UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1184

Protein Folding Simulations in Kink Model

XUBIAO PENG

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2014

ISSN 1651-6214
ISBN 978-91-554-9043-0
urn:nbn:se:uu:diva-232562

Dissertation presented at Uppsala University to be publicly examined in 80101, Ångström Laboratory, Lägerhyddsvägen 1, Uppsala, Friday, 7 November 2014 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Ulrich Hansmann (Department of Chemistry and Biochemistry, University of Oklahoma).

Abstract
Peng, X. 2014. Protein Folding Simulations in Kink Model. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1184. 56 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-554-9043-0.

The structure of protein is essentially important for life activities. Proteins can perform their functions only by specific structures. In this thesis, the kink and multi-kink model for protein description are reviewed. It is shown that most of the loop parts in Protein Databank (PDB) can be described by very limited number of kinks within the experimental precision. Furthermore, by applying the model into two well studied real proteins (myoglobin and villin headpiece HP35), it is shown that the multi-kink model gives correct folding pathway and thermal dynamical properties compared with the experimental results for both proteins. In particular, the kink model is computationally inexpensive compared with other existing models. In the last chapter, a new visualization method for the heavy atoms in the side-chain is presented.

Keywords: protein folding, kink model, soliton

Xubiao Peng, Department of Physics and Astronomy, Theoretical Physics, Box 516, Uppsala University, SE-751 20 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-554-9043-0
urn:nbn:se:uu:diva-232562 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-232562)

Dedicated to my parents

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Hu. S., Krokhotin, A., Niemi, A. J. and Peng, X. (2011) *Towards quantitative classification of folded proteins in terms of elementary functions.* Phys. Rev. E, 83: 041907

II Krokhotin, A., Niemi, A. J. and Peng, X. (2012) *Soliton concepts and protein structure* Phys. Rev. E, 85: 031906

III Krokhotin, A., Niemi, A. J. and Peng, X. (2013) *On the role of thermal backbone fluctuations in myoglobin ligand gate dynamics* J. Chem. Phys., 138: 175101

IV Krokhotin, A., Lundgren, M., Niemi, A. J. and Peng, X. (2013) *Soliton driven relaxation dynamics and protein collapse in the villin headpiece* J. Phys.: Condens. Matter 25: 325103

V Peng, X., Sieradzan, A. K., Scheraga, H. A. and Niemi, A. J. (2014) *Collective motions and structural self-organisation along the myoglobin folding pathway* [Manuscript]

VI Peng, X., Chenani, A., Hu, S., Zhou, Y. and Niemi, A. J. (2014) *A three dimensional visualisation approach to protein heavy-atom structure reconstruction* [Submitted to BMC Struct. Biol.]

Papers not included in this thesis:

VII Niemi, A. J. and Peng, X. (2014), *Folding studies with proteins that have an exceptional complex structure* [Submitted to J. Phys. A]

VIII Sieradzan, A. K., Peng, X. and Niemi, A. J. (2014), *Peierls-Nabarro Barrier and Protein Loop Propagation* [Submitted to Phys. Rev. E]

Contents

1. Introduction

As the "primarily important" part of life, proteins play an important role in cells. Proteins make up around 42%-46% dry weight in one cell [1]. They take part in all kinds of living activities, including building tissues, binding and transporting small molecules, catalyzing the chemical reactions, *etc* [2]. The correct 3D structure of protein is essentially important for life. It is believed that only correctly folded protein can perform the function normally, while misfolding of protein can lead to different kinds of diseases, from neurodegenerative disease to cancer [3, 4].

Based on the shape and solubility, proteins can be generally classified into three different groups: globular proteins, membrane proteins and fibrous proteins. In this thesis, I only focus on the most frequent and water soluble one - globular protein. As an introduction, the protein structure and protein folding problems are described briefly.

1.1 Protein structure

Different from the ordinary linear polymers, protein has a well-organized hierarchy. There are four different levels in protein structure which are shown as Figure 1.1 (a)-(d), respectively.

Figure 1.1 (a) shows that the amino acids are the elements in forming proteins. There are 20 different kinds of amino acids. Each amino acid is abbreviated by one letter or three letters. The amino acids have the same fundamental structures–one amino group and one carboxyl group attached to a carbon atom named C_α atom. The only difference between different amino acids lies in the structures of the side-chains (**R**-groups) which are also connected with the C_α atoms. The **R**-group could be quite simple such as a hydrogen atom in glycine (Gly) or quite complex like the side-chain in tryptophan(Trp). In protein, every two neighboring amino acids form a peptide bond with elimination of an H_2O molecule. The protein is composed of the amino acid residues connected by the peptide bonds. It should be noted that the peptide bond is a partial double bond so that the atoms connecting two neighboring C_α atoms lie in a same plane–the peptide plane. There are two kinds of peptide planes named *cis* and *trans* depending on the dihedral angles ω among atoms C_α-C-N-C_α in the peptide plane. For *trans*-peptide plane, $\omega \approx \pi$; Otherwise, $\omega \approx 0$. As a result, the backbone of protein can be considered as a chain of sequential peptides planes and the side-chain **R**-groups are the branches on it. In this sense,

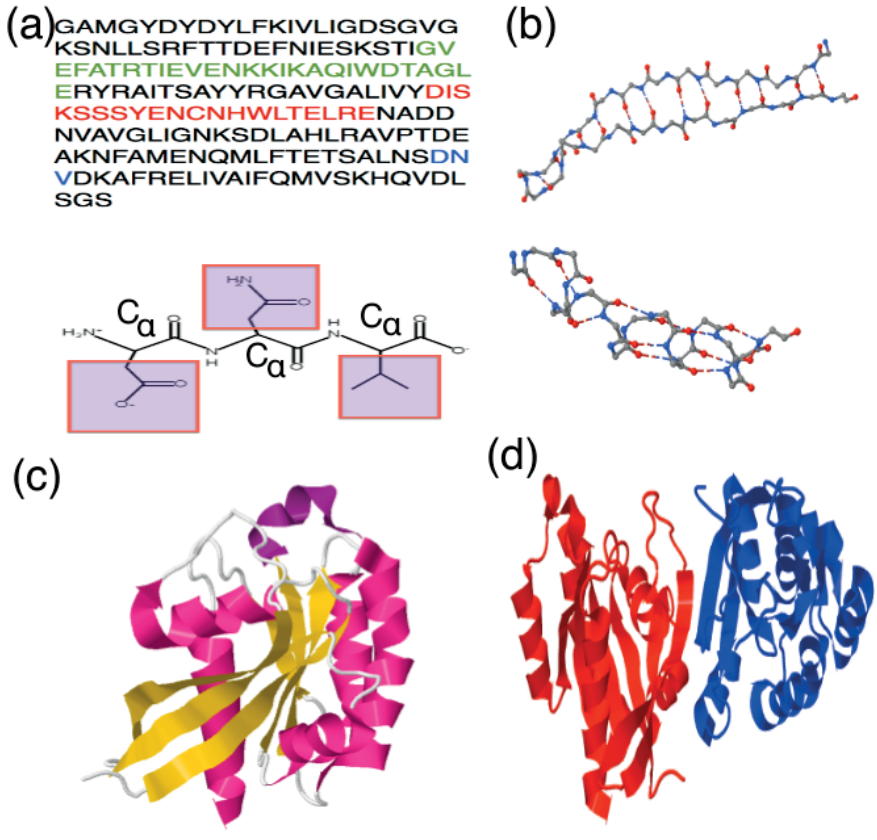

Figure 1.1. The four levels of protein (PDB code 3RWO). (a) The primary structure: the top figure shows the sequence of the protein, while the bottom figure displays the chemical structures corresponding to the three amino acids marked in blue in the top figure. The side-chains are labeled in shades. (b) The secondary structure β-sheet (top) and α-helix (bottom) with hydrogen bonds labeled as dashed line. The β-sheet is corresponding to the sequences marked in green and α-helix corresponding to those in red in subfigure (a). (c) The tertiary structure colored in different secondary structures. (d) The quaternary structures colored in different chains.

the sequence (a string of the amino acid names) is the primary structure of the protein.

Going up from this level, there are some local structures stabilized by particular patterns of hydrogen bonds formed in backbone. These stable structures are called the secondary structures. The most common secondary structures include the β-Sheet and α-Helix as shown in Figure 1.1 (b). Besides, there are different kinds of loops like β-turn in secondary structure level. Different from the β-Sheet and α-Helix, loops usually do not have regular geometric

repetition along the backbone, but in most cases they play an important role in biological function, especially when the protein is binding to some particular ligands.

The tertiary structure is a compact 3D structure in space composed of several secondary structures as shown in Figure 1.1 (c). The side-chains play an important role in forming and stabilizing the tertiary structure. The tertiary structure is a relatively independent unit in protein. Many proteins perform their functions in the tertiary structure.

If the protein contains more than one polypeptide chains, these chains may interact with each other forming the quaternary structure as Figure 1.1 (d). In quaternary structures, the protein-protein interaction plays an important role [5].

1.2 Protein folding problems

The foundation of the protein folding study is the fact that the tertiary structure information is completely encoded in the primary structure. This fact was verified by the remarkable experiment by Christian Anfinsen in 1957 [6, 7]. This experimental evidence told us that it is completely possible to investigate the tertiary structure of protein from the primary sequence information together with solution conditions. Indeed, there are helper proteins known as the chaperones during protein folding process *in vivo*. However, the function of the chaperone seems to provide an isolated environment for protein folding instead of to participate the folding process [8].

The protein folding problems are threefold [9]:

(1) "Physical code of protein folding": To find out the time evolution for the protein folding to its native state from the sequences information and solution conditions.

(2) "The rate mechanism of protein folding": This question is closely related to Levinthal's paradox [10], which states that the rate of protein folding is much faster than all conformations searching. We need to know how the protein knows which conformation not to search.

(3) "Computing protein structures from amino acid sequences": To predict tertiary structure from the primary structure using computer. The solution of this problem can accelerate the discovery of new protein and the drug design.

It should be noted that these three problems are not independent, but closely related to one another. A solution for one of them would definitely benefit the other two.

Many efforts have been made to solve these problems, both theoretically and experimentally. In experiment, many different kinds of protein data bases are built up, including Protein Data Bank (PDB) [11], Protherm [12], KineticDB [13], and so on. Theoretically, many different energy models are

put forward including Gō model [14], different physics-based models (such as AMBER [15, 16, 17], CHARMM [18, 19, 20], OPLS [21, 22], UNRES [23, 24, 25], ASTRO-FOLD [26, 27], GROMACS [28, 29]), knowledge-based models (such as TASSER [30], chunk-TASSER [31], I-TASSER [32, 33], ROSETTA [34]) and the kink model [35, 36]. For the problem (3) on prediction, the computer based protein structure prediction event CASP (Critical Assessment of protein Structure Prediction) has been carried every even summer since 1994 [37].

1.3 Outline

The protein folding problems still remain big challenges for scientists in biology, chemistry, physics and computer science although there have been many different approaches. For example, in the most widely used Molecular Dynamics (MD) simulations, the expensive computational cost and long simulation time are big limitations in their applications.

In this thesis, the kink (multi-kink) model is reviewed and applied on some real proteins, showing the kink model from mathematical physics can describe the protein backbone in a very high precision. The detailed applications show some advantages in kink model:
1. It is calculation inexpensive: The modeling and analysis procedures can be carried on a single processor Mac Desktop. For a protein entry with around 100 residues, it takes *3-4 weeks* for the whole procedure (including parameter training and dynamics simulations). In particular, a complete dynamical folding simulation from a random chain to native structure only takes *minutes* after the proper parameters are found.
2. It gives high accuracy in backbone modeling with root-mean-square deviation (RMSD) less than 1Å.
3. For the proteins checked until now, in particular myoglobin, it gives correct folding intermediate state and folding pathway compared with the experimental result.
However, the limitation is that the parameters in the model have not been derived from the sequence and solution information yet. So far, the kink model is still a native structure based model.

The thesis is organized as follows: In chapter 2, the kink and multi-kink model theory in protein is reviewed. In chapter 3, the universality of the kink model in protein is shown based on Paper I and II. In Chapter 4, the multi-kink model is applied to two proteins–myoglobin and villin headpiece HP35, which is based on Paper III, IV and V. In Chapter 5, the distributions of the heavy atoms in side-chain are studies in Frenet-based frames according to Paper VI.

2. Review on kink model in protein backbone geometry and dynamics

In this chapter, I shall give a review on the theory of the kink model, which will be used in real protein simulations in later chapters. This model was first introduced in the Reference [35] and further theoretically developed in Reference [38, 39] together with Paper III.

2.1 Protein C_α trace

As is stated in Sec.1.1, the protein backbone is composed of sequential peptide planes. Any two neighboring C_α atoms share a peptide plane, in which the angles between any two covalent bonds are almost constant. As a result, the distance between any two neighboring C_α atoms is almost a constant value (3.8 Å for *trans*-peptide, and 2.8 Å for *cis*-peptide plane). For further simplification, one can ignore the peptide planes together with their orientations and directly connect the neighboring C_α atoms by virtual bonds. In this way, we get a highly coarse grained model of protein - the C_α trace. In the C_α trace, the steric constraint

$$|\mathbf{r}_i - \mathbf{r}_k| \geq 3.8 \text{ Å} \quad \text{for } |i-k| \geq 2 \tag{2.1}$$

exists between two non-neighboring C_α atoms along the chain. This relation is mainly due to the steric effect between the peptide planes and is well respected by most of the folded proteins in PDB. The constraint (2.1) will be enforced *throughout* all the simulations in this thesis.

2.2 C_α trace description

As a discrete curve, the C_α trace can be described by a local frame– Discrete Frenet frame (DF-frame) defined as:

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|} \tag{2.2}$$

$$\mathbf{b}_i = \frac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{|\mathbf{t}_{i-1} \times \mathbf{t}_i|} \tag{2.3}$$

$$\mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i \tag{2.4}$$

where $\mathbf{r}_i$ (i = 1, ..., N) is the coordinate of the i-th C_α atoms counting from the N terminals. Apparently, each DF-frame $(\mathbf{t}_i, \mathbf{n}_i, \mathbf{b}_i)$ is determined by every three consecutive C_α atoms $(\mathbf{r}_{i-1}, \mathbf{r}_i, \mathbf{r}_{i+1})$. Between two consecutive DF-frames along the discrete curve, there is a relation:

$$\begin{pmatrix} \mathbf{n}_{i+1} \\ \mathbf{b}_{i+1} \\ \mathbf{t}_{i+1} \end{pmatrix} = \begin{pmatrix} \cos\kappa\cos\tau & \cos\kappa\sin\tau & -\sin\kappa \\ -\sin\tau & \cos\tau & 0 \\ \sin\kappa\cos\tau & \sin\kappa\sin\tau & \cos\kappa \end{pmatrix}_{i+1,i} \begin{pmatrix} \mathbf{n}_i \\ \mathbf{b}_i \\ \mathbf{t}_i \end{pmatrix} \tag{2.5}$$

where

$$\kappa_{i+1,i} \equiv \kappa_i = \arccos(\mathbf{t}_{i+1} \cdot \mathbf{t}_i) \qquad \in [0, \pi] \tag{2.6}$$

$$\tau_{i+1,i} \equiv \tau_i = \nu \arccos(\mathbf{b}_{i+1} \cdot \mathbf{b}_i) \qquad \in [-\pi, \pi) \tag{2.7}$$

with

$$\nu = sgn[(\mathbf{b}_i \times \mathbf{b}_{i+1}) \cdot \mathbf{t}_i]. \tag{2.8}$$

The angle κ_i is the virtual bond angle, and τ_i is the virtual dihedral angle as shown in Figure 2.1.

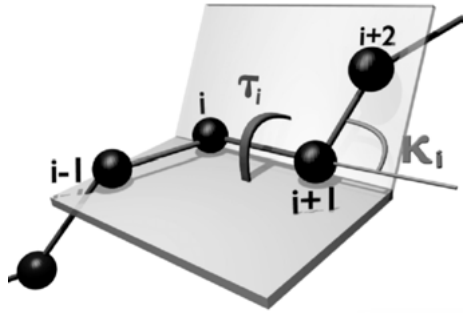

Figure 2.1. Definitions of virtual bond angle (κ_i) and dihedral angle (τ_i).

Assuming all the virtual bond lengths have the same constant values (3.8 Å), the backbone C_α trace can be completely described by these two sets of internal coordinates (κ_i, τ_i). In other words, given an arbitrary initial DF-frame $(\mathbf{t}_1, \mathbf{n}_1, \mathbf{b}_1)$ and two complete sets of angles (κ_i, τ_i), all the remaining DF-frames can be obtained from Eq. (2.5) iteratively. Consequently, the complete C_α trace can be reconstructed by the summation of the tangent vectors:

$$\mathbf{r}_k = \sum_{i=0}^{k-1} |\mathbf{r}_{i+1} - \mathbf{r}_i| \cdot \mathbf{t}_i \approx 3.8 \sum_{i=0}^{k-1} \mathbf{t}_i \tag{2.9}$$

where $\mathbf{r}_k$ is the coordinates of k-th C_α atom.

2.3 Kink model

Geometrically, the C_α trace of a protein only depends on (κ_i, τ_i). It is a natural choice to describe the energy of the protein C_α trace in these angles. Assuming a generic energy function $H(\kappa, \tau)$ with its extremum at $\kappa_i = \kappa_{i0}$ and $\tau_i = \tau_{i0}$. After introducing a small deformation

$$\begin{aligned} \Delta\kappa_i &= \kappa_i - \kappa_{i0} \\ \Delta\tau_i &= \tau_i - \tau_{i0} \end{aligned}$$

one can expand the energy around the extremum point

$$\begin{aligned} H(\kappa_i, \tau_i) \;=\; & H(\kappa_{i0}, \tau_{i0}) + \sum_k \{ \frac{\partial H}{\partial \kappa_k}_{|0} \Delta\kappa_k + \frac{\partial H}{\partial \tau_k}_{|0} \Delta\tau_k \} \\ + \sum_{k,l} \Big\{ & \frac{1}{2} \frac{\partial^2 H}{\partial \kappa_k \partial \kappa_l}_{|0} \Delta\kappa_k \Delta\kappa_l + \frac{\partial^2 H}{\partial \kappa_k \tau_l}_{|0} \Delta\kappa_k \Delta\tau_l + \frac{1}{2} \frac{\partial^2 H}{\partial \tau_k \partial \tau_l}_{|0} \Delta\tau_k \Delta\tau_l \Big\} + \mathscr{O}(\Delta^3) \end{aligned} \tag{2.10}$$

Apparently, the first term is constant and the second term vanishes because of containing the first order derivative at the extremum. For the remaining terms, the symmetry properties of the backbone geometry need to be considered.

Before that, let's first rename variables $(\Delta\kappa, \Delta\tau) \to (\kappa, \tau)$ that are identified as the virtual bond and torsion angles shown in Figure 2.1. A fact should be noticed: The angles κ_i, τ_i in energy function is defined by DF-frames ($\mathbf{t}_i$,$\mathbf{n}_i$,$\mathbf{b}_i$), but in Equation (2.9) the coordinates of C_α atoms only depend on the tangent vector $\mathbf{t}_i$. If the normal plane is rotated by one angle Δ_i around the tangent vector $\mathbf{t}_i$, i.e.,

$$\begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_i \to \begin{pmatrix} \cos\Delta_i & \sin\Delta_i & 0 \\ -\sin\Delta_i & \cos\Delta_i & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_i \tag{2.11}$$

the geometry of the C_α trace is not affected. But the angles κ_i, τ_i should perform a corresponding transformation as [40]

$$\kappa_i\, T^2 \;\to\; e^{\Delta_i T^3} (\kappa_i T^2)\, e^{-\Delta_i T^3} \;=\; \kappa_i (T^2 \cos\Delta_i - T^1 \sin\Delta_i) \tag{2.12}$$

$$\tau_i \;\to\; \tau_i + \Delta_{i-1} - \Delta_i \tag{2.13}$$

Here (T^i) $(i = 1,2,3)$ are three generators of the SO(3) Lie algebra generators,

$$T^1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \quad T^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \quad T^3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

It should be noted that after the rotation on the normal plane, the curvature κ_i is extended to a complex number from a real number. Angles κ_i take real

values only when the rotation angles $\Delta_i = 0$ or $\Delta_i = \pi$. Apparently, $\Delta_k = 0$ is corresponding to the case of DF-frame definition. The rotation $\Delta_k = \pi$ for all the sites $k > i$ (i is some particular site) will reverse all the directions of $\mathbf{b}_k$ and $\mathbf{n}_k$, which leads to the Z(2) gauge transformation in (κ, τ) profiles

$$\begin{array}{rcll} \kappa_k & \rightarrow & -\kappa_k & \text{for all } k \geq i \\ \tau_i & \rightarrow & \tau_i - \pi & \end{array} \tag{2.14}$$

The energy equation (2.10) should be only affected by the geometry of the curve and be independent of the frame transformation (2.11). According to the Reference [38], the energy satisfying such conditions takes the form like

$$H(\kappa,\tau) = \sum_{i=1}^{N-1} (\kappa_{i+1} - \kappa_i)^2 + \sum_{i=1}^{N} \left\{ q\,(\kappa_i^2 - m^2)^2 + \frac{d}{2}\,\kappa_i^2\tau_i^2 - b\kappa_i^2\tau_i - a\tau_i + \frac{c}{2}\tau_i^2 \right\} \tag{2.15}$$

after expanding to the lowest orders.

Since the energy is obtained from the basic properties of the geometry, any energy function of protein, either at all-atom or coarse-grained level, *should* reproduce (2.15) when expanded around the native state.

The $\kappa - \tau$ profiles at extremum can be derived by minimizing the Hamiltonian (2.15) [35, 36]:
First, according to the extremum condition

$$\frac{dH}{d\tau_i} = 0 \tag{2.16}$$

the torsion angles τ_i can be expressed in a function of the bond angles κ_i

$$\tau_i[\kappa_i] = \frac{a + b\kappa_i^2}{c + d\,\kappa_i^2} \;\equiv\; \frac{1 + u\,\kappa_i^2}{v + w\,\kappa_i^2} \tag{2.17}$$

where $u = b/a$ and $v = c/a$ and $w = d/a$.

By inserting equation (2.17) into equation (2.15), the torsion angles τ_i are eliminated and the following equation for the bond angle κ_i is obtained.

$$\kappa_{i+1} = 2\kappa_i - \kappa_{i-1} + \frac{dV_{pot}[\kappa]}{d\kappa_i^2}\kappa_i \quad (i = 1, ..., N) \tag{2.18}$$

where $\kappa_0 = \kappa_{N+1} = 0$ and

$$V_{pot}[\kappa] = -a\left(1 - \frac{uv}{w}\right)^2 \frac{1}{2(v + w\kappa^2)} - \left(a\,\frac{u^2}{2w} + 2qm^2\right)\kappa^2 + q\,\kappa^4 \tag{2.19}$$

The equation (2.18) combined with (2.19) is a Generalized Discrete Nonlinear Schrödinger Equation(GDNLSE).

The first term in the potential energy (2.19) is the generalization of the Vinetskii-Kukhtarev potential contribution [41], while the remaining two terms are the double well potential with extremum point at $\kappa_i = \pm m$.

Note that the parameter a appears in the potential energy Equation (2.19) although it is not in the expression equation (2.17) explicitly. Therefore, it characterizes the relative contributions of the torsion angles τ_i in the total energy (2.15). In protein, the torsion angles τ_i are always more flexible than the bond angle κ_i, which means near the energy minimum the energy contribution from τ_i is much smaller than that from κ_i. So it is reasonable to require the parameter a to be much smaller than parameters q and m.

Consequently, in the protein energy function (2.15) the parameters satisfy $qm^2 > 1$ and $a << 1$, which leads to a result that the first term in the potential (2.19) is neglectable compared with the other two. In this case, the potential energy is bimodal and there is a kink-type solution for equation (2.18) [35, 36], which interpolates the two minima at $\kappa_i = \pm m$. The explicit form of the kink in terms of elementary functions has not been found. But the existence and uniqueness have been proven in Reference [36]. Until now, there are two ways to give the approximation of the kink solution in protein:

1. It has been shown that the solution of the kink can be numerically obtained by an iterative procedure in Reference [36]. For doing this way, the software program *ProPro* has been developed in the following link:

$$\texttt{http://www.folding-protein.org} \tag{2.20}$$

Note in *ProPro* the parameters q and m in energy function (2.15) are split into two sets, named as $q1,\ q2$ and $m1,\ m2$ to represent the asymmetric properties in backbone structure for one kink. Parameters $q1,\ m1$ are used in the first half of the kink and $q2,\ m2$ for second half.

2. Alternatively, an ansatz of κ_i can be used to represent the kink-type solution as

$$\kappa_i = \frac{(\mu_{r1} + 2\pi N_{r1})e^{\sigma_{r1}(i-s_{r1})} - (\mu_{r2} + 2\pi N_{r2})e^{-\sigma_{r2}(i-s_{r2})}}{e^{\sigma_{r1}(i-s_{r1})} + e^{-\sigma_{r2}(i-s_{r2})}} \tag{2.21}$$

$$\tau_i = \frac{a_r}{1 + d_r \kappa_{i-1/2}} \tag{2.22}$$

where the quantity $\kappa_{i-1/2} = \frac{1}{2}(\kappa_i + \kappa_{i-1})$. The parameters s_{r1} and s_{r2} determine the center of the kink-solution, and $\mu_{r1}, \mu_{r2} \in [0, \pi]$ together with N_{r1}, N_{r2} describe the κ values on both sides of the kink far away from the center. N_{r1} and N_{r2} are the covering numbers that determine how many times κ_i covers the fundamental domain $[0, \pi]$ when we traverse the topological kink once. Usually the constraints $s_{r1} = s_{r2} = s_r$ and $N_{r1} = N_{r2} = N_r$ are imposed for simplicity.

It should be noted that the parameters in the ansatz are not the same as the ones in the Hamiltonian (2.15). The relation between these two kinds of parameters is still unclear. At the moment, the ansatz is just a simplified and computationally efficient way to describe the kink (κ, τ) profiles due to its analytical form. It turns out a good approximation in describing the (κ, τ) profile of a kink structure if one *only* cares about the geometry in protein. Because of the lack of energy interpretation in the ansatz, in the study of thermal dynamical properties for protein, only the numerical method can be used.

2.4 Parameters

Both the energy function (2.15) and the ansatz (2.21), (2.22) contain a number of parameters, related to the physical and chemical properties of the protein, even its environment. In principle these parameters can be derived from the knowledge of the physical and chemical properties in protein, together with the environment. But at the moment, this has not yet been achieved. Presently, for a particular protein the proper parameters are directly searched through a Monte Carlo Annealing Simulation (MCAS), which minimizes the RMSD between the model (either reconstructed from the ansatz or from minimizing the energy (2.15)) and real protein. In particular, the MCAS procedure of searching for the parameters in energy (2.15) is embedded in the software *Propro* at (2.20).

2.5 Multi-kink model

Based on the single kink modeling, the multi-kink model for a complete peptide chain (C_α trace) can be constructed as follows:

1. One can calculate the (κ, τ) profiles from the given protein structure and divide the (κ, τ) profiles along the chain into several different kink-like structures by applying Z(2) gauge transformations. The site i in Z(2) gauge transformation (2.14) is usually the center of a certain kink. Two neighboring kinks share a short common secondary structure segment, where κ and τ are almost constant. The center of kink happens where both κ value and τ value have obvious changes. Usually, the κ angles at center have relatively small jump while τ angles fluctuate dramatically.

2. The corresponding kink model for every kink-like structure identified in step 1 can be set up by finding proper parameters either in the ansatz or in the energy (2.15) depending on the modeling purpose.

3. After the individual kink modeling is finished, there are two different ways to deal with the connections: (1) In ansatz modeling, one can directly connect the separate kinks together like playing Lego; (2) If the modeling is based on energy (2.15), the total energy function should be the summation of

all the energies in individual kink modeling and the (κ, τ) profiles should be the ones minimizing the total energy. Note that in the latter case the (κ, τ) profiles at minimum of the total energy is not just a collection of those in minimizing individual kinks because the term $(\kappa_{i+1} - \kappa_i)^2$ at the connection points of two neighboring kinks also appears in the total energy. As a result, during the kinks connection process the retraining of parameters is needed to make the overall (κ, τ) profiles at total energy minimum consistent with those in PDB structure.

In the following chapters related to studying the dynamical properties of proteins, the multi-kink model of protein is constructed numerically, using the programs *GaugeIT* and *ProPro* in (2.20).

2.6 Non-equilibrium dynamics simulation

With the multi-kink modeling on the native state of the protein, one can study the dynamical properties and folding process of the protein. Assuming the unfolding and folding process of a protein is due to the fluctuation of the ambient temperature, the non-equilibrium dynamics can be simulated by Monte Carlo (MC) sampling with the temperature heated up and cooled down. The MC sampling is carried out according to the Glauber algorithm, in detail:

1. At the beginning, the protein is at the native state of the multi-kink model and the MC temperature T_{low} is very low.

2. For the first N_1 steps, the temperature is increased linearly in logarithm scale from T_{low} to T_{high}. In the middle N_2 steps, the temperature is kept on the high temperature T_{high} to let the protein fully thermalized and unfolded; For the last N_3 steps, the temperature is cooled back to T_{low} in the same rate as the heating process.

On each MC step, the perturbation is performed either on one bond angle $\kappa_i \to \kappa_i + 0.015r$ or torsion angle $\tau_i \to \tau_i + 1.5r$, where r is a random number with gaussian distribution whose expected value is 0 and variation is 1. Note different scales on κ and τ perturbation amplitudes are selected to represent the different stiffness for the bond angles and torsion angles in real protein. The perturbed conformation is accepted by the Glauber probability [42, 43, 44]

$$P = \frac{x}{1+x} \quad \text{with} \quad x = \exp\{-\frac{\Delta E}{kT}\} \tag{2.23}$$

where ΔE is the internal energy difference of the multi-kink model between consecutive MC time steps evaluated from the effective internal energy (2.15). Meanwhile, the steric constraint (2.1) is checked for each conformation. Once it is violated, the conformation gets rejected.

3. The heating-cooling cycle described in step 2 is repeated M times to get statistical properties in the unfolding-folding procedure.

Since the temperature is changing at each MC step, the sampled states are non-equilibrium.

2.7 Temperature renormalization

It should be noticed that the parameters in energy function (2.15) do not depend on the temperature explicitly. In particular, in the thermal dynamic simulations as Sec.2.6, the parameters in the energy are fixed. For example, the nearest neighbor coupling term $-J\sum_{i=1}^{N-1}\kappa_{i+1}\kappa_i$ in energy is always normalized to $-2\sum_{i=1}^{N-1}\kappa_{i+1}\kappa_i$ for all temperatures.

In practice this coupling coefficient J should be different with temperature changing. As a result, the MC temperature kT in the Glauber probability (2.23) is not the physical temperature. It relates to the physical temperature t in such a way

$$\frac{2}{kT} \rightarrow \frac{J(t)}{k_B t} \tag{2.24}$$

where k_B is the Boltzmann constant. The coupling coefficient $J(t)$ obeys the renormalization group equation,

$$t\frac{dJ}{dt} = \beta_J(J;a,b,c,d,m,q) \tag{2.25}$$

Assuming that the leading order of β_J only depends on $J(t)$ and the expansion of $J(t)$ at low temperature is

$$J(t) \approx J_0 - J_1 t^{\alpha} + ... \tag{2.26}$$

the equation (2.25) can be simplified as

$$\alpha(J(t) - J_0) = \beta_J(J(t)) \tag{2.27}$$

From equation (2.24), the coupling coefficient $J(t)$ can be expressed in terms of MC temperature kT and physical temperature $k_B t$

$$J(t) \approx \frac{2k_B t}{kT} \tag{2.28}$$

As a result, the renormalization group equation for MC temperature kT is obtained by combining (2.25) with (2.28)

$$t\frac{d}{dt}(\frac{1}{kT}) = -\frac{1}{kT} + \frac{1}{2k_B t}\beta_J(\frac{2k_B t}{kT}) \tag{2.29}$$

After introducing some substitutions

$$\beta_J(\frac{2k_B t}{kT}) = \frac{2k_B t}{kT} + F(\frac{2k_B t}{kT}) \tag{2.30}$$

$$y = \frac{1}{kT} \quad \text{and} \quad x = \frac{1}{2k_B t} \tag{2.31}$$

equation (2.29) becomes

$$\frac{dy}{dx} = -F(\frac{y}{x}) \tag{2.32}$$

The solution is

$$\ln(\lambda x) = -\int^{\frac{y}{x}} \frac{du}{u+F(u)} \tag{2.33}$$

where λ is an integral constant. Furthermore, one can assume the leading nonlinear correction of function $\beta_J(J(t))$ is logarithmic for simplicity, *i.e.,*

$$F(u) = u(\eta - 1 + \alpha \ln u) \tag{2.34}$$

Then the integral part in equation (2.33) can be calculated analytically, and the solution is

$$kT = 2k_B t \exp(\frac{\eta}{\alpha} + (2k_B t/\lambda)^{\alpha}) \tag{2.35}$$

Recall the expression of $J(t)$ in equation (2.26) and (2.28), at low temperature we have the relation

$$kT = \frac{2k_B}{J_0} t. \tag{2.36}$$

Comparing the equation (2.36) and (2.35), the proper parameters η and λ can be selected to satisfy the low temperature limit. As a result, the relation between the MC temperature and physical temperature is shown as

$$kT \approx \frac{2}{J_0} k_B t \exp(\frac{J_1}{J_0} t^{\alpha}). \tag{2.37}$$

3. The universality of the kink structure

In this chapter, I shall show that single kink model can describe the super-secondary structure in protein with a high precision. More generally, the universality of the kink structures in PDB is illustrated. Since only the geometry of the protein backbone is considered in this chapter, the ansatz solution is used during modeling. However, it must be stressed that all the structures modeled by the ansatz can also be done by solving the GDNLSE (2.18) numerically with a similar RMSD precision. This chapter is mainly based on Paper I and II.

3.1 Kink model and super-secondary structure

The (κ, τ) profile analysis in Sec.2.3 indicates that the minimization of energy (2.15) gives a particular pattern of (κ, τ) profiles: The angles (κ, τ) should have constant values at both sides far away from the kink center and vary a lot nearby the center. Fortunately, in protein such structures do exist and are named as super-secondary structures. The super-secondary structure consists of two secondary structures connected by a relatively flexible loop.
For secondary structure, the angles κ and τ take different approximately constant values, such as

$$\begin{cases} \kappa \approx 1.5 \\ \tau \approx 1 \end{cases} \quad \text{For } \alpha\text{-Helix} \tag{3.1}$$

and

$$\begin{cases} \kappa \approx 1 \\ \tau \approx \pi \end{cases} \quad \text{For } \beta\text{-Strand} \tag{3.2}$$

For the sites at loop, the values of the angles (especially for τ) fluctuate quite a lot. This behavior is well consistent with the (κ, τ) kink pattern obtained from the minimization of the energy (2.15). The only difference is that the κ angles in super-secondary structure are always positive, while the κ profile satisfying the equation (2.18) or ansatz (2.21) should have a sign change when it passes by the center of the kink. This difference can be eliminated by a Z(2) gauge (2.14) at some loop site.

Take the super-secondary structure segment site 8-34 from an entry with PDB code 1ABS as an example. The (κ, τ) profile calculated from PDB structure is shown in Figure 3.1 (a), and after Z(2) gauge transformation at site 20

the (κ, τ) profile is shown in Figure 3.1 (b). In Figure 3.1 (b), the (κ, τ) profile fitted from the ansatz is also presented as a comparison, which shows that the ansatz fits the data well by averaging the (κ, τ) fluctuations at the secondary structure.

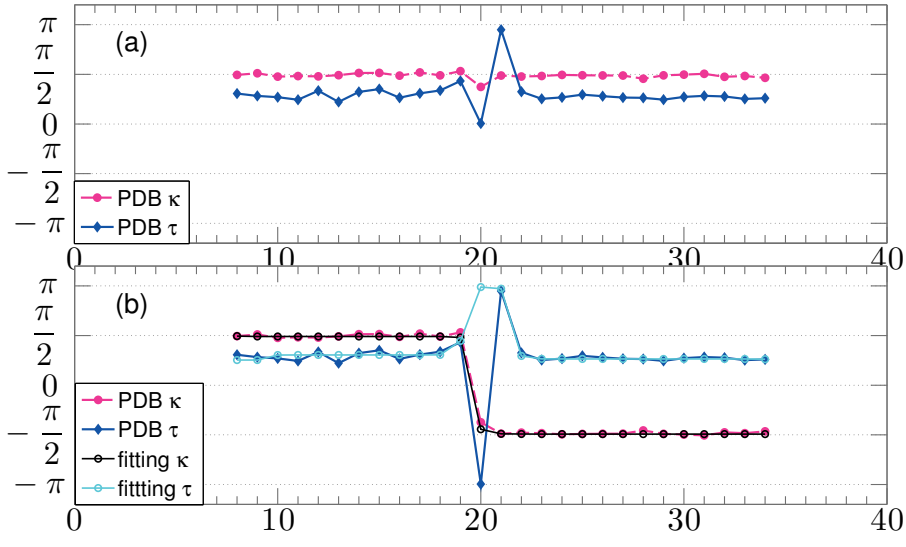

Figure 3.1. The (κ, τ) profiles for a super-secondary structure segment 8-34 in PDB entry 1ABS (a) and the profiles after Z(2) gauge (b). The x-axis denotes the site index for PDB entry 1ABS. In both subfigures, the κ angles calculated from PDB are in red color while τ angles from PDB are in blue. In figure (b), the fittings from the ansatz are also shown in black (for κ) and cyan (for τ).

To see the global fitting result, the 3D discrete curve is constructed from the fitted (κ, τ) profiles (the bond length is taken the constant value 3.8Å) and then the RMSD compared with the PDB structure is calculated. It turns out that the RMSD is as small as 0.4Å. Hence, this particular "helix-loop-helix" structure is well modeled by a corresponding kink.

In fact, the kink model does not only fit the "helix-loop-helix" super-secondary structure, but also other types such as "helix-loop-strand" shown in paper I (a segment with site 398 - 416 in PDB code 3DLK) and "strand-loop-strand"(a multi-kink model on PDB entry 3LL1 has been performed).

A priori it appears that the number of the parameters in describing a folded protein, might be quite large. This is true especially in Gō-like model. But since (2.15) models the super-secondary structure of the C_α-trace in terms of one kink, the number of parameters is often quite small compared with the degree of freedom. In the example of the super-secondary structure 8-34 in 1ABS, there are 27 C_α atoms in all. Assuming the bond length is a constant value 3.8 Å, there are still around 27×2=54 degree of freedom. Its kink model with at most nine parameters can describe this super-secondary struc-

ture as precise as $RMSD = 0.4$Å. Consequently the energy function (2.15) has a substantial predictive capacity compared with the Gō model and its variants [45, 46, 47, 48, 49, 50], where the number of parameters is usually more than the degree of freedom.

3.2 Universality

In a super-secondary structure, the secondary structure parts are quite regular and rigid, the relative orientation between the two secondary structure parts is determined by the loop part in the middle. In Sec.3.1, it has shown that a particular super-secondary structure can be described by one proper kink model with sub-Ångström RMSD precision. In other words, the kink model fits the loop part very well, which connects the neighboring secondary structures.

A question comes out: How many kink models are needed to describe most of the loop structures in PDB with some error tolerance in RMSD?

3.2.1 The experimental precision

To answer this question, a reasonable criteria for the error tolerance in RMSD should be determined first. Since I are doing modeling on given experimental structures in PDB, the precision in the PDB structure data should be considered. If the RMSD between the model and experimental structure is smaller than the experimental uncertainty, the model should be precise enough to describe the experimental structure.

In PDB, there are mainly two experimental methods in determining the protein structures: X-ray crystallography and Nuclear Magnetic Resonance (NMR). The NMR method is relatively new developed technique for protein structure analysis. It can measure the protein structures at room temperature, which is quite useful for those proteins hard to be crystallized. As the NMR experiment gives an ensemble of structures and the kink model currently is fitted to only one structure, it is not possible to use NMR ensemble as a reference currently. Hence in the statistics I only concentrate on the structures measured from X-ray crystallography.

In X-ray crystallography, the overall precision of the measurement is characterized by the resolution due to the equipment. Besides, for each particular atom in protein, there is another uncertainty in the PDB structure named as the Debye-Waller factor (also known as B-factor in PDB). The factor is to characterize the thermal motion uncertainty for each atom in protein structure due to the finite temperature [51]. For the protein crystallographical structures with resolution better than 2.0 Å in PDB, the B-factors of the C_α atoms are usually less than $B_{max} = 35$Å. According to the Debye-Waller relation, the

corresponding fluctuation distance are

$$\sqrt{\langle(\Delta x)^2\rangle_{max}} = \sqrt{\frac{B_{max}}{8\pi^2}} = 0.65\text{Å} \tag{3.3}$$

Based on this result, the criteria of the RMSD precision is set to 0.6 Å, which is a little smaller than the maximum uncertainty (3.3) derived from B-factor. If the RMSD between the kink model and one protein segment is less than this value, the segment is considered well modeled by this kink.

3.2.2 Kink model exhausting search in PDB

After setting up the RMSD precision criteria, it is time to estimate the number of the kinks that are needed in describing the loops in a PDB subset listed in Reference [52]. The protein subset is further confined to the proteins with resolution better than 2.0 Å since the criteria (3.3) is based on the statistics of proteins whose resolution is better than 2.0 Å. With this restriction, there are a total of 3,027 proteins containing 193,640 loop sites in the dataset. Note that the proteins in this dataset are less than 25% homology equivalence. So the selected dataset is large enough to represent the whole PDB dataset.
The detailed strategy for finding the total kink models in this dataset is as follows:

1. First an arbitrary super-secondary motif in the dataset is selected and modeled using a proper kink described in Sec.2.3. Since only the geometrical property is considered without any energy involved here, the kink ansatz equations (2.21) and (2.22) are used for simplification. Obviously, the structure of the kink model contains two parts of secondary structure besides the loop part. To translate the super-secondary-structure-model to a loop-model, the secondary structures are cut off as much as possible from the obtained kink, only leaving the loop structure.

2. Starting from the reduced kink model obtained from step 1, all the segments with RMSD less than 0.5Å compared with the model in the dataset are searched and removed off from the dataset.

3. From the remaining dataset, another different super-secondary motif is selected, and the steps 1 and 2 are repeated. This procedure will not stop until most of the loop sites are covered by the existing kinks.

In this way, a library with 200 kinks is established, covering 92% loop sites in the dataset when the criteria is 0.6 Å. The high percentage indicates that the kink-like structure is quite universal in loops. With very limited number of kinks, most of the loop structures can be described within the experimental uncertainty.

As an application of the 200 kinks library, in Paper II a long loop with 10 residues was modeled by directly joining their corresponding kinks together. After the connection, the RMSD of the loop modeling is as small as 0.31Å.

4. Kink model application to myoglobin and villin headpiece HP35

In this chapter, the multi-kink model is applied on two particular proteins– myoglobin and villin headpiece HP35. From the multi-kink model established for these two proteins, the thermal dynamical properties and the folding pathway are studied.

The analysis of the properties during the folding process shows that the results from multi-kink model are well consistent with the experimental evidences and other theoretical predictions for both proteins.

4.1 Kink model on myoglobin

Myoglobin was studied in Paper I, III and V. However, in Paper I, the multi-kink model was set up from ansatz (2.21) and (2.22), so it has little contributions to the thermal dynamics and folding process due to the lack of energy parameters. As a result, the application on myoglobin is mainly based on Paper III and V.

4.1.1 Myoglobin introduction

Myoglobin is a globular protein that plays a central role in oxygen transport and storage in muscle cells. It is the first protein determined by X-ray crystallography [53]. Myoglobin contains eight helices named from A to H, and these helices are connected together by short loops. Hence, myoglobin has quite typical "helix-loop-helix" super-secondary structures. The myoglobin, especially its folding pathway and substates, have been extensively studies both theoretically and experimentally [54, 55, 56, 57, 58, 59, 60]. The typical super-secondary structures and the extensive experimental evidences are the motivations for studying myoglobin in the viewpoint of multi-kink model.

4.1.2 Structure fitting

So far, there are totally 363 myoglobin entries in PDB. One particular myoglobin entry (PDB code 1ABS) has been selected as our modeling target. The reason for choosing this one lies in the fact that its crystallographical structure is measured near the liquid helium temperature (20K) [61] and hence has the

lowest B-factor among the myoglobin data in PDB. There are 154 residues indexing from 0 to 153 in this entry. However, for simplification, the flexible tails are cut off and only the main segment site 8-149 is left for the multi-kink modeling.

According to the property of the (κ, τ) profiles calculated from PDB entry, the segment is divided into ten different kinks using Z(2) gauge transformations. Note the number of kinks is more than that of the helices since some loop part is split into kink pairs. With this kink identification, I use the software *Propro* to find the proper parameters.

Finally, a ten-kink model is obtained with RMSD=0.81Å compared with the PDB structure. The 3D superimposition is shown as Figure 4.1, where the purple structure is from PDB while the cyan one is from the kink model. The

Figure 4.1. The 3D superimposition of the multi-kink model (cyan) and PDB structure (purple)

(κ, τ) profiles are shown in Figure 4.2, displaying how the kinks are divided along the backbone together with the fitting result between protein entry and the modeling. The Figure 4.2 (a) is the (κ, τ) profiles calculated directly from the PDB structure, while Figure 4.2 (b) gives the (κ, τ) profiles with particular Z(2) gauge transformations. The site where κ changes its sign indicates the centers of the kinks. In Figure 4.2 (b), the (κ, τ) profiles obtained from the multi-kink model are also presented, showing a precise multi-kink model for the particular myoglobin entry.

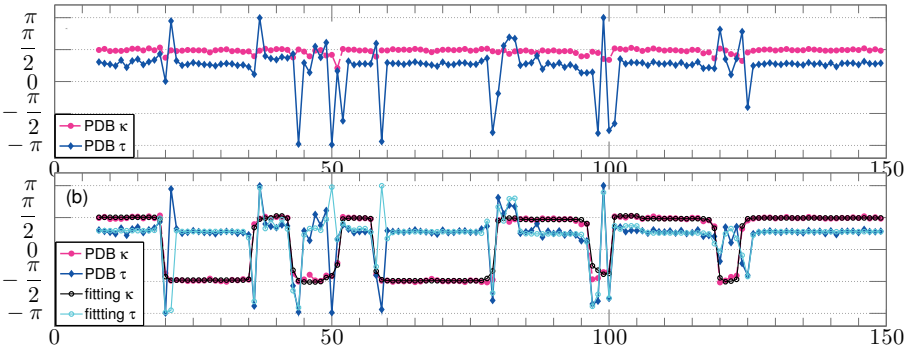

Figure 4.2. The (κ, τ) profiles calculated from PDB structure without Z(2) transformations (a) and after Z(2) transformations (b). In both figures, the red lines are κ and blue lines are τ with x-axis the site indices in PDB. In subfigure (b), the (κ, τ) profiles obtained from multi-kink model are also presented in black (for κ) and cyan (for τ).

4.1.3 Non-equilibrium dynamics simulation on myoglobin

Starting from the multi-kink model obtained for myoglobin and following the methodology described in Sec.2.6, the non-equilibrium heating-cooling simulation for myoglobin is performed on a Mac Desktop. To make the protein fully thermalized and the process to be adiabatic enough, the MC steps and temperatures are set to $N_1 = N_2 = N_3 = 5M$ and $\mathrm{T}_{low} = 10^{-17}$, $\mathrm{T}_{high} = 10^{-4}$. Such a heating-cooling cycles are taken 100 times to get the statistical properties.

The simulation is quite time efficient. Such a complete unfolding and folding cycle takes, no more than 3 min *in silico* time on a single processor. It should be noted that 15M MC steps are not necessary if one just require that the myoglobin unfolds to a random configuration at high temperature and then correctly fold back ($RMSD < 2$Å compared with the native state) with temperature cooled down. According to my test, the fastest unfolding-folding cycle simulation with correctly folding back to native state only needs 550K MC steps, taking around 10s *in silico* time on a single processor. This unfolding-folding simulation time is comparable with the real myoglobin folding time (around 2.5 s).

4.1.4 Simulation result analysis

Based on the simulations, the results were analyzed in several different aspects, including the RMSD and Radius of Gyration (Rg) evolution, Nucleation sites and folding rate ratio estimation. Finally, the gate dynamics is also analyzed during the stage from native state to molten globule state.

Rg and RMSD in heating-cooling cycle

Rg is an important quantity to characterize the overall shape of the protein, especially when there are some phase transitions in structure. According to the definition, Rg is calculated by the formula

$$R_g = \sqrt{\frac{1}{2N^2}\sum_{i,j}(\mathbf{r}_i - \mathbf{r}_j)^2} \tag{4.1}$$

The average values of RMSD and Rg for the 100 repeated heating-cooling cycles are calculated. Figure 4.3 shows the evolution of the average Rg together with its fluctuation during heating-cooling cycle. It displays a clear intermediate state (molten globule) no matter in heating or cooling procedure. The value of Rg at the intermediate state is around 23 Å and at the unfolded state is about 32 Å, which are both consistent with the experimental result [55].

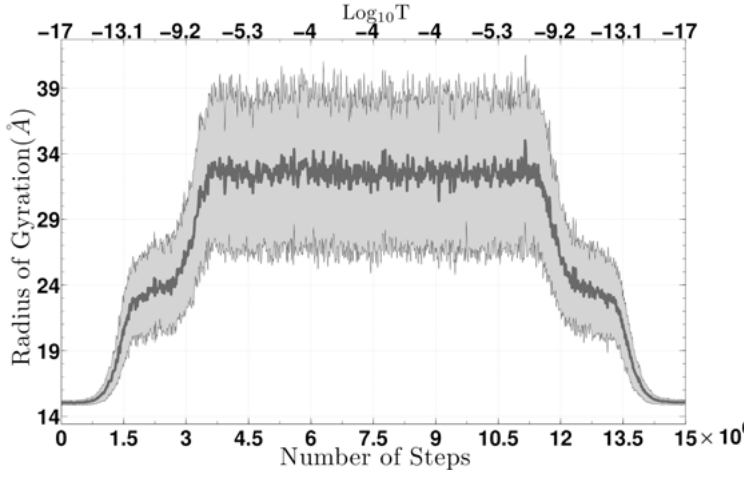

Figure 4.3. The evolution of Rg during the heating-cooling cycles. The darker line is the average values of Rg for 100 heating-cooling cycles while the lighter shaded area is the fluctuation among 100 cycles.

To identify the temperatures at transition points, the susceptibility of the RMSD curve is calculated as

$$\chi_{RMSD}(T) = \frac{\partial RMSD(T)}{\partial \log_{10} T} \tag{4.2}$$

where the $RMSD(T)$ curve can be fitted by a function in Reference [62]

$$<RMSD(T)> = h_1 + h_2 \arctan[h_3(x-x_1)] + h_4 x \arctan[h_5(x-x_2)] - h_6 x \tag{4.3}$$

with $x = \log_{10} T$.

Once the susceptibility is obtained, the transition points can be identified at the peaks of the susceptibility. During the cooling process of myoglobin, two transition points are identified at $T1 = 10^{-8.64}$ and $T2 = 10^{-13.56}$ as shown in Figure 4.4.

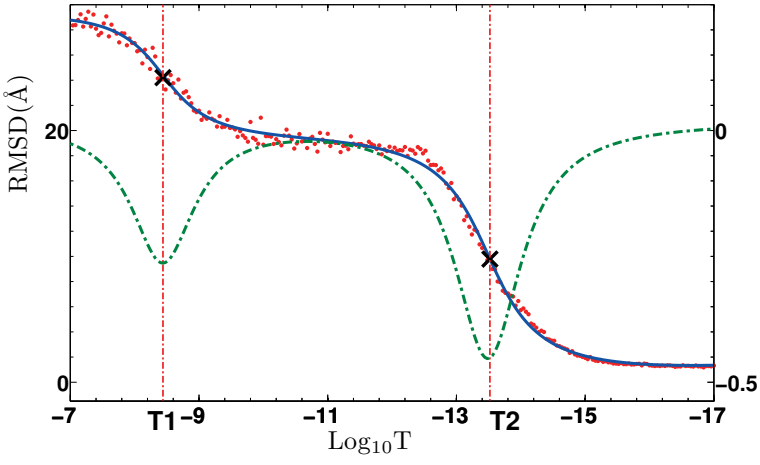

Figure 4.4. The susceptibility of the RMSD evolution during cooling process. The red dots are the average values of Rg at different temperatures for 100 repeated heating-cooling cycles, the blue solid line is the fitting result from equation (4.3), and the green dashed line is the corresponding susceptibility calculated from equation (4.2). The two transition temperatures are labeled as T1 and T2.

For Rg, the corresponding susceptibility χ_g can be defined similar to equation (4.2) with a corresponding fitting function (4.3).

Folding pathway

From the heating-cooling simulation, a clear folding pathway can be identified. Since the geometry of the myoglobin is determined by the angles (κ_i, τ_i), the conformation changes of the myoglobin during folding process, in particular the α-helix nucleation phenomenon, should be reflected on these angles. For this, the evolution of fluctuations in these angles are calculated. Since the torsion angles τ are much more flexible than the bond angles κ, as a consequence the nucleation sites are more sensitive to the fluctuations of the τ angles during the heating and cooling process.

The fluctuation of τ at certain temperature is calculated from the sampled conformations at corresponding temperature by the equation

$$\Delta\tau_i = \sqrt{\frac{\sum_{k=1}^{100}(\tau_{i,k} - \bar{\tau}_i)^2}{100}} \tag{4.4}$$

where i is the PDB site index, k is the index of the 100 cycles, and $\bar{\tau}_i$ is the average τ value of at i-th C_α atom over the 100 cycles. The quantity $\Delta\tau_i$ is

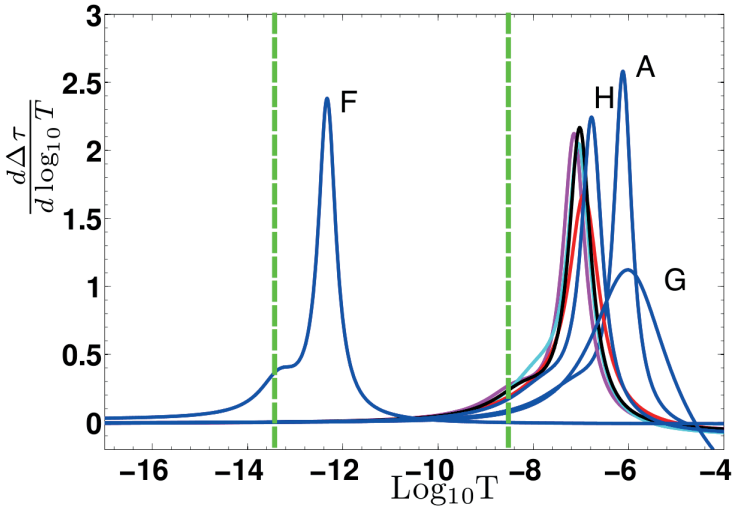

Figure 4.5. The derivative of $\Delta\tau_X$ (4.5) to MC temperature $\log_{10} T$ for the eight helices in myoglobin. The helices A, F, G ,H are noted near their peaks in the plot. The helices B, C, D, E are identified by colors magenta, cyan, red, black, respectively. The green dashed lines denote the peaks of susceptibility of Rg in heat-cooling cycles.

to characterize the stability of the residues along the protein backbone. For a helical segment X, the average τ fluctuation is defined as

$$\Delta\tau_X = \sqrt{\frac{\sum_{i\in X}\Delta^2\tau_i}{\mid X \mid}} \tag{4.5}$$

with $\mid X \mid$ the length of the helical segment X.

The fluctuation $\Delta\tau_X$ for eight helices are shown in Figure 4.5, indicating different stabilities for the eight different helices and the folding pathway in myoglobin.

According to Figure 4.5 and Figure 4 in Paper V, during the folding process the formation of the helices are: $G \rightarrow A$, H, $partial\ D \rightarrow D$, B, C, $E \rightarrow F$. In particular, as temperature is cooled down, $\Delta\tau_i$ at site 105-118 on Helix G first drop down (See Figure 4 (B) in Paper V), which means these sites form a helix nucleus first. These results are fully consistent with the theoretical nucleation prediction [58] as well as the experimental result [59]. The only inconsistence is that all helices except F helix are formed before the molten globule is formed in the simulation because of the over-stability of the secondary structure in multi-kink model.

Folding rate estimation

As is shown in Sec.4.1.4, there are two transition points at $T1 = 10^{-8.64}$ and $T2 = 10^{-13.56}$ in the stages from unfolded to molten globule and molten globule to native state, respectively. Assuming the protein folding rate k is linearly

proportional to the MC temperature T, the ratio of the folding time between unfolded to molten globule and molten globule to native structure can be estimated as

$$\frac{t_{unfolded-molten}}{t_{molten-native}} = \frac{1/k_{unfolded-molten}}{1/k_{molten-native}} = \frac{1/T1}{1/T2} = 1.25 \times 10^{-5}. \tag{4.6}$$

The ratio of the folding time between unfolded-molten and molten-native is just a slightly faster than the experimental result [59], which is from $\frac{1\ ms}{2.5\ s} = 4 \times 10^{-4}$ to $\frac{300\ \mu s}{2.5\ s} = 1.20 \times 10^{-5}$.

Ligand Gate Dynamics

The function of the myoglobin is to bind and transport the oxygen molecules through the HEME group. It is important to know the mechanism of the oxygen transport in myoglobin, especially the role of backbone conformation changes during this process.

In this section, the affection of the backbone conformation on the ligand transportation is monitored by the structure stability near to the HEME group in the heating-cooling cycle. The oxygen transport happens at room temperature or physiological temperature, which is corresponding to the transitions between native state to molten globule state. Before the analysis, three gates surrounding the HEME group for the ligands binding are identified as is shown in Figure 4.6. Each gate is formed by two non-neighboring secondary structures with eight residues long, in detail: Gate1 is composed of 37 (PRO) – 44 (ASP) and 96 (LYS) – 103 (TYR) colored in blue, Gate2 consists of 61 (LEU) – 68 (VAL) and 89 (LEU) – 96 (LYS) colored in cyan, and Gate3 is 25 (GLY) – 32 (LEU) and 106 (PHE) – 113 (HIS) colored in red, as is shown in Figure 4.6. The Gate distance is defined as:

$$d_i = \sqrt{\sum_{n=1}^{n=8} (x_i[n] - y_i[n])^2} \quad (i = 1,2,3) \tag{4.7}$$

where $x_i[n]$ and $y_i[n]$ are the coordinates of n-th C_α atoms on both sides in the i-th Gate. During the heating-cooling cycle, the distances of all three gates increase with temperature heated up and decrease with temperature cooled down. However, the stabilities of three gates are different. To characterize the different stabilities, the ratios between different gate distances are defined

$$\frac{Gate1}{Gate2} = \frac{d_1}{d_2}, \quad \frac{Gate3}{Gate2} = \frac{d_3}{d_2}, \quad \frac{Gate3}{Gate1} = \frac{d_3}{d_1} \tag{4.8}$$

during the heating-cooling cycle.

The evolutions of different gate ratios with MC temperature are shown in Figure 4.7. As shown in Figure 4.7, the distance of Gate3 changes much faster than the other two, which means that Gate3 is the most sensitive one to the

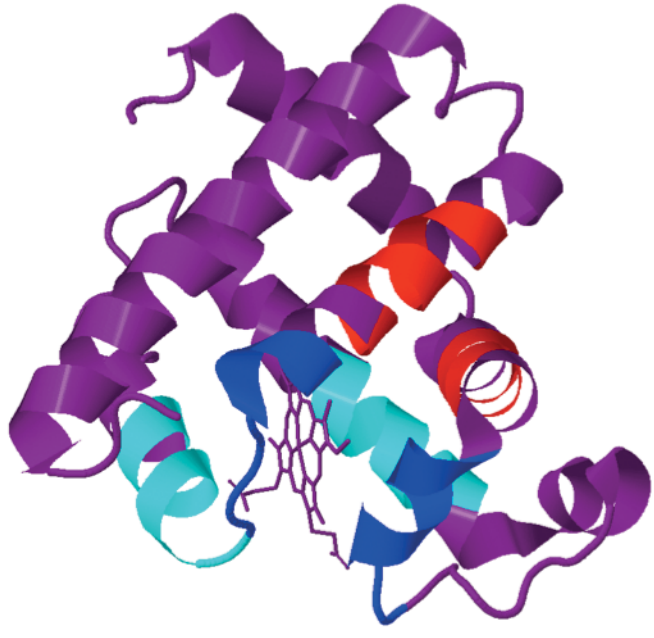

Figure 4.6. The three gates surrounding the HEME group. Gate1 is colored by blue, Gate2 is colored by cyan,and Gate3 is in red. The HEME group lies in the cavity surrounding by these three gates.

temperature changes. It implies that during the ligands transport process, the Gate3 plays an more important role if only the backbone effects are considered. The conclusion is consistent with the experiment that reveals L29 is critical during the ligand transport [63].

4.2 Kink model on villin headpiece HP35

Following the steps of modeling myoglobin, the villin headpiece HP35 was studied. Villin headpiece HP35 is a subdomain with 35 residues, and is usually viewed as the "the hydrogen atom of protein folding" . It has been well studied both experimentally [64, 65] and theoretically [66, 67, 68, 69]. It is also the first protein studied from the viewpoint of kink model [35, 36]. Here I mainly summarize its dynamical properties based on Paper IV.

4.2.1 Structure fitting

The target for HP35 is selected as the PDB entry 1YRF, which is measured at 95K with a resolution 1.07Å[70]. The flexible tail of the entry is cut off

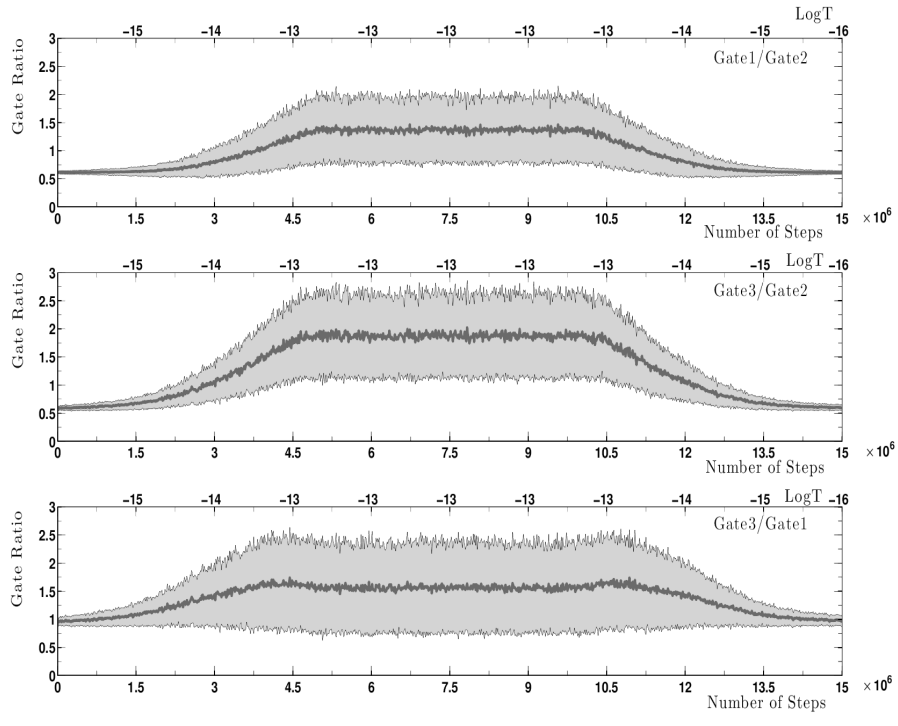

Figure 4.7. The gate ratio evolutions during heating-cooling cycle. The dark lines are the average values for the 100 cycles and the lighter shaded areas denote the fluctuations. The subfigures from top to bottom are for gate ratios $\frac{Gate1}{Gate2}$, $\frac{Gate3}{Gate2}$, $\frac{Gate3}{Gate1}$, respectively.

and only the segment 47–73 is left for modeling. Similar to myoglobin, the parameters in the energy (2.15) are trained by *Propro* and finally give the RMSD between the model and the corresponding PDB entry around 0.51Å. The 3D superimposition in Figure 4.8 shows a good fitting in configuration.

The (κ, τ) profiles calculated from PDB structure indicate that there are two kinks in this small segment as shown in Figure 4.9. It should be reminded that this protein is composed of three helices connected by loops, which gives a good reason for the double-kink profile. In Figure 4.9 (b), the comparison between the PDB structure and the fitted model shows a good consistence.

4.2.2 Non-equilibrium dynamics simulation on villin headpiece

The heating-cooling simulation procedure is the same as myoglobin except the detailed MC parameters. For villin headpiece the MC steps $N_1 = N_2 = N_3 = 1.5M$ and $T_{low} = 10^{-12}$. Some other simulations with different MC steps have been tested, but they give similar results. For the high temperature, several

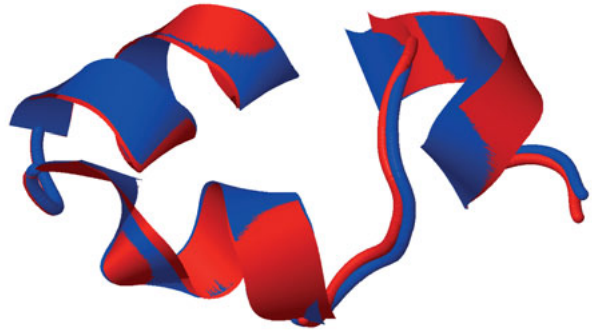

Figure 4.8. The 3D superimposition of the multi-kink model (blue) and PDB structure (red)

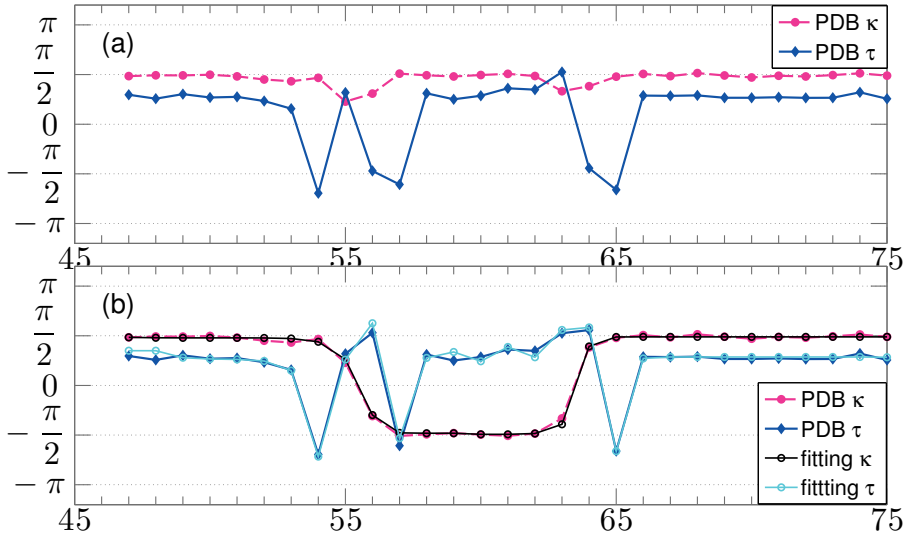

Figure 4.9. The (κ, τ) profiles calculated from PDB structure (top) and multi-kink model fitting on Z(2) gauged (κ, τ) profiles (bottom). In both figures, the red lines are κ and blue lines are τ calculated from PDB structure with x-axis the site indices in PDB. The black line and cyan line in bottom subfigure are the fitted κ angles and τ angles, respectively.

different values are taken from $T_{high} = 10^{-10}$ to $T_{high} = 10^{4}$ to check the upper limit of T_{high} when the protein still folds back with RMSD<1Å.

It is found that when high temperature T_{high} exceeds a temperature $T_{limit} = 10^{-2.4}$, the protein becomes quite difficult to fold back to the native state. In Figure 4.10, the RMSD evolutions under different high temperatures nearby

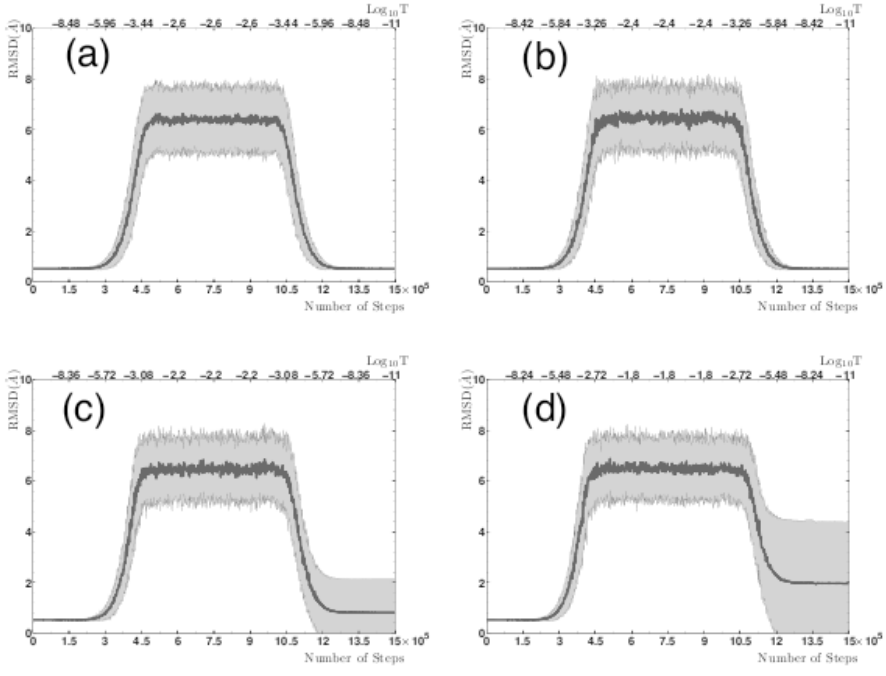

Figure 4.10. The RMSD evolutions during heating-cooling cycle at different high temperatures T_{high}. The darker lines are the average RMSD values for 100 heating-cooling cycles, while the lighter shaded areas are the corresponding fluctuations. (a) $T_{high} = 10^{-2.6}$ (b) $T_{high} = 10^{-2.4}$ (c) $T_{high} = 10^{-2.2}$ (d) $T_{high} = 10^{-1.8}$

$T_{limit} = 10^{-2.4}$ are shown. It can be seen that at $T_{high} = 10^{-2.4}$, the final RMSD value after heating-cooling cycle is still quite small and there is no obvious fluctuations. But at $T_{high} = 10^{-2.2}$, the final RMSD distribution expanded a lot because the misfolding states start to appear. When the high temperature is increased further as $T_{high} = 10^{-1.8}$, even the average value of the final RMSD is larger than 2Å, which means the misfolding states occupying large percentage at the end of the 100 heating-cooling cycles now.

Besides, during the heating-cooling cycles with different T_{high}, no intermediate state is found in all the RMSD evolution pictures. This property is different from the case of myoglobin, where the intermediate state (molten globule) is obvious. However, this result is consistent with the experimental evidence [65], which showed villin headpiece was described well with two-state model.

4.2.3 Folding pathway and misfolding state

In the heating-cooling cycles when the protein can still fold back, the folding path way is directly identified by looking at the 3D conformations during the heating-cooling cycle. Some 3D conformation shots during the cooling process are shown in Figure 4.11. In Figure 4.11, the helices A, B, C are colored in red, green and blue, respectively. The lighter shaded configuration represents the native state of villin headpiece 1YRF. It is easy to see that order of the helices formation is C $\rightarrow$ B $\rightarrow$ A. This conclusion is well consistent with the result from MD simulations [66, 67, 68].

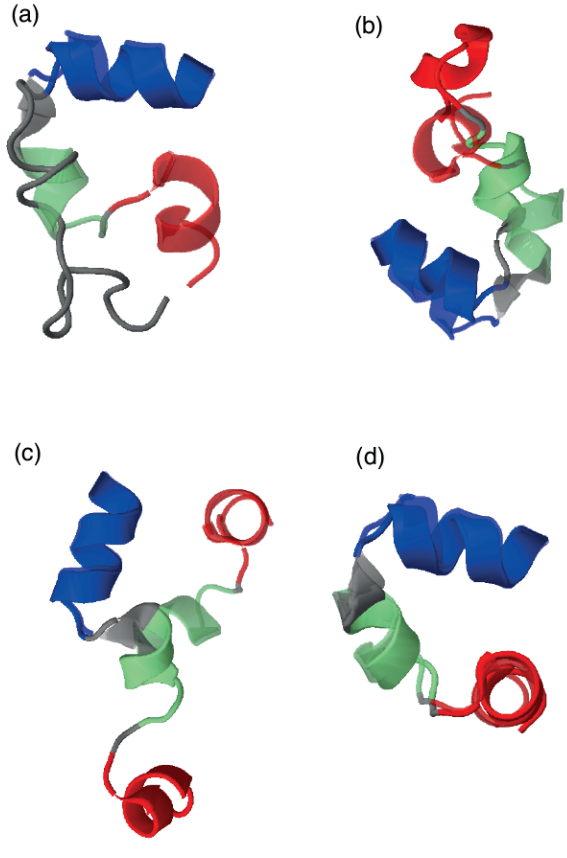

Figure 4.11. The folding pathway obtained from kink-model during the cooling process. The helices A, B, C are colored in red, green and blue, respectively. The native state configuration is colored in lighter shades, while the configurations during the cooling process are in dark colors. The cooling process is in the order of (a) $\rightarrow$ (b) $\rightarrow$ (c) $\rightarrow$ (d).

During the heating-cooling cycle when $T_{high} > 10^{-2.4}$, the misfolding configuration starts to appear. In Figure 4.12, a most frequent misfolding config-

uration from the kink-model in heating and cooling cycle is shown. It should be noticed that this misfolding configuration is consistent with the meta-stable configuration 5 in Reference [69].

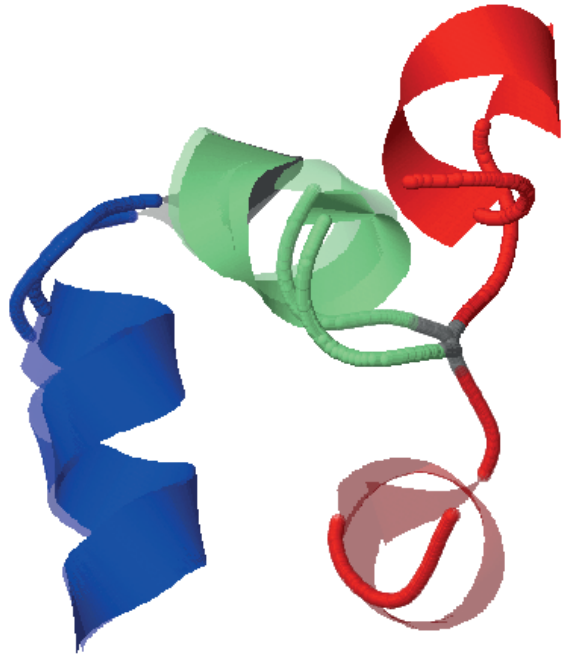

Figure 4.12. A most frequent misfolding configuration from kink model during the heating-cooling cycle. The helices A, B, C are colored in red, green and blue, respectively. The native state configuration is colored in lighter shades, while the misfolding configuration is in dark color.

5. Side-chain visualization in Frenet-Based frames

In previous chapters, we focused on modeling the C_α trace of the protein backbone. The reason for paying much attention on C_α trace modeling is the fact that once the C_α trace is known, the full atom representation of the protein can be obtained by side-chain statistical properties from PDB. Many algorithms have been developed to do the full atom completion in protein, such as PULCHRA [71], NEST [72], MAXSPROUT [73] and SCRAWL [74]. Most of the algorithms are based on the side-chain rotamer libraries. The rotamer library can be backbone independent or dependent [75, 76, 77], showing there are well localized distribution patterns for the side-chain atoms.

In reference [78, 79], it showed the distribution of the C_β atoms in DF-frame as well as C_γ atoms in C_β frame and their advantages in secondary structure dependence. In Paper VI, this method is developed to higher level side-chains and C_β atom distributions are compared with those in the frames defined by REMO [80] and PULCHRA [71].

In this chapter, I want to summarize some other findings in paper VI. Note the dataset for statistics is confined to the X-ray crystallographic protein entries with resolution better than 1.0 Å. Up to the time when statistics was done, there were around 500 entries in the dataset.

5.1 C_β atom rotamer revisit

C_β atom is directly connected on the C_α atom, so it is a natural choice to observe the C_β atom in DF-frame centered on the corresponding C_α atom. The distribution of all C_β atoms are shown in Figure 5.1.1 (a), where the secondary structure has been identified.

5.1.1 C_β atom at terminal

It is frequently presumed that the termini are unstructured and highly flexible. However, according to the distribution of the C_β atoms at terminus in the DF-frame, this is not the case for C_β atoms. By locating the first two and last two C_β atoms along the protein chains in DF-frame, the distribution of the C_β atoms at terminus is shown in Figure 5.1.1 (b). It shows that in the DF-frames the orientations of the two terminal C_β atoms are highly regular. Their positions on the surface of the C_α centered sphere are fully in line with that of all C_β atoms in Figure 5.1.1 (a). In particular, there are very few outliers.

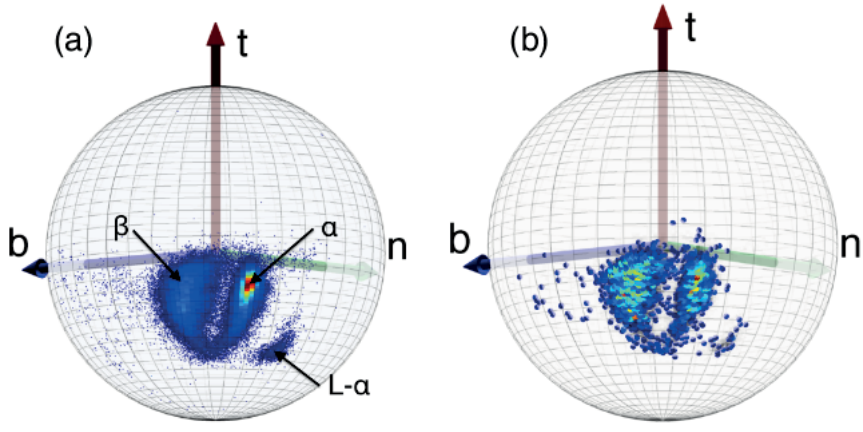

Figure 5.1. The distributions of all C_β atoms (a) and C_β atoms in terminus (b) in DF-frame

5.1.2 C_β and proline

Proline is very special among the 20 kinds of amino acids because its side-chain connects back to the backbone nitrogen atom. This particularity makes the positions of the proline as well as its neighboring amino acids unusual. In particular, the peptide plane of the proline is much easier to form cis-structure than other amino acids. The distributions of C_β for *cis* and *trans* proline in DF-frame are shown in Figure 5.2. As shown in Figure 5.2 (a), the distribution of *trans* proline is consistent with the the mainland of all C_β distribution. However, the *cis*-proline are located outside of the main C_β distribution as shown in Figure 5.2 (b).

Figures 5.3 (a)-(d) shows the distributions of the C_β carbons that are located either *immediately after* or *right before* a proline, where the grey background is the mainland of all C_β distribution. In (a) and (b), the C_β atoms after *trans* and *cis* proline are displayed, respectively. Subfigures (c) and (d) are the distributions of the C_β atoms before *trans* and *cis* proline. By comparison of (a) and (c) it can be concluded that the C_β atoms after *trans*-proline match the background well while before *trans*-proline C_β atoms have high preference in β-strand or loop region. For *cis*-proline in subfigure (b) and (d), since the number is not large enough, the distribution does not have very clear area with high density. However, it seems the C_β atoms after *cis*- proline prefer β-strand region a little and C_β atoms before *cis*-proline definitely stay far away from the mainland of usual C_β atoms distribution.

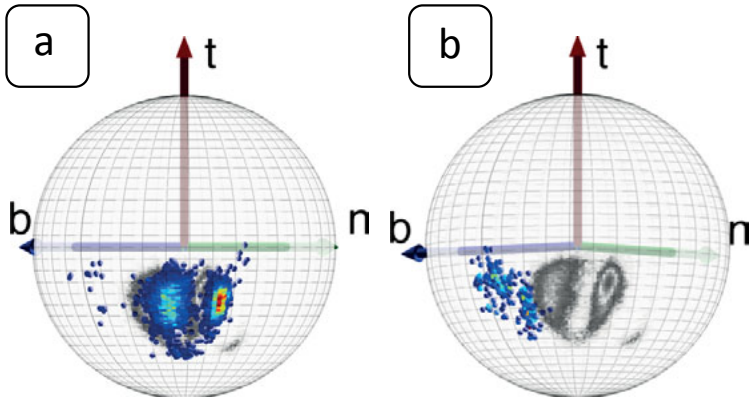

Figure 5.2. The distributions of C_β atoms in DF-frame for *trans*-PRO (a) and *cis*-PRO. The gray background is the mainland of the distribution for all C_β atoms.

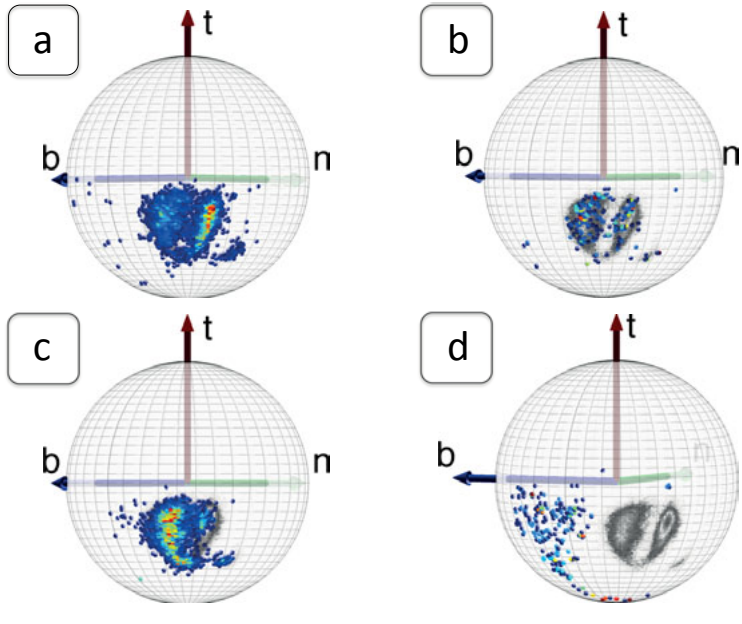

Figure 5.3. The distributions of C_β atoms in DF-frame for the residues neighboring PRO. (a) C_β atom after *trans*-PRO (b) C_β atom after *cis*-PRO (c) C_β atom before *trans*-PRO (d) C_β atom before *cis*-PRO.

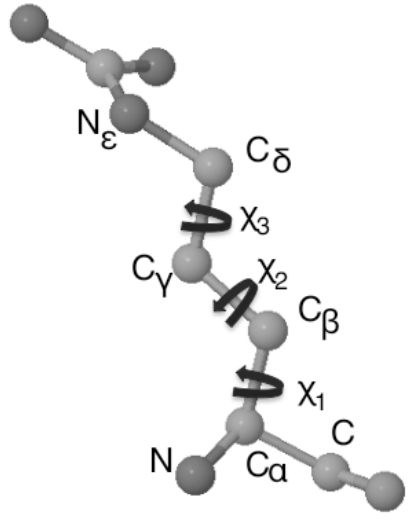

Figure 5.4. Definitions of the dihedral angles χ_1, χ_2, χ_3.

5.2 Higher level rotamers

5.2.1 C_γ atom rotamers

The conventional rotamer libraries are characterized by a set of dihedral angles ($\chi_1, \chi_2, \chi_3, \chi_4$) along the side-chain for different levels of atoms. Each dihedral angle is determined by four consecutive atoms. The definitions of χ_1 and χ_2 are shown as Figure 5.4. Clearly, χ_1 is the dihedral angle defined by N,C_α,C_β,C_γ, while χ_2 is among C_α,C_β,C_γ,C_δ, and similarly for other angles. According to the χ_1 distribution, there are three different rotamers in C_γ: *gauche* $\pm$(g$\pm$) and *trans*(t).

In fact, based on the definition of dihedral angle χ_1, a χ_1-frame can be defined as:

$$\mathbf{t}_{\chi_1} = \frac{\mathbf{r}_\beta - \mathbf{r}_\alpha}{|\mathbf{r}_\beta - \mathbf{r}_\alpha|} \tag{5.1}$$

$$\mathbf{n}_{\chi_1} = \frac{\mathbf{s} - \mathbf{t}_{\chi_1}(\mathbf{s}\cdot\mathbf{t}_{\chi_1})}{|\mathbf{s} - \mathbf{t}_{\chi_1}(\mathbf{s}\cdot\mathbf{t}_{\chi_1})|} \quad \text{where} \quad \mathbf{s} = \mathbf{r}_\alpha - \mathbf{r}_\mathrm{N} \tag{5.2}$$

$$\mathbf{b}_{\chi_1} = \mathbf{t}_{\chi_1} \times \mathbf{n}_{\chi_1} \tag{5.3}$$

with $\mathbf{r}_\alpha$, $\mathbf{r}_\beta$ and $\mathbf{r}_N$ the coordinates of the pertinent C_α, C_β and N atoms, respectively. In this frame, if one looks at the C_γ atom centered on the corresponding C_β atom, the longitude of C_γ atom in this frame is just the χ_1 angle. The distribution of C_γ atom in χ_1 frame is shown in Figure 5.5 (a). As is seen from Figure 5.5 (a), the distribution of C_γ atoms is a narrow circle on the

sphere centered on C_β atom. To make the visualization more clear, a generalized stereographical projection is introduced as follows:
Assuming the latitude angles is θ and longitude is ϕ, the generalized stereographical projection is

$$x + iy = f(\theta)e^{i\phi} \tag{5.4}$$

$$f(\theta) = \frac{1}{1+\exp\{\theta^2\}} \tag{5.5}$$

where x and y are the coordinates of the stereographic plane.
Figure 5.5 (b) shows the result after stereographic projection with rotamers identified. The radius is corresponding to $f(\theta)$ while the polar angle is ϕ. Apparently, this frame is just another form of visualization for the χ_1 rotamers and the rotamers of $g\pm$ and t has no clear secondary structure dependence.

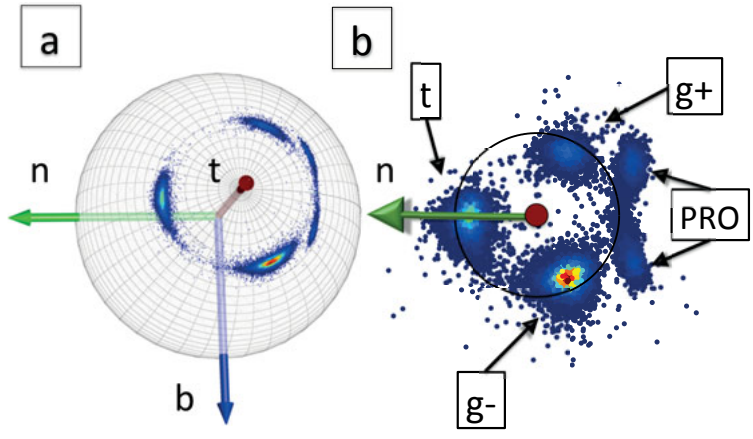

Figure 5.5. The distribution of C_γ atoms in χ_1 frame on a C_β centered sphere (a) and the corresponding stereographic projection (b). The different rotamers $t,g\pm$ together with PRO are identified in stereographic projection (b).

However, there is an alternative frame named C_β frame defined as:

$$\mathbf{t}_\beta = \frac{\mathbf{r}_\beta - \mathbf{r}_\alpha}{|\mathbf{r}_\beta - \mathbf{r}_\alpha|} \tag{5.6}$$

$$\mathbf{n}_\beta = \frac{\mathbf{t}_\beta \times \mathbf{t}_\alpha}{|\mathbf{t}_\beta \times \mathbf{t}_\alpha|} \tag{5.7}$$

$$\mathbf{b}_\beta = \mathbf{t}_\beta \times \mathbf{n}_\beta \tag{5.8}$$

If one is located at each C_β atom, and observes the corresponding C_γ atom in the C_β frames, each C_γ rotamer in χ_1 frame is split into two parts corresponding to Helix and Strand, respectively. Figure 5.6 shows the result in C_β

frames after the stereographical projection (5.5), where both the conventional rotamers and the secondary structures are specified.

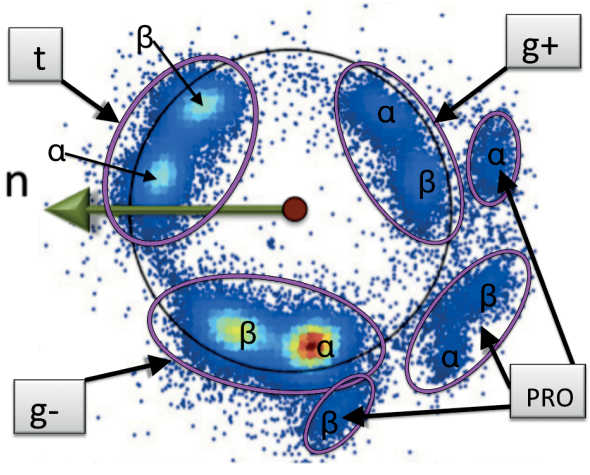

Figure 5.6. The distribution of C_γ atoms in C_β frame. Both the rotamers including *t*,*g*±,PRO and the secondary structures are labeled in the figure.

5.2.2 C_δ and higher level rotamers

For even higher side-chain levels, a uniform definition is provided. Assuming the atom level we are studying is corresponding to χ_i (i=2,3,4), and this dihedral angle is defined by four consecutive atoms A,B,C,D. In this notation, the distribution of atom D is what we want to study, and atoms A,B,C,D are in the ascending order of the distances to the backbone. The corresponding χ_i-frame can be defined as:

$$\mathbf{t}_{\chi_i} = \frac{\mathbf{r}_C - \mathbf{r}_B}{|\mathbf{r}_C - \mathbf{r}_B|}$$

$$\mathbf{n}_{\chi_i} = \frac{\mathbf{t}_{\chi_i} \times \mathbf{t}_{\chi_{i-1}}}{|t_{\chi_i} \times \mathbf{t}_{\chi_{i-1}}|} \quad \text{where} \quad \mathbf{t}_{\chi_{i-1}} = \frac{\mathbf{r}_B - \mathbf{r}_A}{|\mathbf{r}_B - \mathbf{r}_A|}$$

$$\mathbf{b}_{\chi_i} = \mathbf{t}_{\chi_i} \times \mathbf{n}_{\chi_i}$$

with $\mathbf{r}_A$, $\mathbf{r}_B$, $\mathbf{r}_C$ the coordinates of atoms A, B and C, respectively. With this χ_i frame, the conventional χ_i angles are the longitudes with a clockwise rotation of 90°. Similar to the χ_1 frame, the side-chain distribution dose not depend on the secondary structure explicitly in these χ_i frames.
Correspondingly, another alternative frame for observing atom D can be defined as:

$$\mathbf{t}_C = \frac{\mathbf{r}_C - \mathbf{r}_B}{|\mathbf{r}_C - \mathbf{r}_B|} \tag{5.9}$$

$$\mathbf{n}_C = \frac{\mathbf{t}_C \times \mathbf{t}_\alpha}{|\mathbf{t}_C \times \mathbf{t}_\alpha|} \tag{5.10}$$

$$\mathbf{b}_C = \mathbf{t}_C \times \mathbf{n}_C \tag{5.11}$$

where $\mathbf{t}_\alpha$ is the tangent vector of the DF-frame on the corresponding C_α atom. Since these frames contains the information on the corresponding DF-frames, they are given a name of Frenet-based frames. Located at atom C observing the atom D in this frame, one should be able to get the secondary structure dependent rotamers automatically.

As an example for these frames, the distributions of the non-aromatic C_δ atoms are shown in Figure 5.7 and 5.8 . It should be noted that starting from δ level, the chemical structure of the side-chain can be classified into two kinds depending on whether there is benzyl-like structure in the side-chain. If there is a benzyl-like structure in the side-chain, it is aromatic; otherwise, it is non-aromatic. The distributions for the C_δ atoms in aromatic and non-aromatic are completely different. Here only the non-aromatic C_δ atoms distributions are shown. Figure 5.7 shows the C_δ atom distribution in χ_2 frame while Figure 5.8 shows the distribution in its corresponding Frenet-based frame – C_γ frame. In Figure 5.8 the secondary structures are clearly identified in subfigure (b) (for α-Helix) and (c) (for β-strand). Similar distributions for even higher levels are obtained in Paper VI.

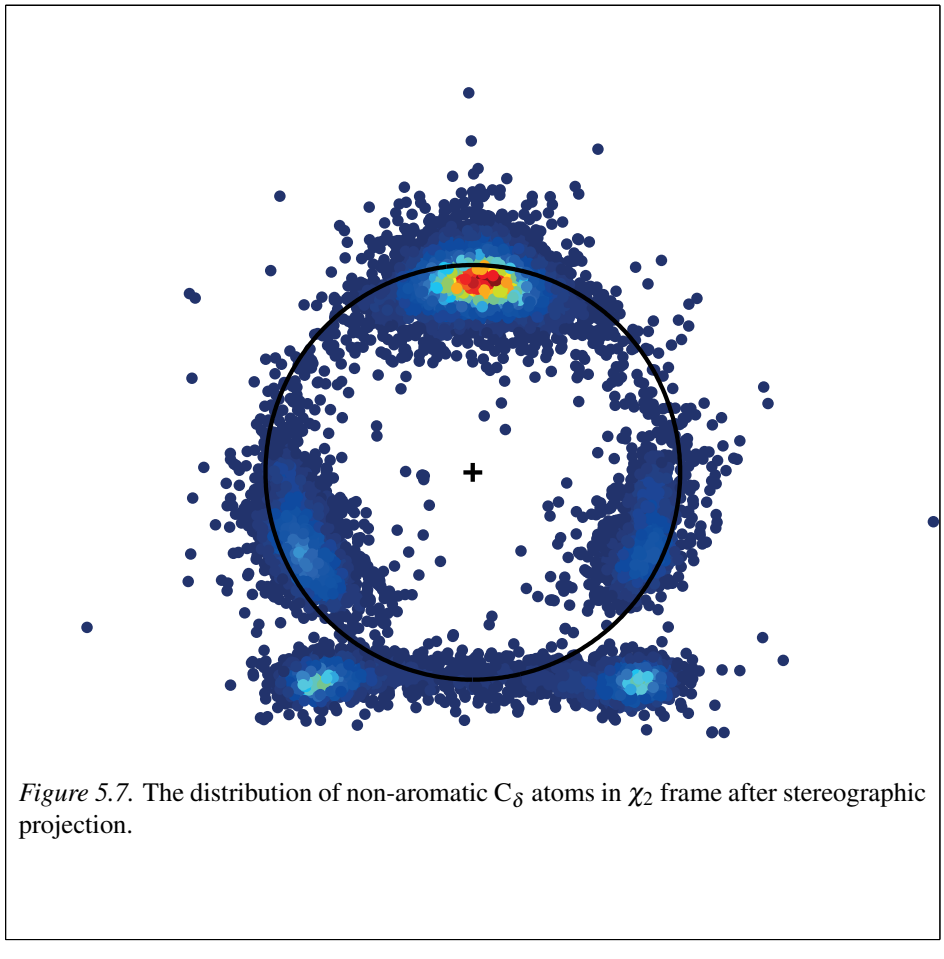

Figure 5.7. The distribution of non-aromatic C_δ atoms in χ_2 frame after stereographic projection.

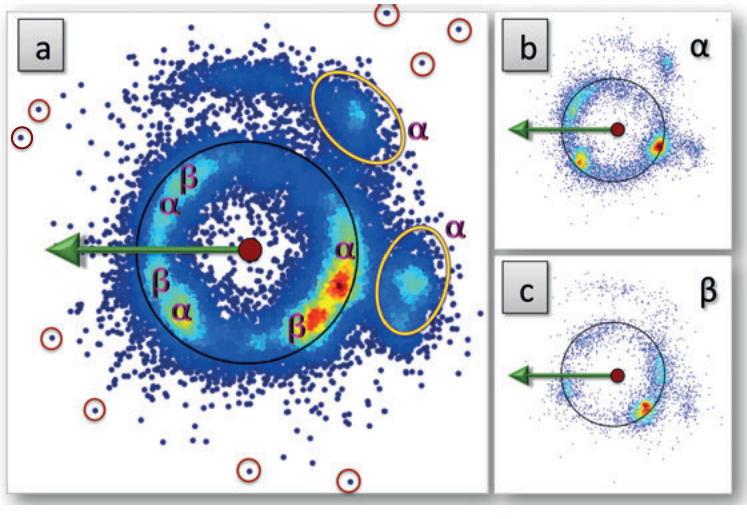

Figure 5.8. The distribution of non-aromatic C_δ atoms in its Frenet-based frame after stereographic projection. (a) The distribution for all non-aromatic C_δ atoms. The secondary structures are labeled in the figure, and some outliers are encircled. (b) The distribution for non-aromatic C_δ atoms on helix (c) The distribution for non-aromatic C_δ atoms on β-strand.

6. Conclusion

In this thesis, the theory of the multi-kink model in protein backbone (C_α trace) modeling is reviewed. It shows that most loops in PDB can be described by limited number of kinks. From the multi-kink model, two specific proteins–myoglobin and villin headpiece HP35 are modeled with RMSD<1Å precision. Based on the structure modeling, the dynamical properties and folding pathway are studied for both proteins, showing highly consistent with the experimental results. In the last chapter, the distributions of the side-chain atoms in Frenet-based frames are studied. With the Frenent-based frames, the distributions of the side-chains at different levels can be both localized and secondary structure dependent.

It turns out the multi-kink model is an effective model in describing particular protein and finding its folding pathway. Recall the three problems in the protein folding discussed in Sec.1.2, at the moment the multi-kink model are mainly approaching the second problem, i.e., the folding pathway of protein. The applications in myoglobin shows that the multi-kink model have great advantages in the protein folding pathway sampling. For the first and third problems, the physical interpretation of the parameters in energy function need to be further clarified. In particular, the relations between the parameters in energy (2.15) and sequence information need to be known for the tertiary structure prediction problem.

Acknowledgements

First of all I would like to thank my supervisor Antti. Thank you for your great help and guidance on my PhD study, especially the protein folding research. It is your profound insight in protein folding and physics that gives me lots of confidence on my research. During the time working with you, I learned quite a lot from you on every aspect including thinking, writing, reporting, collaborating and so on. I really enjoyed that time very much.

I also appreciate every member in our *Folding Proteins* group over these years: Andrey, Adam, Si, Martin, Shuangwei, Jin, Yan, Fan, Nora, Yifan, Alireza, Daniel, Yanzhen, Jiaojiao. It is a great pleasure to have discussions and collaborations with you all. In particular, I need to thank Adam, Si, Shuangwei for reading through and commenting my thesis. Thank Maxim Chernodub and Stam Nicolis for the helpful discussions in Tours.

I would also like to thank the collaborator Harold A. Scheraga from Cornell University for the advice in the myoglobin research. Thank Nevena Litova from Institute for Nuclear Research and Nuclear Energy for the collaboration in calculating the kadanoff plot for protein kink model. Thank Jianfeng He from Beijing Institute of Technology for working on protein hIAPP together. Thank Alexandr Nasedkin and Jan Davidsson from department of chemistry for the collaboration on the SAXS study. I look forward to our further collaborations later.

I also need to thank my master supervisor Molin Ge for leading me to the scientific research world and always offering me best advices no matter on the research or on the life.

Many thanks to the faculty members, postdocs and graduate students in the department. Thank you for creating so excellent atmosphere for working in the department as well as the great Friday fika time. I would miss the time working together.

Thank all of my friends for the happy time we have spent together. Thank you for the days that we traveled, played and drank together. Thank you for your caring and sharing in life.

Finally I would like to thank my family especially my parents for caring about every detailed thing in my life. Thank you very much for your complete love and understanding on me.

Summary in Swedish

Proteiner är arbetshästarna i cellen. De kan endast utför sina funktioner för korrekt vekta strukturer. Proteiners felveckning kan leda till många sjukdomar såsom Alzheimers, Parkinson, de nya varianten av CJD och typ II diabetes. Därför är den korrekta 3D-strukturen för proteinet i grunden viktigt för livet.

Proteinvecknings problem är att studera mekanismen bakom hur proteinet vecker sig in i sina rätta tillstånd samt att förutsäga 3D-strukturer från aminosyrorna (sekvens) information.

Mycket har gjorts för att lösa problemen från olika områden, exempelviss biology, kemi, fysik, matematik samt datorvetenskap. Trots detta är problemen olösta, speciellt veckningens hastighet jämförelse med in vivo.

I denna avhandling, använder jag oss av the kink (multi-kink) modellen ursprungligen från matematiska fysiken i beskrivningen av protein C_α trace. Modellens introduceras genom minimeringen av proteinets genetiska energi funktion, vilket endast baseras på symmetrin i geometrin.

Med väldigt begränsade antal kinks (200 kinks), kan de flesta loop strukturerna i PDB beskrivas med RMSD prescision som är mindre än experimentell osäkerhet. Med multi-kink modellen, utförde jag detaljerade modellering samt dynamiska simulationer på riktiga proteiner. Särskilt användes modellen systematiskt i myoglobin (och villin headpiece HP35) simulering och resultaten jämfördes med experimentella värden.

Som komplement till C_α trace modellering studeras distrubtionen av side-chain i Frenet-based frame i den sista kapitlen. Det visar sig att Frenet-based frames är väl lokaliserade och beror på den sekondära strukturen automatiskt. Så en sådan visualisering av side-chain distrubtionen kan vara en viktig komplement side-chain rotamer libraries.

Genom att jämföra multi-kink modellen med Molekyl Dynamiska metoder, finner jag att simuleringar med multi-kink modellen är mycket mer tids effektiv. Den enda begränsning med multi-kink modellen är att jag inte kan derivera parametrarna från sekvens informationen och lösnings tillstånd för tillfället. Hittills är kink-modellen endast struktur baserad modellering och kan inte ge någon information om 3D-strukturen endast från sekvens informationen.

Till skillnad från GO-liknande modeller, parametrarna inom multi-kink modellen är färre än frihetsgraderna. Utifrån denna fördelen samt kink-modellens universalitet inom PDB, förtydlar jag att the kink (multi-kink) modellen har

förmågan att förutspå 3D-strukturer utgående från sekvens information. När parametrarna i kink-modellen kan bli bestämda utifrån sekvens och miljö information, skulle denna metod bli ett kraftfullt verktyg inom bestämning av protein strukturer, samt dennas dynamiska egenskaper.

References

[1] P. Hellung-Larsen and A. P. Andersen. Cell volume and dry weight. *J. Cell Biol.*, 92:319–324, 1989.

[2] G. Petsko and D. Ringe. *Protein Structure and Function*. New Science Press Ltd, 2007.

[3] C. Soto. Unfolding the role of protein misfolding in neurodegenerative diseases. *Nat. Rev. Neurosci.*, 4:49–60, 2003.

[4] A. N. Bullock, J. Henckel, and A. R. Fersht. Quantitative analysis of residual folding and dna binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy. *Oncogene*, 19:1245–1256, 2000.

[5] J. Janin, R. P. Bahadur, and P. Chakrabarti. Protein–protein interaction and quaternary structure. *Q. Rev. Biophys.*, 41:133–180, 2008.

[6] M. Seal, F. H. White Jr., and C. B. Anfinsen. Reductive Cleavage of Disulfide Bridges in Ribonuclease. *Science*, 125:691–692, 1957.

[7] C. B. Anfinsen. Principles that Govern the Folding of Protein Chains. *Science*, 181:223–230, 1973.

[8] A. L. Fink. Chaperone-Mediated Protein Folding. *Physiol. Rev.*, 79:425–449, 1999.

[9] K. A. Dill and J. L. MacCallum. The Protein-Folding Problem, 50 Years On. *Science*, 338:1042–1046, 2012.

[10] C. Levinthal. Mossbauer spectroscopy in biological systems. In J. T. P. DeBrunner and E. Munck, editors, *Proceedings of a Meeting held at Allerton House,Monticello, IL*, pages 22–24. University of Illinois Press, Urbana, 1969.

[11] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[12] K. A. Bava, M. Gromiha, H. Uedaira, K. Kitajima, and A. Sarai. Protherm, version 4.0: Thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, 32:D120–D121, 2004.

[13] N. S. Bogatyreva, A. A. Osypov, and D. N. Ivankov. Kineticdb: a database of protein folding kinetics. *Nucleic Acids Res.*, 37:D342–D346, 2009.

[14] Y. Ueeda, H. Taketomi, and N. Gō. Studies on protein folding, unfolding, and fluctuations by computer simulation. ii. a. three-dimensional lattice model of lysozyme. *Biopolymers*, 17:1531–1548, 1978.

[15] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Jr. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784, 1984.

[16] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, Jr. K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[17] Y. Duan and P. A. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282:740–744, 1998.

[18] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. Charmm: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.

[19] E. Neria, S. Fischer, and M. Karplus. Simulation of activation free energies in molecular systems. *J. Chem. Phys.*, 105:1902–1921, 1996.

[20] Jr. A. D. MacKerell, D. Bashford, M. Bellott, Jr. R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, III W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586–3616, 1998.

[21] W. L. Jorgensen and J. Tirado-Rives. The opls potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110:1657–1666, 1988.

[22] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the opls all- atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118:11225–11236, 1996.

[23] A. Liwo, J. Lee, D. R. Ripoll, J. Pillardy, and H. A. Scheraga. Protein structure prediction by global optimization of a potential energy function. *Proc. Natl. Acad. Sci., U. S. A.*, 96:5482–5485, 1999.

[24] A. Liwo, M. Khalili, and H. A. Scheraga. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci. U.S.A.*, 102:2362–2367, 2005.

[25] S. Ołdziej, C. Czaplewski, A. Liwo, M. Chinchio, M. Nanias, J.A. Vila, M. Khalili, Y. A. Arnautova, A. Jagielska, M. Makowski, H. D. Schafroth, R. Kaźmierkiewicz, D. R. Ripoll, J. Pillardy, J.A. Saunders, Y.K. Kang, K.D. Gibson, and H.A. Scheraga. Physics-based protein-structure prediction using a hierarchical protocol based on the unres force field: Assessment in two blind tests. *Proc. Natl. Acad. Sci. U.S.A.*, 102:7547–7552, 2005.

[26] J. L. Klepeis and C. A. Floudas. Astro-fold: a combinatorial and global optimization frame-work for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys. J.*, 85:2119–2146, 2003.

[27] J. L. Klepeis, Y. Wei, M. H. Hecht, and C. A. Floudas. Ab initio prediction of the three-dimensional structure of a de novo designed protein: a double-blind case study. *Proteins*, 58:560–570, 2005.

[28] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. Gromacs: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.*, 91:43–56, 1995.

[29] E. Lindahl, B. Hess, and D. van der Spoel. Gromacs 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.*, 7:306–317, 2001.

[30] Y. Zhang and J. Skolnick. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. U. S. A.*, 101:7594–7599, 2004.

[31] H. Zhou and J. Skolnick. Ab initio protein structure prediction using chunk-tasser. *Biophys. J.*, 5:1510–1518, 2007.
[32] S. Wu, J. Skolnick, and Y. Zhang. Ab initio modeling of small proteins by iterative tasser simulations. *BMC Biol.*, 5:17, 2007.
[33] Y. Zhang. Template-based modeling and free modeling by i-tasser in casp7. *Proteins*, 69 Suppl 8:108–117, 2007.
[34] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.*, 268:209–225, 1997.
[35] M. Chernodub, S. Hu, and A. J. Niemi. Topological solitons and folded proteins. *Phys. Rev. E*, 82:011916, 2010.
[36] N. Molkenthin, S. Hu, and A. J. Niemi. Discrete Nonlinear Schrödinger Equation and Polygonal Solitons with Applications to Co llapsed Proteins. *Phys. Rev. Lett.*, 106:078102, 2011.
[37] Moult J. A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.*, 15:285–289, 2005.
[38] S. Hu, Y. Jiang, and A. J. Niemi. Energy functions for stringlike continuous curves, discrete chains, and space-filling one dimensional structures. *Phys. Rev. D*, 87:105011, 2013.
[39] A. Krokhotin, M. Lundgren, and A. J. Niemi. Solitons and collapse in the λ-repressor protein. *Phys. Rev. E*, 86:021923, 2012.
[40] S. Hu, M. Lundgren, and A. J. Niemi. Discrete Frenet frame, inflection point solitons, and curve visualization with application s to folded proteins. *Phys. Rev. E*, 83:061908, 2011.
[41] V. O. Vinetskii and N. V. Kukhtarev. Theory of the conductivity induced by recording holographic grating in nonmetallic materials. *Sov. Phys. Solid State*, 16:2414–2415, 1975.
[42] R. J. Glauber. Time-Dependent Statistics of the Ising Model. *Journ. Math. Phys.*, 4:294–307, 1963.
[43] A. B. Bortz, M. H. Kalos, and J. L. Lebowitz. A new algorithm for monte carlo simulation of ising spin systems. *Journ. Comput. Phys.*, 17:10–18, 1975.
[44] B. A. Berg. *Markov Chain Monte Carlo Simulations And Their Statistical Analysis*. World Scientific, Singapore, 2004.
[45] J. Karanicolas and C. L. Brooks III. The origins of asymmetry in the folding transition states of protein l and protein g. *Protein Sci.*, 11:2351–2361, 2002.
[46] M. Cieplak and T. X. Hoang. Universality classes in folding times of proteins. *Biophys. Journ.*, 84:475–488, 2003.
[47] L. Huang and E. I. Shakhnovich. Is there an en route folding intermediate for cold shock proteins? *Protein Sci.*, 21:677–685, 2012.
[48] J. H. Meinke and U. H. E. Hansmann. Protein simulations combining an all-atom force field with a go term. *J. Phys. Condens. Matter*, 19:285215, 2007.
[49] M. Cieplak and T. X. Hoang. Folding of proteins in go models with angular interactions. *Physica A*, 330:195–205, 2003.
[50] W. G. Noid. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.*, 139:090901, 2013.
[51] G. A. Petsko and D. Ringe. Fluctuations in protein structure from x-ray diffraction. *Annu. Rev. Biophys. Bioeng.*, 13:331–371, 1984.

[52] PDBselect - selection of a representative set of PDB chains. `http://bioinfo.tg.fh-giessen.de/pdbselect`, 2011.

[53] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181:662–666, 1958.

[54] M. Dametto and A. E. Cárdenas. Computer simulations of the refolding of sperm whale apomyoglobin from high-temperature denaturated state. *J. Phys. Chem. B*, 112:9501–9506, 2008.

[55] D. Eliezer, P. A. Jennings, P. E. Wright, S. Doniach, K. O. Hodgson, and H. Tsuruta. The radius of gyration of an apomyoglobin folding intermediate. *Science (New York, NY)*, 270:487–488, 1995.

[56] T. Uzawa, S. Akiyama, T. Kimura, S. Takahashi, K. Ishimori, I. Morishima, and T. Fujisawa. Collapse and search dynamics of apomyoglobin folding revealed by submillisecond observations of α-helical content and compactness. *Proc. Natl. Acad. Sci. U.S.A*, 101:1171–1176, 2004.

[57] T. Uzawa, C. Nishimura, S. Akiyama, K. Ishimori, S. Takahashi, H. J. Dyson, and P. E. Wright. Hierarchical folding mechanism of apomyoglobin revealed by ultra-fast h/d exchange coupled with 2d nmr. *Proc. Natl. Acad. Sci. U.S.A*, 105:13859–13864, 2008.

[58] R. R. Matheson Jr and H. A. Scheraga. A method for predicting nucleation sites for protein folding based on hydrophobic contacts. *Macromolecules*, 11:819–829, 1978.

[59] P. A. Jennings and P. E. Wright. Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Science*, 262:892–896, 1993.

[60] R. Elber and M. Karplus. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science*, 235:318–321, 1987.

[61] I. Schlichting, J. Berendzen, G.N. Phillips Jr., and R.M. Sweet. Crystal structure of photolysed carbonmonoxy-myoglobin. *Nature*, 371:808–812, 1994.

[62] M. Chernodub, M. Lundgren, and A. J. Niemi. Elastic energy and phase structure in a continuous spin Ising chain with applications to chiral homopolymers. *Phys. Rev. E*, 83:011126, 2011.

[63] J. S. Olson, J. Soman, and G. N. Phillips Jr. Ligand pathways in myoglobin: A review of trp cavity mutations. *IUBMB Life*, 59:552–562, 2007.

[64] J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter. Sub-microsecond protein folding. *J. Mol. Biol.*, 359:546–553, 2006.

[65] J. Kubelka, W. A. Eaton, and J. Hofrichter. Experimental tests of villin subdomain folding simulations. *J. Mol. Biol.*, 329:625–630, 2003.

[66] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How fast-folding proteins fold. *Science*, 334:517–520, 2011.

[67] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. U.S.A*, 110:5915–5920, 2013.

[68] J. S. Yang, S. Wallin, and E. I. Shakhnovich. Universality and diversity of folding mechanics for three-helix bundle proteins. *Proc. Natl. Acad. Sci. U.S.A*, 105:895–900, 2008.

[69] A Jain and G Stock. Identifying metastable states of folding proteins. *J. Chem. Theor. Comput.*, 8:3810–3819, 2012.

[70] T. K. Chiu, J. Kubelka, R. Herbst-Irmer, W. A. Eaton, J. Hofrichter, and D. R. Davies. High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein. *Proc. Natl. Acad. Sci. U.S.A*, 102:7517–7522, 2005.

[71] P. Rotkiewicz and J. Skolnick. Fast procedure for reconstruction of full- atom protein models from reduced representations. *J. Comput. Chem.*, 29:1460–1465, 2008.

[72] D. Petrey, Z. Xiang, C. L. Tang, L. Xie, M. Gimpelev, T. Mitros, C. S. Soto, S. Goldsmith-Fischman, A. Kernytsky, A. Schlessinger, I. Y. Y. Koh, E. Alexov, and B. Honig. Using Multiple Structure Alignments, Fast Model Building, and Energetic Analysis in Fold Recognition and Homology Modeling. *Proteins*, 53(Suppl 6):430–435, 2003.

[73] L. Holm and C. Sander. Database Algorithm for Generating Protein Backbone and Side-chain Co-ordinates from a CCLTrace. *J. Mol. Biol*, 218:183–194, 1991.

[74] G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack Jr. Improved prediction of protein side-chain conformations with scwrl4. *Proteins*, 77:778–795, 2009.

[75] S. C. Lovell, J. Word, J. S. Richardson, and D. C. Richardson. The penultimate rotamer library. *Proteins*, 40:389–408, 2000.

[76] R. L. Dunbrack Jr. and M. Karplus. Backbone-dependent rotamer library for proteins application to side-chain prediction. *J. Mol. Biol.*, 230:543–574, 1993.

[77] M. S. Shapovalov and R. L. Dunbrack. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19:844–858, 2011.

[78] M. Lundgren, A. J. Niemi, and F. Sha. Protein loops, solitons, and side-chain visualization with applications to the left-handed helix region. *Phys. Rev. E*, 85:061909, 2012.

[79] M. Lundgren and A. J. Niemi. Correlation between protein secondary structure, backbone bond angles, and side-chain orientations. *Phys. Rev. E*, 86:021904, 2012.

[80] Y. Li and Y. Zhang. REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins: Struct., Funct., Bioinf.*, 76:665–676, 2009.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1184

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-232562

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2014