UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1054

Dynamics of Discrete Curves with Applications to Protein Structure

SHUANGWEI HU

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2013

ISSN 1651-6214
ISBN 978-91-554-8694-5
urn:nbn:se:uu:diva-199987

Dissertation presented at Uppsala University to be publicly examined in Å10132, Ångströmlaboratoriet, Lägerhyddsvägen 1, Uppsala, Monday, September 2, 2013 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract
Hu, S. 2013. Dynamics of Discrete Curves with Applications to Protein Structure. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1054. 41 pp. Uppsala. ISBN 978-91-554-8694-5.

In order to perform a specific function, a protein needs to fold into the proper structure. Prediction the protein structure from its amino acid sequence has still been unsolved problem. The main focus of this thesis is to develop new approach on the protein structure modeling by means of differential geometry and integrable theory. The start point is to simplify a protein backbone as a piecewise linear polygonal chain, with vertices recognized as the central alpha carbons of the amino acids. Frenet frame and equations from differential geometry are used to describe the geometric shape of the protein linear chain. Within the framework of integrable theory, we also develop a general geometrical approach, to systematically derive Hamiltonian energy functions for piecewise linear polygonal chains. These theoretical studies is expected to provide a solid basis for the general description of curves in three space dimensions. An efficient algorithm of loop closure has been proposed.

Keywords: Frenet equations, integrable model, folded proteins, discrete curves

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

ISSN 1651-6214
ISBN 978-91-554-8694-5
urn:nbn:se:uu:diva-199987 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-199987)

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., Lundgren, M. & Niemi, A. J. (2011). *Discrete Frenet frame, inflection point solitons, and curve visualization with applications to folded proteins.* Physical Review E, 83(6), 061908. [arXiv:1102.5658]

II Hu, S., Jiang, Y., & Niemi, A. J. (2013). *On energy functions for string-like continuous curves, discrete chains, and space-filling one dimensional structures*. Physical Review D, 87(10), 105011 [arXiv:1210.7371]

III Hu, S., & Niemi, A. J. (2013). *On bifurcations in framed curves and chains*. [Submitted to Nonlinearity]

IV Hinsen, K., Hu, S., Kneller, G., & Niemi, A. J. (2013). *On coarse grained representations and the problem of protein backbone reconstruction* [Manuscript]

Reprints were made with permission from the publishers.

Papers not included in this thesis:

5. Chernodub, M., Hu, S., & Niemi, A. J. (2010). *Topological solitons and folded proteins*. Physical Review E, 82(1), 011916. [arXiv:1003.4481]

6. Molkenthin, N., Hu, S., & Niemi, A. J. (2011). *Discrete Nonlinear Schrödinger Equation and Polygonal Solitons with Applications to Collapsed Proteins*. Physical Review Letters, 106(7), 078102. [arXiv:1009.1078]

7. Hu, S., Krokhotin, A., Niemi, A. J., & Peng, X. (2011). *Towards quantitative classification of folded proteins in terms of elementary functions*. Physical Review E, 83(4), 041907. [arXiv:1011.3181]

Contents

1. Introduction and motivation

One of my friends, an experimental biologist, once time commented on my work of theoretical modeling, "if a protein structure can be predicted solely from its chemical sequence on computer, I am going to lose my job." Truly I agreed with her real meaning that the most reliable way to determine a protein structure is by direct experimentation. However, compared with the sequence measurement it is much more difficult to experimentally determine a protein structure, let alone the observation of protein motion on tiny time-scale.

Theoretical methods can help to match such a gap, giving us the insight on the thorough behaviour of protein systems that are usually impossible even by the fine-tuned experimental instruments. On the other hand, it is a necessity and challenge to make sense of the experiment data, for purpose of finding the most general law behind the complexity of molecular biology, and for the purpose of aiding new drug design. This thesis is to cast some novel viewpoints on some of this complexity, by integrating methods from both differential geometry and integrable systems into the modeling of protein structures.

The start point of our methodology is to simplify the protein backbone chain as a space curve in three dimensions. In terms of space curves one can model many problems in physics, such as a polymer chain [28, 13, 14, 15, 16], a vortex filament in a fluid [30, 25, 46], and in a superfluid [52] and many other applications. Some abstract objects can also described by space curves. Interesting cases include a spin vector in the magnetic spin chain model being taken as the tangent vector to some space curve [17], and even the $n-$body problem being approximated by smooth closed space curves [4].

It deserves attention to study the possible connections between the moving space curves and integrable evolution equations [21]. An integrable system has several excellent features that allows one to make global analysis about it. For example, a system governed by integrable evolution equation has an infinite number of integrals of motion. Mathematical structure of an integrable system is also associated with a Lax pair and interesting solutions such as solitons [17] . We found that these solitons solution describes the buckling of the curve and is then applicable to model the helix-loop-helix motif in protein structure [6, 31].

In this thesis, we shall first give an introduction of protein structure and differential geometry of curves. Then we make the geometric analysis of protein backbones and find a relation connecting both $NC_\alpha C$ backbone and C_α-only backbone. Analogue with Ramachandran plot, the distribution of virtual bond angle and torsion of C_α-only backbone has clearly defined the regular features

of protein secondary structures such as α helix and β-sheet. Since the curve to describe protein structure is polygonal style and thus lots of our efforts have been focused on the nontrivial discretization that preserve the integrable structure of the moving curves. The guide principle lies on the observation that the conserved quantities must be invariant under local frame rotations. This geometric principle not only helps us systematically derive Hamiltonian energy functions of curves but also inspires us an efficient algorithm for loop closure in protein structure modeling.

2. Protein structure and modeling

Proteins are involved in almost all cellular functions, from specific binding of other molecules, to enzymes catalysis of chemical reactions, to molecular switches for controlling cellular processes, to structural support elements of living systems [12]. To perform a given function most proteins need to fold into the a unique three-dimensional structure known as the native state. Incorrect folding can have disastrous consequences such as resulting prions or Alzheimer's disease. Consequently, the structure determination plays the central role in protein research and has many biological/medical applications.

Proteins can be generally divided into three classes: globular proteins, membrane proteins and fibrous proteins. From herein we concentrate on globular proteins, the most frequent class.

Since solution of protein folding problem has been pursued over several decades, there have exit many insightful concepts and observations in this area and new progress is on-going. In this chapter, we only give a very short introduction of protein structure and modeling. We here won't touch on interesting aspects of many closely related topics, e.g. the landscape and funnel, ϕ-values analysis, CASP(Critical Assessment of Techniques for Protein Structure Prediction). These topics and others can be found, for example, among two representative reviews by Ken A. Dill *et al* [10, 11]. Introduction of protein structures can be found in the book given by Petsko and Ringe [38].

2.1 Protein structure

Once synthesized, a protein chain is much loosy and irregular. But it quickly folds to a structure of a well-organized hierarchy, which arranges from primary level to secondary level, and to tertiary level, as illustrated in Fig. 2.1.

Proteins are linear heteropolymers, composed of a set of twenty amino acids (also called residues). Each of these twenty amino acid is denoted by a Roman letter. The composition of the amino acid letter forms a a sequence called the primary structure of the protein. By structure, all of the twenty amino acids share a common backbone, which consists of the amino group (NH_2), the alpha carbon C_α and the caboxylic acid group (COOH). Amino acids differ by the side chains attaching to C_α and thus have different chemical properties. By losing a water molecule, two neighboring amino acids are covalently connected together through peptide bonds (C(O)NH). Continuing this way, a long

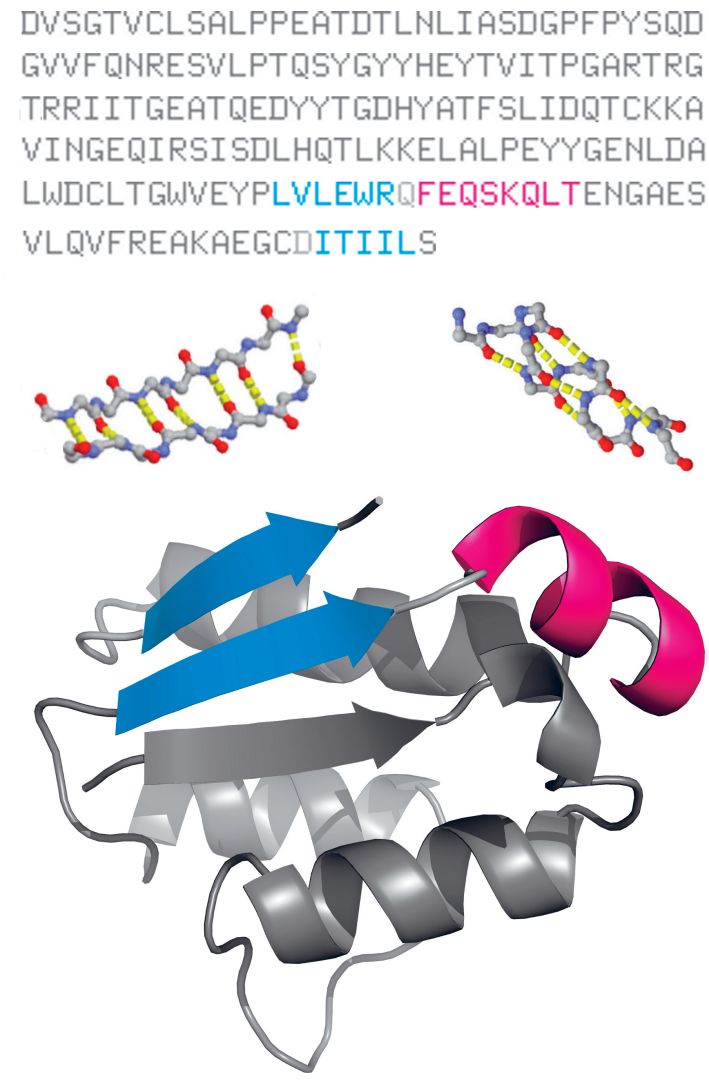

Figure 2.1. Protein structure hierarchy (PDB 1ay7 chain B). On the top is shown the primary structure which is the amino acid sequence. On the middle are shown two typical regular secondary structure elements, i.e. β-sheet (left) and α-helix (right). It is worthy to notice the hydrogen bonding pattern in both. On the bottom is the tertiary structure, which is the compact packing of the secondary structure elements with the help of irregular loops. Highlight are the β-sheet (in blue) and α-helix (in red), respectively.

polypeptide chain is synthesized, as exhibited in the early stage of protein formation in cell.

Frequently, local segment of around $3-30$ amino acids form regular structures, either α-helices or β-sheets. It deserves notice of the regular pattern of hydrogen bonds in both α-helix and β-sheet (middle plot on Fig. 2.1). In an α-helix, the C=O group of the nth residue accepts a hydrogen bond from the N-H group of the $n+4$th residue. In β-sheets, two or more strands that may be distant along the protein sequence are dragged closely by hydrogen bonds between them. Other regular secondary elements such as left-handed helices (L-helices) are also observed, but not so often.

In a folded protein, the α-helices and β-sheets organize into a compact object, called the tertiary structure of the protein. An amino acid sequence in nature adopts a unique tertiary structure, ensuring the stability of its function. While two similar sequence can share the structural resemblance, it also frequently happens that many proteins have similar structures but low sequence similarity.

In many cases, proteins don't function by themselves. Rather they prefer to forming a complex with other proteins. These corporative complexes may preserve their complementarity under evolutionary pressure.

2.2 Protein structure modeling

It has been several decades for people to look for ways of simulating the protein folding on a computer and predicting the structure of a protein from its amino acid sequence. In the 1950's Anfinsen and coworkers proposed that many proteins have a unique three-dimensional structure, corresponding to a minimum of free energy [2]. How to calculate protein free energy and how to efficiently find the global minimum of this energy are then the two main tasks in protein structure prediction and modeling.

An energy function differs according to the framework within which the folding forces are tackled. Quantum mechanics has the advantage of accurate description but are computationally very intensive, even for simulations of even very small peptides. As a tradeoff, the study of protein is usually done within the semi-empirical classical mechanical methods. They are empirical in the sense that a good design and clever guess is often necessary. One main goal of my thesis project is to model a more realistic energy function from differential geometry and integrable theory.

It is convenient for the classical energy functional to associate with the traditional minimization methods, such as force-field molecular dynamics (MD) and Monte Carlo (MC) simulations. The MD simulation method is based on Newton's second law,

$$\mathbf{F}_i = m_i \frac{d^2\mathbf{r}_i}{dt^2},\ \mathbf{F}_i = -\frac{\partial}{\partial \mathbf{r}_i} E\left(\{\mathbf{r}_i\}\right), \tag{2.1}$$

where $\mathbf{F}_i$ is the force exerted on the atom i, m_i its mass, $\mathbf{r}_i$ the atom coordinate, and $E(\{\mathbf{r}_i\})$ the free energy of the protein. Integration of the equation of motion for each atom in the system then results in a trajectory that describes the dynamics of a protein. From this trajectory, we can then get the average values of properties of interest, based on the ergodic hypothesis, which states that the time average equals the ensemble average. MD simulations can be time consuming and computationally expensive. However, with the advance in the speed of computers, and new advanced methods, it is now possible to determine the structure of the native state directly from the sequence of small proteins [11].

Rather than modeling the dynamics of a system, the goal of an MC simulation is to capture statistical (thermodynamical) properties of a system by a stochastic search. While the types of moves in an MD simulation are strictly dictated by Newton's laws of physics, there is no such restriction on the moves in an MC simulation. The only requirement is that the simulation is not biased, which can be ensured by enforcing detailed balance and ergodicity. As a result, MC simulation potentially enhances the scope of simulations in terms of size and timescale, and is therefore widely applied for *ab initio* protein structure prediction. The popular implementation is done with the framework of Markov Chain MC, in which the equilibrium generates the Boltzman distribution of the protein system.

2.2.1 Representation

In the protein modeling, the choice of representation matters a lot since it affects both the design of energy function and the searching strategy. Because of the large degrees of freedom in protein system, the theoretical investigation becomes extremely difficult for the full atom representation. Therefore, reduced model or coarse-grained representation is preferred, with the trade-off between computational cost and accuracy.

Traditionally, MD studies of protein take all atoms into account explicitly. For MC studies and MD simulations in recent time the coarse-graining techniques are widely used. It have been shown that it is allowed to recover the fine-grained representation from coarse-grained representation in protein systems without significant loss of information [23, 7].

Among the simplified representations a common choice is to only take the heavy atoms of the protein backbone (N, C_α, C) (sometimes to also include the side chain as a pseudo atom). In convention, the dihedral angles between the plane defined by C-N-C_α and the plane defined by N-C_α-C is called Ψ. And the dihedral angle between the plane defined by N-C_α-C and the plane defined by C_α-C-N is called Φ. In the now being considered backbone representation, only these two angle Ψ, Φ (called Ramachandran angles) are assumed to be the dynamical variables, while others, including the backbone lengths, covalent

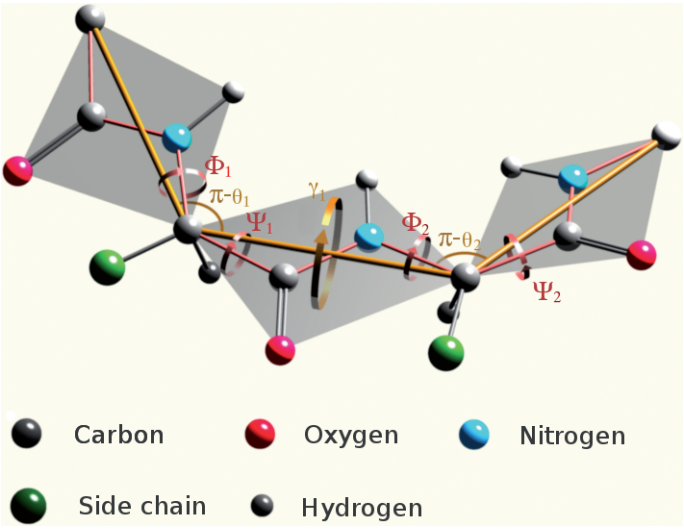

Figure 2.2. Two representations of protein structure. For the N-C_α-C backbone representation (in red), covalent bond angles and lengths are largely invariant while varying Ramachandran angles (Ψ, Φ) give the protein different configurations. Shown in grey is the peptide plane (including O, H atoms) which is determined by the ideal torsion angle $\omega = \pi$. For the coarse-grained representation of C_α-backbone (in gold), the bond length of C_α-C_α is fixed while bond angle θ and torsion angle γ are dynamical variables.

bond-angles and the ω torsion angle are taken as fixed parameters, since they display only insignificant fluctuations from their average values. Yet we found that this small fluctuations matter the modeling of the long chain structure of proteins. For example, replacing the bond angles or the ω torsion angle by their own average values, commonly yields protein structures that are different from the native structures. In Paper IV, we address this problem in details.

Another coarse-grained representation is the C_α-backbone where each residue is represented by its C_α atom only. Now bond angle θ (defined over three consecutive C_α atoms) and torsion angle γ (defined over four consecutive C_α atoms) are the dynamical variables while bond length between two neighboring C_α atoms is fixed (3.806Å). Though more coarse-grained, the C_α-backbone with uniform bond length surprisingly reproduce the protein structure, regardless of the beneath fluctuation of covalent bond-angles and the ω torsion angle that make trouble in the NC_αC backbone representation.

We will discuss more about these backbone representations mathematically in next chapter.

3. Differential geometry of curves

This chapter reviews some basic definitions and results concerning the differential geometry of curves, both continuous ones and discrete ones. These results will be immediately applicable to the analysis of protein structures, whose backbones are represented by curves. Yet the application of differential geometry of curves is very general and much beyond the scope of this thesis.

3.1 Frenet frame and equations

There is an intuitive way of thinking the differential geometry of curves: we can take a curve as the trajectory of a moving particle. This dynamic perspective suggests a local frame at each point on the trajectory. Neighboring frames are not necessarily the same and thus the connection between them will of the key interest for the learning of the trajectory shape.

Among many possibilities of defining the local frame, Frenet frame is a natural choice. Consider a space curve $\mathbf{r}(s)$ parameterized by its arc length s, such that

$$|\partial_s \mathbf{r}(s)|^2 = 1. \tag{3.1}$$

The arc length s denotes a kind of time the moving particle has moved. Define the tangent vector $\mathbf{t} = \partial \mathbf{r}/\partial s \equiv \mathbf{r}'(s)$, the normal vector $\mathbf{n}(s) = \mathbf{t}'(s)/|\mathbf{t}'(s)|$ and the binormal vector $\mathbf{b}(s) = \mathbf{t}(s) \times \mathbf{n}(s)$ (Fig. 3.1). These three unit vectors

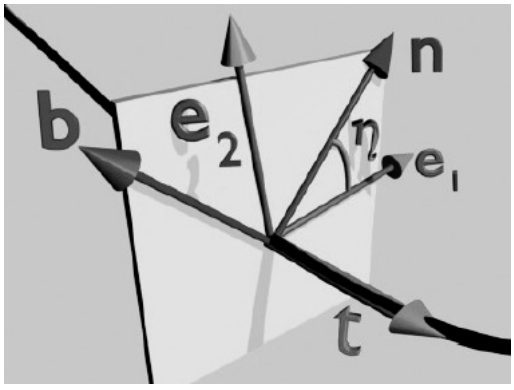

Figure 3.1. Frenet frame along a continuous curve. Adapted from Paper I.

form an orthogonal frame, called Frenet frame, at any point on the curve and satisfy the Frenet equations [19, 48]

$$\frac{d}{ds}\begin{pmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{pmatrix}=\begin{pmatrix}0 & \kappa & 0\\ -\kappa & 0 & \tau\\ 0 & -\tau & 0\end{pmatrix}\begin{pmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{pmatrix}\equiv\mathbf{Q}(s)\begin{pmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{pmatrix}, \tag{3.2}$$

where κ is called the curvature and τ the torsion. The curvature κ describes how different the curve is from a linear line and the torsion τ denotes the deviation of the curve from planarity. The shape of the curve is uniquely decided by both functions $\kappa(s)$ and $\tau(s)$, up to a global translation and/or rotation.

This differential system in Eq. (3.2) has the ordered exponential solution, with the form

$$\begin{pmatrix}\mathbf{t}(s)\\ \mathbf{n}(s)\\ \mathbf{b}(s)\end{pmatrix} = \mathbf{U}(s)\begin{pmatrix}\mathbf{t}(0)\\ \mathbf{n}(0)\\ \mathbf{b}(0)\end{pmatrix}, \tag{3.3}$$

$$\begin{aligned}\mathbf{U}(s) &\equiv \mathscr{P}\left(\exp\left\{\int_0^s \mathbf{Q}(s')\,ds'\right\}\right)\\ &= 1+\int_0^s ds'\mathbf{Q}(s')+\int_0^s ds'\int_0^{s'} ds''\mathbf{Q}(s')\,\mathbf{Q}(s'')+\cdots\end{aligned} \tag{3.4}$$

It is important to notice that $\mathbf{U}(s)$ is an orthogonal transformation matrix, i.e. $\mathbf{U}^T=\mathbf{U}^{-1}$, which follows the exponentiation construction and the skew-symmetric Frenet matrix of $\mathbf{Q}(s)$, i.e. $\mathbf{Q}^T=-\mathbf{Q}$. The orthogonality ensures that the Frenet basis are always orthogonal and normalized, i.e.

$$|\mathbf{t}(s)|^2+|\mathbf{n}(s)|^2+|\mathbf{b}(s)|^2=|\mathbf{t}(0)|^2+|\mathbf{n}(0)|^2+|\mathbf{b}(0)|^2. \tag{3.5}$$

The Taylor expansion of $\mathbf{U}(s)$ comes a closed form when $\kappa(s)$ and $\tau(s)$ are constant or piecewise constant. The former case corresponds to a helix while the latter corresponds to a so-called polyhelix. Supposed $\kappa(s)$ and $\tau(s)$ are constant within Δs, we then have

$$\begin{aligned}\mathbf{U} &= e^{\mathbf{Q}\delta}\\ &= \exp\begin{pmatrix}0 & \kappa\delta & 0\\ -\kappa\delta & 0 & \tau\delta\\ 0 & -\tau\delta & 0\end{pmatrix}=\begin{pmatrix}a & b & c\\ -b & d & e\\ c & -e & f\end{pmatrix},\end{aligned} \tag{3.6}$$

$$a=\frac{\tau^2+\kappa^2\cos(q\delta)}{q^2}, b=\frac{\kappa}{q}\sin(q\delta), c=(1-\cos(q\delta))\frac{\kappa\tau}{q^2},$$

$$d=\cos(q\delta), e=\frac{\tau}{q}\sin(q\delta), f=\frac{\kappa^2+\tau^2\cos(q\delta)}{q^2}, q=\kappa^2+\tau^2.$$

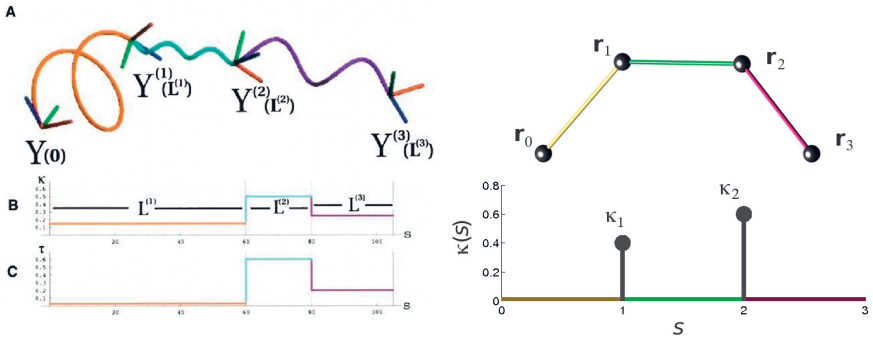

Figure 3.2. Left: the polyhelix model of protein structure ([22]). C_α atoms are connected by helicies. The curvature and torsion functions are piece-wise constant. Right: the polygon model of protein structure. C_α atoms are connected by straight lines. The curvature and torsion (not shown here) profiles are a series of impulse functions located at C_α atoms.

3.2 Discretize the curve

There are lots of different schemes to discretize the curve, for different purposes. For example, computer scientists utilize the discrete curve for real time applications. What they focus on is the simulation speed and physical correctness rather than the curve structure itself [51]. Here we review several approaches of curve model of protein chains.

The first one is the polyhelix model of protein structure, taking the expression (3.7) on its basis [22]. In this representation, C_α atoms are connected by a serie of connected helices, whose curvatures, torsions and arc length are nonlinearly fitted. The values of curvature and torsion can be used to characterize the structure preferences.

There is, however, an simpler choice of representation, the impulse function for both curvature and torsion profile along the arc length, that is, having nonzero values only at vertices. This implies the polygonal representation of protein structure. In Fig. 3.2 we show the comparison between these two approaches. The advantage of impulse profile of curvature and torsion lies on the fact that arc length equals the distance between two consecutive C_α atoms, i.e. $s = i\delta, i = 0,\ldots,N$ (δ is the bond length and N is the number of C_α atoms). So the ordered exponential matrix only involve once $\mathbf{Q}$ to switch from one C_α atom to its neighbor,

$$\mathbf{U}\left((i+1)\,\delta\right) = e^{\mathbf{Q}(i\delta)}. \tag{3.7}$$

From the Trotter-Suzuki formula [49], one can show that

$$e^{\mathbf{Q}} = e^{\kappa \mathrm{T}_3 + \tau \mathrm{T}_1} = e^{\kappa \mathrm{T}_3} e^{\tau \mathrm{T}_1} + \mathscr{O}\left(\kappa\tau\right), \tag{3.8}$$

where T_1 and T_3 are generators of SO(3) rotation, i.e. $\left(\mathrm{T}_i\right)_{jk} = \varepsilon^{ijk}$. It comes a surprise to see that the above approximation expression becomes *exact* when

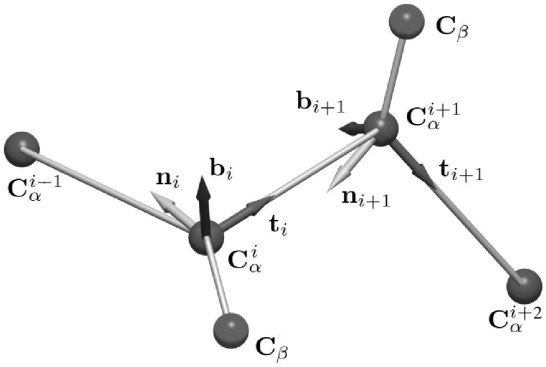

Figure 3.3. Discrete Frenet frame along the C_α-backbone of a protein chain. C_β atoms are shown for reference.

we choose a special discretization of Frenet frame, as following

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|}, \mathbf{b}_i = \frac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{|\mathbf{t}_{i-1} \times \mathbf{t}_i|}, \mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i,$$
$$\begin{pmatrix} \mathbf{t} \\ \mathbf{n} \\ \mathbf{b} \end{pmatrix}_i = e^{\kappa_i \mathbf{T}_3} e^{\tau_i \mathbf{T}_1} \begin{pmatrix} \mathbf{t} \\ \mathbf{n} \\ \mathbf{b} \end{pmatrix}_{i-1}. \tag{3.9}$$

In Fig. 3.3 we illustrate this discrete Frenet frames along the protein C_α backbone..The above formula can be regarded as two-point difference of the continuous equations if we compare it with the continuous case (3.2). The discretized curvature κ_i and torsion τ_i have now the geometrical meaning of bond angle and torsion angle (see Fig. 2.2). In fact, the derive of Eq. (3.9) is straightforward. Firstly two consecutive discrete Frenet frames are related by an SO(3) rotation, which can be specified by three Euler angles. Since of the orthogonality of $\mathbf{b}_i$ with $\mathbf{t}_{i-1}$, one Euler angle can be excluded. The other two remaining angles are further identified as bond angle and torsion angle. We have discussed this approach in more details in Paper I. We draw attention to [18] since it has early utilized the same formula (3.9) for modeling the long chain of polymer molecules.

In Ref. [40, 41, 42] they proposed another approach which is essentially the three-point difference scheme, which makes the computation more complex and parameters less geometrical meaning. In Ref. [27] , they used the discretization scheme by means of Cayley transform [45]

$$\mathbf{U} \approx \left(1 + \frac{\delta}{2}\mathbf{Q}\right)\left(1 - \frac{\delta}{2}\mathbf{Q}\right)^{-1}. \tag{3.10}$$

It serves a good approximation if δ is small. When it comes to the application on proteins, such a scheme has no direct geometric meaning of curvature and torsion.

3.3 Protein backbone geometry

Here we illustrate the applicability of differential geometry on protein chain. There are more than one possibility we can go. Here as the main example, we focus on one geometric relation between $NC_\alpha C$ backbone and C_α-only backbone. Then we mention several other applications.

3.3.1 Geometric relation between $NC_\alpha C$ backbone and C_α-only backbone

In Subsection 2.2.1, we have shown two representation of protein backbone structure. One is $NC_\alpha C$ backbone and the other is C_α-only backbone (also see Fig. 2.2). The dynamical variables of the former representation are the Ramachandran angles (Ψ,Φ), while the dynamical variables of the latter are the virtual bond and torsion angles (θ,γ). Here we show how these two sets of angles are related, by applying the differential geometry theory in last section.

Firstly denote the frame matrix as $\mathscr{F}_i = (\mathbf{t}_i, \mathbf{n}_i, \mathbf{b}_i)$ and form a new 4×4 matrix as following

$$\mathscr{G}_i = \begin{pmatrix} \mathscr{A}_i & \mathbf{r}_i \\ \mathbf{0} & 1 \end{pmatrix}. \tag{3.11}$$

Then the discrete Frenet equations (3.9) can be rewritten in the new form

$$\mathscr{G}_{i+1} = \mathscr{G}_i \begin{pmatrix} \mathscr{R}_x(\tau_i) & \Delta_i \mathbf{e}_x \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \mathscr{R}_z(\kappa_{i+1}) & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}, \tag{3.12}$$

where $\mathscr{R}_x(\cdot)$ and $\mathscr{R}_z(\cdot)$ are the rotation matrix around the x,z-axis respectively, and the unit vector $\mathbf{e}_x = (1,0,0)^T$. This above formula has the advantage that it has compact form and is thus easily implemented in the numerical way.

We use the convention for indexing the atoms and the peptide planes as shown in Table 3.1. Then we set the initial point at $C_{\alpha,k}$ and define the local frame as

$$\mathbf{r}_{3k-2} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{t}_{3k-2} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{n}_{3k-2} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{b}_{3k-2} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Since of the planarity of peptide bond ($\omega = \pi$), we can simplify the formula () into blocked form. Let us do it step by step, as following

$$C_{\alpha,k} \quad : \quad \mathscr{G}_{3k-2} = \begin{pmatrix} (\mathbf{t}_{3k-2}, \mathbf{n}_{3k-2}, \mathbf{b}_{3k-2}) & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix} = \mathscr{I}, \tag{3.13}$$

$$C_k \quad : \quad \mathscr{G}_{3k-1} = \mathscr{G}_{3k-2} \begin{pmatrix} \mathscr{R}_x(\Psi_k) & \Delta_1 \mathbf{e}_x \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \mathscr{R}_z(\kappa_2) & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}, \tag{3.14}$$

$$N_{k+1} \quad : \quad \mathscr{G}_{3k} = \mathscr{G}_{3k-1} \begin{pmatrix} \mathscr{R}_x(\omega) & \Delta_2 \mathbf{e}_x \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \mathscr{R}_z(\kappa_3) & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}, \tag{3.15}$$

N-C_α-C backbone atom	$C_{\alpha,k}$	C_k	N_{k+1}
index i	$3k-2$	$3k-1$	$3k$
Bond angle $\kappa_i\,(^\circ)$	κ_1 =68.9	κ_2 =63.4	κ_3 =58.6
Torsion angle τ_i	Ψ_k	$\omega=\pi$	Φ_{k+1}
Bond length $\Delta_i\,(\text{Å})$	$\Delta_{1,k}$ =1.525	$\Delta_{2,k}$ =1.330	$\Delta_{3,k}$ =1.460
Virtual C_α bond angle	θ_k	-	-
Virtual C_α torsion angle	γ_k	-	-
Virtual C_α bond length (Å)	Δ_α =3.806	-	-

Table 3.1. *The convention for indexing the atoms and the peptide planes, and the values of bond/torsion angles and bond lengths along the protein $NC_\alpha C$ backbone. Here k denotes the amino acid indexing. We have assumed the chain start from its N-terminus.*

$$
\begin{aligned}
C_{\alpha,k+1} \quad : \quad & \mathscr{G}_{3k+1} = \mathscr{G}_{3k} \begin{pmatrix} \mathscr{R}_x(\Phi_{k+1}) & \Delta_3 \mathbf{e}_x \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \mathscr{R}_z(\kappa_1) & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix} \\
& = \begin{pmatrix} \mathscr{A}_{3k+1} & \mathbf{r}_{3k+1} \\ \mathbf{0} & 1 \end{pmatrix}, \qquad (3.16) \\
\mathscr{A}_{3k+1} \quad = \quad & \mathscr{R}_x(\Psi_k)\mathscr{R}_z(\kappa_2)\mathscr{R}_x(\omega)\mathscr{R}_z(\kappa_3)\mathscr{R}_x(\Phi_{k+1})\mathscr{R}_z(\kappa_1), \qquad (3.17) \\
\mathbf{r}_{3k+1} \quad = \quad & (\Delta_1 + \Delta_2 \mathscr{R}_x(\Psi_k)\mathscr{R}_z(\kappa_2) + \Delta_3 \mathscr{R}_x(\Psi_k)\mathscr{R}_z(\kappa_2)\mathscr{R}_x(\omega)\mathscr{R}_z(\kappa_3))\,\mathbf{e}_x \\
= \quad & \mathscr{R}_x(\Psi_k)\,\mathbf{p}, \qquad (3.18) \\
\mathbf{p} \quad \equiv \quad & \begin{pmatrix} \Delta_1 + \Delta_2\cos\kappa_2 + \Delta_3\cos(\kappa_2-\kappa_3) \\ \Delta_2\sin\kappa_2 + \Delta_3\sin(\kappa_2-\kappa_3) \\ 0 \end{pmatrix}. \qquad (3.19)
\end{aligned}
$$

Note that this expression of $\mathbf{p}$ involves only the constant parameters, reflecting the repeating unit of polypeptide chain. Geometrically, it is the relative directional vector for one alpha carbon to its neighbor. Its norm is then the bond length of the alpha carbons, i.e. $|\mathbf{p}| = \Delta_\alpha = 3.806$Å.

For the next peptide bond ($C_{\alpha,k+1} - C_{k+1} - N_{k+2} - C_{\alpha,k+2}$), we can readily get $\mathscr{G}_i$ matrix at $C_{\alpha,k+2}$ ($i = 3k+4$):

$$
\begin{aligned}
\mathscr{G}_{3k+4} \quad = \quad & \begin{pmatrix} \mathscr{A}_{3k+1} & \mathscr{R}_x(\Psi_k)\,\mathbf{p} \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \mathscr{A}_{3k+4} & \mathscr{R}_x(\Psi_{k+1})\,\mathbf{p} \\ \mathbf{0} & 1 \end{pmatrix} \\
= \quad & \begin{pmatrix} \mathscr{A}_{3k+1}\mathscr{A}_{3k+4} & \mathscr{R}_x(\Psi_k)\,\mathbf{p} + \mathscr{A}_{3k+1}\mathscr{R}_x(\Psi_{k+1})\,\mathbf{p} \\ \mathbf{0} & 1 \end{pmatrix}. \qquad (3.20)
\end{aligned}
$$

For this formula we can read out the position of $C_{\alpha,k+2}$,

$$\mathbf{r}_{3k+4} = (\mathscr{R}_x(\Psi_k) + \mathscr{A}_{3k+1}\mathscr{R}_x(\Psi_{k+1}))\,\mathbf{p}. \qquad (3.21)$$

Similarly we can get the position of next alpha carbon $C_{\alpha,k+3}$,

$$\mathbf{r}_{3k+7} = (\mathscr{R}_x(\Psi_k) + \mathscr{A}_{3k+1}\mathscr{R}_x(\Psi_{k+1}) + \mathscr{A}_{3k+1}\mathscr{A}_{3k+4}\mathscr{R}_x(\Psi_{k+2}))\,\mathbf{p}. \qquad (3.22)$$

Now we have four consecutive alpha carbons which is enough to calculate the virtual bond and torsion angles . The unit tangent vector $\mathbf{T}_k$ that points from $C_{\alpha,k}$ towards $C_{\alpha,k+1}$ is constructed using

$$\mathbf{T}_k = \frac{\mathbf{r}_{3k+1} - \mathbf{r}_{3k-2}}{|\mathbf{r}_{3k+1} - \mathbf{r}_{3k-2}|} = \frac{\mathscr{R}_x(\Psi_k)\,\mathbf{p}}{\Delta_\alpha}. \tag{3.23}$$

Similarly, the next two tangent vectors are

$$\mathbf{T}_{k+1} = \frac{\mathbf{r}_{3k+4} - \mathbf{r}_{3k+1}}{\Delta_\alpha} = \frac{\mathscr{A}_{3k+1}\mathscr{R}_x(\Psi_{k+1})\,\mathbf{p}}{\Delta_\alpha}, \tag{3.24}$$

$$\mathbf{T}_{k+2} = \frac{\mathbf{r}_{3k+7} - \mathbf{r}_{3k+4}}{\Delta_\alpha} = \frac{\mathscr{A}_{3k+1}\mathscr{A}_{3k+4}\mathscr{R}_x(\Psi_{i+2})\,\mathbf{p}}{\Delta_\alpha}, \tag{3.25}$$

The binormal and vectors of the virtual C_α atoms are computed as in the standard way

$$\mathbf{B}_k = \frac{\mathbf{T}_{k-1} \times \mathbf{T}_k}{|\mathbf{T}_{k-1} \times \mathbf{T}_k|}, \mathbf{N}_k = \mathbf{B}_k \times \mathbf{T}_k. \tag{3.26}$$

Thus the cosine value of virtural bond angle is

$$\begin{aligned}
\cos\theta_k &= \mathbf{T}_{k-1} \cdot \mathbf{T}_k \\
&= \frac{\mathbf{p}^T \mathscr{R}_z(\kappa_2)\mathscr{R}_x(\omega)\mathscr{R}_z(\kappa_3)\mathscr{R}_x(\Phi_i)\mathscr{R}_z(\kappa_1)\mathscr{R}_x(\Psi_i)\,\mathbf{p}}{\Delta_\alpha^2} \\
&= \frac{1}{14.4856}\,(4.72711 - 3.36108\cos\Phi_k - 0.475692\cos\Phi_k\cos\Psi_k \\
&\quad -4.49331\cos\Psi_k + 1.32138\sin\Phi_k\sin\Psi_k)\,.
\end{aligned} \tag{3.27}$$

When $\Phi, \Psi = \pm\pi$, we get the minimum value of $\theta_{\min} = 33.3^\circ$. When $\Phi, \Psi = 0$, we get the maximum value of $\theta_{\max} = 104.4^\circ$. This result consists with the statistics of PDB data (c.f. Fig. 3.4). We can also calculate the virtual torsion angle , as following

$$\begin{aligned}
\cos\gamma_{k+1} &= \mathbf{B}_k \cdot \mathbf{B}_{k+1} = \frac{(\mathbf{T}_{k-1} \times \mathbf{T}_k) \cdot (\mathbf{T}_k \times \mathbf{T}_{k+1})}{|(\mathbf{T}_{k-1} \times \mathbf{T}_k)|\,|(\mathbf{T}_k \times \mathbf{T}_{k+1})|} \\
&= \frac{(\mathbf{T}_{k-1} \cdot \mathbf{T}_k)(\mathbf{T}_k \cdot \mathbf{T}_{k+1}) - |\mathbf{T}_k|^2 (\mathbf{T}_{k-1} \cdot \mathbf{T}_{k+1})}{\sin\theta_k \sin\theta_{k+1}} \\
&= \cot\theta_k \cot\theta_{k+1} - \csc\theta_k \csc\theta_{k+1}\,(\mathbf{T}_{k-1} \cdot \mathbf{T}_{k+1})\,,
\end{aligned} \tag{3.28}$$

and the involved vector product has the explicit form as

$$\Delta_\alpha^2 \mathbf{T}_{k-1} \cdot \mathbf{T}_{k+1} = \cos\Phi_{k+1}(\cos\Psi_{k+1}(0.0522293 - 0.474024\sin\Psi_k \sin\Phi_k) + \sin\Psi_k(-4.49331\sin\Psi_{k+1} - 3.34929\sin\Phi_k) + \cos\Psi_k(1.6119\cos\Psi_{k+1} - 1.32138\sin\Psi_{k+1}\sin\Phi_k + 11.3892) + 0.369034) + \cos\Phi_k(\cos\Psi_{k+1}(1.1461 - 0.171248\sin\Psi_k\sin\Phi_{k+1}) - 1.20998\sin\Psi_k\sin\Phi_{k+1} + 0.103157\sin\Psi_{k+1}\sin\Phi_{k+1} + \cos\Psi_k(0.0371361\cos\Psi_{k+1} - 0.474024\sin\Psi_{k+1}\sin\Phi_{k+1} - 0.0390685) + \cos\Phi_{k+1}(-0.475692\sin\Psi_k\sin\Psi_{k+1} + \cos\Psi_k(0.170647\cos\Psi_{k+1} + 1.20573) - 0.0371361\cos\Psi_{k+1} - 0.262391) - 1.20573) + 0.350782\cos\Psi_{k+1}\cos\Psi_k - 0.369034\cos\Psi_k - 1.6119\cos\Psi_{k+1} + 0.108524\sin\Psi_k\sin\Phi_k - 11.4292\sin\Psi_k\sin\Phi_{k+1} - 0.475692\cos\Psi_{k+1}\cos\Psi_k\sin\Phi_k\sin\Phi_{k+1} - 3.36108\cos\Psi_k\sin\Phi_k\sin\Phi_{k+1} - 0.103157\sin\Psi_k\cos\Psi_{k+1}\sin\Phi_k - 1.61758\sin\Psi_k\cos\Psi_{k+1}\sin\Phi_{k+1} + \sin\Psi_{k+1}\sin\Phi_{k+1}(-4.47755\cos\Psi_k + 1.31675\sin\Psi_k\sin\Phi_k - 0.145083) + 1.69578.$$

In similar way we can compute the sign of the torsion angle. Though the above result is somewhat complicated, the general relation can be read out as

$$\theta_k = \theta_k(\Phi_k, \Psi_k), \quad \gamma_{k+1} = \gamma_{k+1}(\Phi_k, \Psi_k, \Phi_{k+1}, \Psi_{k+1}). \tag{3.29}$$

It indicates that if we are given four C_α atoms, or equivalently $(\theta_k, \theta_{k+1}, \gamma_{k+1})$, then we need one more condition to determine two sets of Ramachandran angles $(\Phi_k, \Psi_k, \Phi_{k+1}, \Psi_{k+1})$. This extra condition is enough to decide the rest part of protein chain by iterating the above relation. So in principle there is essentially only one different degrees of freedom between $NC_\alpha C$ representation and C_α-only representation. Yet in reality it is not true. Even small derivation of covalent bond lengths and angles from their average values will result in a large difference of ending position in the above iterative construction. What is addressed in Paper IV is the same issue from another aspect.

We can use the formula to calculate the virtual bond and torsion angles of typical secondary elements. The result is shown in Table 3.2. These regularly occurring elements exhibit as either major or minor peaks on the Ramachadran plot (Fig. 3.4(a)). Equivalently, we do the statistics of virtual bond and torsion angles, as shown on Fig. 3.4(b-d). As expected, such a distribution also reflect the secondary structure preferences. On the plots, we distinguish two virtual bond angles θ_1 and θ_2 at two consecutive C_α atoms, as different pairings with the virtual torsion angle γ (see Fig. 2.2). That is, θ_1 of the first three C_α atoms is plotted again either γ or θ_2 of the last three C_α atoms. Surprisingly, the $\theta_1 - \theta_2$ distribution plot is not symmetric. This feature is also shown by the difference between $\theta_1 - \gamma$ and $\theta_2 - \gamma$ plots. In the future we shall analyze the structure clusters by combing all these three subplots.

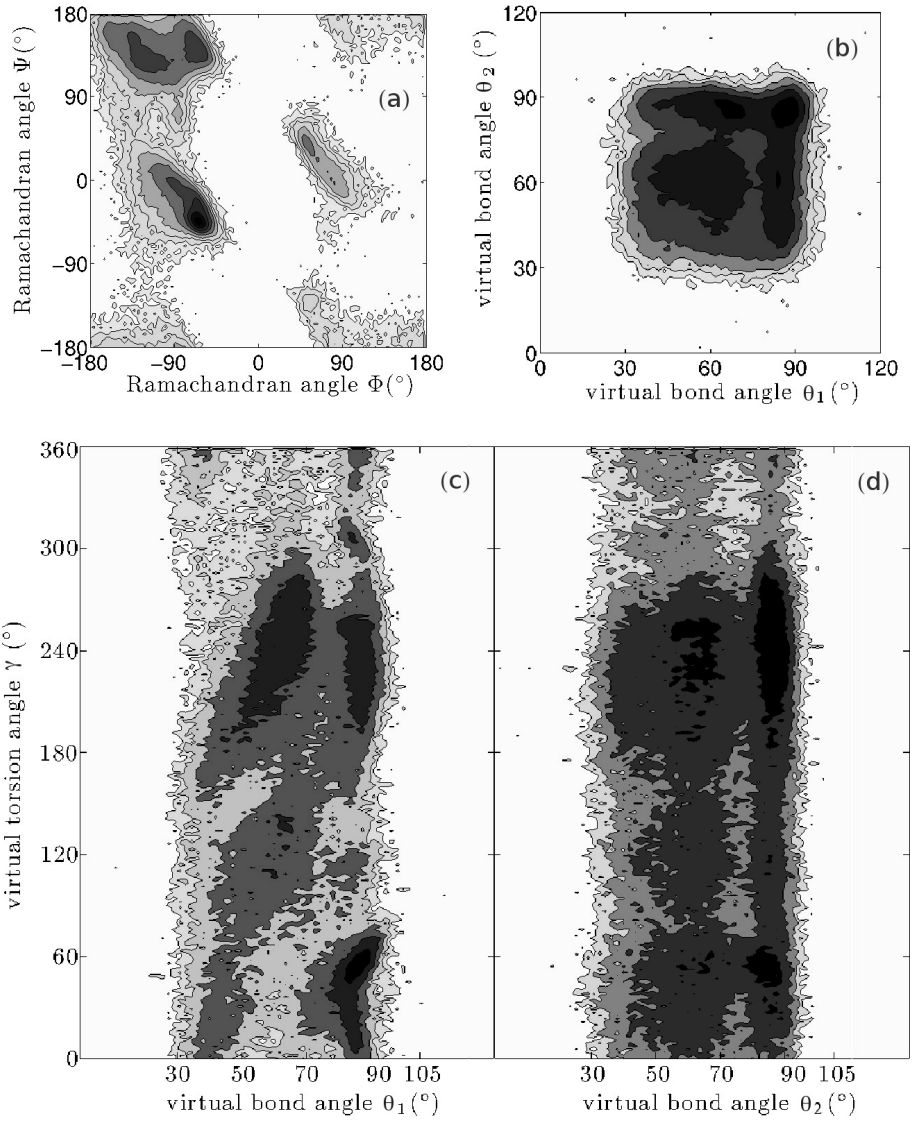

Figure 3.4. (a) Distribution of Ramachandran angles Φ, Ψ. (b-d) Distribution of virtual bond angle θ_1, θ_2 and torsion angle γ. Statistics is done over a pruned set of PDB structures, the same as in Paper IV. Density is in logarithmic scale. On subplots (c-d) the torsion angle τ has been shifted by 360° if it is negative, for the purpose of better showing the clusters. It is interesting to compare the clusters with the ideal value in Table 3.2.

	α-helix	β-sheet	polyproline helix	L-helix
(Φ,Ψ)	$(-63^\circ,-43^\circ)$	$(-120^\circ,130^\circ)$	$(-75^\circ,150^\circ)$	$(50^\circ,50^\circ)$
(θ,γ)	$(88^\circ,50^\circ)$	$(55^\circ,-168^\circ)$	$(60^\circ,-106^\circ)$	$(89^\circ,-56^\circ)$

Table 3.2. *The values of geometric values of four regular secondary structures.*

3.3.2 Remark on three other applications to protein geometry

We finally remark on three more applications of differential geometry on protein structure modeling. The first one is a soliton model of protein backbone. Lots of our effort has been focused on this direction from the very beginning of the project. In Paper I and in [36, 6, 31] we have extensively addressed on this idea. The topological soliton has been identified as the helix-loop-helix motif in protein, in contrast with the classical Davydov soliton which representing an excitation propagation along the α-helix self-trapped amide I [9, 8, 29].

The second application is to address the question that to what extent the various angles and bond lengths can be replaced by their average values. For protein structure modeling and prediction people construction of coarse grained force fields which take only a subset of the full set of atomic degrees of freedom as dynamically active variables. Along the main chain these dynamical variables are identified as Ramachandran angles (Φ,Ψ). In Paper IV, we show that a coarse graining where a subset of angular variables is replaced by uniform values, commonly yields geometrically incorrect protein structures. Related work also shows that the cases peptide bonds strongly deviating from planarity (ω other than π or 0) is not rare, instead, "conserved but not biased toward active sites" [3].

The third possible application comes to structural alignment of protein chains. Though lots of relating work has been done on this subject (see, for example, review [20]), there is still need to design more reliable, fast and automatic algorithms. In most existing methods, the core idea is to find a good definition of similarity for structure segment. Here we utilize the discrete Frenet equations to propose a new definition of similarity measurement. Still refer to Fig. 2.2. For four consecutive C_α atoms, we have two virtual bond angles θ_1,θ_2 and one virtual torsion angles. We can use them to form some sort of hyperspherical coordinates on S^3, e.g.

$$\mathbf{u} = (\cos\theta_1, \sin\theta_1\cos\theta_2, \sin\theta_1\sin\theta_2\cos\gamma, \sin\theta_1\sin\theta_2\sin\gamma)\,. \tag{3.30}$$

Then the similarity would be simply the dot product between structure segments from each chain. By evaluating the similarity between successive fragments, we can further calculate a distance matrix. Based on this distance matrix, methods like dynamical programming would find optimal alignment.

4. Time evolution of space curves

In this chapter we are interested in a typical model of a moving space curve, the local induction motion. It was originally constructed to describe the motion of vortices, which are regions of fluid flow that rotate around a central axis. Other typical examples of a moving space curve in three dimensions include the winds surrounding hurricanes, vortex filaments in superfluids and especially a protein chain in the cell.

The local induction model has been well known to support an integrable structure, allowing one to take global analysis about the system behavior. For example, a system governed by integrable evolution equation has an infinite number of integrals of motion, which can be tackled via the inverse scattering transform method [17]. Mathematical structure of an integrable system is also associated with a Lax pair and interesting solutions such as solitons. A soliton emerges when nonlinear interactions combine elementary constituents into a localized collective excitation that are stable with respect to weak perturbations and behave like a particle with invariant shape and velocity. When two solitons collide, they first merge into one and then separate into two with the same shape and velocity as before the collision.

Here we firstly review the integrable motion of a continuous curve. Then we emphasize on the discrete version of the model, which preserves the integrable structure. The theoretical study has provided benefits to the applications for learning bifurcation behavior of closed curves (addressed in Paper III), and for designing algorithm of loop closure.

4.1 Integrable motion of continuous curves

Under the local induction approximation, Levi-Civita and Da Rios discovered on 1906 [44] a time evolution equation for closed curves as the model of very thin isolated filament $\mathbf{r} = \mathbf{r}(s,t)$ in incompressible fluid,

$$\frac{d\mathbf{r}}{dt} = \frac{d\mathbf{r}}{ds} \times \frac{d^2\mathbf{r}}{ds^2}. \tag{4.1}$$

This equation is also known as the smoke ring flow. Associating the curve $\mathbf{r}$ with Frenet frame $(\mathbf{t}, \mathbf{n}, \mathbf{b})$, we can transform Eq. (4.1) into

$$\frac{d\mathbf{r}}{dt} = \mathbf{t} \times \kappa\mathbf{n} = \kappa\mathbf{b}. \tag{4.2}$$

Sometimes the above equation is also called binormal curvature flow, as its form indicates. In 1972 Hasimoto [2] discovered that by introducing a map

$$\psi(s) = \kappa(s) \exp(i \int_0^s \tau ds') \tag{4.3}$$

the local induction motion is in fact equivalent to the nonlinear Schrödinger (NLS) equation

$$i\partial_t \psi = -\partial_s^2 \psi - \frac{1}{2} |\psi|^2 \psi, \tag{4.4}$$

which is a completely integrable Hamiltonian system. From Eq. (4.1), it is easy to observe that the arc-length is differentially conserved by the flow:

$$\partial_t |\partial_s \mathbf{r}|^2 = 2\partial_s \mathbf{r} \cdot \partial_s (\partial_s \mathbf{r} \times \partial_{ss} \mathbf{r}) = 2\partial_s \mathbf{r} \cdot (\partial_s \mathbf{r} \times \partial_{sss} \mathbf{r}) = 0. \tag{4.5}$$

In similar way one can show that the total squared curvature, $\int |\partial_{ss} \mathbf{r}|^2 ds = \int \kappa^2 ds$ is also conserved. In fact, there are infinitely many more conserved quantities, i.e. the system is completely integrable. In Paper II, we systematically derive these conserved quantities.

The corresponding Poisson structure of NLS equation is

$$\begin{aligned} H &= \int ds \left(|\partial_s \psi|^2 - \frac{1}{4} |\psi|^4 \right), \\ \left\{ \psi(s), \bar{\psi}(s') \right\} &= i\delta(s - s'). \end{aligned} \tag{4.6}$$

In terms of curvature and torsion, the above equation translates into

$$\begin{aligned} H &= \int ds \left((\partial_s \kappa)^2 + \kappa^2 \tau^2 - \frac{1}{4} \kappa^4 \right), \\ \left\{ \kappa(s), \tau(s') \right\} &= \frac{1}{2\kappa(s)} \frac{\partial}{\partial s} \delta(s - s'). \end{aligned} \tag{4.7}$$

Equation of motion is then followed

$$\partial_t \kappa = -2(\partial_s \kappa)\tau - \kappa \partial_s \tau, \tag{4.8}$$

$$\partial_t \tau = \frac{\partial}{\partial s} \left(\frac{\partial_s^2 \kappa - \kappa \tau^2 + \frac{1}{2} \kappa^3}{\kappa} \right). \tag{4.9}$$

4.2 Integrable motion of polygonal curves

We expect a suitable discretization scheme would preserve the integrable structure. Lattice Heisenberg model (LHM) [17] is such a natural discretization of local induction equation. The integrable LHM model has the Poisson structure

as

$$H = -\frac{2}{\delta^2}\sum_i \log\left(1+\mathbf{t}_i\cdot\mathbf{t}_{i+1}\right),$$
$$\left\{\mathbf{t}_i^a, \mathbf{t}_j^b\right\} = -\varepsilon^{abc}\mathbf{t}_i^c\,\delta_{ij}. \tag{4.10}$$

The corresponding Heisenberg flow is as following

$$\begin{aligned}\frac{d\mathbf{t}_i}{dt} &= \{H,\mathbf{t}_i\} = -\mathbf{t}_i\times\frac{\partial H}{\partial \mathbf{t}_i}\\ &= \frac{2}{\delta^2}\left(\frac{\mathbf{t}_i\times\mathbf{t}_{i+1}}{1+\mathbf{t}_i\cdot\mathbf{t}_{i+1}} - \frac{\mathbf{t}_{i-1}\times\mathbf{t}_i}{1+\mathbf{t}_{i-1}\cdot\mathbf{t}_i}\right)\\ &= \frac{2}{\delta^2}\left(\tan\frac{\kappa_{i+1}}{2}\mathbf{b}_{i+1} - \tan\frac{\kappa_i}{2}\mathbf{b}_i\right).\end{aligned} \tag{4.11}$$

In the last line, relations from discrete Frenet equations (Eq. (3.9)) have been introduced. One feature of the model is its preservation of the end-to-end distance of the curve

$$\frac{d}{dt}\sum_{i=1}^N \mathbf{t}_i = \frac{2}{\delta^2}\sum_{i=1}^N\left(\tan\frac{\kappa_{i+1}}{2}\mathbf{b}_{i+1} - \tan\frac{\kappa_i}{2}\mathbf{b}_i\right) = 0. \tag{4.12}$$

Here the periodic condition of the closed curve is assumed, i.e. $\mathbf{t}_i = \mathbf{t}_{N+i}$. The conclusion is also true for an open curve. This feature can be regarded as the discrete analogue of length conservation as in the continuous case (see Eq. (4.5)).

The curve is then reconstructed as $\mathbf{r}_i = \mathbf{r}_{i-1} + \delta\mathbf{t}_{i-1}$. In the explicit way the flows on $\mathbf{r}_i$ reads

$$\frac{d\mathbf{r}_i}{dt} = \frac{2}{\delta}\frac{\mathbf{t}_{i-1}\times\mathbf{t}_i}{1+\mathbf{t}_{i-1}\cdot\mathbf{t}_i} = \frac{2}{\delta}\tan\frac{\kappa_i}{2}\mathbf{b}_i. \tag{4.13}$$

Compared with Eq. (4.2), we can define the discrete curvature as

$$\kappa(s = i\delta) \to \frac{2}{\delta}\tan\frac{\kappa_i}{2}. \tag{4.14}$$

The denotation may be a bit confusing but κ_i on the right-hand side is bond angle. When $\kappa_i \to \pi$, the discrete curvature diverges; in consequence, Hamiltonian in terms of squared discrete curvature doesn't prefer super-bending structure. This is particularly useful to model the C_α structure since the bond angle there is limited to $[33.3^\circ, 104.4^\circ]$.

Here we would like to start from the Poisson bracket in Eq. (4.10) to derive the Poisson brackets between bond/torsion angles, for the purpose of future application of more general interactions. Since one bond angle involve two consecutive tangent vectors while one torsion angle involve three, there are seven non-vanishing brackets, i.e. $\{\kappa_i, \kappa_{i+1}\}$, $\{\kappa_{i-2}, \tau_i\}$, $\{\kappa_{i-1}, \tau_i\}$, $\{\kappa_i, \tau_i\}$,

$\{\kappa_{i+1},\tau_i\}$, $\{\tau_{i-2},\tau_i\}$ and $\{\tau_{i-1},\tau_i\}$. Since the calculation takes the same skill for each bracket, here only the details of computing $\{\kappa_i,\kappa_{i+1}\}$ is given, as following

$$\begin{aligned}
\{\cos\kappa_i,\cos\kappa_{i+1}\} &= \{\mathbf{t}_i\cdot\mathbf{t}_{i-1},\mathbf{t}_i\cdot\mathbf{t}_{i+1}\} \\
&= \mathbf{t}_i\cdot\{\mathbf{t}_{i-1},\mathbf{t}_i\cdot\mathbf{t}_{i+1}\}+\{\mathbf{t}_i,\mathbf{t}_i\cdot\mathbf{t}_{i+1}\}\cdot\mathbf{t}_{i-1} \\
&= \mathbf{t}_i\cdot\left(\mathbf{t}_{i-1}\times\frac{\partial(\mathbf{t}_i\cdot\mathbf{t}_{i+1})}{\partial\mathbf{t}_{i-1}}\right)+\left(\mathbf{t}_i\times\frac{\partial(\mathbf{t}_i\cdot\mathbf{t}_{i+1})}{\partial\mathbf{t}_i}\right)\cdot\mathbf{t}_{i-1} \\
&= 0+\mathbf{t}_i\times\mathbf{t}_{i+1}\cdot\mathbf{t}_{i-1} \\
&= \mathbf{t}_{i-1}\times\mathbf{t}_i\cdot\mathbf{t}_{i+1} \\
&= \sin\kappa_i\sin\tau_{i+1}\sin\kappa_{i+1}. \qquad (4.15)
\end{aligned}$$

So we get

$$\{\kappa_i,\kappa_{i+1}\}=\sin\tau_{i+1}. \qquad (4.16)$$

In the similar way, we can calculate the other brackets. The results are summarized as following

$$\begin{aligned}
\{\kappa_i,\kappa_{i+1}\} &= \sin\tau_{i+1}, \\
\{\kappa_{i-2},\tau_i\} &= -\cos\tau_{i-1}\csc\kappa_{i-1}, \\
\{\kappa_{i-1},\tau_i\} &= \cot\frac{\kappa_{i-1}}{2}+\cos\tau_i\cot\kappa_i, \\
\{\kappa_i,\tau_i\} &= -\cot\frac{\kappa_i}{2}-\cos\tau_i\cot\kappa_{i-1}, \qquad (4.17) \\
\{\kappa_{i+1},\tau_i\} &= \cos\tau_{i+1}\csc\kappa_i, \\
\{\tau_{i-1},\tau_i\} &= \csc\kappa_{i-1}\left(\sin\tau_i\cot\kappa_i+\sin\tau_{i-1}\cot\kappa_{i-2}\right), \\
\{\tau_{i-1},\tau_{i+1}\} &= \sin\tau_i\csc\kappa_{i-1}\csc\kappa_i.
\end{aligned}$$

At the same time, the Hamiltonian is translated into (the factor $\frac{2}{\delta^2}$ has been rescaled to be one)

$$H=-\sum_i\log\left(1+\mathbf{t}_i\cdot\mathbf{t}_{i+1}\right)=-2\sum_i\log\cos\frac{\kappa_i}{2}. \qquad (4.18)$$

And the equation of motion (4.11) becomes

$$\frac{d\kappa_i}{dt}=\tan\frac{\kappa_{i-1}}{2}\sin\tau_i-\tan\frac{\kappa_{i+1}}{2}\sin\tau_{i+1}. \qquad (4.19)$$

$$\begin{aligned}
\frac{d\tau_i}{dt} &= \cos\tau_i\left(\cot\kappa_i\tan\frac{\kappa_{i-1}}{2}-\cot\kappa_{i-1}\tan\frac{\kappa_i}{2}\right) \\
&\quad+\tan\frac{\kappa_{i+1}}{2}\cos\tau_{i+1}\csc\kappa_i-\tan\frac{\kappa_{i-2}}{2}\csc\kappa_{i-1}\cos\tau_{i-1}. \qquad (4.20)
\end{aligned}$$

It is also straightforward to check the Jacobi identity.

Define the discrete Hasimoto map as

$$\psi_i = \tan\frac{\kappa_i}{2} e^{i\vartheta_i},\ \vartheta_i = \frac{1}{2}\left(\sum_{k=1}^{i}\tau_k - \sum_{k=i+1}^{N}\tau_k\right). \tag{4.21}$$

Combining both Eq. (4.19) and Eq. (4.20) we otain the LNS_2 equation [17] (a factor of 2 has been rescaled)

$$i\frac{d\psi_i}{dt} = -\left(\psi_{i+1} - 2\psi_i + \psi_{i-1}\right) - |\psi_i|^2\left(\psi_{i+1} + \psi_{i-1}\right). \tag{4.22}$$

The corresponding Poisson structure is

$$H = -\sum_i \left(\psi_i\bar{\psi}_{i+1} + \bar{\psi}_i q_{i+1}\right) = -2\sum_i \tan\frac{\kappa_i}{2}\tan\frac{\kappa_{i+1}}{2}\cos\tau_{i+1}, \tag{4.23}$$

$$\left\{\psi_i, \bar{\psi}_j\right\} = i\left(1 + |\psi_i|^2\right)\delta_{ij}, \left\{\psi_i, \psi_j\right\} = \left\{\bar{\psi}_i, \bar{\psi}_j\right\} = 0. \tag{4.24}$$

Thus we have established the equivalence between the LHM and LNS_2 models, by a direct calculation in terms of the discrete Frenet equations. More interesting results are presented in Paper II. Similar work from different approaches can be found in [1, 26, 24].

4.3 Binormal flow algorithm for loop closure

The theoretical study of local induction model in previous section has shown an interesting feature that the equation of motion preserves the end-to-end distance (see Eq. (4.12)). This feature inspired us to devise an algorithm for loop closure in protein structure modeling. Loop closure problem requires to efficiently construct a protein segment for matching two fixed endings. This problem arises either in homology modeling or in *de novo* structure prediction (for example, see review [34]).

We try to circumvent the loop closure problem by two general steps. The first step is to generate a segment of given end-to-end distance (defined by the two fixed target points), from an arbitrary configuration. The second step is to globally move the segment to bridge the fixed target points. This global move is simply a translation and a rotation. But if there is steric hindrance between the segment and the rest of the protein structure, further deformation of the segment is possible under the motion of Eq. (4.12) that preserve end-to-end distance. So in principle, the second step is always solvable and thus not considered here. In sequel we focus on the first step, that is, how to generate a segment of a given end-to-end distance. Our method doesn't need to depend on the physical energy, so it is essentially a sampling strategy.

The dynamical variables we choose are the tangent vector defined on the bond C_α-C, denoted as $\mathbf{t}_{1,k}$, k is the index of residue (see Table 3.1), and on

the bond C-N, denoted as $\mathbf{t}_{2,k}$. The tangent vector on the bond N-C_α (denoted as $\mathbf{t}_{3,k}$) will be taken as dummy variable so that its motion is passive to keep ω invariant. Apparently, in order to preserve the bond angle $\kappa_1, \kappa_2, \kappa_3$, the motion of a tangent vector be restricted on the cone defined by the bond angle. Mathematically the general equation of motion would be

$$\frac{d\mathbf{t}_i}{d\xi} = c_i \mathbf{t}_i \times \mathbf{t}_{i-1}, \tag{4.25}$$

which preserves the bond angle for arbitrary difference step $\delta\xi$,

$$\cos\kappa_i^{\text{new}} = \mathbf{t}_{i-1} \cdot \left(\mathbf{t}_i + \delta\xi \frac{d\mathbf{t}_i}{d\xi} \right) = \mathbf{t}_{i-1} \cdot (\mathbf{t}_i + \delta\xi c_i \mathbf{t}_i \times \mathbf{t}_{i-1}) = \cos\kappa_i. \tag{4.26}$$

The variable ξ is not the real time but denote the searching time of loop closure in the configuration spacetime. Similar with Eq. (4.2), the motion of $\mathbf{t}_i$ is along the binormal vector direction, giving the method the name of *binormal motion algorithm* (BMA). In practice, we start from the N-terminal and change the tangent vectors $\mathbf{t}_i$'s in the sequential way or in the random way. When in the sequential way, we first update the beginning tangent vector according to Eq. (4.25) which has an analytical solution using Rodrigue's rotation. This rotation has to globally apply on all the atoms in the rest part of the chain. Moving along the chain we can change the tangent vectors (only $\mathbf{t}_1$'s and $\mathbf{t}_2$'s) in the same manner, until the end. When in the random way, tangent vector at random site is updated and rotation has to be done over the rest part of the chain. By this way, we can remove the bias of larger changing for the beginning part as in the sequential way.

It is worthy to notice that the tangent vector in Eq. (4.25) isn't necessarily normalized, in other words, the equation automatically preserves the bond length. This observation in fact indicates the separation of the radial and angular part of the bond motion. The bond vibration along radial direction can be further treated by a different way.

The motion according to Eq. (4.25) is arbitrary, similar to the pivot move in Monte Carlo approach. However it can be a directed move if we associate the coefficient c_i with some target function (not necessarily the physical interaction). For example, in loop closure problem we propose a choice as

$$c_i = \mathbf{B}_i \cdot \mathbf{t}_{i-1}, \;\; \mathbf{B}_i = -\frac{2}{\delta}\left(1 - \frac{d_0}{|\mathbf{R}_e|}\right)\mathbf{R}_e, \tag{4.27}$$

where $\mathbf{R}_e = \mathbf{r}_N - \mathbf{r}_1 = \sum_{i=1}^{N} \mathbf{t}_i$ is the end-to-end vector of the moving segment, and d_0 is the distance between two target points. Intuitively, this equation tries to minimize the difference between d_0 and $|\mathbf{R}_e|$. The vector $\mathbf{B}_i$ is the analogue of magnetic field as in Landau-Lifshitz spin model [17].

Some local interactions can be introduced as well to model the Ramachandran torsion angles preferences. For instance, we can write a local interaction as $f_k = f_k(\cos\Psi_k, \cos\Phi_k)$ which is assumed to depend on the amino acid

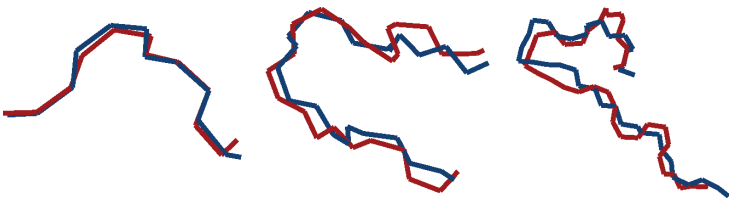

Figure 4.1. Three typical loop closure results with lowest RMS distance, generated from 5000 sampling of BMA. Left: 1qnrA_195-198; middle: 1ctqA_144-151; right: 1f74A_11-22. For each pair, in red is the native structure read from PDB. In blue is the simulation result.

residue type. The exact form can be derived from the statistics over some known structure dataset. Then the explicit form of $\mathbf{B}_i$ is computed as follows,

$$\begin{aligned}
\mathbf{B_{1,k}} &= -\frac{\partial f_k}{\partial \mathbf{t}_{1,k}} = -\frac{\partial f_k}{\partial \cos\Phi_k}\frac{\partial \cos\Phi_k}{\partial \mathbf{t}_{1,k}}, \\
&= \frac{\partial f_k}{\partial \cos\Phi_k}\csc\kappa_3 \csc\kappa_1 \mathbf{t}_{2,k-1}, \qquad (4.28) \\
\mathbf{B_{2,k}} &= -\frac{\partial f_k}{\partial \mathbf{t}_{2,k}} = -\frac{\partial f_k}{\partial \cos\Psi_k}\frac{\partial \cos\Psi_k}{\partial \mathbf{t}_{2,k}}, \\
&= \frac{\partial f_k}{\partial \cos\Psi_k}\csc\kappa_1 \csc\kappa_2 \mathbf{t}_{3,k-1}. \qquad (4.29)
\end{aligned}$$

Again the similar calculation techniques as in Eq. (4.15) are used here.

In Table 4.1 we show a parallel simulation (no consideration of local interaction), as comparison with the classical approach of Cyclic Coordinate Descent (CCD) method [5]. In Fig. 4.1 are shown three typical loop closure results with lowest RMS distance, generated from 5000 sampling of BMA. Overall, the performance of BMA is similar with or better than the CCD result, especially for the longer chain. There are two reasons for this improvement. The first is due to the reduce search space of the configurations. Let us remember BMA focus on the end-to-end distance instead of ending position. This simplification reduce the degenerate configurations related by a translation or rotation. The second reason lies on the intrinsic rotation feature of the present algorithm. When moving, each tangent vector rotates around the end-to-end vector. This results a twisting configuration that matches the global feature of loops.

Of course we shall remind that we still need a further rotation to move the generated segment toward the target positions. Those samples of segment that have other than minimum RMS distance might be more consistent with the rest part of the whole protein chain. Combined with an energy function the present algorithm would model the loop structure in more realistic way. This shall be the direction of the future work.

Length 4		
Loop	Min RMSD (Å)	
	CCD	BFA
1dvjA_20-23	0.606	0.564
1dysA_47-50	0.676	0.330
1eguA_404-407	0.675	0.379
1ej0A_74-77	0.337	0.322
1i0hA_123-126	0.616	0.336
1id0A_405-408	0.671	0.276
1qnrA_195-198	0.491	0.260
1qopA_44-47	0.627	0.547
1tca_95-98	0.393	0.825
1thfD_121-124	0.495	0.347
Avg. min RMSD	0.559	0.419

length 8		
Loop	Min RMSD (Å)	
	CCD	BFA
1cruA_85-92	1.753	1.133
1ctqA_144-151	1.344	0.859
1d8wA_334-341	1.506	1.063
1ds1A_20-27	1.581	1.530
1gk8A_122-129	1.684	1.117
1i0hA_145-152	1.351	1.482
1ixh_106-113	1.605	1.483
1lam_420-427	1.604	1.297
1qopB_14-21	1.849	1.488
3chbD_51-58	1.659	1.106
Avg. min RMSD	1.594	1.256

Length 12		
Loop	Min RMSD (Å)	
	CCD	BFA
1cruA_358-369	2.538	1.527
1ctqA_26-37	2.487	1.506
1d4oA_88-99	2.487	1.838
1d8wA_46-57	4.827	1.914
1ds1A_282-293	3.042	1.789
1dysA_291-302	2.478	1.624
1eguA_508-519	2.137	1.804
1f74A_11-22	2.715	1.511
1q1wA_31-42	3.378	1.686
1qopA_178-189	4.568	1.740
Avg. min RMSD	3.050	1.694

Table 4.1. *Minmum RMSD from X-ray structure in 5000 trials per loop. We take the same test loops as in CCD [5]. The loop is considered to be closed if the difference between the moving end-to-end distance and the target distance is less than 0.08Å. The maximum search step is limited to 5000. The difference step $\delta\xi$ is taken as 0.02.*

5. Concluding remark

Now let us have a very general overview of what we have done. We utilize the transfer matrix formalism to consecutively map the discrete Frenet frame from one vertex of discrete curve to its neighbor. This intrinsically discrete approach enables us to conveniently describe curves, the backbone of folded proteins for example, for which the continuum limit has a nontrivial Hausdorff dimension. Integrability study and bifurcations analysis further provides us the insights on the time evolution of the curves. The theoretical study has inspired us the applications on protein structure modeling, such soliton model of protein backbone, and the binormal motion algorithm of loop closure.

Though our the framework remains within a homopolymer representation of the protein structure, our work with the folded protein suggests that protein folding is probably subject to the geometric constraints, at least at the early stage when the non-covalent interactions have not yet dominated the driving force. The further introduction of amino acid specific information, even within a hydrophobic-hydrophilic scheme, should be compatible with the shape geometry of the protein backbone. The competition between them would probably help to successfully discriminate the correct target fold, as well as to suitably describe the dynamics of folding process. These considerations would be the directions of the future work.

Loop modeling would be of great interest since most of protein functional sites lie on the loop region, instead of on the regular α-helices or β-sheets. Our work have shown two aspects of loop structure. One is the soliton model of main chain, in which the helix-loop-helix motif has been identified with soliton solution of the gauge-invariant energy functional. On the other hand, we have devised an efficient algorithm for loop closure sampling. We hope we can combing both approaches and then provide better modeling of loop structure. One possibility would be the consideration of evolution, in order to efficiently distinguish the conserved sites that might be related with functions.

On even more general scope, it is always helpful to keep in mind those big problems, such as what's the drive force and mechanism of protein folding, how to design a protein sequence from given structure, what happen exactly inside of the diseases like Alzheimer's. The work in this thesis might be only very tiny step towards the answer of these big problems, but that's enough. Maybe what I have done is no better than a folding game on the computer. Yet life has no ending. I hope there is always something to fold.

Acknowledgments

First of all I would like to thank my supervisor Antti J. Niemi for his great support and for lots of inspiring discussion together. Every time I suffer frustration or doubt on research, he always showed his optimistic encouragement. Those moments, together with the exciting discovery, label our research pathway. His considerable guidance within and beyond this thesis, I sincerely acknowledge.

I also appreciate every member in our *Folding Proteins* group for working together and sharing good time together. If I calculate the path integral of my graduate work, I can't ignore your inspiring discussions in the every stage of this work. I am haunted by lots of memory when I am now writing down your names: Martin, Di, Yan, Nora, Xubiao, Fan, Andrey, Yifan, Jin and Alireza.

Collaborators including Maxim Chernodub from Tours, Ying Jiang from Shanghai, Gerald Kneller and Konrad Hinsen from Orléans also deserve to be thanked for the helpful discussions we have had. Specially thank Ying for his warm hospitality when I visited Shanghai.

I am also grateful to those with whom I worked before coming to Uppsala. Special thanks to my master supervisor Molin Ge, for his providing me my first real-world research experience and to Jing Sheng for showing me the fantastic world of polymer entanglement. Meanwhile thank my postdoc advisor Alessandra Carbone for giving me the chance of continuing to play with protein.

Many thanks also to all faculty members, postdocs and graduate students at my dual academic affiliations. Life in Uppsala and in Tours are much different but both are great places to work and I enjoyed the time having you around. Working, staying or even having lunch or fika with you have been my good time to memorize.

Thank you as well to all my lovely friends for loyalty, enthusiasm and laughters. You know I miss you, for sharing time with me, for your intelligence sparkles, for the days we traveled together, for the invitation to homemake diners, for the national sailing competition we beat, for together taking the crazy boat race on Valborg celebration. Feel lucky to meet you all.

Thank to Google and Wikipedia. I know this thank will definitely get learned.

Finally, my warmest thanks go to my family for all their love, moral and support. Even though they never completely understand my project, they still provide me complete love. I love you with all my heart.

Summary in Swedish

För att utföra en viss funktion, så behöver ett protein veckas ihop till korrekt struktur. Huvudfokus i denna avhandling är att, med hjälp av differentialgeometri och generella koncept som gaugeinvarians och solitoner, skapa teoretiska modeller av veckade proteiners struktur.

Ett proteins ryggrad kan sägas vara en styckvist linjär flersidig kedja, där hörnen är de centrala C_α kolatomerna i aminosyrorna. Konstruktionen kan beskrivas av Frenetekvationer, vilket förtydligar gaugestrukturen och leder till en effektiv energifunktional. En speciell, topologiskt stabil lösning kallad soliton har kunnat relateras till helix-loop-helixmotivet i proteinstrukturen. Parametrarna som karaktäriserar hur ett protein veckar sig är globala på den sekundära nivån, och därmed definierade bortom alla detaljer och komplexitet av aminosyror och deras interaktioner. Huvudkedjans veckning till fullständiga protein byggs därmed upp genom att multipla solitoner monteras ihop. Vi har funnit att modelleringen av ett antal biologiskt aktiva protein återskapar den ursprungliga strukturen med experimentell noggrannhet.

Motiverade av dessa framgångsrika tillämpningar av kurvteori för simulering av proteinstrukturer har vi retroaktivt undersökt de teoretiska egenskaperna hos kurvor i tre dimensioner. Först härleder vi Hamilton-energifunktioner för både kontinuerliga kurvor och diskretiserade kedjor, inom ramen för invariansprincipen och integrabla hierarkier. För kontinuerliga kurvor finner vi att en Weyldual existerar för den icke-linjära Schrödingerekvationshierarkin, vilket även relaterar till energidensiteter relevanta för strängar i tre rumsdimensioner. Vi föreslår att denna ytterligare hierarki också är integrerbar, och undersöker detta explicit till den första icke-triviala ordningen. En korrekt diskretiserad version av den icke-linjära Schrödingerekvationen, som bevarar integrabiliteten, har granskats för jämförelse med de kontinuerliga motsvarigheterna.

Vi undersöker även bifurkationen genom tidsutvecklingen hos en sluten ramad kurva. Vi argumenterar att dessa strängar uppför sig som ett band, men med en utökad repertoar som inkluderar inflektionspunkter: självlänkningstalet är inte en global topologisk invariant, utan gör diskontinuerliga hopp när perestrojka inträffar.

Vi hoppas att vårt teoretiska angrepssätt kan ge en systematisk grund för den generella beskrivningen av både kontinuerliga och diskreta sträng-lika konfigurationer i tre dimensioner, i synnerhet rumsutfyllande strukturer. Baserat på detta tillvägagångssätt och den fortsatta introduktionen av aminosyrors speci-

fika information, förväntar vi oss en mer realistisk modellering av proteinstrukturer i framtiden.

References

[1] Ablowitz, M. J., & Ladik, J. F. (1976). *Nonlinear differential-difference equations and Fourier analysis.* Journal of Mathematical Physics, 17, 1011.

[2] Anfinsen, C. B. (1956). *The limited digestion of ribonuclease with pepsin.* Journal of Biological Chemistry, 221(1), 405-412.

[3] Berkholz, D. S., Driggers, C. M., Shapovalov, M. V., Dunbrack, R. L., & Karplus, P. A. (2012). *Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites.* Proceedings of the National Academy of Sciences, 109(2), 449-453.

[4] Buck, G. (1998). *Most smooth closed space curves contain approximate solutions of the n-body problem.* Nature, 395(6697), 51-53.

[5] Canutescu, A. A., & Dunbrack, R. L. (2003). *Cyclic coordinate descent: A robotics algorithm for protein loop closure.* Protein Science, 12(5), 963-972.

[6] Chernodub, M., Hu, S., & Niemi, A. J. (2010). *Topological solitons and folded proteins.* Physical Review E, 82(1), 011916.

[7] Clementi, C. (2008). *Coarse-grained models of protein folding: toy models or predictive tools?.* Current Opinion in Structural Biology, 18(1), 10-15.

[8] Dauxois, T., & Peyrard, M. (2006). *Physics of solitons.* Cambridge University Press.

[9] Davydov, A. S. (1973). *The theory of contraction of proteins under their excitation.* Journal of Theoretical Biology, 38(3), 559-569.

[10] Dill, K. A. (1999). *Polymer principles and protein folding.* Protein Science, 8(06), 1166-1180.

[11] Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). *The protein folding problem.* Annual Review of Biophysics, 37, 289.

[12] Dobson, C. M., & Karplus, M. (1999). *The fundamentals of protein folding: bringing together theory and experiment.* Current Opinion in Structural Biology, 9(1), 92-101.

[13] Edwards, S. F. (1965). *The statistical mechanics of polymers with excluded volume.* Proceedings of the Physical Society, 85(4), 613.

[14] Edwards, S. F. (1966). *The theory of polymer solutions at intermediate concentration.* Proceedings of the Physical Society, 88(2), 265.

[15] Edwards, S. F. (1967). *Statistical mechanics with topological constraints: I.* Proceedings of the Physical Society, 91(3), 513.

[16] Edwards, S. F. (1967). *The statistical mechanics of polymerized material.* Proceedings of the Physical Society, 92(1), 9.

[17] Faddeev, L. D., Takhtajan, L. A., & Reyman, A. G. (2007). *Hamiltonian methods in the theory of solitons.* Springer.

[18] Flory, P., & Volkenstein, M. (1969). *Statistical mechanics of chain molecules.* Biopolymers, 8(5), 699-700.

[19] Frenet, F. (1852). *Sur les courbes à double courbure.* Journal de Mathématiques pures et Appliquées, 437-447.

[20] Hasegawa, H., & Holm, L. (2009). *Advances and pitfalls of protein structural alignment.* Current Opinion in Structural Biology, 19(3), 341-348.
[21] Hasimoto, H. (1972). *A soliton on a vortex filament.* Journal of Fluid Mechanics, 51(3), 477-485.
[22] Hausrath, A. C., & Goriely, A. (2006). *Repeat protein architectures predicted by a continuum representation of fold space.* Protein Science, 15(4), 753-760.
[23] Heath, A. P., Kavraki, L. E., & Clementi, C. (2007). *From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes.* Proteins: Structure, Function, and Bioinformatics, 68(3), 646-661.
[24] Hoffmann, T. (2000). *On the equivalence of the discrete nonlinear Schrödinger equation and the discrete isotropic Heisenberg magnet.* Physics Letters A, 265(1), 62-67.
[25] Holm, D. D., & Stechmann, S. N. (2004). *Hasimoto transformation and vortex soliton motion driven by fluid helicity.* arXiv preprint nlin/0409040.
[26] Izergin, A. G., & Korepin, V. E. (2009). *A lattice model related to the nonlinear Schrödinger equation.* `arXiv:0910.0295.`
[27] Kats, Y., Kessler, D. A., & Rabin, Y. (2002). *Frenet algorithm for simulations of fluctuating continuous elastic filaments.* Physical Review E, 65(2), 020801.
[28] Kratky, O., & Porod, G. (1949). *Röntgenuntersuchung aufgelöster Fadenmoleküle.* Recueil, 68, 1106.
[29] Manton, N., & Sutcliffe, P. M. (2004). *Topological solitons.* Cambridge University Press.
[30] Moffatt, H. K., & Ricca, R. L. (1991). *Interpretation of invariants of the Betchov-Da Rios equations and of the Euler equations.* In The Global Geometry of Turbulence (pp. 257-264). Springer US.
[31] Molkenthin, N., Hu, S., & Niemi, A. J. (2011). *Discrete Nonlinear Schrödinger Equation and Polygonal Solitons with Applications to Collapsed Proteins.* Physical Review Letters, 106(7), 078102.
[32] Morris, A. L., MacArthur, M. W., Hutchinson, E. G., & Thornton, J. M. (1992). *Stereochemical quality of protein structure coordinates.* Proteins: Structure, Function, and Bioinformatics, 12(4), 345-364.
[33] Myers, J. K., & Oas, T. G. (2001). *Preorganized secondary structure as an important determinant of fast protein folding.* Nature Structural & Molecular Biology, 8(6), 552-558.
[34] Lau, N., Oxley, A., & Nayan, M. Y. (2012, September). *Protein folding and loop closure: Some bioinformatics challenges.* In AIP Conference Proceedings (Vol. 1482, p. 701).
[35] Nakayama, K. (2007). *Elementary vortex filament model of the discrete nonlinear Schrödinger equation.* Journal of the Physical Society of Japan, 76(7).
[36] Niemi, A. J. (2003). *Phases of bosonic strings and two dimensional gauge theories.* Physical Review D, 67(10), 106004.
[37] Pace, C. N. (1990). *Measuring and increasing protein stability.* Trends in Biotechnology, 8, 93-98.
[38] Petsko, G. A., & Ringe, D. (2004). *Protein structure and function.* Sinauer Associates Inc.
[39] Ptitsyn, O. B. (1995). *How the molten globule became.* Trends in biochemical

sciences, 20(9), 376.
[40] Rackovsky, S., & Scheraga, H. A. (1978). *Differential Geometry and Polymer Conformation. 1. Comparison of Protein Conformations.* Macromolecules, 11(6), 1168-1174.
[41] Rackovsky, S., & Scheraga, H. A. (1980). *Differential geometry and polymer conformation. 2. Development of a conformational distance function.* Macromolecules, 13(6), 1440-1453.
[42] Rackovsky, S., & Scheraga, H. A. (1981). *Differential geometry and polymer conformation. 3. Single-site and nearest-neighbor distribution and nucleation of protein folding.* Macromolecules, 14(5), 1259-1269.
[43] Scheraga, H. A., Khalili, M., & Liwo, A. (2007). *Protein-folding dynamics: overview of molecular simulation techniques.* Annu. Rev. Phys. Chem., 58, 57-83.
[44] Da Rios, Da Rios, L. S. (1906). *On the motion of an unbounded fluid with a vortex filament of any shape.* Rend. Circ. Mat. Palermo, 22, 117-135.
[45] Rudin, W. (1987). *Real and Complex Analysis.* 1987. Cited on, 156.
[46] Shashikanth, B. N., & Marsden, J. E. (2003). *Leapfrogging vortex rings: Hamiltonian structure, geometric phases and discrete reduction.* Fluid Dynamics Research, 33(4), 333-356.
[47] Tobin, R. (1996). *Molecular collapse: the rate-limiting step in two-state cytochrome c folding.* Proteins: Structure, Function, and Genetics, 24, 413-426.
[48] Spivak, M. (1970). *A Comprehensive Introduction to Differential Geometry II.* Publish or Perish, Inc., Boston, 1970.
[49] Suzuki, M. (1976). *Generalized Trotter's formula and systematic approximants of exponential operators and inner derivations with applications to many-body problems.* Communications in Mathematical Physics, 51(2), 183-190.
[50] Tozzini, V. (2005). *Coarse-grained models for proteins.* Current Opinion in Structural Biology, 15(2), 144-150.
[51] Weissmann, S., & Pinkall, U. (2010). *Filament-based smoke with vortex shedding and variational reconnection.* ACM Transactions on Graphics (TOG), 29(4), 115.
[52] Volovik, G. E., & Volovik, G. E. (2009). *The universe in a helium droplet* (Vol. 117). Oxford University Press.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1054

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.

Distribution: publications.uu.se
urn:nbn:se:uu:diva-199987

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2013