Structure*. 2nd ed. 1999, New York: Garland Publishing.
9. http://en.citizendium.org/wiki/Protein_structure

10 http://ekhidna.biocenter.helsinki.fi/

11. Chen, Y., et al., *Protein folding: then and now.* Arch Biochem Biophys, 2008. **469**(1): p. 4-19.
12. Dill, K.A., et al., *The protein folding problem.* Annu Rev Biophys, 2008. **37**: p. 289-316.
13. Zwanzig, R., A. Szabo, and B. Bagchi, *Levinthal's paradox.* Proc Natl Acad Sci U S A, 1992. **89**(1): p. 20-2.
14. Dill, K.A. and H.S. Chan, *From Levinthal to pathways to funnels.* Nat Struct Biol, 1997. **4**(1): p. 10-9.
15. Daggett, V. and A.R. Fersht, *Is there a unifying mechanism for protein folding?* Trends Biochem Sci, 2003. **28**(1): p. 18-25.
16. Nolting, B. and K. Andert, *Mechanism of protein folding.* Proteins, 2000. **41**(3): p. 288-98.
17. Fersht, A.R., *Nucleation mechanisms in protein folding.* Curr Opin Struct Biol, 1997. **7**(1): p. 3-9.
18. Maity, H., et al., *Protein folding: the stepwise assembly of foldon units.* Proc Natl Acad Sci U S A, 2005. **102**(13): p. 4741-6.
19. Dill, K.A., et al., *The protein folding problem: when will it be solved?* Curr Opin Struct Biol, 2007. **17**(3): p. 342-6.
20. Ozkan, S.B., et al., *Protein folding by zipping and assembly.* Proc Natl Acad Sci U S A, 2007. **104**(29): p. 11987-92.
21. Voelz, V.A. and K.A. Dill, *Exploring zipping and assembly as a protein folding principle.* Proteins, 2007. **66**(4): p. 877-88.

22. Duan, Y. and P.A. Kollman, *Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.* Science, 1998. **282**(5389): p. 740-4.
23. Pitera, J.W. and W. Swope, *Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins.* Proc Natl Acad Sci U S A, 2003. **100**(13): p. 7587-92.
24. Vila, J.A., D.R. Ripoll, and H.A. Scheraga, *Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures.* Proc Natl Acad Sci U S A, 2003. **100**(25): p. 14812-6.
25. Alder, B.J., *Studies in Molecular Dynamics. I. General Method.* The Journal of chemical physics, 1959. **31**(2): p. 459.
26. Dokholyan, N.V., et al., *Discrete molecular dynamics studies of the folding of a protein-like model.* Fold Des, 1998. **3**(6): p. 577-87.
27. Jang, H., C.K. Hall, and Y. Zhou, *Folding thermodynamics of model four-strand antiparallel beta-sheet proteins.* Biophys J, 2002. **82**(2): p. 646-59.
28. Dill, K.A., et al., *Principles of protein folding--a perspective from simple exact models.* Protein Sci, 1995. **4**(4): p. 561-602.
29. Tanaka, S. and H.A. Scheraga, *Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins.* Macromolecules, 1976. **9**(6): p. 945-50.
30. Miyazawa S, J.R., *Estimation of effective interresidue contact energies from protein crystal structures; quasi chemical approximation.* Macromolecules, 1985. **18**: p. 534-552.
31. Miyazawa, S. and R.L. Jernigan, *Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading.* J Mol Biol, 1996. **256**(3): p. 623-44.
32. Sippl, M.J., *Knowledge-based potentials for proteins.* Curr Opin Struct Biol, 1995. **5**(2): p. 229-35.
33. Kocher, J.P., M.J. Rooman, and S.J. Wodak, *Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches.* J Mol Biol, 1994. **235**(5): p. 1598-613.
34. Bahar, I. and R.L. Jernigan, *Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation.* J Mol Biol, 1997. **266**(1): p. 195-214.
35. Jernigan, R.L. and I. Bahar, *Structure-derived potentials and protein simulations.* Curr Opin Struct Biol, 1996. **6**(2): p. 195-209.
36. Gunn JR, M.A., Friesner RA, *Hierarchical Algorithm for computer modeling of protein tertiary strcture: folding of myoglobin to 6.2 A resolution.* Journal of Physical Chemistry, 1994. **98**: p. 702-711.
37. Casari, G. and M.J. Sippl, *Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds.* J Mol Biol, 1992. **224**(3): p. 725-32.
38. Tobi, D. and R. Elber, *Distance-dependent, pair potential for protein folding: results from linear optimization.* Proteins, 2000. **41**(1): p. 40-6.
39. Mukherjee, A., P. Bhimalapuram, and B. Bagchi, *Orientation-dependent potential of mean force for protein folding.* J Chem Phys, 2005. **123**(1): p. 014901.

40. Melo, F. and E. Feytmans, *Novel knowledge-based mean force potential at atomic level.* J Mol Biol, 1997. **267**(1): p. 207-22.
41. Skolnick, J., et al., *Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct?* Protein Sci, 1997. **6**(3): p. 676-88.
42. Keskin.O, I.B., R.Jernigan., *Emprical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions.* Protein Sci, 1998. **7**: p. 3578-86.
43. Erman B, D.K., *Gaussian model of protein folding.* Journal of Chemical Physics, 2000. **112**(2): p. 1050-6.
44. Erkip A, E.B., Seok Chaok, Dill K, *Parameter optimization for the Gaussian model of protein folding.* Polymer, 2002. **43**: p. 495-501.
45. Haliloglu T., B.I., Erman B., *Gaussian Dynamics of Folded Proteins.* Physical Review Letters, 1997. **79**(16): p. 3090-93.
46. F.MacGregor, S.Y.a.J., *Principal-Component Analysis of Multiscale Data for Process Monitoring and Fault Diagnosis.* AIChE Journal, 2004. **50**(11): p. 2891-2903.
47. KRESTA J.V, M.J.F.a.M.T.E., *Multivariate Statistical Monitoring of Process Operating Performance.* The Canadian Journal of Chemical Engineering, 1991. **60**: p. 35-47.
48. Nomikos P., M.J.F., *Monitoring Batch Processes Using Multiway Principal Component Analysis.* AIChE Journal, 1994. **40**(8): p. 1361-75.
49. Rannar S, M.J., Wold S, *Adaptive batch monitoring using hierarchical PCA.* Chemometrics and intelligent laboratory systems, 1998. **41**: p. 73-81.
50. Geladi P., a.B.K., *Partial Least Squares Regression: A tutorial.* Analytica Chimica Acta, 1986. **185**(1): p. 1-17.
51. Nomikos P., M.J.F., *Multivariate SPC Charts for Monitoring Batch Processes.* Technometrics, 1995. **37**(1): p. 41-59.

VITA

Sefer Baday was born in 1984, Kars, Turkey. He had been Trabzon Yomra Fen Lisesi for high school education. He received his BS degree in 2006, in Chemical Engineering at Boğaziçi University. He worked as a teaching and research assistant in Koç University during 2006-2008.

.