UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 921

Bending, Twisting and Turning

Protein Modeling and Visualization from a Gauge-Invariance Viewpoint

MARTIN LUNDGREN

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2012

ISSN 1651-6214
ISBN 978-91-554-8338-8
urn:nbn:se:uu:diva-172358

Dissertation presented at Uppsala University to be publicly examined in Å80101, Ångströmlaboratoriet, Lägerhyddsvägen 1, Uppsala, Friday, May 25, 2012 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

Lundgren, M. 2012. Bending, Twisting and Turning: Protein Modeling and Visualization from a Gauge-Invariance Viewpoint. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 921. 68 pp. Uppsala. ISBN 978-91-554-8338-8.

Proteins in nature fold to one dominant native structure. Despite being a heavily studied field, predicting the native structure from the amino acid sequence and modeling the folding process can still be considered unsolved problems. In this thesis I present a new approach to this problem with methods borrowed from theoretical physics. In the first part I show how it is possible to use a discrete Frenet frame to define the discrete curvature and torsion of the main chain of the protein. This method is then extended to the side chains as well. In particular I show how to use the discrete Frenet frame to produce a statistical distribution of angles that works in similar fashion as the commonly used Ramachandran plot and side chain rotamers. The discrete Frenet frame displays a gauge symmetry, in the choice of basis vectors on the normal plane, that is reminiscent of features of Abelian-Higgs theory. In the second part of the thesis I show how this similarity with Abelian-Higgs theory can be translated into an effective energy for a protein. The loops of the proteins are shown to correspond to solitons so that the whole protein can be constructed by gluing together any number of solitons. I present results of simulating proteins by minimizing the energy, starting from a real line or straight helix, where the correct native fold is attained. Finally the model is shown to display the same phase structure as real proteins.

Keywords: protein folding, discrete frenet frame, solitons, protein visualization

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

ISSN 1651-6214
ISBN 978-91-554-8338-8
urn:nbn:se:uu:diva-172358 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-172358)

List of papers

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

I U. H. Danielsson, M. Lundgren and A. J. Niemi, Gauge field theory of chirally folded homopolymers with applications to folded proteins, *Phys. Rev. E* **82**, 021910 (2010)

II M. N. 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)

III S. Hu, M. Lundgren and A. J. Niemi, Discrete Frenet frame, inflection point solitons, and curve visualization with applications to folded proteins, *Phys. Rev. E* **83** 061908 (2011)

IV M. Lundgren, A. J. Niemi and F. Sha, Protein loops, solitons and side-chain visualization with applications to the left-handed helix region, `arXiv:1104.2261 [q-bio.BM]`. Submitted to Phys. Rev. E.

V M. Lundgren and A. J. Niemi, Backbone covalent bond dynamical symmetry breaking and side-chain geometry of folded proteins, `arXiv:1109.0423 [q-bio.BM]`. Submitted to Phys. Rev. E.

Reprints were made with permission from the publishers.

Contents

Part I: Protein Geometry and Visualization

Part II: Protein Modeling

1. Introduction and background

I take a look out my window. There I see the trucks coming and going, delivering goods from halfway around the world. Everything we need to keep our lives going. I see a helicopter passing high above. Transporting people, desperately in need of aid, to the hospital. In the far distance I can see a construction crane adding concrete blocks to a rising building. All over the city the same thing. We truly live in the age of the machines. The epitome of which is the machine I am now turning to, to write this text.

Yet, despite our knowledge of how to build machines for all possible intents and purposes, when I look into the mirror I see something that we cannot fully explain. Hidden in our bodies are machines, so powerful yet still so simple, performing all the tasks necessary for life. True to our nature as curious beings we cannot help but try to explore and explain one of the few frontiers of scientific knowledge remaining.

1.1 A quick guide to protein folding

The first real piece in this puzzle was uncovered by Watson and Crick in 1953 [59]. They found the blueprints; the structure of DNA. However, DNA is just a carrier of information, the real workers are the proteins. Over the following two decades the full process of how the DNA code was transcribed to RNA and then into an amino acid sequence was discovered. Now, the major question remaining was how the amino acid sequence folded into its functional form, the protein. Unfortunately we still do not know the full answer to that question 40 years later.

1.1.1 The protein folding problems

One remarkable property of proteins was first discovered in an experiment by Christian Anfinsen in 1957 [53, 2]. What he found was that a denatured protein outside the cell would, when the environment allowed, fold back into its native functional state. This is a really remarkable property. Proteins are able to self-assemble without help from any exterior mechanisms.

To be fair, Anfinsen's experiment was made outside the cell, in vitro. For folding in the cell, in vivo, there are helper proteins, chaperones. However, it seems their main purpose is not to guide the folding but to act to isolate the protein from the crowded environment of the cell, to prevent errors in the folding process [19].

Anfinsen's discovery is the central part of protein folding. It tells that the initial configuration is not important to the end result. The protein should be able to fold; regardless of which random unfolded state it starts from and the only guide is the internal free energy. It also tells some important things to expect from the energy of the protein.

First of all, since the protein always folds by itself into one unique shape the natural explanation is that the native state is the state with the lowest free energy. Furthermore, there has to be a sizable energy difference between the native state and any other state. If this energy gap was too small then there would be oscillations between different states because small random thermal fluctuations from the environment would be enough to push it up to a slightly higher energy state. This is not observed. Finally the native state cannot be too topologically complicated. For the protein to fold there has to be an accessible pathway from the unfolded state to the native state. It is not obvious that a randomly composed protein would have all of these properties naturally. Most likely the process of evolution has, during billions of years, selectively chosen proteins that fold easily and have stable native states.

What is the protein folding problem then? The fact is that proteins do fold into one unique native structure. Still, three questions needs to be answered before we can say that we really understand the protein folding process fully. The first question is why does it fold? The second question is what is the target? Finally, the third question is how does it get there? Neither of these questions has been fully answered. A more detailed description of the questions and of our current level of knowledge requires more information of what proteins really look like. That I will give in the next section, but first a few words on another question that hides behind the others. Why? Why do we need this knowledge, beside the obvious reason of scientific curiosity?

Proteins are everywhere in our bodies. They are involved in everything that happens. It is then, of course, of terrible importance that they work properly. I have said that the proteins always fold into their functional native state. That is not entirely true. Sometimes something goes wrong somewhere in the process. Proteins that have misfolded can no longer perform their proper function. On top of that they might aggregate with other misfolded proteins to form an insoluble plaque. Normally the body has cleaning mechanisms to remove these things before it gets dangerous, but sometimes these mechanisms are not working properly.

Alzheimer's disease, although the exact cause is unknown, is associated with a build-up of insoluble plaques of misfolded proteins in the brain. Prions became known with the outbreak of mad cow disease and its human form Creutzfeldt-Jakob disease. They are essentially proteins that have folded into the wrong configuration. What is unique about prions is that they are able to, unlike viruses and bacteria, transmit a disease without any genetic code for self-replication. They do it by influencing the folding of other proteins of the same kind, making them fold into a similar shape as the prion. Other diseases are also related to the folding of proteins, like Parkinson's disease, cystic fibrosis and even normal allergies. If we had a better knowledge of the protein folding process, then we would gain a better knowledge of these diseases and it could help the process of finding cures.

Another field where a detailed description of the folding process would be vital is protein design and redesign. In this case the goal is to create a new, or modify an existing, protein to perform a particular task. The reason for designing a protein could be designing a new drug or creating a catalyst for some particular reaction. The number of possibilities are almost limitless.

1.1.2 The structure of proteins

The protein consists of multiple levels of structure, all shown in Fig. 1.1. Deep down there is the primary structure, the building blocks. In essence, a protein is a long chain where each link is an amino acid, commonly called residues. There are only 20 different types of amino acids used in proteins. All amino acids share a common part, which makes up the main chain of the protein. The side chain is where the different types of amino acids differ.

The side chain has very different properties for different amino acids. For example arginine has a long, positively charged side chain, aspartic acid has a short negatively charged side chain and cysteine has a side chain with a sulphur atom able to form disulfide bonds with other cysteines. This difference in the side chains of the amino acids is what makes each type of protein unique. Hence, the final shape of the protein and how to get there must somehow be encoded in the amino acid sequence. Cracking that code has, however, proved hard.

The next level of structure in the proteins, the secondary structure, is local structure defined by the pattern of hydrogen bonds between the amino acids. The two most common types of secondary structure are the α-helix and the β-sheet. The α-helix is a helical pattern where the carbonyl oxygen of the ith residue forms a hydrogen bond with the amine hydrogen of the i+4th residue while β-sheets are formed by two or more straight and flat conformations, called β-strands, interacting and forming hydrogen bonds between them in a regular pattern.

The hydrogen bonds, that are the basis of the secondary structure, are formed between atoms of the main chain. The difference between the different residues has no effect on the hydrogen-bonding pattern. Still, some residues prefer to be in one or the other secondary structure. For example proline, with its ring shape, cannot exist inside a helix without distorting it. Valine does not like helices either, but the reason here is steric constraints because of the shape of the side chain. Similarly the other amino acids have different propensities for

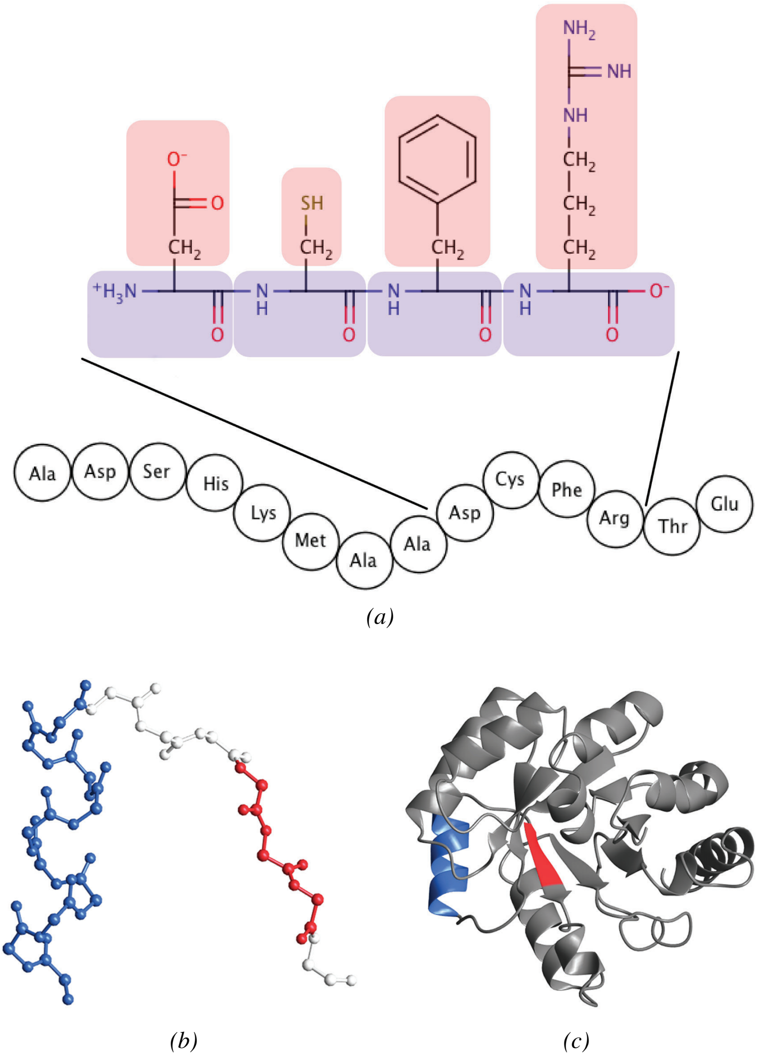


Figure 1.1: The different structure levels of the protein. (a) shows how the primary structure is constructed from a string of amino acids, each consisting of a main chain part (blue) and a side chain part (red). (b) shows the secondary structure, displaying an alpha helix (blue) and a beta strand (red) connected by a short loop. The side chains are not shown here. (c) shows the tertiary structure where the helix and strand from (b) are included.

different kinds of secondary structure. However, this is not enough to predict the secondary structure of a protein based on the primary structure to more than about 80% accuracy. The secondary structure is not totally dependent on the primary structure either. Some sequences that form an alpha-helix in one protein forms a beta-strand in another because of the local environment.

The next structure level is the tertiary structure. This is the full three dimensional structure of the proteins, describing the coordinates for all the atoms. The tertiary structure defines the function of the protein, since the shape is tailored for it to perform one specific task. The tailoring could be a cavity in the protein structure formed in a way to bind a specific molecule to catalyze a specific reaction, or it could be a structure that is formed in such a way as to be able to grab hold to a DNA strand.

One feature of the tertiary structure is that it is very compact. In fact the density is more or less as high as possible. There are two main reasons why the proteins are so dense, the hydrophobic effect and hydrogen bonding. First of all, the hydrophobic effect is caused by hydrophobic side chains being exposed to water. Since hydrophobic molecules cannot form hydrogen bonds with water, it is more favorable for them to cluster together, because of entropy reasons. This will cause an unfolded protein to quickly fold into a compact shape with a hydrophobic core. One of the problems with misfolded proteins is that they might have a lot of hydrophobic residues at the surface. This will cause them to aggregate with other misfolded proteins, forming insoluble plaques. Proteins put in a nonpolar environment will denature very rapidly.

Even though the side chain may be hydrophobic the main chain still contains a carbonyl oxygen and an amide hydrogen. It is energetically unfavorable for those if they are not bonded to anything. Hence, the formation of secondary structure with a regular hydrogen bonding pattern. In the native structure, essentially all atoms that can form hydrogen bonds also form bonds [16].

It has been debated whether the protein folding process is driven by one particular force or should be seen as the sum of many forces. This latter approach was the earlier interpretation, where the folding process was an intricate mix of the hydrophobic effect, hydrogen bond formation, charged interactions, van der Waals interactions, disulfide bond formation etc. The question is still open with some arguing that the hydrophobic effect is the most important part [46, 38], while others argue that hydrogen bonding is the most important part [50, 64]. The order at which the different levels of structure form is also disputed. Is the secondary structure formed before, after, or at the same time as the protein collapses into a compact shape?

1.1.3 Folding pathways

The microscopic forces affecting the protein folding process are known. The question why proteins fold is, to a large extent, answered. It then seems that the other two questions, how they fold and what the target is, would be easy to answer as well. Simply construct an energy function containing all the information about the different types of interactions. A simple simulation would then show the folding pathways and the final native state. Unfortunately it is not that simple.

In 1968 Cyrus Levinthal realized something that would later be named Levinthal's paradox. The number of possible misfolded configurations of the average protein is astronomically large. In this immense jungle of faulty configurations, how can the folding protein find the one native state? The typical protein folds in a matter of microseconds to milliseconds so there has to be some way for it to search the huge configuration space in a selective manner. There are, however, some indications of how this mystery is solved.

Folding time can be measured experimentally. Results have shown that the folding time is correlated with the amount of local connections in the native state [23]. That is, if there are a lot of interactions between residues that are close on the chain then the folding will be fast. This will be the case for structures rich in α-helices. On the other hand, structures with a lot of long-range interactions, like β-sheet rich structures, will take longer time to fold. The conclusion from this is that the protein starts off the folding by searching for local conformations.

There are three main theories of what the folding process looks like. First there is the framework model, which says that the secondary structure is the first to form [47]. When that has formed, interactions between the hydrophobic residues will cause the protein to collapse into a compact shape, followed by a slow search for the right tertiary structure. The second model, the hydrophobic collapse model, is the opposite [38]. The folding starts with the hydrophobic collapse. While in the framework model the tertiary structure depended on the secondary structure, in the hydrophobic collapse model it is the other way around. Finally the third option is the nucleation condensation model [3]. This model says that secondary and tertiary structure are formed concurrently. The folding starts in a nucleus, with the correct local structure formed, and spreads from there to the unfolded parts. The true description of the folding might very well be some combination of these, where different proteins using different pathways.

1.1.4 Simulating folding

The first question anyone who tries to simulate protein folding should ask himself or herself is what level of detail is necessary? The most detailed approach is to do the full quantum-mechanical calculations. While it has the benefits of giving the best results and not requiring any parameter fitting, it is totally unfeasible to use it on the scale of a full protein. The level of detail is simply too high, and even the best computers available today would not be able to produce a result in a reasonable timeframe. Despite this, quantum mechanical approaches has some uses in simulating smaller molecules and for simulating a protein binding to a ligand, by letting the active sites be described in full detail and use a more coarse-grained approach for the rest of the protein.

The most common approach is to use all-atom methods. Here, the role of electrons is simplified. Instead the covalent bonds are explicitly defined beforehand and effective energy terms are introduced for the long-range interactions. These methods require a large set of parameters to specify things like the optimal bond angle and van der Waals forces between all different combinations of atoms. A small simplification of the all-atom method is the united-atom method, where the methyl hydrogen atoms are also treated implicitly.

An issue is the treatment of the surrounding solvent. We saw the hydrophobic effect had a large effect on the folding process. Simply running the simulation in a vacuum would not work very well. That would leave a lot of polar residues on the surface of the protein in an energetically unfavorable unbonded state, while in reality they would form hydrogen bonds with the water molecules. There are two ways around this. Either introduce the water molecules explicitly in the simulation. This would increase the number of atoms needed to be simulated a lot. The other way is to introduce water implicitly as a mean-field effect, where the average effect of the water per unit area and the area of the protein exposed to the water is taken into account.

An all-atom approach is regularly combined with a physical force-field, where the force field is the description of the potential energy of the system, i.e. an energy function and the set of parameters used. Typically the parameters are constructed from either experimental results or quantum chemical calculations.

There are two main methods of simulation, molecular dynamics and the Monte Carlo method. The latter is a stochastic method where random changes in the protein are evaluated based on their effect on the energy of the system. Monte Carlo methods are generally a good way of finding the minima of a potential. However, it is not obvious that the way it gets there is in any way physical. For example, there is no inherent time-scale. For detailed simulations, where the pathway to the target state is sought after, molecular dynamics is commonly used instead.

In molecular dynamics simulations, contrary to Monte Carlo methods, the folding is deterministic. Knowing the potential energy of the system, the force and acceleration on each atom can be calculated through

$$m\mathbf{a} = \mathbf{F} = -\nabla U \tag{1.1}$$

where m is the atom mass, **a** the acceleration, **F** the force and U the potential energy of the atom. Knowing the acceleration of all atoms in the system is enough to describe their movement completely. However, it is not possible to solve this multi-body system in any analytical way so the main approach is to run computer simulations with a discretized version of this equation with a time-step of the order of 10^{15} seconds. The computer power of today then

allows for simulations of a small folding protein over timescales up to the order of a few milliseconds [40]. This correlates with the real folding time for these smaller, often alpha-helix rich, structures.
All-atom molecular dynamics simulations have given some good results. In [40] 12 different small proteins were folded to an RMSD (root-mean-squared deviation) between 1 and 4 Å from their native state. However, longer proteins still poses a major challenge. Not only would they need more computer-time because of the increased number of atoms. The folding time of the larger proteins is larger, even on the order of seconds for some proteins, so the simulation would have to cover a longer time-interval as well. Longer proteins would also lead to more problems with Levinthal's paradox, because the number of possible conformations grows immensely.

There are ways to increase the speed of molecular dynamics simulations by introducing things like cut-offs for long range interactions, but to really be able to simulate larger systems a more coarse-grained approach is necessary. There are many ways to do this, depending on the level of detail needed [56]. The simplest version is the one-bead model, which means that the entire amino acids is represented by one site, normally the central carbon. The problem with one-bead models is how to encode all the properties of the amino acid, like size and geometry in just one point. One way to introduce the direction of the side chain into the system is to change into a two-bead model. In this case there is a second site at the center of the side chain for each residue, a variant of which is the United-Residue method [49]. By adding the main chain atoms, in a four-bead model, the hydrogen-bonding can be described explicitly and would improve the ability to study the formation of the secondary structure. In the other direction even more coarse-grained models exists, where the protein is treated as composed of larger rigid blocks that are interacting with each other.

Low bead models are often combined with Gō-like models, which means they have a bias to fold to the known native state. The reason for this is that the potential of the most coarse-grained models is not good enough to find the correct native state without a nudge in the right direction. Gō-like models were introduced to see if it was possible to find a pathway from an unfolded state to the native state if the native state was known. They only take into account the interactions between sites in contact in the native structure. All other interactions can be lumped together as friction [9]. Modern approaches have improved the Gō-like models and used them for different problems, but they are still limited to cases where the native interactions are dominating [56]. In the case of misfolding, where non-native interactions play an important role, the friction approximation is not valid.

1.2 Outline

The protein folding problem still poses a major challenge for scientists. The detailed all-atom simulations are still far away from being able to simulate anything but the smallest proteins. In this thesis I will present a novel approach to the protein folding problem, using methods learned from theoretical physics.

The thesis will be divided into two parts. In the first part I will show how to use a discrete version of the Frenet frame to create a geometrical representation of the proteins. Chapter 2 will be about the geometry of the main chain and will largely be based on Paper III. Chapter 3 will extend this to the side chains and is based on Paper IV and V.

The second part is about defining an energy for the proteins and running simulations. In Chapter 4 I will talk about solitons and how to use solitons to model proteins. Although I was part of the creation of the model with Paper I, I was not directly involved in introducing solitons and developing the techniques to use them, which was mainly done in [10] and [42]. However, solitons are an important part of the model and take part in Papers IV-V as well as in more recent work that are still to be published. Chapter 5 is based on Paper II and describes a detailed investigation of the phase structure of the model.

Part I:
Protein Geometry and Visualization

2. Main chain geometry

The starting point of most protein models is the geometrical representation and the common approach is to simply use the coordinates of the atoms. A step in the simulation of a protein would then involve shifting the coordinates of one or more of the atoms slightly. This will change the distance between atoms and thus change the energy of the system. However, in most of the usual energy functions the energy does not just depend on the positions directly. Many of the potential terms deal with the angles of the bonds. Fortunately these angles can be calculated if you know the surrounding atom positions.

Another way of doing this is to start from the angles instead. One step of the simulation would then involve a bend or a twist at one or more points. If all the angles and bond lengths are known the coordinates can be calculated uniquely. This is a common approach in systems where local coordinates are easier to use than global coordinates, i.e. where it is easier to relate one point in the system to its neighbors than to an outside observer.

In our model we use a coarse-grained description of the protein with the position of the amino acid given only by the central carbon, the C_α. This means the protein can be seen as a discrete, piecewise linear, one-dimensional curve, where the vertices correspond to the central carbons. To describe this we use the curvature and torsion of the curve to create a discrete Frenet frame at each site.

In this chapter I will describe the local coordinates we use, starting from the continuous Frenet frame, and moving on to a discrete frame. This chapter will, in many ways, follow the procedure in Paper III.

2.1 The Frenet frame

The most common way of describing a continuous and differentiable curve, $\mathbf{c}(s)$, in $\mathbb{R}^3$ is by using the Frenet equation [34]:

$$\frac{d}{ds}\begin{pmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{pmatrix} = \begin{pmatrix}0 & \kappa & 0\\ -\kappa & 0 & \tau\\ 0 & -\tau & 0\end{pmatrix}\begin{pmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{pmatrix} \tag{2.1}$$

where

$$\mathbf{t} = \frac{\dot{\mathbf{c}}}{||\dot{\mathbf{c}}||} \equiv \frac{1}{||\dot{\mathbf{c}}||}\frac{d\mathbf{c}(s)}{ds}$$

is the unit tangent vector, pointing in the direction of the curve,

$$\mathbf{b} = \frac{\dot{\mathbf{c}} \times \ddot{\mathbf{c}}}{||\dot{\mathbf{c}} \times \ddot{\mathbf{c}}||}$$

is the unit binormal vector,

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

is the unit normal vector,

$$\kappa = \frac{||\dot{\mathbf{c}} \times \ddot{\mathbf{c}}||}{||\dot{\mathbf{c}}||^3}$$

is the curvature and finally

$$\tau = \frac{(\dot{\mathbf{c}} \times \ddot{\mathbf{c}}) \cdot \dddot{\mathbf{c}}}{||\dot{\mathbf{c}} \times \ddot{\mathbf{c}}||^2}$$

is the torsion. A right-handed Frenet frame is now formed by the three vectors ($\mathbf{t}$,$\mathbf{n}$,$\mathbf{b}$) at each point of the curve. With no loss of generality the curve can be parametrized by arc-length so that $||\dot{\mathbf{c}}|| = 1$ which simplifies the equations [34].
From (2.1) and the following equations it can be seen that if $\mathbf{c}(s)$ is known then the Frenet frame can be calculated throughout the curve. However, it is also easy to see that if the curvature and torsion for the curve is known, then the Frenet frames can also be calculated, but only up to a global rotation and translation.

2.1.1 The discretized Frenet equations

The problem with the Frenet frame is that it is not well defined for a straight curve, or for a kink. In a straight segment $\kappa = 0$, and $\mathbf{n}$ and $\mathbf{b}$ have no well defined direction. The conventional approach is to only define the Frenet equations where the curve behaves in a good way. In our case this is not possible since we are working with a discrete curve. In Paper I and II we solved this by working with a simple discretized version of the Frenet equations

$$\begin{aligned}\mathbf{t}_{i+1} &= \frac{\mathbf{t}_i + \delta \kappa_i \mathbf{n}_i}{\sqrt{1+\delta^2\kappa_i^2}} \\ \mathbf{n}_{i+1} &= \frac{\mathbf{n}_i + (-\kappa_i \mathbf{t}_i + \tau_i \mathbf{b}_i)}{\sqrt{1+\delta^2\left(\kappa_i^2+\tau_i^2\right)}} \\ \mathbf{b}_{i+1} &= \mathbf{t}_{i+1} \times \mathbf{n}_{i+1}\end{aligned} \tag{2.2}$$

where δ is the distance between two vertices on the curve, i.e. the distance between two consecutive C_α, which can, to a good approximation, be taken as the constant value 3.8 Å. If $\mathbf{r}_i$ is the position of the C_α of the ith amino acid

then ($\mathbf{t}_i$,$\mathbf{n}_i$,$\mathbf{b}_i$) is the local Frenet frame at $\mathbf{r}_i$, and $\mathbf{t}_i$ points in the direction of $\mathbf{r}_{i+1}$ so that $\mathbf{r}_{i+1} = \mathbf{r}_i + \delta \mathbf{t}_i$. Letting $\delta \to 0$ gives (2.1) back. As we discovered there are several drawbacks with using this method. The most obvious is that a curve is not allowed to turn more than 90° at each site. This can be seen from (2.2) as it is impossible to find a value for κ_i such that $\mathbf{t}_{i+1}$ has a component in the direction of $-\mathbf{t}_i$. Furthermore a 90° turn would require $\kappa \to \infty$ which is impractical since turns in proteins are often close to this value.
In the standard Frenet equations in 3-dimensions the curvature is always taken to be positive. However, in (2.2) the curvature needs to be defined on $\mathbb{R}$ to cover all (forward-facing) directions. The idea to define the curvature for negative values as well proved useful even when not strictly necessary for geometric reasons, as we will see in the next section.

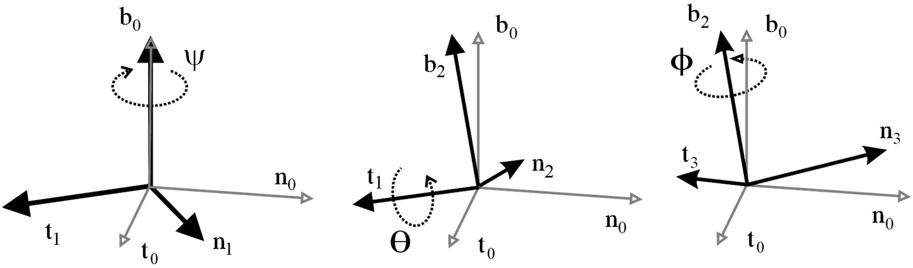


Figure 2.1: Rotation of DF-frame by use of Euler angles.

2.1.2 The discrete Frenet equations

The discretized Frenet equations proved too problematic to work with so in Paper III we introduced a fully discrete version of the Frenet equations, based on rotations with Euler angles. Similar methods are common in other areas, such as robotics [54], aeronautics [32] and in virtual reality [58].

If the positions of the vertices, $(\mathbf{r}_0 \dots \mathbf{r}_n)$, are known, where n is the number of vertices, then the unit tangent vector can be introduced as

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

If the vertices $\mathbf{r}_{i+1}$, $\mathbf{r}_i$ and $\mathbf{r}_{i-1}$ are not located on a straight line then the unit binormal vector can be defined as

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

and the unit normal vector as

$$\mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i.$$

In this way a Discrete Frenet frame (DF-frame) can be defined at every site except at $\mathbf{r}_0$ and $\mathbf{r}_n$. To get the discrete version of (2.1) (DF-equation) we need a transfer matrix that maps the frame at vertex i to the frame at vertex i+1,

$$\begin{pmatrix} \mathbf{t}_{i+1} \\ \mathbf{n}_{i+1} \\ \mathbf{b}_{i+1} \end{pmatrix} = T_{i+1,i} \begin{pmatrix} \mathbf{t}_i \\ \mathbf{n}_i \\ \mathbf{b}_i \end{pmatrix}. \tag{2.3}$$

The DF-frame at $i+1$ is related to the DF-frame at i by an SO(3) rotation, that is most easily parametrized in the terms of Euler angles. With the angles ϕ, θ and ψ as in Fig. 2.1 the transfer matrix, $T_{i+1,i}$ looks like

$$T_{i+1,i} = \begin{pmatrix} c_\psi c_\phi - c_\theta s_\psi s_\phi & c_\psi s_\phi + c_\theta s_\psi c_\phi & s_\theta s_\psi \\ -s_\psi c_\phi - c_\theta c_\psi s_\phi & -s_\psi s_\phi + c_\theta c_\psi c_\phi & s_\theta c_\psi \\ s_\theta s_\phi & -s_\theta c_\phi & c_\theta \end{pmatrix}_{i+1,i} \tag{2.4}$$

where s_ϕ is short for $\sin\phi$ and c_ϕ is short for $\cos\phi$ and similarly for the other angles. The angles should here be treated as bond variables, defined between the vertices.

This expression can be further simplified by realizing that the definition of $\mathbf{b}_i$ requires that $\mathbf{b}_{i+1} \cdot \mathbf{t}_i = 0$ for all i, which gives the condition that

$$\sin\theta \sin\phi = 0$$

for all bonds. Setting $\theta = 0$ would make $\mathbf{b}_i$ constant and essentially make the curve planar. This is not desirable, so based on this we conclude that the correct interpretation is that $\phi = 0$ everywhere. With that condition the DF-equation simplifies to

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

From the DF-equation the procedure for calculating the two remaining angles is straightforward. The discrete bond angle, ψ, is given by

$$\psi_{i+1,i} = -\mathrm{sign}\left(\mathbf{n}_{i+1} \cdot \mathbf{t}_i\right) \arccos\left(\mathbf{t}_{i+1} \cdot \mathbf{t}_i\right) \tag{2.6}$$

where $\mathrm{sign}\left(\mathbf{n}_{i+1} \cdot \mathbf{t}_i\right) \leq 0$ for all i, and the discrete torsion angle θ is given by

$$\theta_{i+1,i} = -\mathrm{sign}\left(\mathbf{b}_{i+1} \cdot \mathbf{n}_i\right) \arccos\left(\mathbf{b}_{i+1} \cdot \mathbf{b}_i\right) \tag{2.7}$$

From a geometrical viewpoint the bond angle, $\psi_{i+1,i} \in [0,\pi]$, measures the angle between $\mathbf{t}_i$ and $\mathbf{t}_{i+1}$ on the plane determined by the three vertices $\mathbf{r}_i$, $\mathbf{r}_{i+1}$ and $\mathbf{r}_{i+2}$. The torsion angle, $\theta_{i+1,i} \in [-\pi,\pi]$, measures the angle between the two planes defined by the vertices $(\mathbf{r}_{i-1}, \mathbf{r}_i, \mathbf{r}_{i+1})$ and $(\mathbf{r}_i, \mathbf{r}_{i+1}, \mathbf{r}_{i+2})$, as seen in Fig. 2.2.

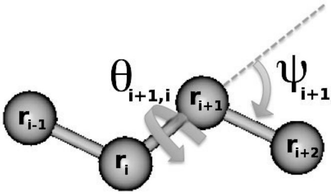


Figure 2.2: The bond angle $\psi_{i+1,i}$ and torsion angle $\theta_{i+1,i}$

2.1.3 Curve reconstruction

A common situation is that the bond and torsion angles are known and the coordinates are unknown. This situation can arise when, like in our model, the energy of the system depends on the angles and energy minimization procedures are used to find the optimal angles. Then the coordinates of the curve can be reconstructed in the following way. First, the angles are not sufficient by themselves. The DF-frame has to be known at one point to be able to use equation (2.5) and the location of this point has to be known as well. Fortunately, the choice of both frame and location is rather arbitrary. Different choices for these only introduce a global translation and rotation to the protein.

In our simulations the following has proved to be good choices for the initial frame

$$\mathbf{r}_0 = \delta \begin{pmatrix} -\cos\psi_{1,0} \\ \sin\psi_{1,0} \\ 0 \end{pmatrix}$$

$$\mathbf{r}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\mathbf{t}_0 = \begin{pmatrix} \cos\psi_{1,0} \\ -\sin\psi_{1,0} \\ 0 \end{pmatrix}$$

$$\mathbf{t}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \mathbf{n}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \mathbf{b}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{2.8}$$

where δ is the bond-length assumed to be constant $= 3.8$ Å. Repeated use of equation (2.5) together with the fact that $\mathbf{r}_{i+1} = \mathbf{r}_i + \delta\mathbf{t}_i$ will give the coordinates for the whole curve.

2.2 Gauge properties

The choice of using the DF-frame to describe the curve has so far been arbitrary. In this section I will try to show that there are deeper reasons to select that particular framing. In physical systems it is always important to recognize symmetries and conserved properties. Since we know that we can describe the protein using only the backbone angles θ and ψ (in our coarse-grained model) we should also be able to write an energy for the system in terms of those two angles.

2.2.1 The continuous case

Suppose that we made a rotation in the normal plane of the DF-frame so that

$$\begin{pmatrix} \mathbf{n} \\ \mathbf{b} \end{pmatrix} \rightarrow \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix} = \begin{pmatrix} \cos\eta & -\sin\eta \\ \sin\eta & \cos\eta \end{pmatrix} \begin{pmatrix} \mathbf{n} \\ \mathbf{b} \end{pmatrix} \tag{2.9}$$

where η is some arbitrary angle and $\eta = 0$ would give back the standard Frenet frame, then this would transform the bond and torsion angles as well. However, the energy of the system should not depend on our choice of basis axis on the normal plane. The energy has to be invariant under this transformation. Equation (2.9) acts on the curvature and torsion as following

$$\tau \rightarrow \tau - \dot{\eta} \tag{2.10}$$

$$\kappa T^3 \rightarrow \kappa \left(T^3 \cos\eta - T^2 \sin\eta \right) \tag{2.11}$$

where T^2 and T^3 belongs to the adjoint basis of SO(3) Lie algebra

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

The transformations (2.10) and (2.11) can be recognized from the Abelian Higgs model, where the rotation corresponds to a U(1) gauge transformation, κ corresponds to the complex scalar field ϕ and τ corresponds to the spatial components of the U(1) gauge field A_1 [44]. Since we require that the energy of the system being invariant of this gauge transformation a natural starting point is to use the following gauge invariant free energy from Abelian Higgs theory

$$F = \int_0^L ds \left\{ |(\partial_s + iA_1)\,\phi|^2 + \lambda \left(|\phi|^2 - \mu^2 \right)^2 \right\}. \tag{2.12}$$

This concept will be developed further in chapter 4.

2.2.2 The discrete case

The discrete version of (2.10) and (2.11) can be shown to be

$$\theta_{i+1,i} \to \theta_{i+1,i} + \Delta_i - \Delta_{i+1} \tag{2.13}$$

$$\psi_{i+1,i} T^3 \to \psi_{i+1,i} \left(T^3 \cos\Delta_{i+1} - T^2 \sin\Delta_{i+1} \right) \tag{2.14}$$

where we can see that $\theta_{i+1,i}$ transforms in a similar way as τ so we can denote this the discrete torsion, $\theta_{i+1,i} \equiv \tau_{i+1,i}$. Similarly we can denote $\psi_{i+1,i}$ the discrete curvature $\psi_{i+1,i} \equiv \kappa_{i+1,i}$. This notation will be used in Chapter 4 and 5.

Of particular interest is the transformation that sends

$$\mathbf{b}_i \to -\mathbf{b}_i$$

$$\mathbf{n}_i \to -\mathbf{n}_i$$

i.e. a rotation of the basis axes on the normal plane by 180°. This rotation can be achieved by setting $\Delta_{i+1} = \pi$ and $\Delta_k = 0$ for all $k \neq i+1$. The transformation would act on the angles in the following way:

$$\begin{aligned} \theta_{i+1,i} &\to \theta_{i+1,i} - \pi \\ \theta_{i+2,i+1} &\to \theta_{i+2,i+1} + \pi \\ \psi_{i+1,i} &\to -\psi_{i+1,i}. \end{aligned} \tag{2.15}$$

Note here that the transformation acts on two different θ-values because the transformation (2.13) tells that both $\theta_{i+1,i}$ and $\theta_{i+2,i+1}$ depends on Δ_{i+1}.

Normally the Frenet equation uses only positive curvature. From equation (2.6) it follows that ψ-angles calculated from a known curve will always be positive as well. However, there is nothing in the DF-equations that prevents it from working with negative ψ so we can easily extend the range of ψ to $[-\pi, \pi]$. The extension of the range also means that ψ and θ covers every direction twice, so there should be a $\mathbb{Z}_2$-symmetry between negative and positive ψ values. This is exactly the symmetry we have found in (2.15).

2.3 Main chain statistics

The protein data bank (PDB) [6] is the collection of all known protein structures and at the time of writing it has about $80,000$ entries. While this is a very small subset of all existing proteins in nature, it is still a vast amount of data. A first test of the usability in describing the protein main chain in terms of ψ and θ angles would be to look at their statistical distribution in known proteins. Regular patterns should appear for regular structure, like the different kinds of secondary structure.

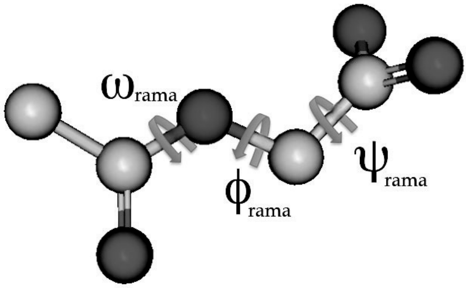


Figure 2.3: Torsion angles of the full protein main chain.

2.3.1 Ramachandran plot

Before going into the results of our method it is good to take a look at the conventional way of looking at statistical angle distributions. The regular way is to look at the torsion angles of the main chain, as they are shown in Fig. 2.3 [29]. The peptide bond angle, ω_{rama} is fixed at $\omega_{rama} = \pi$ except in some rare occasions in prolines where $\omega_{rama} = 0$ [60]. The bond angles can also be considered fixed so the whole main chain can be encoded in the two remaining torsion angles ϕ_{rama} and ψ_{rama}. The plot of ϕ_{rama} and ψ_{rama} angles is called a Ramachandran plot (Fig. 2.4).

Different amino acids can fill different regions in the Ramachandran plot. The two most distinct are proline and glycine. Proline has a cyclic structure, which makes it very rigid and it can only exist in certain distinct regions of the Ramachandran plot. Glycine on the other hand, which is essentially lacking a side chain, is very free and can exist in regions of the Ramachandran forbidden for the other residues [29].

The Ramachandran plot can be used for structure validation, where the ϕ_{rama} and ψ_{rama} values of a new structure can be plotted. A large number of points in regions that are supposed to be forbidden would indicate that the structure is of low quality [28].

2.3.2 DF-plot and stereographic projection

The Ramachandran angles ϕ_{rama} and ψ_{rama} are defined on a torus, which makes them easy to plot on a plane. The angles in the DF-equations, on the other hand, are defined on a sphere (where we have restricted the range of ψ to only positive values to avoid degeneracies in the plot). While it is hard to make a straightforward interpretation of the Ramachandran plot just by looking at it, the plot of the angles from the DF-equations (the DF-plot) has a very clear interpretation. Think of a sphere with its center at the site i, the north pole in the direction of site i+1 and the prime meridian in the direction of the

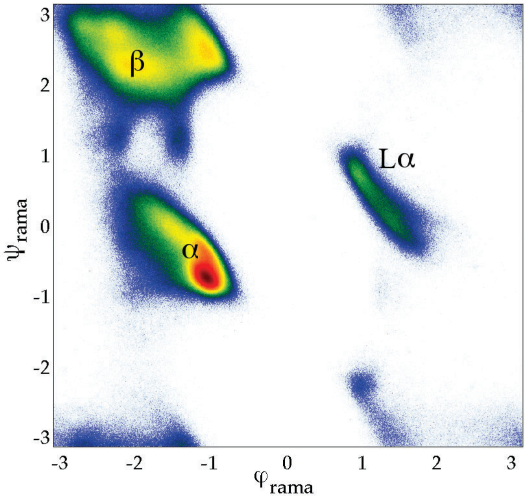


Figure 2.4: The Ramachandran plot for the proteins in the PDB with less than 2 Å resolution. The empty regions are either sterically forbidden or energetically unfavorable. The dense regions are denoted after which kind of structure they represent. α stands for α-helices, β for β-sheets and Lα stands for left-handed α helices. Long left-handed helices are very uncommon but single residues can still bend in a way similar to a left-handed helix.

site i-1. The point on this sphere given by $(\psi_{i+1,i}, \theta_{i+1,i})$ would then give the direction towards site i+2.

Plotting the data from the whole PDB database in the same manner as the Ramachandran plot produces a circular pattern around the north pole of the sphere. For visual purposes it is better to show this in a stereographic projection rather than on a sphere. $\psi_{i+1,i}$ and $\theta_{i+1,i}$ can be transformed to polar stereographic coordinates, r and φ, through

$$r = \frac{1 - \cos\psi}{\sin\psi}$$

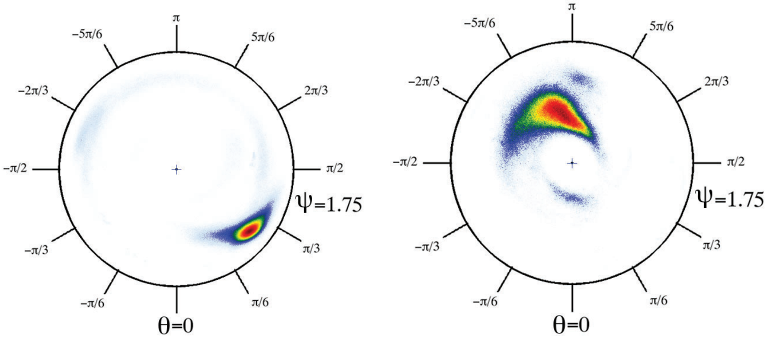
$$\varphi = \theta \tag{2.16}$$

where $r = 0$ would correspond to the north pole and the south pole would be at $r \to \infty$. There are two forbidden regions (best seen in Fig. 2.6) inside and outside the circular pattern. The outside region is forbidden because of steric constraints, while the inner region is perhaps not technically forbidden but energetically very unfavorable.

2.3.3 Secondary structure analysis

A first example of using the DF-plot would be to see how the different secondary structures of the proteins manifest themselves in the plot. The expectation is to find a similar pattern as in the Ramachandran plot with well-defined regions corresponding to different kinds of secondary structure. The difference is that, while the Ramachandran angles are defined on a site, $\psi_{i+1,i}$ and $\theta_{i+1,i}$ are defined between site i and $i+1$.

The figures in this section will all be based on the PDB database, using all items containing proteins with less than 2 Å resolution. The choice of resolution is to get as high quality figures as possible. The result will be the same even with a more refined set of data.

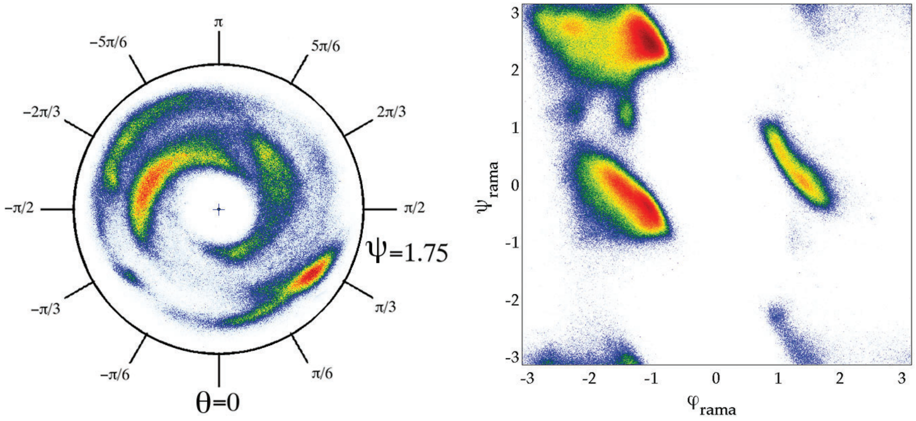


(a) *(b)*

Figure 2.5: Stereographic DF-plot showing the angles between two residues denoted in the PDB database as α-helices (a) and β-sheets (b).

In Fig. 2.5 the DF-plot of α-helices and β-sheets can be seen. A deeper analysis shows that the values for α-helices are centered around $(\psi,\theta) \sim \left(\frac{\pi}{2},1\right)$ and β-sheets are centered around $(\psi,\theta) \sim (1,-\pi)$. The third region in Fig. 2.4, the Lα region, will be treated more carefully in chapter 3.

Fig. 2.6 shows the loop regions of the protein as both a Ramachandran and a DF-plot. In the DF-plot two distinct rings can be seen; one with β-like behavior ($\psi \sim 1$) and one with α-like behavior ($\psi \sim \frac{\pi}{2}$). In the Ramachandran plot the angles at one site is used, but in the DF-plot the angles are defined on the bond between two sites. For regular structure like helices and sheets this would make no big difference. We expect to see roughly the same thing, i.e. small, well defined, densely populated areas. However, in the loop regions there might be a difference in what information is obtainable in the two plots, but the features of the DF-plot have not been studied extensively yet.



(a) *(b)*

Figure 2.6: a) Stereographic DF-plot showing the angles between two residues denoted in the PDB database as having no secondary structure. b) Ramachandran plot of residues with no secondary structure.

3. Side chain geometry

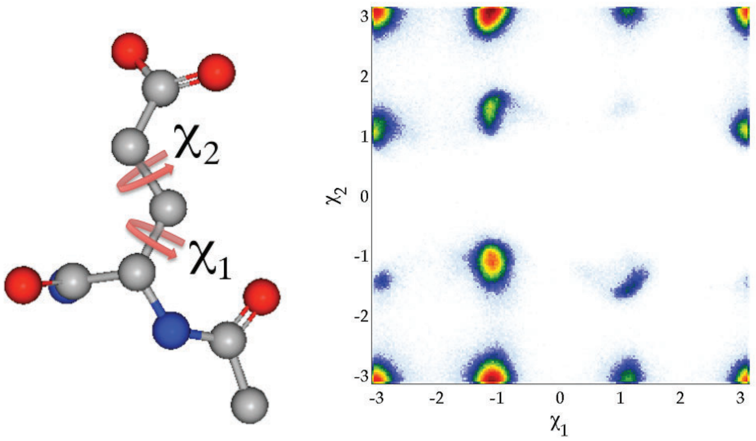
The previous chapter introduced a way to describe the main chain of the protein in terms of the angles ψ and θ. However, to be complete the description also has to include the side chains. In this chapter I will show different methods for describing and visualizing the atoms of the side chains. I will also do a summary of the work in paper IV to analyze the Lα region from Fig. 2.4. Finally I will show an analysis of the Villin headpiece, which is one of the most studied smaller proteins, in terms of main chain and side chain angles.

3.1 Side chain rotamers

A good start to a chapter on side chain geometry is to look at the conventional methods. First of all the position of the C_β, the first atom of the side chains of all amino acids, except for glycine that lacks a side chain. The conventional way of looking at it is that the position of the C_β is fixed if the location of the C, N and C_α of the main chain is known because of the tetrahedrical symmetry of the bonds around the C_α. This is a notion that has been increasingly criticized in later studies [5], and we investigate this symmetry in Paper V and show that, contrary to the older notion, the angles have secondary structure dependence.

Similar to the conventional description of the main chain, the side chain bond angles are assumed to be more or less fixed and the side chain is assumed to be fully described by the torsion angles [18]. The side chains can look very different for different amino acids so the definitions of the torsion angles needed to describe the side chain are different as well. As an example look at Fig. 3.1. In the left part of this figure glutamic acid is shown with the first two torsion angles on the side chain, χ_1 and χ_2. The right part of the figure is called a Janin plot. It shows χ_1 versus χ_2, and the plot really shows the three different low energy conformations, the rotamers, for the side chain torsion angles; *trans*, *gauche+* and *gauche-* with the angles π, $+\frac{\pi}{3}$ and $-\frac{\pi}{3}$ respectively. Certain combinations of these are more common than others. In Fig. 3.1 the combination χ_1 in *trans* and χ_2 in *gauche-* is very common while χ_1 and χ_2 in *gauche+* is more or less forbidden because of steric clashes.

Statistical probabilities for the different rotamers of all amino acids have been collected in rotamer libraries [31]. They can be used for recreating the side chain if the angles are unknown. Different kinds of libraries exists. Some only depend on the type of residue, others depend on the secondary structure and still others depend on the local environment of the residue.



(a) *(b)*

Figure 3.1: a) Glutamic acid with the first two side chain rotamer angles shown. b) Janin plot for glutamic acid.

3.2 DF-frame and side chains

The problem with using the tetrahedral symmetry to define the position of the C_β and the torsion angles to define the rest of the side chain is that both of these are essentially local descriptions. They give very little information about the direction of the side chain in comparison with the main chain and surrounding residues. The Janin plot also has the same problem as the Ramachandran plot in that it is hard to make a straight-forward geometrical interpretation of what it means.

3.2.1 The C_β position

With the DF-frame defined at each residue, the logical way of visualizing the position of the C_β would be to plot it in that frame. Fig. 3.2a shows the distribution of possible directions to the C_β. The center of the sphere is located at the C_α and the radius is the same as the typical length from C_α to C_β. In this plot the location of the C_α of the residues before and after are included as well, with the proper distances. This can give an idea of the direction of the side chain, and possible interactions, with regards to nearby residues.

In Fig. 3.2b the different directions are categorized according to secondary structure. There are two big regions corresponding to the two main secondary structures connected by a loop region in a horse-shoe shaped pattern. In the lower right there is also an isolated island corresponding to the Lα region in the Ramachandran plot, although this plot does not include any glycines because of their lack of a C_β. This island will be discussed further later in this chapter.

From Fig. 3.2b we see that the position of the C_β is clearly dependent on the secondary structure. If you only regard it as fixed by the tetrahedral symmetry N-C_α-C-C_β, which is basically true, then you risk missing the bigger picture, i.e. the secondary structure dependent rotation of the tetrahedron.

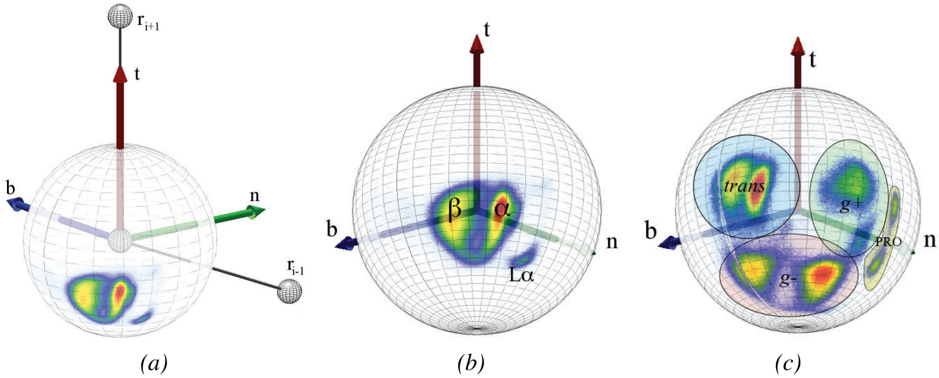


Figure 3.2: a) The distribution of C_β atoms in the DF-frame. The position of the next and previous residue is shown for comparison. b) The direction of the C_β depends heavily on the secondary structure. c) The position of the C_γ in the DF-frame. The three rotamers can be seen, but also a split depending on the secondary structure not seen in the Janin plot. Note that the cyclical shape of proline separates it from the other amino acids.

3.2.2 The C_γ position and further

Why stop at the C_β? Most side chains are longer than that and for most of them the next atom after the C_β is the C_γ. The DF-plot for C_γ can be seen in Fig. 3.2c. The position of the C_γ in relation to the C_β is only determined by χ_1 as the bond angle is more or less fixed. Hence, we expect to see the three rotamers and they are easily found in the figure. However, similar to the location of the C_β, the position of the C_γ depends on the secondary structure. A fact which is not seen in the Janin plot.

Proline with its typical cyclical structure bends in a completely different way than the other residues. The proline spots can be seen in the right part of Fig. 3.2c.

This method can be used to plot the positions of atoms further out in the side chain as well. However, the problem is that the difference between residues gets larger and the number of possible configurations grows the further you go in the side chain. Already in the C_γ picture there are eight distinct spots. Eventually the lack of data will make it impossible to go further.

3.3 C_β frame

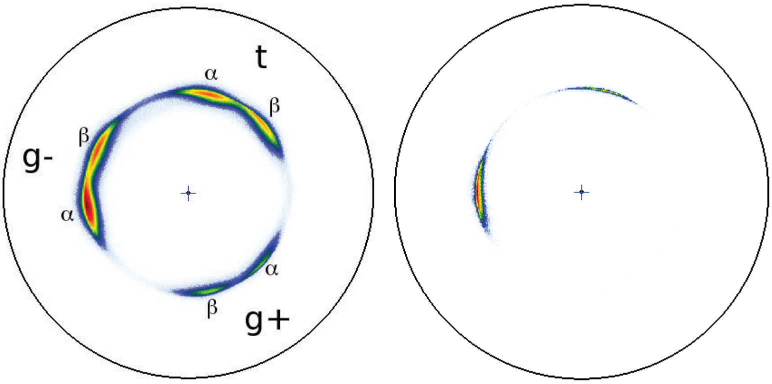
Another frame that can be useful is the C_β frame. It is defined by rotating the DF-frame so that north is in the direction of the line from C_α to C_β and moving the center to the C_β. Formally the definition is

$$\begin{aligned} \mathbf{s}_i &= \frac{\mathbf{r}_{C_\beta,i} - \mathbf{r}_{C_\alpha,i}}{\left|\mathbf{r}_{C_\beta,i} - \mathbf{r}_{C_\alpha,i}\right|} \\ \mathbf{p}_i &= \frac{\mathbf{s}_i \times \mathbf{t}_i}{|\mathbf{s}_i \times \mathbf{t}_i|} \\ \mathbf{q}_i &= \mathbf{s}_i \times \mathbf{p}_i \end{aligned} \tag{3.1}$$

where $\mathbf{r}_{C_\beta,i}$ is the location of the C_β, $\mathbf{r}_{C_\alpha,i}$ the location of the C_α and $\mathbf{t}_i$ the tangent vector of the DF-frame at i. Then $(\mathbf{s}_i, \mathbf{p}_i, \mathbf{q}_i)$ forms an orthonormal frame. This frame is mainly used together with the stereographic projection using the same procedure as in 2.3.2.

In Fig. 3.3a the position of the C_γ is shown in the stereographic C_β frame. In this figure the three-fold symmetry of the rotamers is more evident than in the DF-frame, but the difference between α-like behavior and β-like behavior is still evident.

Continuing from the C_β frame, other frames, like the C_γ frame, can be defined in a similar manner.



(a) (b)

Figure 3.3: a) The position of the C_γ in the stereographic C_β frame. The rotamers are denoted. b) The position of the C_γ for residues with C_β in the Lα island.

3.4 Investigating the Lα island

In Paper IV we made an effort to understand the small Lα island, seen in Fig. 3.2b. It can easily be shown that the residues in the island are the same that are in the Lα region in the Ramachandran plot (Fig. 2.4), except for glycine that lacks a C_β and hence will not show in the DF-plot.

In Fig. 3.3b we can see the C_γ-angles of the sites in the island. From this figure it follows that the most common configuration is χ_1 in *trans* followed by g-, while $g+$ is almost nonexistant. A first step in analyzing these two regions is to look at the distribution of residues shown in Fig. 3.4.

In the figure we can see that there is a large difference in the distribution of residues between the g- conformation and the *trans* conformation. The *trans* conformation is entirely dominated by the two residues asparagine and aspartic acid. Visual inspection of the structure indicates that the reason for this specificity might be interactions between the carbonyl groups in the main chain and the side chain, which would agree with the result in [15].

The g- conformation is much less residue specific, even though asparagine and aspartic acid still dominates. Visual inspection of the structure yields no obvious result even though it might be possible that carbonyl-carbonyl interactions still has some effect.

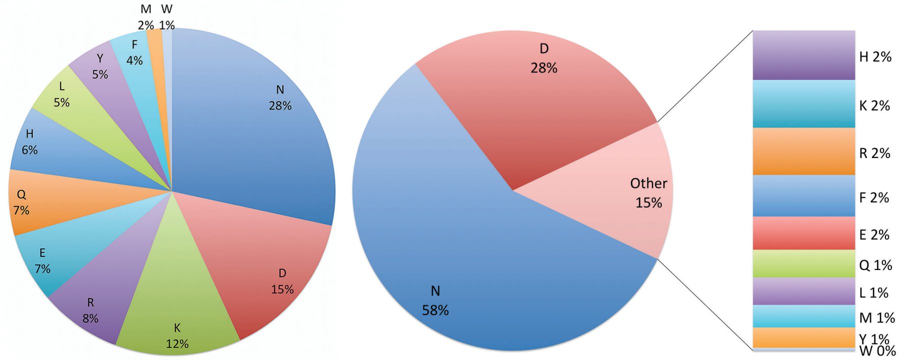
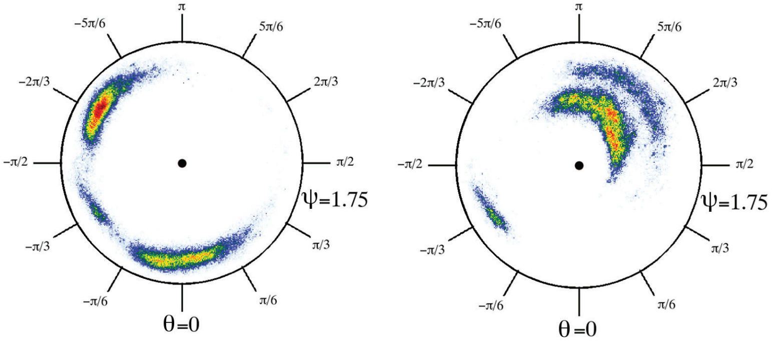


Figure 3.4: The distribution of the different amino acids in the g- island (left) and *trans* island (right).

Another thing to consider is, what does a residue in the Lα island mean for the curvature of the main chain? The Lα region in the Ramachandran plot has been shown to correspond to left handed helices [48]. Continuous left handed helices are forbidden by steric constraints unless they are solely composed of glycine. Hence, we expect to find residues in the Lα island as isolated parts of loops connecting regular secondary structure. The DF-plot is defined on the bond between two sites so it is only possible to plot the situation before and after the island. This is shown in Fig. 3.5.

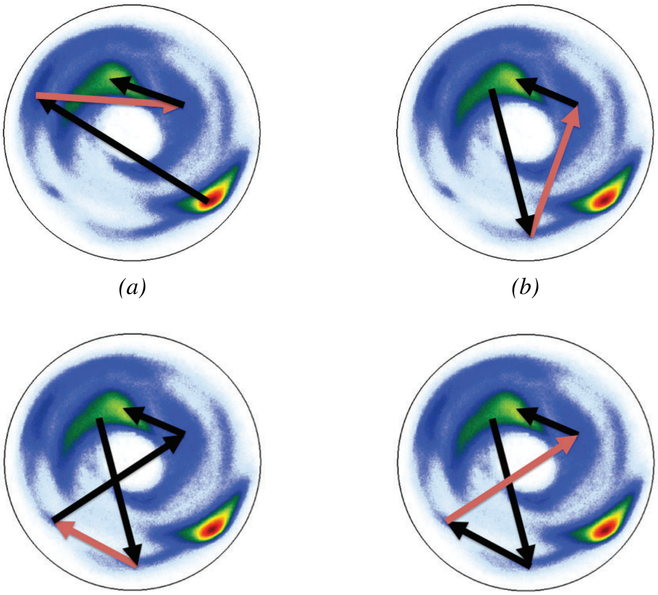
Finally the different regions in Fig. 3.5 can be analyzed to see how they connect to each other and what type of loop this corresponds to. We have identified four different types of loops, shown in Fig. 3.6. One of which, 3.6a, connects an α-helix with a β-strand, while the other three are different types of β-loop-β configurations. There is also the possibility of a combination of



(a) *(b)*

Figure 3.5: The distribution of DF-angles a) before a site in the Lα island and b) after the Lα island.

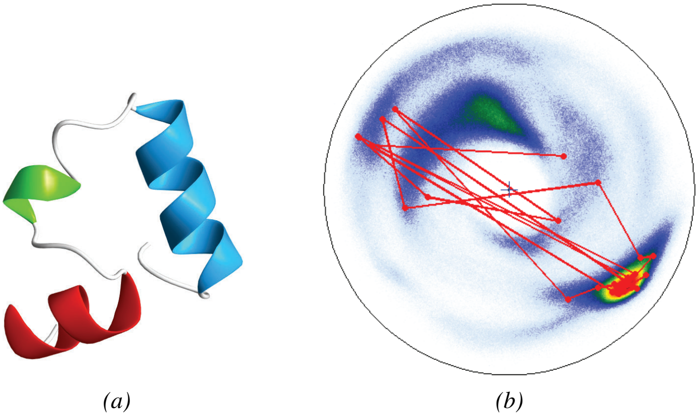
3.6c and 3.6d, where two sites in the Lα island coincide in the same loop, or the even more rare occasion with three sites where the middle one keeps ψ and θ constant at $\left(\frac{\pi}{2}, -\frac{\pi}{3}\right)$. Both 3.6c and 3.6d usually contains glycines in the loops as well, to reduce steric problems, but this has not been studied in detail.



(a) *(b)*

(c) *(d)*

Figure 3.6: Four different types of loops containing sites in the Lα island. The transition caused by that site is marked with a bright red arrow.



(a) *(b)*

Figure 3.7: a) The structure of chicken villin headpiece based on 1YRF in the PDB. b) The trajectory of 1YRF in the DF-plot with the density distribution for all proteins shown in the background.

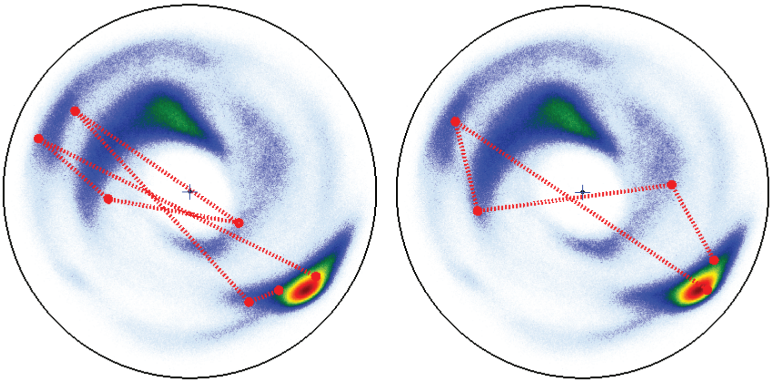
3.5 Chicken villin headpiece (1YRF) geometry

To end the part of the thesis about geometry I would like to utilize the methods developed to study one particular protein, the chicken villin headpiece (1YRF in the PDB) [11]. This is a simple, small protein that has been studied for many different purposes, using many different methods.

1YRF is a short, 35 residue, protein with a 3-helix structure as seen in Fig. 3.7a. Using the methods developed in chapter 2 it is possible to construct the DF-frame along the main chain and calculate the angles ψ and θ. The DF-plot of 1YRF in Fig. 3.7b shows heavy concentration in the α-helical region, as expected, while in the irregular loop regions the angles varies. The regular structure can be easily modelled. The real challenge is to find a good model to describe the loops.

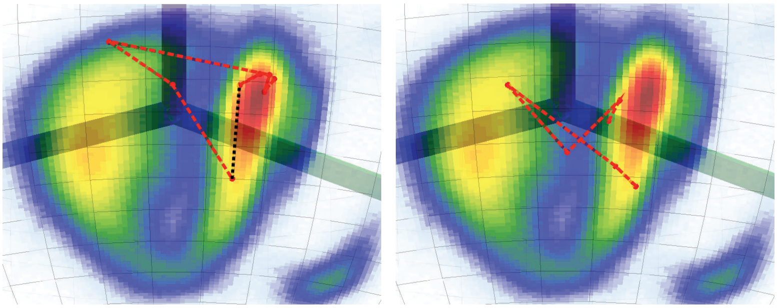
Notable is that the loop regions are not inherently irregular. Fig. 3.8 shows the two helix-loop-helix motifs in 1YRF. There are repeating patterns, like the tendency for the helix to end by shifting θ by π to the opposite side of the plot. The other points also seems to coincide with higher density areas, indicating that there is some kind of order to the loops. Part 2 of this thesis will deal with the question of how to classify loops, and how to predict and simulate them.

Finally the C_βs in the loop regions can be viewed in the DF-frame (Fig. 3.9). This plot is limited in its ability to study loop regions since it does not show glycines, which are very common in irregular structure because of their flexibility. Still, it displays the same pattern as the DF-plot in that the loop region corresponds to a transition from one stable, stationary state, to another. In this case from one α-helix to the next, through an irregular loop region that passes through the β side of the figure.



(a) *(b)*

Figure 3.8: The loop regions of 1YRF in the DF-plot. a) shows the region from site 50 to 57 in the PDB file and b) shows 60 to 65.



(a) *(b)*

Figure 3.9: The C_βs of the loop regions of 1YRF in the DF-plot. a) shows the first loop. The first loop contains a glycine which is not seen in this plot, so the residue before and after the glycine has been connected with a black dotted line to show that it is not a real connection. b) shows the second loop.

Part II:
Protein Modeling

4. A soliton model of proteins

The first part of this thesis introduced a model for how to describe the geometry of proteins in terms of DF-frames. However, it also introduced a few concepts that were left hanging without any deeper explanation of their use. In 2.2.1 the connection to Abelian-Higgs models were hinted at. In 2.2.2 a $\mathbb{Z}_2$ symmetry was introduced, only to be left out of the rest of the chapter. In this chapter these will be critical pieces in the task of solving the question posed in 3.5, how to model the loop regions of the protein. The central concept in solving the problem will be the mathematical structure known as solitons.

In this chapter I will show how solitons can be found in the structure of the main chain of proteins and how they can be described by a variant of the discrete non-linear Schrödinger equation.

In this part of the thesis I will use a slightly different notation than in the first part to get a better agreement with published papers. Instead of the angles (ψ, θ) I will talk about the discrete curvature and torsion (κ, τ). They are equivalent definitions as discussed in 2.2.2.

4.1 Introduction to solitons

Solitons were first described by the Scotsman John Scott Russell in 1834, when he saw, in the Union Canal in Scotland, a wave that was "*assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed*" [51]. This wave that Scott Russell saw was later shown to be accurately described by an exact solution of the Korteweg-de Vries equation [35]. During the last century solitons have been found in many different fields, from superconductivity [13] to cosmic strings and magnetic monopoles [41]. They have also proved useful for data transmission in fiber optics [13].

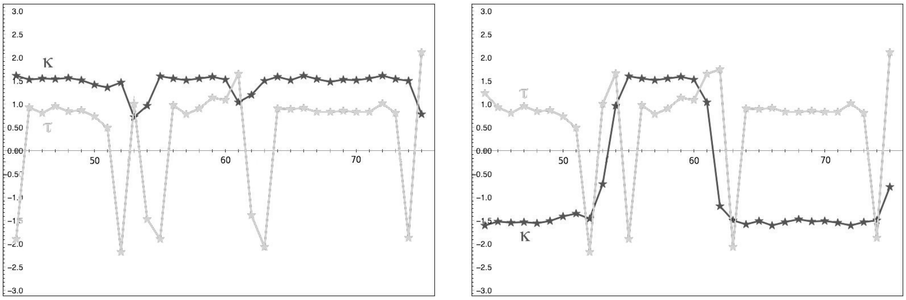
Generally a soliton is a solitary wave, which is localized, non-dispersive and interacts with other solitons without losing energy. In essence, they are waves that in many ways behaves like particles. The non-dispersive nature of solitons is what makes them good for long distance data transmissions. Contrary to ordinary waves the soliton cannot be described as local perturbations around some ground state. Solitons are inherently non-linear phenomena and emerge when non-linear interactions combine elementary constituents into a collective excitation, and can be shown to be the solution to many different kinds of nonlinear equations.

In biological molecules the concept of solitons was first introduced by Alexander Davydov as a way to describe lossless energy transfer in α-helices [14]. At the time it solved a big problem about how the energy from ATP-hydrolysis is transferred through the protein. Later, similar methods have been applied to DNA as well [63].

4.1.1 A soliton view of the main chain

The idea of interpreting the main chain of proteins as a sequence of soliton-(anti)soliton pairs was first introduced in [10]. Here I will return to the protein 1YRF that was discussed in 3.5. In Fig. 4.1a the curvature and torsion of 1YRF is plotted. The fluctuations of the torsion indicate the locations of the loop regions, but the soliton structure is hidden.

However, using the $\mathbb{Z}_2$ gauge transformation defined in 2.2.2, as in Fig. 4.1b, shows a different picture emerging. Here κ has the appearance of a topological soliton-(anti)soliton pair in a double-well potential. The two solitons are located in the loop regions of the protein and interpolate between the two ground-states, i.e. $\kappa = \frac{\pi}{2}$ and $\kappa = -\frac{\pi}{2}$. Note here that the chirality of the helices is the same regardless of the sign of κ.



(a) *(b)*

Figure 4.1: a) The curvature and torsion of 1YRF, with standard positive curvature, along the chain. b) Curvature and torsion after gauge transformation.

The basis of our model is that all proteins can be described as a sequence of solitons. Recent studies have shown that a large subset of the proteins in the PDB can indeed be accurately described by a combination of just a few hundreds of model solitons [37]. However, there is currently no good model to uniquely identify the locations of the solitons and this can be a tricky matter if the data is noisy.

4.2 Abelian Higgs model

The last section showed that proteins display a structure that is very reminiscent of solitons. The task is then to find an equation that can properly describe them. In 2.2.1 the connection was made to Abelian Higgs models and the following energy was suggested as a starting point

$$F = \int_0^L ds \left\{ |(\partial_s + iA_1)\phi|^2 + \lambda \left(|\phi|^2 - \mu^2 \right)^2 \right\}. \tag{4.1}$$

Higgs models and spontaneous symmetry breaking have gained fame lately with the construction of LHC and the search for the Higgs boson [24]. A three-dimensional version of this model is known as the Ginzburg-Landau model of superconductivity [41]. In our case we are interested in a discrete one-dimensional version, to work with a model featuring discrete one-dimensional curves. The following energy function was first introduced in Paper I, and later extended in [10] and [42]

$$E = -\sum_{i=1}^{N-1} 2\kappa_{i+1}\kappa_i + \sum_{i=1}^{N} \left\{ 2\kappa_i^2 + c\left(\kappa_i^2 - m^2\right)^2 + b\kappa_i^2\tau_i^2 \right\}$$

$$+ \sum_{i=1}^{N} \left\{ d\tau_i + e\tau_i^2 + q\kappa_i^2\tau_i \right\}. \tag{4.2}$$

Here the first and second sums comes from the discretization of (4.1). In the third sum the first term is the one-dimensional Chern-Simons term, which through the sign of d controls the chirality of the structure; the second term is a Proca mass term and the third is a regulator. More terms can be added to the energy (4.2) only as long as they are gauge invariant.

The parameters (b, c, d, e, m, q) are global in the sense that they are specific to one particular type of supersecondary structure, i.e. soliton, but do not depend directly on the local structure of the amino acids.

One issue with this energy is that the curvature and torsion are angles and defined modulo 2π. This periodicity is not taken into account in the energy, which means that two identical curves could have different energies.

4.2.1 The nonlinear Schrödinger equation

In [42] it was shown that the energy (4.2) could be directly related to generalized discrete nonlinear Schrödinger equations (GDNLS) and I will here summarize the results. The discrete nonlinear Schrödinger equation have been used in many diverse fields [33], like in the study of polarons in molecular crystals [26], Bose-Einstein condensates [17] and in the field of proteins to study Davydov's soliton [52].

To see the correspondence with the GDNLS equation τ can be removed from (4.2) by using the equation of motion, so that

$$\tau_i = -\frac{1}{2}\frac{d+q\kappa_i^2}{e+b\kappa_i^2}. \tag{4.3}$$

With τ substituted out, the remaining equation of motion looks like

$$\kappa_{i+1} - 2\kappa_i + \kappa_{i-1} = \frac{dU\left[\kappa_i\right]}{d\kappa_i^2} \tag{4.4}$$

where

$$U\left[\kappa_i\right] = -\left(\frac{bd-eq}{2b}\right)^2 \frac{1}{e+b\kappa_i^2} - \left(\frac{q^2+8bcm^2}{4b}\right)\kappa_i^2 + c\kappa^4. \tag{4.5}$$

If the parameters of the equation are chosen in such a way that the potential $U\left[\kappa_i\right]$ has two separate local minima, then there exists a dark soliton solution that interpolates between the two minima [25]. Typically these minima will be located in the vicinity of $\kappa = \pm m$. The parameter m can then be adjusted to fit typical values for proteins, i.e. $m \approx \pm\frac{\pi}{2}$ for α-helices and $m \approx \pm 1$ for β-sheets.

To solve (4.4) it is sufficient to find a fixed point of

$$\kappa_i^{(n+1)} = \kappa_i^{(n)} - \varepsilon\left\{\kappa_i^{(n)}\frac{dU\left[\kappa_i^{(n)}\right]}{d\left(\kappa_i^{(n)}\right)^2} - \left(\kappa_{i+1}^{(n)} - 2\kappa_i^{(n)} + \kappa_{i+1}^{(n)}\right)\right\} \tag{4.6}$$

where $\kappa_i^{(n)}$ is the nth iteration of some initial configuration and ε is an arbitrary, but sufficiently small, constant.

4.3 Building proteins

The geometrical tools defined in the geometry part of the thesis, the concept of solitons and the energy defined in the last section is sufficient to create a phenomenological model for proteins. The only thing remaining is to combine them.

In 4.1.1 it was shown that a sequence of solitons could be used to describe the curvature of a protein. Further, in 4.2 an energy was constructed, with six free parameters, that exhibits soliton solutions. Six parameters should then be enough to describe one particular type of soliton. However, the solitons in 4.2 are symmetric while real proteins might be asymmetric, e.g. a loop taking an α-helix to a β-strand. To deal with this situation the parameters m and c are allowed to be asymmetric as well so that

$$m = \begin{cases} m_1 & N < N_A \\ m_2 & N \geq N_A \end{cases}$$

$$c = \begin{cases} c_1 & N < N_A \\ c_2 & N \geq N_A \end{cases} \tag{4.7}$$

where N_A is some site inside the soliton, generally, but not necessarily, the center of the soliton.

4.3.1 Amino acid size

The components introduced so far have not taken the size of the amino acids into account. This is an important factor, especially when considering that self-intersections should be forbidden during folding. For this reason we have added the following self-avoiding condition to our model

$$\left|\mathbf{r}_i - \mathbf{r}_j\right| > z \tag{4.8}$$

for all ($\mathbf{r}_i$,$\mathbf{r}_j$), where $\mathbf{r}_i$ and $\mathbf{r}_j$ are the locations of site i and j respectively and z a parameter with a typical value of 3.8 Å. The drawback with this method is that it can not give 100 % certainty of no self-intersections occurring. Since the model deals with the angles a small change in one angle in one site may lead to a large shift further along the curve.

Another thing that is important, but not taken into consideration in the present model is the difference between the sizes of the different side chains. Many different models to treat the side chains exists, as described in 1.1.4, with different accuracy [55].

In paper V we introduced an energy function to describe the location of the C_β, which could potentially be used in the future to add the geometrical restrictions of the side chains into the model. I will discuss this energy more in Section 4.5.

4.3.2 Treatment of the solvent

The different ways of treating the solvent were outlined in 1.1.4. Our model is too coarse-grained to use explicit solvent so a more implicit approach is necessary. Conventional methods of implicit solvent rely on that the enormous amount of solvent molecules can be treated in a mean field theory. For example, the Coloumb force in a continuous medium between two charges q_i and q_j can be described as

$$F = \frac{1}{\varepsilon_r}\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}^2}$$

where ε_r is the dielectric constant of the solvent. Protein models in general use more complete models, like the generalized Born model [45], and take into account the exposed surface area as well [1].

In our model we take a different approach and instead assume that the effect of the solvent implicitly affects the soliton parameters.

4.3.3 Parameter fitting

The parameters in the equation (4.2) cannot be derived from first principles. Instead they have to be fitted by comparison to real solitons from proteins. For given parameters the κ values for the soliton solution can be found by starting from a step-function and then running the iteration in equation (4.6) until it converges and τ can then be calculated through equation (4.3).

For geometrical reasons it is reasonable to define τ_i between κ_{i-1} and κ_i, so that in equation (4.3)

$$\kappa_i \to \frac{\kappa_i + \kappa_{i-1}}{2}.$$

This is equivalent to the energy in equation (4.2) if the same substitution is made in the terms $b\kappa^2\tau^2$ and $q\kappa^2\tau$.
There are two ways to compare the simulated (κ_i,τ_i) values to the real soliton in question. Either calculate the (κ_i,τ_i) values for the real soliton by the procedure in 2.1.2, or calculate the curve corresponding to the simulated (κ_i,τ_i) values by the procedure in 2.1.3 and then look at the RMSD between the two curves. A simple Metropolis Monte Carlo procedure [22] is then enough to adapt the parameters to give as good fit as possible to the real soliton.

While the method just described works well for a single soliton, finding parameters for a whole protein, which generally is a multi-soliton configuration, is more complicated. Solitons corresponds to helix-loop-helix motifs and a larger protein can be constructed by gluing two solitons together inside a helix. However, there is no rule as to how much of the helix should be included in each soliton, i.e. where they should be glued together. Empirical tests have shown that it is often convenient to let one of the solitons describe most of the helix.

4.3.4 Future development

There are several obvious areas where an improved model would be good. I have already mentioned the question of the size of the side chains, where a better self-avoiding calculation would be helpful, but the most important improvement would be if a method was found to directly relate the parameter values to the amino acid sequence instead of using a fitting procedure.

Another thing that could be interesting is a deeper study of solitons in proteins, to see how they form and annihilate. It would be really good to find a way to uniquely define where they are located in the real proteins as well, instead of more or less putting them in by hand.

So far the method has only really been tested on structures rich in α-helices, and α-helices are inherently local objects because their stability depends on formation of hydrogen bonds within the helix. β-sheets, on the other hand, are stabilized by forming hydrogen bonds between β-strands that may be far away on the chain. It is an open question if our method, where non-local interactions are not treated in an explicit manner, can handle β-sheets.

4.4 Simulations

The study of simulating proteins using our model is still in its infancy. So far, the main simulations have been done by fitting parameters to one particular protein and then seeing if its possible to get to the correct shape of the protein by minimizing the energy, starting from either a straight line or one long helix.

We have found two different methods to create solitons from a helix. The easy way is to put them there by hand, by making gauge transformations at the correct positions to start with. The gauge transformations do not affect the shape of the curve but it takes the curvature much closer to the target.

The other method has to do with introducing an anti-ferromagnetic parameter to create a soliton-(anti)soliton pair to shift κ from $\frac{\pi}{2}$ to $-\frac{\pi}{2}$ and changing the shape of the double well potential to make the shift easier. In general that method is harder to work with and has a much higher risk of getting stuck in the wrong configuration. In the energy the addition of the anti-ferromagnetic term would correspond to adding a parameter a to the first sum in equation (4.2) so that the energy changes to

$$E = \sum_{i=1}^{N-1} 2a_{i+1,i}\kappa_{i+1}\kappa_i + \dots$$

with the standard ferromagnetic value $a = -1$ used for all sites except the center of the solitons where $a = 1$. The method also requires the parameter c to start at $c = 0$. Both of these parameters then slowly attain their correct equilibrium value during the early stages of the simulation.

4.4.1 Monte Carlo

The energy minimization is conducted through a Markovian Monte Carlo algorithm, where at each step the angle is perturbed at one point on the chain. The resulting state is then kept by the following, standard heat bath [20, 7], probability

$$P = \frac{x}{1+x} \quad \text{with} \quad x = \exp\left\{-\frac{\Delta E}{k_B T}\right\} \tag{4.9}$$

with ΔE being the energy difference between the new and the old state. All simulations are run at low temperature, so kT is chosen in such a way that the results belong to the collapsed phase. This will be discussed further in the next chapter. Note that there is no direct relation between the Monte Carlo temperature T and the real temperature.

There is also no direct relation between the number of Monte Carlo steps needed and the time of folding. Since one step only changes an angle, the actual movement of the chain could be very different for two different steps, and hence the time needed would be different. It might still be possible to relate the two concepts seeing that a folding that requires few Monte Carlo steps would probably be quick in real life as well. One possibility is to assign an average time for each Monte Carlo step, where the time depends on the average number of points affected by the move and hence the length of the chain.

4.4.2 1YRF revisited

Returning to the chicken villin 1YRF discussed in Section 3.5 and 4.1.1 for a final time, it is time to look at how well the model can describe it. 1YRF was first modeled in Paper V and later folded from a straight curve in [36]. The former showed that it was possible to find parameters to describe 1YRF with an accuracy of about 0.4 Å. But it also showed that the average distance between the modeled C_α and the real C_α was less than the Debye-Waller fluctuations [57] as given in the PDB file by the B-factors [66].

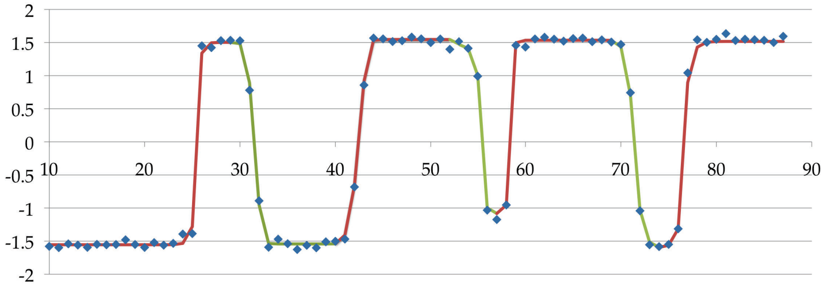
In [36] it was shown that it was possible to fold 1YRF starting from a straight line, by introducing an anti-ferromagnetic parameter. The resulting folding pathway had remarkable similarities to a different Monte Carlo study [65], taking a much more detailed approach, in the order and rate of helix formation.

4.4.3 λ-repressor 1LMB

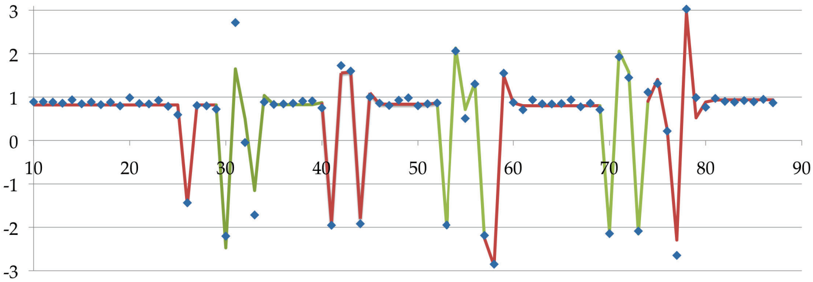
The lambda repressor 1LMB [4] is a small, fast folding protein consisting of 84 residues. It has been used in several protein folding studies and was recently folded to 1.8 Å accuracy with an all-atom simulation [40, 8].

Fig. 4.2 shows the resulting curvature and torsion from fitting the parameters after 1LMB, with the more disordered end regions cut away. The alternating color of the line shows the regions of the different solitons. The fit is really good for the most parts, with the exception of the torsion in the second soliton. The parameters used here gives an RMSD of 0.51 Å but this is a work in progress so further optimization of the parameters might very well give a better result.

Attempts to refold the curve from a long helix, using energy minimization, has shown that this is indeed possible by gauge-transforming the helix. Recent tests have shown that it is likely that it is possible to fold it by the method of pair-formation as well.



(a)



(b)

Figure 4.2: a) The curvature of 1LMB (dots) compared to simulated curvature. The alternating colors of the line indicate the location of the solitons. b) The torsion of 1LMB (dots) compared to the simulated torsion. The resulting simulated curve has a total RMSD of 0.51 Å compared to 1LMB.

4.5 Side chain energy

In Fig. 3.2b the statistical location of the C_β was shown to be shaped like a horse-shoe. Letting the spherical angle θ denote the angle from the tangent direction and ϕ denote the angle from the normal direction it should be possible to define an energy in terms of (θ,ϕ) to describe the loop behavior of the side chain, as seen in Fig. 3.9.

In Paper V such an energy is introduced, and I will refer the motivation of it to that paper, where θ and ϕ are independent of each other and only related through their mutual dependence on κ. The energy is as following

$$E_\theta = \sum_{i=1}^{N}\left\{\frac{d_\theta}{2}\kappa_i^2\theta_i^2 - b_\theta\kappa_i^2\theta_i - a_\theta\theta_i + \frac{c_\theta}{2}\theta_i^2\right\}$$

$$E_\phi = \sum_{i=1}^{N}\left\{\frac{d_\phi}{2}\kappa_i^2\phi_i^2 - b_\phi\kappa_i^2\phi_i - a_\phi\phi_i + \frac{c_\phi}{2}\phi_i^2\right\} \qquad (4.10)$$

and analogous to equation (4.3) the equations of motion gives the values for θ and ϕ as

$$\begin{aligned}\theta_i &= \frac{a_\theta + b_\theta \kappa_i^2}{c_\theta + d_\theta \kappa_i^2} \\ \phi_i &= \frac{a_\phi + b_\phi \kappa_i^2}{c_\phi + d_\phi \kappa_i^2}\end{aligned} \tag{4.11}$$

where we can fix $c_\theta = c_\phi = 1$, without any loss of generality.
This method of modeling the side chain was tested by applying it to the test protein 1YRF, adapting parameters to fit the position of both the C_α and C_β. The parameters we found were able to model the protein with an accuracy of less than 0.4 Å.

5. Scaling laws

Polymers has been studied since the early 19th century, but it was only in the 70s that it was shown, by de Gennes, that polymers in solution should be considered a critical system [21]. Critical systems can be divided into universality classes that enable the calculation of critical properties for an entire class of physical systems using a single representative of the model [61, 62]. In the case of polymers there are three different universality classes, or phases; the self-avoiding random walk (SARW), the ordinary random walk (RW) and the collapsed phase [12]. A regular protein lives in the collapsed phase, while if you heat it up it goes through a phase-transition to the SARW phase. During folding the process is the opposite since there it goes from the SARW to the collapsed phase [30].

The easiest way to see this phase transition is to boil an egg. The egg white is rich in proteins and at room temperature it is a transparent liquid. However, if you boil it then the proteins go through a phase transition to an unfolded (SARW) state that is solid and opaque.

A complete model for proteins needs to model this phase behavior. In this chapter I will summarize the findings of Paper II where we studied the phase structure of our model to compare it with real proteins.

5.1 Compactness index

The different universality classes are characterized by different values of so-called critical exponents that describe the scaling behavior of the proteins in the limit where the number of residues becomes large. The most commonly used critical exponent in the study of proteins is the compactness index, ν, which tells how the radius of the protein, R_g, depends on the number of residues, L. The best way to measure the radius is to use the radius of gyration

$$R_g^2 = \frac{1}{2L^2}\sum_{i,j}(\mathbf{r}_i - \mathbf{r}_j)^2 \tag{5.1}$$

where $\mathbf{r}_i, \mathbf{r}_j$ are the locations of the residues, but other measures like the end-to-end distance are also possible. For large L the radius of gyration goes as

$$R_g^2 \approx R_0^2 L^{2\nu}\left(1 + \beta_1 L^{-\Delta_1} + \ldots\right) \tag{5.2}$$

where ν and Δ_1 are the universal critical exponents [43]. However, the form factor, R_0, and the amplitude of the leading finite size corrections, β_1, are not. The correction terms $\beta_i L^{-\Delta_i}$ are generally small and I will ignore them here. The previously mentioned phases in proteins has the following mean field values for the compactness index ν

$$\nu = \begin{cases} 3/5 & \text{SARW}, \\ 1/2 & \text{RW}, \\ 1/3 & \text{collapsed}. \end{cases} \tag{5.3}$$

The compactness index is the inverse of the Hausdorff dimension of the protein so the collapsed phase would correspond to a space-filling curve. Another theoretical value for ν is worth mentioning. A straight line, or a long straight helix both behaves like a one-dimensional system, i.e. $\nu = 1$.

In practice, for proteins the SARW phase corresponds to high temperature or good solvent and the collapsed phase corresponds to low temperature or bad solvent. These two phases are separated at the Θ temperature which is also the location of the RW phase. In Paper I we measured ν for the proteins in the PDB. The resulting value was $\nu = 0.378$, as seen in figure 5.1, which fits well with other studies [27].

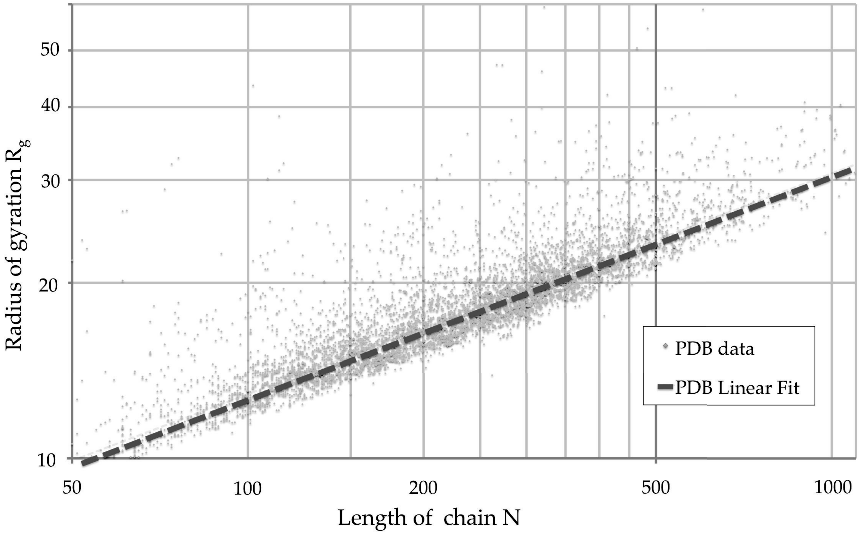


Figure 5.1: The length of the chain versus the radius of gyration for the proteins in the PDB in a Log-Log plot. The fitted line gives a value $\nu = 0.378$. Adapted from paper I.

5.2 Simulations

The basic idea of Paper II was to make simulations using constant parameters but varying the length, i.e. number of residues, and temperature of the model to see how the energy and compactness index changed. The model used for the simulations is similar to the one in Paper I in that it uses the discretized version of the Frenet equations as defined in Section 2.1.1 and a simplified version of the energy in equation (4.2):

$$E = \sum_{n.n} a \left\{1 - \cos\left[\omega\left(\kappa_i - \kappa_j\right)\right]\right\} + \sum_i \left\{ b\kappa_i^2\tau_i^2 + c\left(\kappa_i^2 - m^2\right)^2 + d\tau_i \right\} \qquad (5.4)$$

where the parameters are global and the first sum extends only over nearest neighbors. With all parameters constant the end result only depended on the length of the curve and the Metropolis temperature.
The energy was minimized using a Monte Carlo simulation as described in Section 4.4.1, starting from a straight line ($\kappa = 0$, $\tau = 0$). For each temperature value between 10 and 16 different lengths were used, and for each length at least 200 curves were generated. Each curve was run for 11000 multiplied by the length number of Monte Carlo steps. This value was chosen to give the curve enough time to settle down while still keeping the simulation time for the longer curves reasonably low.

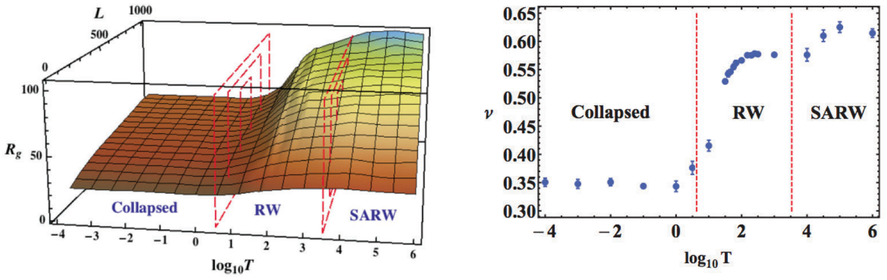
5.2.1 The radius of gyration

The first aim of with the simulations was to see if the compactness index followed the same pattern as in real proteins. The measured radius of gyration is shown in Fig. 5.2a as a function of length and temperature. The data can be fitted to equation (5.2) to get values for ν for different temperatures. The resulting values for the compactness index can be found in Fig. 5.2b.

The asymptotic value for the compactness index at low temperatures is $\nu = 0.35$, which is close enough to the mean field value $\frac{1}{3}$ to show that this is the collapsed phase. The compactness index has no temperature dependence in the collapsed phase. This stands in bright contrast to the RW phase where there is a rapid transition from $\nu = 0.4$ to $\nu = 0.58$ as temperature increases. At the center of this transition the value of ν is very close to $\nu = 0.5$, i.e. the mean field value at the Θ temperature for the RW phase.

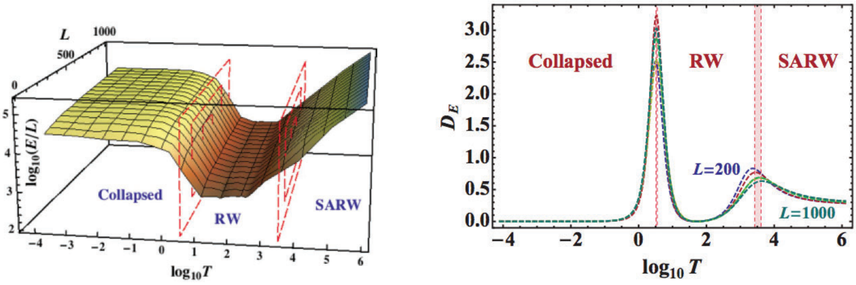
The transition from RW to SARW is harder to see. When $T \to \infty$ the curve has to behave like a self-avoiding random walk. The asymptotic value here is $\nu = 0.62$, which is slightly higher than the mean field value, but consistent with the fact that the mean field value is approached from above when length increases [39].

Low temperature tests have shown that the collapsed phase is not attainable if certain conditions are not met. Setting the parameter d in the energy (5.4) to zero, or removing the self-avoidance from the system produces results with compactness index $\nu = 0.5$. Setting $c = 0$ would instead just produce a straight line with $\nu = 1$.



(a) *(b)*

Figure 5.2: a) The radius of gyration as a function of temperature and length in a Log-Log plot. The three different phases are denoted. b) The compactness index ν as a function of temperature. The red lines mark the different phases. Note that the definitions of the lines have changed from the version in Paper II to agree with (a) and Fig. 5.3.



(a) *(b)*

Figure 5.3: a) The specific energy as a function of temperature and length L. The three different phases are denoted. b) The values for the critical temperatures are given by the maximum of the function (5.5) for different values of L.

5.2.2 Energy and critical temperature

The resulting energy for the simulated curves can be seen in Fig. 5.3a. The plot clearly shows three different regions, here tentatively labeled collapsed, RW and SARW after the different phases. In the collapsed and RW phase the specific energy ($\frac{E}{L}$) shows a weak dependence on L while for high temperatures this dependence disappears. In the limit of infinite temperature the energy has no effect and the curve will only be subject to random fluctuations.

The specific energy of the collapsed phase is much higher than for the RW phase. This is a phenomena that is caused by starting at a high-energy state (straight line). When folding the curves in the collapsed phase gets stuck in high energy conformations while the larger thermal vibrations in the RW phase prevents this from occurring. From an energy point of view the transition from collapsed to RW phase is when the first sum in equation (5.4) becomes irrelevant and the transition from RW to SARW is when the central bump of the double well potential becomes irrelevant.

To find values for the critical temperatures where the phases change it is possible to look at the maxima of the squared logarithmic derivative of the specific energy with respect to temperature

$$D_E(T,L) = \left[\frac{\partial \log E(T,L)}{\partial \log T}\right]^2. \tag{5.5}$$

The plot of this function can be seen in Fig. 5.3b. From this function the critical values for the (Metropolis) temperature is shown to be $\log T_{c1} = 0.53$ and $\log T_{c2} = 3.52$. These values are used to define the phase boundaries in Fig. 5.2 and 5.3.

Finally we have found that in the three phases the energy can be approximated with the following functions

$$E_{coll} = C_{coll} L \ln \frac{L}{L_0^{coll}} \tag{5.6}$$

$$E_{RW} = C_{RW}(T) L \left[1 - \left(\frac{L}{L_0^{RW}(T)}\right)^{-\gamma(T)}\right] \tag{5.7}$$

$$E_{SARW} = C_0 T^{\alpha} L. \tag{5.8}$$

In the collapsed phase the energy follows a logarithmically corrected linear law, with parameters essentially independent of temperature. The energy for the RW phase is easiest to understand by using the mean field value $\nu = 0.5$ to transform it into

$$E(R_g, T) = C_{RW} \left(\frac{R_g}{R_o}\right)^2 \left[1 - \left(L_0^{RW}\right)^{\gamma} \left(\frac{R_g}{R_0}\right)^{-2\gamma}\right] \tag{5.9}$$

which is Hooke's law with a correction term.

Finally the SARW phase, where the energy simply increases linearly with the length of the chain and the temperature dependence can be described by a simple power law. In the transition region between the RW and SARW phases both (5.7) and (5.8) can be used to describe the energy.

6. Epilogue

In the introduction I wrote about the protein folding problem and how it relates to diseases like Alzheimer's. While reading the thesis it is easy to lose track of the bigger picture and only look at the details. In Chapter 2-5 I have hardly mentioned folding and definitely not mentioned Alzheimer's so how does it all fit together?

The conventional ways of doing all-atom simulations with molecular dynamics are based on ideas that were developed in the 70's. By making the parameter sets increasingly more complex and spending millions on custom built supercomputers it is possible to get incredible results. Still, without an enormous breakthrough in computer technology they will always run into a hard limit. The fact that proteins in general are rather large. In practice, to have any hope of simulating larger proteins, a more coarse-grained approach is necessary. Many different methods have been proposed and one of them is described in this thesis. History will tell if this approach will be of any help to solve the protein-folding problem.

The basic idea is to describe the proteins not in terms of their primary or secondary structure but in terms of their transitions, the loops. Each loop would then correspond to a one- or multi-soliton solution of an energy function derived from very general physical principles. The symmetries of the system is what really determines the possible energy terms, not the microscopical interactions.

The geometrical view of proteins described in Chapter 2 and 3 is totally independent on the energy and has nothing to do with if a protein can be described in terms of solitons or not. It is purely geometrical and my hope is that it can be used as a complement to Ramachandran plots and side chain rotamer libraries, for example in structure quality assessments and classification of loops.

As for the future prospects of the model I have already detailed a few logical improvements to the model in Chapter 4. The model is still in its infancy and there are a lot of aspects that we simply do not understand. The simulations are often more tests of the model or proof of concept rather than fully utilizing the model to find new results. I am sure that the practical usability of the model will improve in the future.

With this I finish my contribution to the field of protein folding. It may not be the solution to once and for all get rid of the problem, but in the current situation where no such solution can be seen on the horizon, all small contributions are valuable. Maybe sometime in the future it will be the combination of many new ideas like this one that will increase our knowledge of proteins to the level where we will be able to do something about the diseases, caused by misfolded proteins, that plagues humanity.

Acknowledgements

There were times when I thought that I would never stand here today with a finished thesis. Too many times have I doubted myself. So, I would like to say a few words of thanks to the people who made me who I am today.

First I would like to thank my supervisor Antti. You brought me into the field of proteins. It was a new field for both of us and not always easy to get acceptance for what we were doing. Yet, even when I had my doubts you were always optimistic and believed in our work, and more often than not it turned out you were right in the end.

I would also like to thank my collaborators over these years, Shuangwei, Xubiao, Maxim, Ulf, Fan, Yan, Nora, and especially Andrei for reading through and commenting on the thesis. It has been a great pleasure working with you all. We started out from nothing and look how far we have come.

Thanks to all my fellow Ph.D. students, postdocs and master students. You have all done your part in making the department a great place to work. To the senior staff I would just like to say that, even though I did not have much contact with most of you, you all contributed to giving the department an aura of scientific intellectuality that would seep down to the rest of us.

To all my friends, you know who you are, you all have a share in keeping me sane these years, whether it was by inviting me to poker and beer night, taking a break from physics to talk about football during lunch breaks, discussing the perils of the increasingly complex modern society or astonishing me with yet another beautiful newly-written song.

Finally I would like to thank my dear family and especially my parents for creating an environment at home where curiosity is rewarded and there is always room for discussions. You have always believed in me and supported me, even when my own belief in myself was fading.

Summary in Swedish

Böjande, vridande och vändande
Proteinmodellering och visualisering ur en gaugeinvariant synvinkel.

"Men du, hur mycket ska man äta då?" Frågan gör mig ställd. Försiktigt får jag ur mig några ord om att det där kanske inte riktigt är mitt expertområde. För många betyder proteiner en stor burk med pulver som man ska ta för att bli stor och stark. För mig har den världen alltid känts främmande. Protein har en helt annan betydelse för mig. De är kroppens grovjobbare. Maskinerna som får allting att fungera.

Proteinerna skapas alla på samma sätt, genom att DNA först översätts till RNA för att det sedan i ribosomerna ska bildas en kedja av aminosyror. Denna kedja kommer att vecka ihop sig till ett protein, i precis den form som krävs för att utföra sin funktion. Så långt beter sig alla proteiner likadant men den funktion som de utför kan se helt olika ut för olika proteiner. Vissa är enzymer och ägnar sig åt att katalysera kemiska reaktioner, andra transporterar saker från platsen de finns till platsen de behövs och åter andra bygger upp struktur. Allt i ett så sammanvävt intrikat system att man nästan blir religiös av hur bra det fungerar.

Men, nu är det inte alltid som det fungerar så perfekt. Ibland blir det något fel som gör att proteinerna inte formar sig riktigt som de ska. Den felaktiga formen gör att de inte löser sig så bra i vatten. Istället klumpar de ihop sig till så kallad plack. Det här kan man se i hjärnan på personer som lider av Alzheimers eller Creutzfeldt–Jakobs sjukdom.

De proteiner som finns i allt liv har utvecklats under årmiljonerna för att fylla sin funktion till fulländning, men ibland vill vetenskapen göra små ändringar eller utnyttja naturens egna metoder. Det handlar då framför allt om specialtillverkade proteiner som används i medicinska syften. När man utvecklar dessa vill man kunna räkna ut vilka aminosyror man behöver för att få en viss form på proteinet. Det är formen det hela hänger på. Det är den som styr funktionen.

Här är grunden till proteinveckningsproblemet. Man kan relativt enkelt skapa en kedja av olika kombinationer av de 20 aminosyror som normalt förekommer i proteiner. Att utifrån den bestämma form på det veckade proteinet och därmed dess funktion har visat sig vara ett formidabelt problem. Om man kunde förstå mer om hur proteiner veckas så skulle man också lära sig mer om varför det ibland blir fel. Då kanske man skulle kunna hitta ett botemedel till sjukdomar såsom Alzheimers och Creutzfeldt-Jakobs. Man skulle också kunna designa nya proteiner för specifika syften.

Anledningen till att man inte kan beskriva proteinveckning i detalj idag handlar inte om brist på kunskaper om de kemiska reaktionerna eller vilka atomer som ingår i proteinet. Problemet är att proteiner i allmänhet är så stora att modellerna fallerar på grund av för mycket data. Tusentals atomer som alla påverkar varandra med olika krafter samverkar till att i princip alltid leda proteinet till en enda unik slutgiltig form. Vill man försöka simulera alla dessa interaktioner så måste man, trots den enorma utveckling som skett för datorer de senaste 40 åren, begränsa sig till små korta proteiner. Precis som en meteorolog inte tar hänsyn till varje liten luftmolekyl för att förutsäga vädret så skulle en lösning på proteinveckningsproblemet kunna vara att inte ta hänsyn till alla små mikroskopiska krafter utan istället försöka beskriva det effektiva resultatet av att alla dessa krafter samverkar. Det är den väg vi har valt.

Istället för att bry oss om detaljerna så betraktar vi proteinet som en kantig kurva med de centrala kolatomerna i aminosyrorna som noder. Det enda som finns kvar av utsträckningen på aminosyrorna är kravet att två noder inte får vara närmare varandra än avståndet till närmaste grann-noderna på kurvan. En komponent till behövs för att få en modell för proteiner, nämligen en energi. När ett protein bildas så har det en hög energi, och allting strävar efter att sänka sin egen energi tills dess det råder jämvikt med omgivningen. Det svåra är att veta hur man ska definiera en energi utan att ta hänsyn till alla små krafter mellan atomer.

Vad vi har gjort är att använda generella fysikaliska principer om symmetri för att sätta upp en energi för hur kurvan böjer sig och vrider sig. Den allra lägsta energin får en punkt på kurvan om den kröker sig precis som en α-helix, som är en av de vanligaste byggstenarna i ett protein. Det lägsta energitillståndet för hela kurvan är då att den formar sig till en enda lång helix. Så ska inte riktiga proteiner se ut och det kommer inte kurvan att göra heller. Orsaken till det stavas solitoner.

I vår modell finns det två sätt för kurvan att kröka sig. Antingen har den positiv eller negativ krökning. Om man bara tittar på kurvan kan man inte se någon skillnad på en helix som bildats genom positiv eller negativ krökning. Utgår man från kurvan som en rak linje så kommer vissa delar av linjen börja kröka sig först. Varje nod på kurvan påverkar sina grannar så om en nod börjar kröka åt ett visst håll så kommer den att dra med sig omkringliggande noder. Det här pågår tills dess att hela kurvan är uppdelad i områden med krökning i positiv eller negativ riktning.

Där de här områdena möts så kommer det finnas övergångsregioner där krökningen förändras från positiv till negativ helix. Matematiskt kallas det här för en soliton. Sett utifrån kommer det se ut som att kurvan från ena ändan formar sig som en helix för att sedan gå över i en oregelbunden region innan nästa helix börjar. Det här mönstret upprepar sig sedan genom hela kurvan och det är samma mönster som man kan se i riktiga proteiner. Det jag inte har nämnt här, och som vi heller inte undersökt i någon större utsträckning än, är de andra återkommande byggstenarna i proteiner, β-flak (β-sheet) och andra mindre vanliga typer. I princip kan man beskriva dem på samma sätt, men med en annan krökning, även om β-flak kan vara svåra att ta med i modellen på grund av att de är mer beroende av kontakter med andra delar av proteinet.

Varje sådan här soliton kan beskrivas med några få parametrar men eftersom modellen inte tar hänsyn till alla små krafter, utan handlar om vad som händer när alla dessa samverkar, så är det väldigt svårt att härleda värden för dessa parametrar ur kunskapen om de underliggande aminosyrorna. Det som återstår är att ta ett riktigt protein att jämföra med och försöka anpassa parametrarna för att ge så bra överensstämmelse som möjligt. Det finns resultat som tyder på att det bara finns ett begränsat antal olika typer av solitoner. I så fall skulle det räcka med att bygga upp ett bibliotek över de möjliga typerna och problemet skulle förenklas till att handla om att från aminosyresekvensen förutsäga vilken soliton som skulle vara aktuell i det fallet.

Avslutningsvis så vill jag bara säga att de här modellerna är väldigt nya och otestade på många plan. Mycket återstår att göra och det är mycket som vi ännu inte förstår, men skulle det någon gång i framtiden visa sig att solitoner är en återvändsgränd så kan ändå de geometriska metoderna i första delen av avhandlingen stå på egna ben. Jag tror att det här synsättet har en framtid och kommer att öka vår förståelse för hur proteiner fungerar. Det kommer inte att lösa proteinveckningsproblemet en gång för alla, men i det läge vi är i idag, när ingen lösning står för dörren, blir alla bidrag värdefulla.

References

[1] M. am Busch, A. Lopes, N. Amara, C. Bathelt, and T. Simonson. Testing the coulomb/accessible surface area solvent model for protein stability, ligand binding, and protein design. *BMC Bioinformatics*, 9(1):148, 2008.

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

[3] M. Arai and K. Kuwajima. Role of the molten globule state in protein folding. In C. Robert Matthews, editor, *Protein folding mechanisms*, volume 53 of *Advances in Protein Chemistry*, pages 209 – 282. Academic Press, 2000.

[4] L. J. Beamer and C. O. Pabo. Refined 1.8 Å crystal structure of the λ repressor-operator complex. *Journal of Molecular Biology*, 227(1):177 – 196, 1992.

[5] D. S. Berkholz, M. V. Shapovalov, R. L. Dunbrack Jr., and P. A. Karplus. Conformation dependence of backbone geometry in proteins. *Structure*, 17(10):1316 – 1325, 2009.

[6] 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.

[7] A.B. Bortz, M.H. Kalos, and J.L. Lebowitz. A new algorithm for monte carlo simulation of ising spin systems. *Journal of Computational Physics*, 17(1):10 – 18, 1975.

[8] G. R. Bowman, V. A. Voelz, and V. S. Pande. Atomistic folding simulations of the five-helix bundle protein λ_{6-85}. *Journal of the American Chemical Society*, 133(4):664–667, 2011.

[9] J. D. Bryngelson and P. G. Wolynes. Intermediates and barrier crossing in a random energy model (with applications to protein folding). *The Journal of Physical Chemistry*, 93(19):6902–6915, 1989.

[10] M. Chernodub, S. Hu, and A. J. Niemi. Topological solitons and folded proteins. *Phys. Rev. E*, 82:011916, Jul 2010.

[11] 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. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7517–7522, May 2005.

[12] J.D. Cloizeaux and G. Jannink. *Polymers in Solution: Their Modelling and Structure*. Oxford Classic Texts in the Physical Sciences. Oxford University Press, 2010.

[13] T. Dauxois and M. Peyrard. *Physics Of Solitons*. Cambridge University Press, 2006.

[14] A.S. Davydov. *Solitons in molecular systems*. Mathematics and its applications (Kluwer Academic Publishers).: Soviet series. Kluwer Academic Publishers, 1991.

[15] C. M. Deane, F. H. Allen, R. Taylor, and T. L. Blundell. Carbonyl-carbonyl interactions stabilize the partially allowed ramachandran conformations of asparagine and aspartic acid. *Protein Eng*, 12(12):1025–1028, December 1999.

[16] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl. The protein folding problem. *Annual Review of Biophysics*, 37(1):289–316, 2008. PMID: 18573083.

[17] J. C. Eilbeck and M. Johansson. The Discrete Nonlinear Schrödinger Equation - 20 Years On. In Luis Vazquez, Robert S. Mackay, and Maria P. Zorzano, editors, *Localization and Energy Transfer in Nonlinear Systems: Proceedings of the Third Conference, June 17-21, 2002, San Lorenzo De El Escorial Madrid*. World Scientific Publishing Company, 2002.

[18] R. A. Engh and R. Huber. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallographica Section A*, 47(4):392–400, Jul 1991.

[19] A L Fink. Chaperone-mediated protein folding. *Physiological Reviews*, 79(1):425–449, 1999.

[20] R. J. Galuber. Time-dependent statistics of the ising model. *Journal of Mathematical Physics*, 4(2):294–307, 1963.

[21] P.G. Gennes. *Scaling concepts in polymer physics*. G - Reference, Information and Interdisciplinary Subjects Series. Cornell University Press, 1979.

[22] H. Gould, J. Tobochnik, and W. Christian. *An introduction to computer simulation methods: applications to physical systems*. Pearson Addison Wesley, 2007.

[23] V. Grantcharova, E. J. Alm, D. Baker, and A. L. Horwich. Mechanisms of protein folding. *Current opinion in structural biology*, 11(2):70–82, 2001.

[24] G. G. Hanson. Searching for the higgs. *Physics*, 2:106, Dec 2009.

[25] M. Herrmann. Heteroclinic standing waves in defocusing dnls equations: variational approach via energy minimization. *Applicable Analysis*, 89(10):1591–1602, 2010.

[26] T. Holstein. Studies of polaron motion: Part 1. the molecular-crystal model. *Annals of Physics*, 8(3):325 – 342, 1959.

[27] L. Hong and J. Lei. Scaling law for the radius of gyration of proteins and its dependence on hydrophobicity. *Journal of Polymer Science Part B: Polymer Physics*, 47(2):207–214, 2009.

[28] R. W.W. Hooft, C. Sander, and G. Vriend. Objectively judging the quality of a protein structure from a ramachandran plot. *Computer applications in the biosciences : CABIOS*, 13(4):425–430, 1997.

[29] S. Hovmöller, T. Zhou, and T. Ohlson. Conformations of amino acids in proteins. *Acta Crystallographica Section D*, 58(5):768–776, May 2002.

[30] K. Huang. *Lectures on statistical physics and protein folding*. World Scientific, 2005.

[31] R. L. Dunbrack Jr. Rotamer libraries in the 21st century. *Current Opinion in Structural Biology*, 12(4):431 – 440, 2002.

[32] E. W. Justh and P. S. Krishnaprasad. A simple control law for uav formation flying. Technical Report TR 2002-38, Institute for Systems Research, 2002.

[33] P.G. Kevrekidis and R. Carretero-González. *The Discrete Nonlinear Schrödinger Equation: Mathematical Analysis, Numerical Computations and*

Physical Perspectives. Springer tracts in modern physics. Springer, 2009.

[34] W. Klingenberg. *A course in differential geometry*. Graduate texts in mathematics. Springer-Verlag, 1978.

[35] D. J. Korteweg and G. de Vries. Xli. on the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. *Philosophical Magazine Series 5*, 39(240):422–443, 1895.

[36] A. Krokhotin, M. Lundgren, and A. J. Niemi. Soliton driven relaxation dynamics and universality in protein collapse. `arXiv:1111.2028 [physics.bio-ph]`.

[37] A. Krokhotin, A. J. Niemi, and X. Peng. Soliton concepts and protein structure. *Phys. Rev. E*, 85:031906, Mar 2012.

[38] Y. Levy and J. N. Onuchic. Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.*, 35:389–415, 2006.

[39] B. Li, N. Madras, and A. Sokal. Critical exponents, hyperscaling, and universal amplitude ratios for two- and three-dimensional self-avoiding walks. *Journal of Statistical Physics*, 80:661–754, 1995. 10.1007/BF02178552.

[40] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How fast-folding proteins fold. *Science (New York, N.Y.)*, 334(6055):517–520, October 2011.

[41] N. Manton and P.M. Sutcliffe. *Topological Solitons*. Cambridge Monographs on Mathematical Physics. Cambridge University Press, 2004.

[42] N. Molkenthin, S. Hu, and A. J. Niemi. Discrete nonlinear schrödinger equation and polygonal solitons with applications to collapsed proteins. *Phys. Rev. Lett.*, 106:078102, Feb 2011.

[43] B. G. Nickel. One-parameter recursion model for flexible-chain polymers. *Macromolecules*, 24(6):1358–1365, 1991.

[44] A. J. Niemi. Phases of bosonic strings and two dimensional gauge theories. *Phys. Rev. D*, 67:106004, May 2003.

[45] A. Onufriev. *Continuum Electrostatics Solvent Modeling with the Generalized Born Model*, pages 127–165. Wiley-VCH Verlag GmbH & Co. KGaA, 2010.

[46] C. N. Pace, B. A. Shirley, M. McNutt, and K. Gajiwala. Forces contributing to the conformational stability of proteins. *The FASEB Journal*, 10(1):75–83, 1996.

[47] O.B. Ptitsyn and A.A. Rashin. A model of myoglobin self-organization. *Biophysical Chemistry*, 3(1):1 – 20, 1975.

[48] C. Ramakrishnan and G.N. Ramachandran. Stereochemical criteria for polypeptide and protein chain conformations: Ii. allowed conformations for a pair of peptide units. *Biophysical Journal*, 5(6):909 – 933, 1965.

[49] A. V. Rojas, A. Liwo, and H. A. Scheraga. Molecular dynamics with the united-residue force field: Ab initio folding simulations of multichain proteins. *The Journal of Physical Chemistry B*, 111(1):293–309, 2007. PMID: 17201452.

[50] G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan. A backbone-based theory of protein folding. *Proceedings of the National Academy of Sciences*, 103(45):16623–16633, 2006.

[51] J.S. Russell. *Report on waves: made to the meetings of the British Association in 1842-43*. 1845.

[52] A. Scott. Davydov's soliton. *Physics Reports*, 217(1):1 – 67, 1992.

[53] M. Sela, F. H. White, Jr., and C. B. Anfinsen. Reductive Cleavage of Disulfide

Bridges in Ribonuclease. *Science*, 125:691–692, April 1957.

[54] B. Siciliano and O. Khatib. *Springer Handbook of Robotics*. Gale virtual reference library. Springer, 2008.

[55] W. Sun and J. He. From isotropic to anisotropic side chain representations: Comparison of three models for residue contact estimation. *PLoS ONE*, 6(4):e19238, 04 2011.

[56] V. Tozzini. Coarse-grained models for proteins. *Current Opinion in Structural Biology*, 15(2):144–150, April 2005.

[57] I. Waller. Zur frage der einwirkung der wärmebewegung auf die interferenz von röntgenstrahlen. *Zeitschrift für Physik A Hadrons and Nuclei*, 17:398–408, 1923. 10.1007/BF01328696.

[58] W. Wang, B. Jüttler, D. Zheng, and Y. Liu. Computation of rotation minimizing frames. *ACM Trans. Graph.*, 27(1):2:1–2:18, March 2008.

[59] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171:737–738, April 1953.

[60] W.J. Wedemeyer, E. Welker, and H.A. Scheraga. Proline cis-trans isomerization and protein folding. *Biochemistry*, 41(50):14637–44, 2002.

[61] K. G. Wilson. Renormalization group and critical phenomena. i. renormalization group and the kadanoff scaling picture. *Phys. Rev. B*, 4:3174–3183, Nov 1971.

[62] K. G. Wilson. Renormalization group and critical phenomena. ii. phase-space cell analysis of critical behavior. *Phys. Rev. B*, 4:3184–3205, Nov 1971.

[63] L.V. Yakushevich. *Nonlinear physics of DNA*. Wiley-VCH, 2004.

[64] J. S. Yang, W. W. Chen, J. Skolnick, and E. I. Shakhnovich. All-atom ab initio folding of a diverse set of proteins. *Structure*, 15(1):53–63, January 2007.

[65] J. S. Yang, S. Wallin, and E. I. Shakhnovich. Universality and diversity of folding mechanics for three-helix bundle proteins. *PNAS*, 105(3):895–900, JAN 22 2008.

[66] Z. Yuan, T. L. Bailey, and R. D. Teasdale. Prediction of protein b-factor profiles. *Proteins: Structure, Function, and Bioinformatics*, 58(4):905–912, 2005.

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2012

Distribution: publications.uu.se
urn:nbn:se:uu:diva-172358