Bacterial flagellar filament protein unfolding/refolding mechanisms studied by molecular dynamics simulations

CHNG CHOON-PENG

A thesis presented for the degree of
Doctor of Science

Supervising professor: KITAO AKIO

Department of Computational Biology
Graduate School of Frontier Sciences
The University of Tokyo
Japan

June 2008

Dedicated to

My parents, who missed having me by their side.

Abstract

In the self-assembly process of the flagellar filament, the micrometer-long 'propeller' of the bacterial flagellum, subunit proteins called flagellin are polymerized into the growing filament. Flagellin from *S. typhimurium*, the only available flagellin structure at the time of this study, has two highly conserved helical filament-core domains (D0, D1) and two Hypervariable Region (HVR) domains (D2a/b, D3) rich in β-strands that will be exposed on the filament surface in the assembled form. Flagellins synthesized in the bacterial cytoplasm have to travel through a channel in the center of the filament leading from the cytoplasm to a cavity or 'refolding chamber' under the filament cap. As the 20 Å diameter channel is too narrow for folded flagellin, this implies the coupling of unfolding/refolding processes to the protein transport. The study of these processes forms the focus of this thesis. It is known that cells contain machinery powered by ATP to unfold proteins by mechanical forces. The form flagellin takes during transport should be related to how it is unfolded. To investigate the preferred mechanical unfolding pathway of flagellin, force-probe molecular dynamics simulations have been used. Lower unfolding forces are associated with unraveling flagellin from its adjacently-located termini (producing a fully extended polypeptide chain) as compared to stretching flagellin along its length. After reaching the 'refolding chamber' at the distal end of the channel, flagellin has to be refolded before it can be assembled. Thermal unfolding simulations that probe spontaneous refolding suggest that persistent three-stranded β-sheets in the denatured state of HVR domains might constitute folding initiation sites to guide refolding. Volume estimates indicated that the 'chamber' might accommodate only either denatured HVR domains or filament-core domains at any one time, suggesting a two-step refolding process with HVR domains folding and exiting the 'chamber' first. Insights into this natural nanoscale transport system might form the basis for future bionanotechnology applications.

Acknowledgements

First and foremost, many sincere thanks to my thesis supervisor Prof KITAO Akio who accepted me as a Research Student in April 2005. I have since greatly benefitted from his guidance and criticisms, especially about the craftsmanships of effective scientific writing and presentations. I also appreciate his giving me the freedom to drive the direction of my research (though I've made many wrong turns as a result :)

I wish to thank all members of MolDes lab, past and present, for their companionship. In particular, I thank Dr JOTI Yasumasa (assistant professor) for his management of in-house computational and storage resources. I thank the wonderful secretaries (past and present) for their administrative assistance. I also thank postdoctoral research fellows Dr YANG Lee-Wei from Taiwan for his friendship and many sound advices, and Dr TAKEMURA Kazuhiro for his informal lessons in the Japanese language. Finally, I'll treasure the friendship of Mr NISHIMA "Walter" Wataru, who has served time-and-again as my interpreter and helped me with my daily life's challenges.

The computational requirements for this thesis research were partly provided by supercomputers at the Research Center for Computational Science (Okazaki Research Facilities, National Institutes of Natural Sciences) and the Japan Atomic Energy Agency. I gratefully acknowledge the computer time granted.

My stay in Japan was made possible by a graduate scholarship (April 2005 to September 2008) from the Japan Ministry of Education, Culture, Sports, Science and Technology (MEXT). Thanks also goes to the Japan Student Services Organization for accommodation arrangements when I first arrived in Japan.

Lastly, many thanks to M. Imran at the University of Durham who prepared this wonderful LyX thesis template and made my writing so much easier! And last but not least, thanks to the LyX developer community for this wonderful piece of Open-Source software.

Contents

Chapter 1

Introduction

"All my life through, the new sights of Nature made me rejoice like a child."

– Marie Curie (1867–1934)

1.1 Bacterial flagellum: a self-assembled nanomachine

To find food, a bacterium needs to move about. The more efficient it does so, the better it would survive. Bacterial movement or motility relies on the working of the bacterial flagellum. Depending on the species, some bacteria have several of these growing out of its outer-most membrane whereas others have only one or two (Fig. 1.1). In bacteria with many flagella, rotating them in the same direction causes bundling and swimming in a certain direction. If the nutrient gradient increases in that direction, the bacterium continues swimming. Otherwise, flagella rotation is reversed and the bundle dissolves, resulting in tumbling. The bacterium then starts to swim again in a randomly chosen new

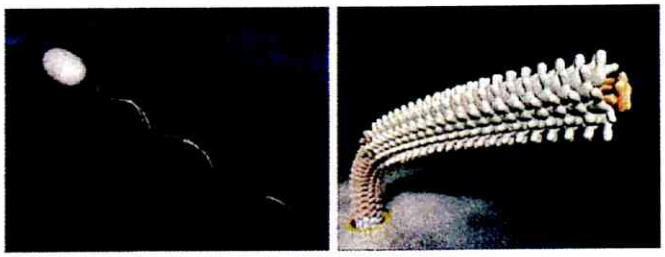


Figure 1.1: **The bacterial flagellum.** (*Left*) *A* swimming bacterium with several flagellar filaments bundled together. (*Right*) The extracellular portion of a flagellum, showing the hook (brown) and filament (beige) with the pentameric filament cap protein (orange). Figure taken from the Protonic Nanomachine website: http://www.fbs.osaka-u.ac.jp/eng/labo/09a.html.

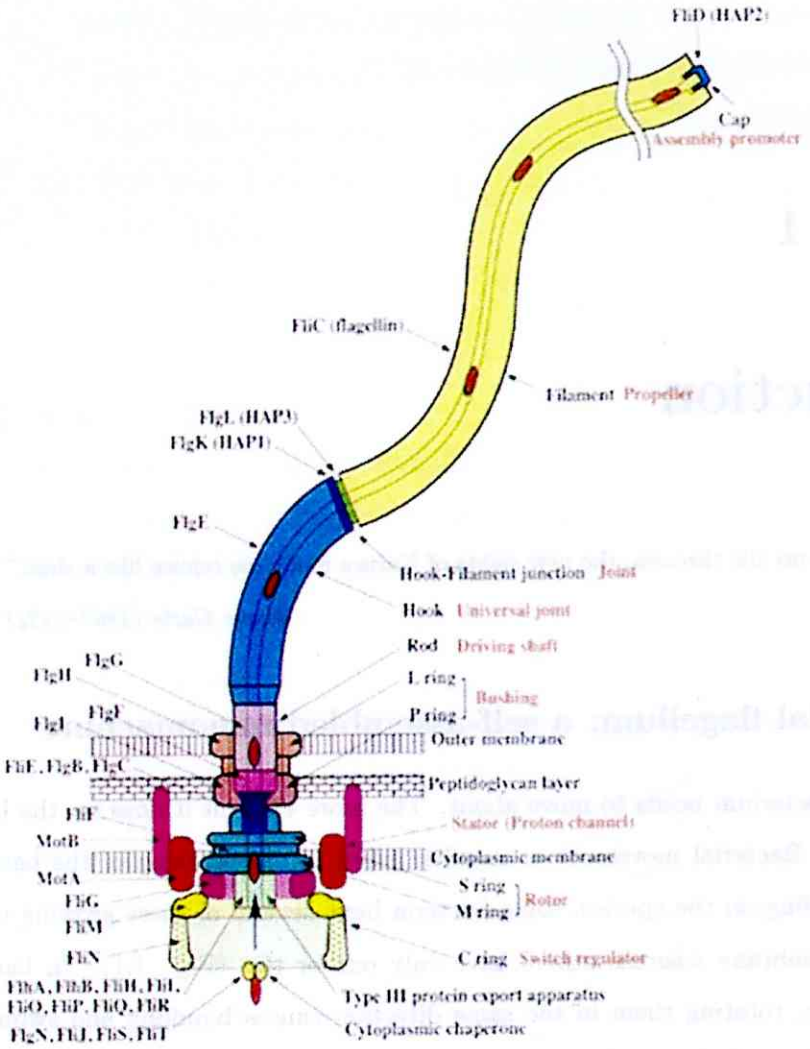


Figure 1.2: ***Salmonella* flagellum components.** Many proteins are involved in the assembly of flagellum parts. Rotation of the rotor is transmitted through the rod and hook segments to the long filament. Figure taken from the Protonic Nanomachine website.

direction in the hope of going in the right direction this time.

Each component of the flagellum is self-assembled from proteins (Fig. 1.2): a basal body that anchors the flagellum to the bacterial membranes, the flagellar rotary motor surrounding the basal body, the micrometer-long tube-like filament outside the bacterium, and the flagellar hook which connects the filament to the basal body and enables several flagellum to point towards the same side of the bacterium to form a bundle for swimming [Macnab, 2003]. The self-assembled nature of bacterial flagellum makes it an interesting object of study for potential bio-nanotechnological applications. How Nature manages to construct this nanomachine without external help is still not fully understood. Its seemingly intricate construction have even led some to argue that the flagellum could not

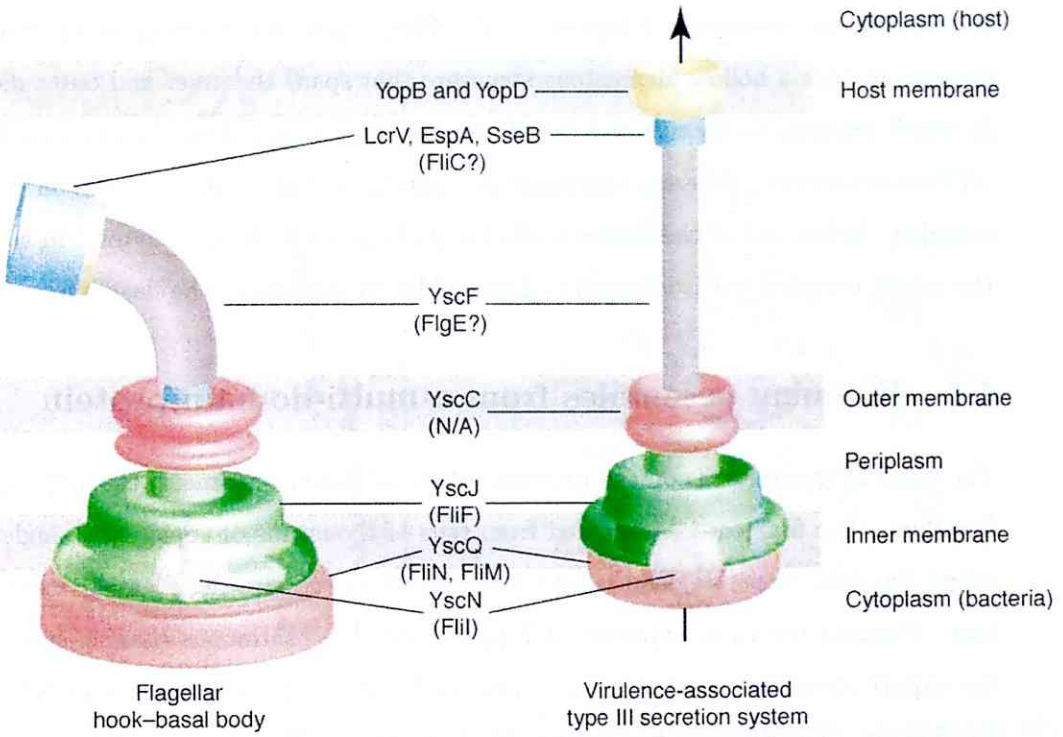


Figure 1.3: **Flagellum vs needle.** The bacterial flagellum (*left*) is morphologically similar to the needle complex (*right*). Homologous proteins are labelled (those from flagellar within brackets). Flagellin (FliC) is homologous to the needle extension (blue). The needle translocation pore (yellow) forms on the eukaryotic host membrane to facilitate virulent protein entry. Figure reprinted from *Trends in Biochemical Sciences*, **31**, Calvin K. Yip and Natalie C.J. Strynadka, "New structural insights into the bacteiral type III secretion system", 223–230, copyright 2006, with permission from Elsevier.

have evolved by natural selection, but is a result of intelligent design (the work of god). Such creationist arguments have been refuted by Pallen and Matzke, who pointed to the sequence and structural similarities between flagellar and non-flagellar proteins [Pallen and Matzke, 2006].

1.2 Type III secretion systems: flagellum and needle complex

Figure 1.3 shows a comparison of the bacterial flagellum with the needle complex. Both bacterial flagellum and the pathogenic needle complex are examples of the type III secretion system (T3SS) [Desvaux et al., 2006] and can both be expressed on a pathogenic bacterium. Whereas flagellum provides motility, the needle complex is used to inject proteins into eukaryotic host for infection. The organization of the needle complex is reviewed in [Moraes et al., 2008]. The flagellar export system and the needle complex share a common ancestor

but evolved independently [Gophna et al., 2003]. Common features of the two export systems include a hollow filamentous structure that spans the inner and outer membrane in which proteins to be secreted travel through the central channel, and that both are ATPase-dependent. Whereas the flagellum contains a motor, this is absent in the needle complex. At the end of the channel, effector proteins enter the host cytosol in the case of the needle complex but is retained and assembled in the case of the flagellum.

1.3 Filament assembles from a multi-domain protein

The focus of this thesis is on the tube-like bacterial flagellar filament, the *propeller* of the flagellum. The filament is assembled from tens of thousands of copies of a single protein called *flagellin*. Fresh flagellin is added to the tip of the growing filament instead of its base. Flagellin has to be exported and move through a continuous channel starting from the export apparatus located in the basal body along the hollow filament towards the growing tip. The filament cap protein Hook-Associated-Protein 2 (HAP2) bound to the tip assists in the polymerization of flagellin monomers into filament. Experiment has shown that addition of HAP2 to a HAP2-gene deficient bacterium prevented flagellin leakage and induced filament growth [Ikeda et al., 1993].

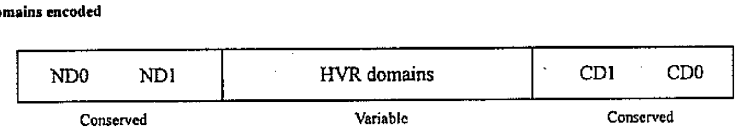


Figure 1.4: **Flagellin domain organization.** Schematic showing domains and subdomains in the polypeptide chain encoded by flagellin gene *fliC*. In the folded protein, ND0 and CD0 segments come together to form the D0 domain. Same for the D1 domain.

Flagellin is a multi-domain protein. The layout of its domains is shown schematically in Fig. 1.4. Its polypeptide chain is folded back such that the N- and C-terminus are next to each other in the tertiary structure. There are two terminal-proximal domains (D0, D1) that are rich in α-helices and which forms the inner and outer tubes of the filament (filament-core domains), respectively (Fig. 1.5 *a*). Due to the structural importance, these are highly conserved across bacterial species. The middle segment of the polypeptide chain contains the so-called Hypervariable Region (HVR) domains. The HVR segment of the flagellin gene *fliC* varies greatly among bacteria species, both in length and amino-acid composition. This could be a way for the filament to adapt to different operating

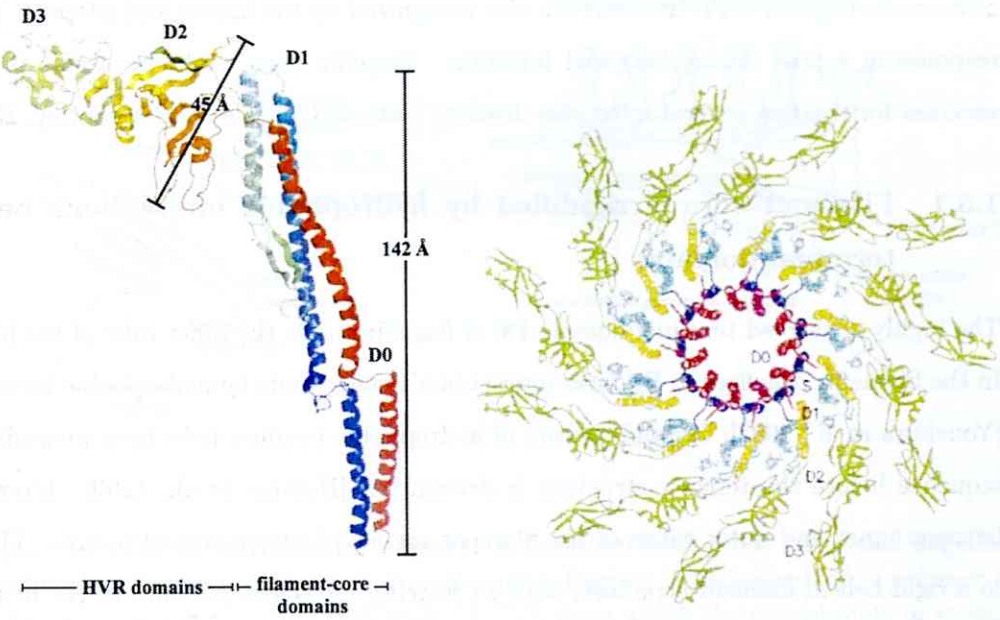


Figure 1.5: **Flagellin topology and assembly.** (*a*) The tertiary structure of the 494-residue *S. typhimurium* flagellin as the flagellar filament subunit, with ND0 in blue and CD0 in red. Cartoon rendered from PDB code 1UCU using molecular visualization software PyMOL [DeLano, 2002]. (*b*) Quaternary structure of flagellin. End-on view of the C_α backbone filament model from the distal end (from the cap, say). 'S' in the figure represent the 'spoke' linker region connecting D0 to D1. Figure reprinted by permission from Macmillan Publishers Ltd: *Nature* (2003, **424**, 643–650), copyright 2003.

environments.

In *Salmonella typhimurium*, a Gram-negative bacteria that multiplies in the human gastrointestinal tract and could cause gastroenteritis but causes a disease resembling typhoid in mice, the HVR segment code for two domains (D2, D3). In contrast, *C. crescentus* flagellin has much reduced HVR domains but motility is not affected [Trachtenberg and DeRosier, 1988]. The only 3D structure of bacterial flagellin available in the Protein Data Bank is from *S. typhimurium*. Figure 1.5 *a* shows a cartoon representation of the flagellin structure as found in the filament. The HVR domains have a high percentage of β-sheets for mechanical stability. HVR domains are exposed on the filament surface and have been mutated or substituted to allow for construction of filament exposing certain protein domains on its surface. By genetically engineering cysteine-rich loops in the HVR domains, nanotube bundles have been created via formation of disulphide-bonds between the cysteine loops [Kumara et al., 2006]. Biologically-inspired nanotubes have many applications in nanotechnology, such as for packaging small drug molecules or enzymes for delivery.

Salmonella flagellin HVR domains are also recognized by the innate and adaptive immune response of a host during bacterial infection. Flagellin thus might be useful as part of vaccines for treating several infectious diseases [Salazar-Gonzalez and McSorley, 2005].

1.3.1 Filament core is rigidified by hydrophobic interactions between terminal domains

The highly conserved terminal domain D0 of flagellin forms the inner tube of the filament. In the filament, α-helices in D0 form inter-subunit coiled-coils by hydrophobic interactions [Yonekura et al., 2003]. Heptad repeats of hydrophobic residues have been identified from sequence before the tertiary structure is determined [Homma et al., 1990]. Interactions between inner and outer tubes of the filament are also hydrophobic in nature. These led to a rigid helical filament structure, with 11 flagellin molecules per turn of the helix (Fig. 1.5 *b*).

1.4 Flagellin transport involves distinct unfolding/refolding phases

The initial aim of this research was to simulate the transport of flagellin through a modeled segment of the filament. I think the transport of proteins could be as important as self-assembly (which might occur on a time scale that is too long to simulate at the atomistic level) for the creation of nano-structures. However, the diameter of the flagellar channel turned out to be only about 20 Å as determined from the atomic model of the filament [Yonekura et al., 2003]. Main-chain and side-chain atoms of residues near the C-terminus (specifically Gln484, Asn488 and Arg494) from each assembled flagellin stick into the channel. Hence, the channel inner surface is hydrophilic in nature. The "sealed-up" nature of the inner tube surrounding the channel suggests that it is unlikely for the channel to expand in response to passing flagellin. This channel is too narrow to accommodate flagellin in its natively folded Γ-shape or even after some partial unfolding and bending about the D1-D2 junction, since the D2 cross-section of around 45 Å is still too large (Fig. 1.5 *a*). Hence, flagellin has to be substantially unfolded before transport. As a result, I switched my research theme to investigate the unfolding and refolding phases that accompany the transport phase.

In Figure 1.6, I show a cartoon for bacterial flagellin export. Flagellin in isolation

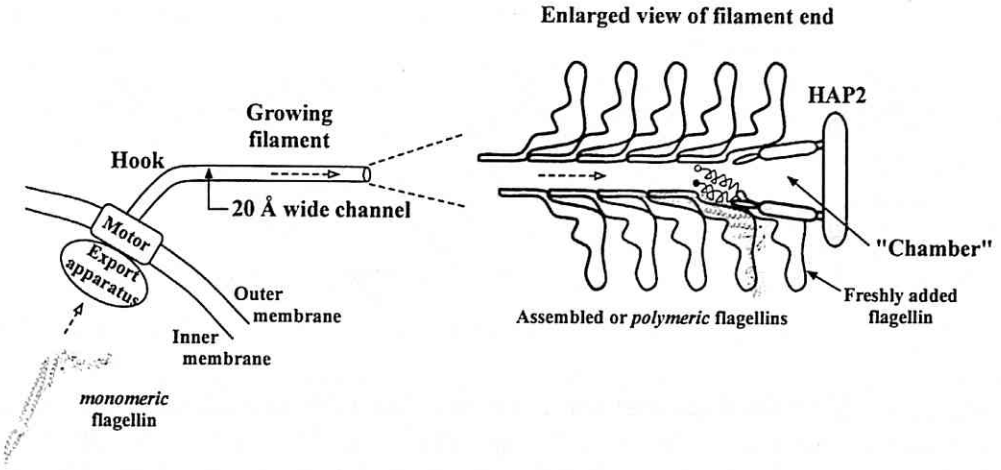


Figure 1.6: **Bacterial flagellin export.** A flagellum with a few assembled flagellin near the tip of the filament is shown in this schematic diagram. The assembled or *polymeric* form of flagellin has been solved (PDB code 1UCU), from which the cytoplasmic or *monomeric* form with disordered termini helices has been obtained from our molecular dynamics simulation in solvent. The conformation of flagellin during transport through the channel is still unknown but is suggested to be highly unfolded due to the narrow channel cross-section. Refolding then takes place in the "chamber" before assembly with the help of HAP2 chaperone. The newly added flagellin shown still has disordered termini helices, with filled circle representing the N-terminus.

(denoted monomeric flagellin; see Chapter 4) is somehow unfolded by the export apparatus, located at the mouth of the continuous flagellar channel inside the basal body (Fig. 1.2). Flagellin then travels through the 20 Å wide channel until it reaches a cavity under the filament cap, formed by the final round of assembled flagellins. There, flagellin has to refold unaided and add into the filament tip by interacting with HAP2 [Yonekura et al., 2000]. But what mechanism could be employed by the export apparatus to unfold flagellin? Put more generally, how do cells unfold proteins?

1.5 Many physiological processes involve unfolding proteins by force

The way proteins are unfolded in the cellular environment is different from spontaneous *in vitro* unfolding by solvent denaturant or high temperatures. Cellular machines such as proteasomes unfold proteins by pulling on the proteins' exposed N- or C-terminus, that is by means of ATP-powered mechanical forces [Prakash and Matouschek, 2004]. Many protein translocation systems such as Sec translocase or the mitchondrial import

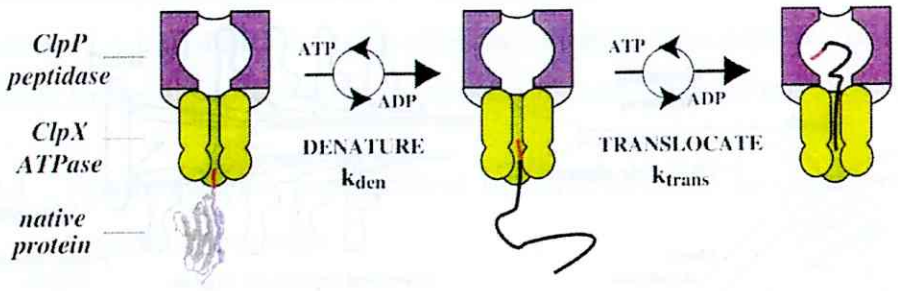


Figure 1.7: **Protein degradation machine.** The ATP-powered mechanical denaturation and translocation of a natively folded protein by the bacterial ClpXP proteasome. The polypeptide terminus segment in red is the degradation tag that is used for recognition by the ATPase. Rate constants for denaturation and translocation phases are denoted by k_{den} and k_{trans} respectively. Figure reprinted from *Cell*, **114**, "Linkage between ATP consumption and mechanical unfolding during the protein processing reactions of an AAA+ degradation machine", 511–520, copyright 2003, with permission from Elsevier.

systems also involve unfolding proteins by mechanical means. Unfolding rate under force is much larger than spontaneous unfolding because such unfoldases catalyze the unfolding "reaction" by changing the unfolding pathway [Lee et al., 2001]. For example, the very stable monomeric β-barrel Green Fluorescent Protein (GFP) that resists denaturation by 6 M urea was rapidly degraded when recruited to the bacterial ClpAP proteasome when tagged at the C-terminus with the ClpAP recognition sequence [Weber-Ban et al., 1999].

Proteasomes are large cylindrical complexes that contains a degradation chamber accessible only through narrow 10–15 Å channels. Hence only unfolded polypeptides can be threaded into the chamber. Access to the channel is controlled by so-called regulatory particle (containing hexameric ATPases) that recognizes, unfolds and translocates proteins tagged for degradation. The ClpXP bacterial proteasome with separate denaturation and translocation cycles is shown in Fig. 1.7. The rate of ATP consumption by the ClpX (same family as ClpA) ATPase during denaturation is four times slower than during translocation of the unfolded polypeptide [Kenniston et al., 2003]. The local stability of the protein near the degradation tag is correlated to the amount of ATP needed for denaturation. However, unfolding of titin (a β-sheet protein found in muscles) variants of different stabilities do not change the rate of ATP hydrolysis. This suggests that the protein could slip or dissociate transiently from ClpXP complex to prevent stalling the molecular machine if unfolding does not occur during a cycle of ATP hydrolysis. A uniform unfolding force was hence repeatedly applied during denaturation, with more ATP molecules consumed

for more stable proteins [Kenniston et al., 2003].

For the flagellar export system, controversy still exists on the true origin of the mechanical forces. There are some suggestions that the ATPase FliI, that shares significant structural similarity to the α and β subunits of the hexameric F_1-ATPase [Imada et al., 2007], might act as an unfoldase. The FliI homolog in the needle complex in pathogenic bacteria, IncV, could induce the unfolding of the secreted protein SptP [Akeda and Galán, 2005]. Hence, FliI might work in a similar fashion as the unfoldase for flagellar export proteins. However, no experimental studies of FliI action has been undertaken to our knowledge. On the other hand, there are also suggestions that a Proton Motive Force (combination of proton concentration and electric gradients) through the transport channel might be sufficient [Minamino and Namba, 2008, Paul et al., 2008] though cross-talk with other unfoldases has yet to be ruled out as pointed by Galán [Galán, 2008]. In view of the above, we would only assume some mechanical force is available to unfold flagellar proteins but do not identify the source in this thesis.

1.6 Aims of this thesis

1. Determine the likely form that the bacterial flagellar filament protein (flagellin) adopts during transport through the filament channel.

2. Determine the mechanism whereby flagellin spontaneously refold in the cavity under the filament cap.

3. Generalize the findings from *Salmonella typhimurium* flagellin to flagellin homologs from other species and other flagellar export proteins.

Learning how proteins can be made mechanically strong yet easy to unfold/refold for long distance transport, independent of the amino acid sequence, might be useful in bionanotechnology. We can thus design proteins that can recover their function after being threaded through narrow channels in an unfolded state.

1.7 Why simulations?

The movement of flagellar proteins inside the filament channel cannot be visualized by spectroscopic means. Computer simulations, in particular those based on Molecular Dynamics (MD), could potentially provide a "window" to the process and a guess at the form

flagellin adopts during transport. Mechanical unfolding simulations were carried out to reveal the unfolding pathways and compare the mechanical effort involved in two models of the transport form we proposed in this thesis. Results obtained from simulations should be compared to single-molecule measurements when they become available. For the folding of flagellin, molecular simulations could also suggest the pathway at atomic-resolution. The complexity of folding such a large protein in the computer is overcome by performing folding *in reverse*, using high temperature MD to mimic the unfolding process. Again, results await validation by future experiments.

Nevertheless, the large and multi-domain nature of flagellin pose challenges to thermal and mechanical unfolding simulations in terms of the high computation costs and in the analysis of simulation data. Certain approximations, such as the lack of an explicit representation of solvent during mechanical unfolding, has to be taken in view of computation resource limitations.

1.8 How this thesis is organized

This thesis is organized into six parts. This chapter provides an introduction and motivation for the research. Chapter 2 presents some bioinformatics-based analysis on HVR domains from flagellin homologs. In chapter 3, I will introduce the simulation methods and relevant theory. Chapter 4 presents a model of the monomeric form of bacterial flagellin based on the polymeric form (the subunit of the flagellar filament). This model is used for mechanical and thermal unfolding simulation studies, presented in Chapters 5 and 6. Conclusions and outlook of this research is the focus of the last chapter, Chapter 7.

Chapter 2

Bioinformatics analysis of flagellin HVR

2.1 Multiple sequence alignment of HVR amino-acid sequences

From the multiple sequence alignment (MSA) of flagellin homologs by Beatson [Beatson et al., 2006], I have extracted out the corresponding HVR segments of the following homologs (number of residues in HVR indicated in brackets): 1UCU_FliC (241), HELFE_FlaA (258), BURCE_FliC (250), AQUPY_FlaA (245), RHOSH_FliC (238).[1]

These sequences are aligned using the multiple-sequence alignment program ClustalW 1.83 (a heuristic progressive alignment algorithm) running from ch.EMBnet.org. Parameters include BLOSUM scoring matrix, gap opening penalty of 10 (and 5) and gap extension penalty of 0.05. A "pretty" output is shown in Fig. 2.1. Columns are framed in blue by ESPript if $> 70\%$ of the residues have similar physico-chemical properties, though these may not be meaningful due to the small number of sequences aligned. Poor alignment is obtained with default parameters, with hydrophobic and polar residues are included in the same aligned column. Even with a reduced gap opening penalty of 5 to reflect the large sequence divergence, no significant improvement results (Fig. 2.1 *b*). Alignment using T-COFFEE gave a more-gapped MSA (not shown) similar to that from ClustalW with the smaller gap penalty. The low sequence conservation indicates that the available HVR domain 3D structures from 1UCU could not serve as good templates for homology

[1]HELFE is short-hand for *Helicobacter felis*, BURCE for *Burkholderia cepacia*, AQUPY for *Aquifex pyrophilus* and RHOSH for *Rhodobacter sphaeroides*.

modeling efforts.

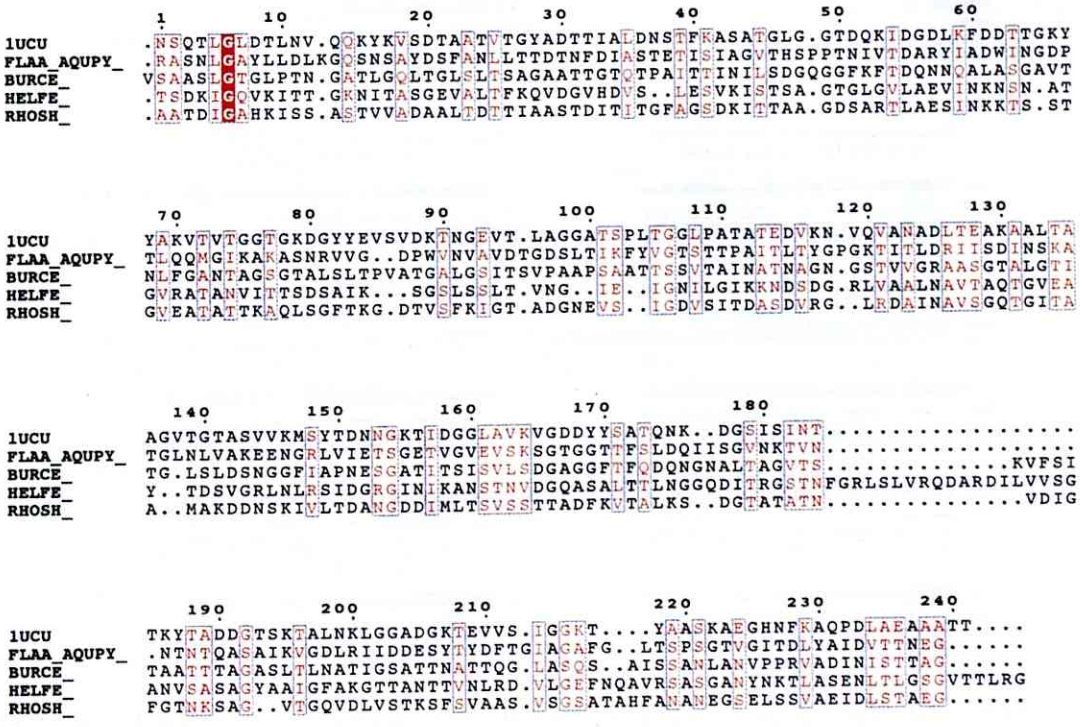
2.2 Secondary structure predictions for flagellin homologs

The secondary structures encoded by the HVR sequences are predicted by PSIPRED v2.6 [Jones, 1999] implemented in the webserver [McGuffin et al., 2000]. PSIPRED combines neural network predictions with a MSA derived from a Position-Specific-Iterated-BLAST database search. Three additional homologs with smaller HVR sequence are included in the predictions: AZOBR_lafl (151), YERPE_LafA (135) and VIBPA_FlaE (121).[2] From Figures 2.2 and 2.3, we can see that a high percentage of β-strands are predicted to be encoded by the HVR sequences (except for YERPE which has more α-helices).

Incidentally, the X-ray structure of a flagellin homolog with very small HVR segment was solved to 2 Å resolution by a group at Kyoto University [Maruyama et al., 2008]. The secondary structures predicted by PSIPRED and Jpred3 [Cole et al., 2008] (turning off PDB search) are shown in Fig. 2.4. Both methods predicted the location of the central α-helix quite well and shows comparable accurracies for β-strands. The good predictions for 1UCU and 2ZBI give confidence that current secondary-structure prediction algorithms have become quite accurate, though not perfect. From the predictions, it is thus likely that most HVR domains (exposed on the filament surface) contain at least one β-sheet which could provide the required mechanical stability.

[2] AZOBR for *Azospirillum brasilense*, YERPE for *Yersinia pestis* and VIBPA for *Vibrio parahaemolyticus*.

(a) Gap opening penalty of 10



(b) Gap opening penalty of 5

Figure 2.1: **Multiple sequence alignments.** ESPript 2.2 [Gouet et al., 2003] formatted multiple sequence alignment of HVR segments in selected flagellin homologs (residues 163 to 403 for 1UCU) by ClustalW.

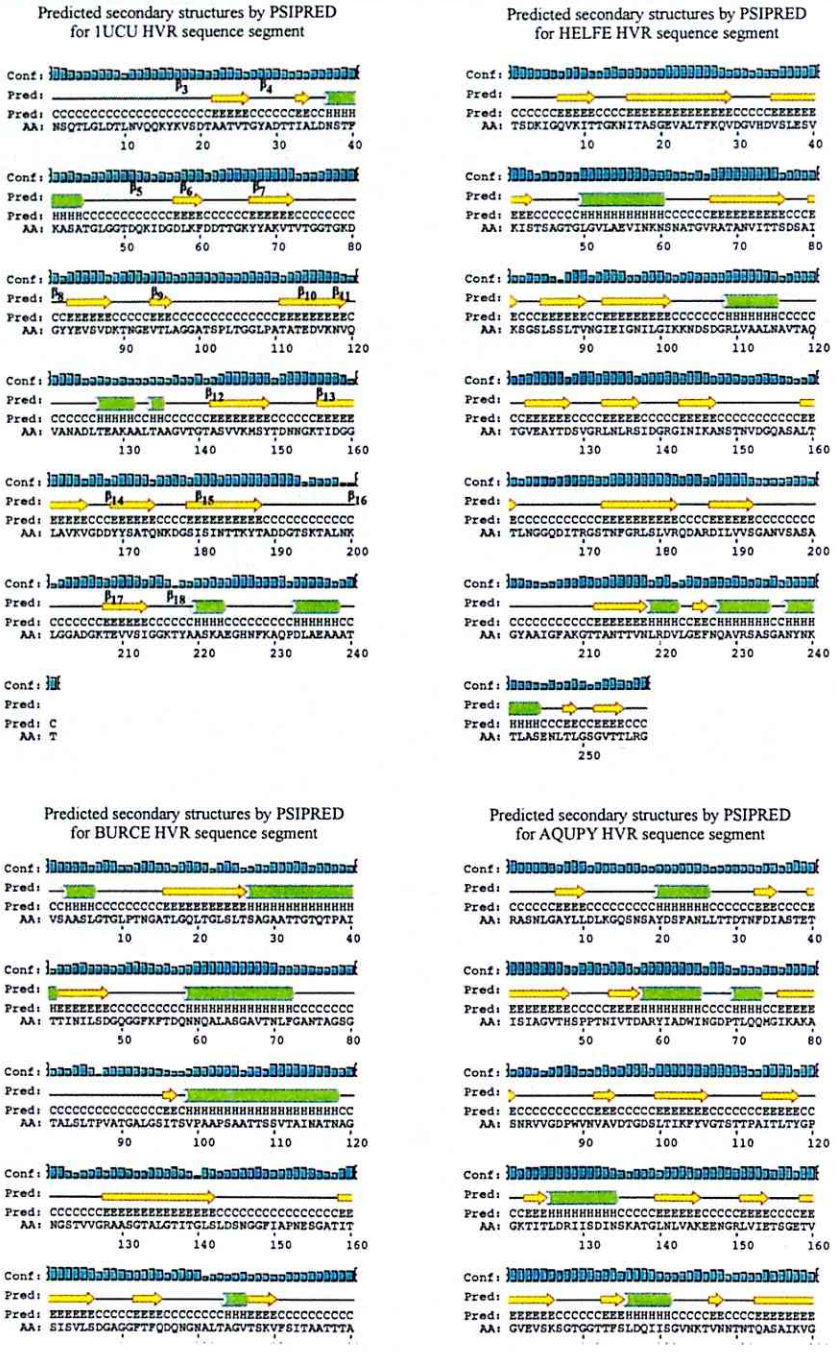


Figure 2.2: **PSIPRED predictions** of secondary structures in selected flagellin HVR homologs. The starting position for β-strands from the solved 1UCU structure are indicated onto the PSIPRED prediction to give an idea of the accuracy.

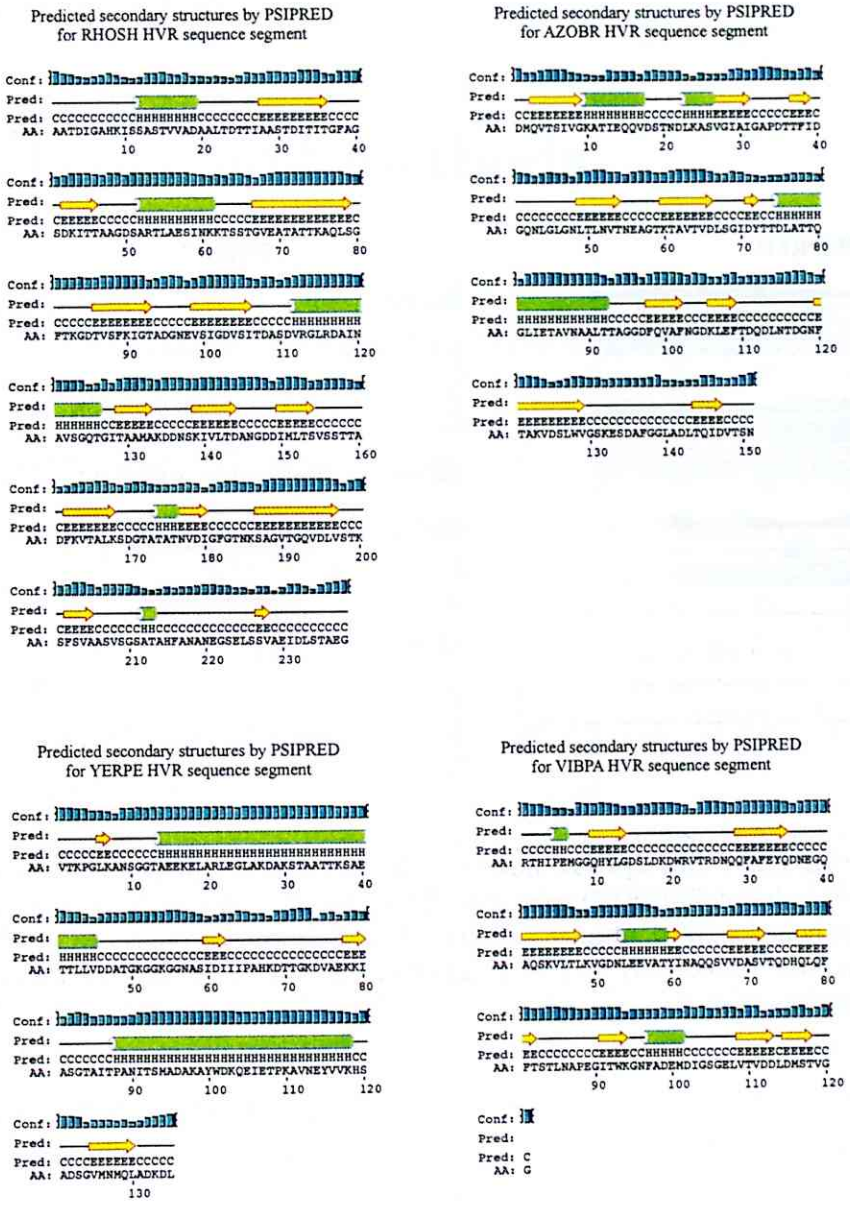


Figure 2.3: Figure 2.2 continued.

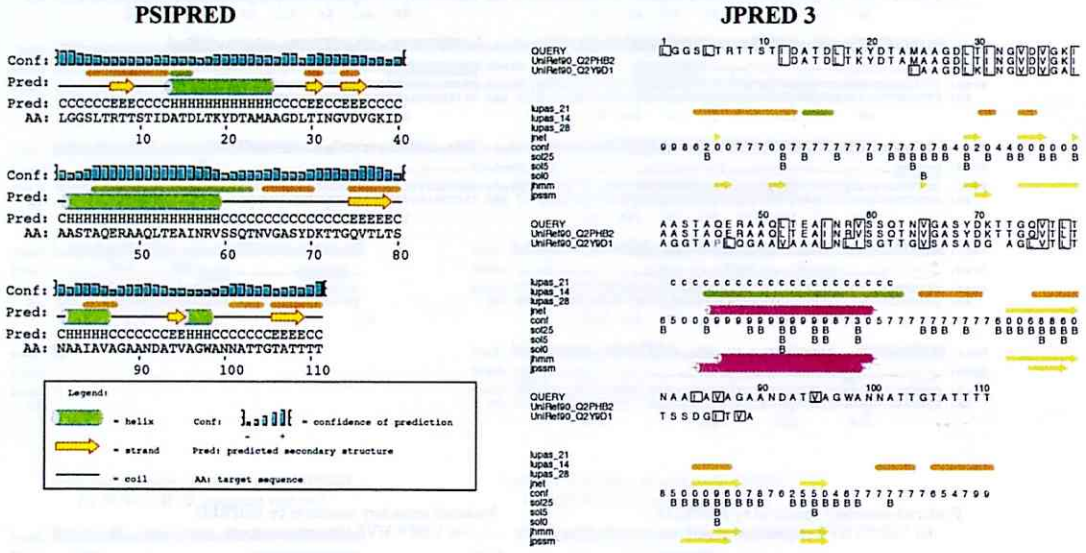


Figure 2.4: **PSIPRED and Jpred3 have similar accuracy.** Comparing sec-structure prediction accuracies of PSIPRED and Jpred3 for the HVR domain from the recently solved flagellin homolog 2ZBI. Orange and green bars indicate β-strands and α-helices in the solved structure respectively and are overlaid to give a sense of prediction accuracies. Jnet predictions are used.

Chapter 3

Theory and methods

3.1 Molecular Dynamics (MD) simulations

The current paradigm of molecular biology is that sequence dictates structure and structure dictates function. However, oftentimes it is the dynamic behavior of a protein that determines its function [Henzler-Wildman and Kern, 2007]. Enzymes, for instance, often have active sites whose access is regulated by large-scale collective motions of portions of the protein. A famous example is that of HIV-1 protease, the enzyme that cleaves nascent polyproteins into functional proteins for viral assembly and replication [Kohl et al., 1988]. It is a homo-dimer with the active site covered by a pair of "flaps" formed by loops that has to open wide enough for the substrate to bind.

Although spectroscopic methods such as Nuclear Magnetic Resonance (NMR) can give us an indication of the conformational flexibility of a protein, simulations allow us to follow the dynamics at the atomic level. Molecular dynamics (MD) simulations, first introduced to study simple liquid behavior by representing the interaction of atoms by a collision of hard spheres [Alder and Wainwright, 1957], have since been extended to the study of biomolecules (proteins, nucleic acids) with more complex interactions and extensively used to complement experimental studies of biomolecular function [Dodson et al., 2008]. Essentially, MD simulations "breathed life" into static biomolecular structures deposited in the Protein Data Bank. MD allows us to observe the dynamics of biomolecules under equilibrium conditions of temperature and pressure, for example.

In this section, I will introduce the theoretical background, solvent models and practical aspects of MD simulation. I will conclude this section by mentioning the successes and limitations of MD simulations. In the following two sections, I will introduce the variants

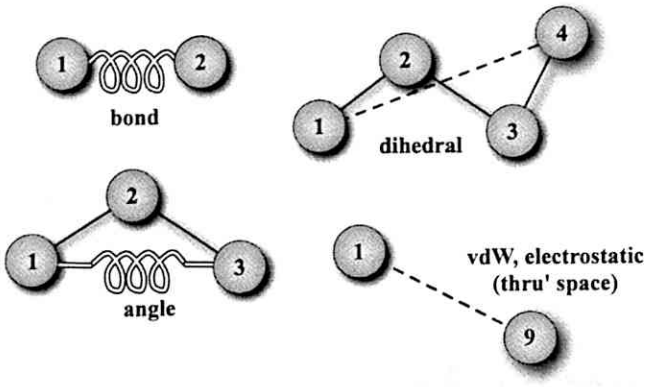


Figure 3.1: **Molecular mechanics force-field.** Schematic representation of *bonded* (bond stretching, angle bending, dihedral angle rotating) and *non-bonded* (vdW and electrostatic) interactions between atoms. Bond (1-2) and angle (1-3) interaction terms are typically represented using spring-like functions. The mathematical expressions are given in the text.

of MD simulations that I have employed in this thesis, namely force-probe and high-temperature MD.

3.1.1 Theoretical background

In a nutshell, a bio-molecular dynamics simulation consist of two components: (i) a definition of how atoms in the biomolecule (and surrounding solvent) interact with each other, and (ii) how the position of the atoms change with time.

Part (i) is known as classical molecular mechanics (MM) and describes the inter-atomic interactions between atoms. More accurately, MM only treats interactions between atomic nuclei and ignores any dynamics of electrons. A *force-field* includes the mathematical expressions for the potential energy terms and the associated adjustable parameters. Refinement methods as part of structure determination by X-ray diffraction or NMR involves use of such a MM force-field. Typical force-fields include *bonded* inter-atomic interactions to maintain the bond between two atoms, the angle between three atoms and dihedral angle between four atoms (angle between the plane formed by atoms {1, 2, 3} and that by atoms {1, 4, 3}). They also inevitably include *non-bonded* interactions such as van der Waals (between induced charges) and electrostatic (between fixed charges) for atoms separated by three or more bonds. Figure 3.1 shows a schematic diagram of these interactions between atoms. The intra-molecular potential energy function in the AMBER [Case et al., 2004] force-field is a summation over the following terms (pairwise additive):

$$
\begin{aligned}
E_{bond}(r) &= \sum_{bonds} k_r \left(r - r_{eq}\right)^2 \\
E_{angle}(\theta) &= \sum_{angles} k_\theta \left(\theta - \theta_{eq}\right)^2 \\
E_{dihedral}(\phi) &= \sum_{dihedrals} \sum_{n=1,2,3} \frac{V_n}{2} \left(1 + cos\left[n\phi - \gamma\right]\right) \\
E_{nb}(R_{ij}) &= \sum_{i<j}^{atoms} \left\{ \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \right) + \left(\frac{q_i q_j}{\epsilon R_{ij}} \right) \right\}
\end{aligned}
$$

The use of harmonic functions to represent bond stretching and angle bending terms is common among MM force-fields. Adjustable parameters are the spring or force-constants k_r, k_θ with corresponding equilibrium bond lengths r_{eq} and angles θ_{eq}. A three-term Fourier series is used for the torsional angle term with parameter γ being the phase of the cosine function: the angle at which the first maximum occurs. The first part of the non-bonded energy E_{nb} is the van der Waals (vdW) interaction represented by the 6-12 Lennard-Jones functional form, consisting of a long-range attraction (power of 6) and a short-range repulsion (power of 12) between atoms i and j separated by distance R. The second part is the Coulomb term describing electrostatic interactions between fixed charges q_i, with ϵ the dielectric constant of the medium. The atomic charges (actually *partial* charges since their absolute values could be less than one electronic charge unit) were assigned by best-fitting the resulting electrostatic potentials to those obtained from quantum mechanics [Leach, 2001]. Although more complex functions can be used to describe each of these energy terms, the above are usually adopted as a compromise between speed and accuracy [Jorgensen and Tirado-Rives, 2005].

The forces between atoms can now be computed by numerically differentiating the potential functions, which allows us to then solve the Newton's equation of motion to get how the positions and velocities of all the atoms change with time. The forces are then updated and the cycle repeats. This is part (ii), the dynamics aspect of MD. The changes are often followed at very small discrete time intervals, typically 1-femtosecond. This captures the fastest motion in the system (bond vibrations involving H-atoms that occur over several femtoseconds) and ensures the stability of the numerical solutions to the equation of motion, meaning that unphysically large energies do not occur. A simulation lasting one nanosecond would thus require a million update cycles. Depending on the

size of the biomolecular system, simulations of tens to hundreds of nanoseconds might be needed in order to observe the equilibrium dynamics.

3.1.2 Representation of solvent

Because biologically processes occur in solution, an accurate representation of the surrounding solvent is required for realistic MD simulations of biomolecular systems. There are two main categories of solvent representation: explicit and implicit.

In the explicit solvent case, a model of the solvent molecule is used. For the rigid water models such as TIP3P (Transferable Intermolecular Potentials, 3-Point), the bonds between oxygen and hydrogen atoms as well as the angle between them is fixed [Jorgensen and Tirado-Rives, 2005]. Each of the water nuclei contains a partial charge. Inter-molecular interaction involves vdW and Coulomb terms which are compatible with those from common MM force-fields. For instance, simulations using the AMBER force-field are usually performed with the TIP3P water model.

In the implicit solvent case, no solvent molecules are represented in the simulation. Instead, their contribution to the system energy (the free energy of transferring the biomolecule from vaccum into the solvent) is estimated and added to the MM energy between atoms of the biomolecule. Because free-energy [1] is a thermodynamic state function, the change is path-independent. We can thus make use of a thermodynamic cycle (Fig. 3.2) to compute ΔG_{solv} as the sum of two parts: the energy cost of removing charges in vacuum and replacing them in solvent ΔG_{elec} and the cost of solvating a neutral molecule $\Delta G_{nonelec}$. The calculation of ΔG_{elec} is the more time-consuming part due to the long-ranged nature of electrostatic interactions. The most accurate yet computationally-intensive way is to solve the Poisson-Boltzmann (PB) equation for the electrostatic potential at the position of the atomic charges $\phi(\mathbf{r}_i)$ and computing the difference in electrostatic potential energy $\sum_i q_i \left[\phi(\mathbf{r}_i) - \phi(\mathbf{r}_i)_{vac}\right]$. A cheaper alternative is by the Generalized Born (GB) approximation which involves estimating the effective born radius of each atom (details in Appendix A). The $\Delta G_{nonelec}$ is often approximated to be proportional to the total solvent accessible surface area (SA) of the molecule (by computing the amount of the spherical surface each atom of the molecule presents to the exterior). Thus, implicit solvent models are often

[1] Usually referring to Gibbs free energy $G = H - TS$ where H is the enthalpy, T the temperature in Kelvins and S the entropy. A chemical reaction is favorable if G becomes lower in the process, since G is lowest for a system reaching thermodynamic equilibrium at constant temperature and pressure.

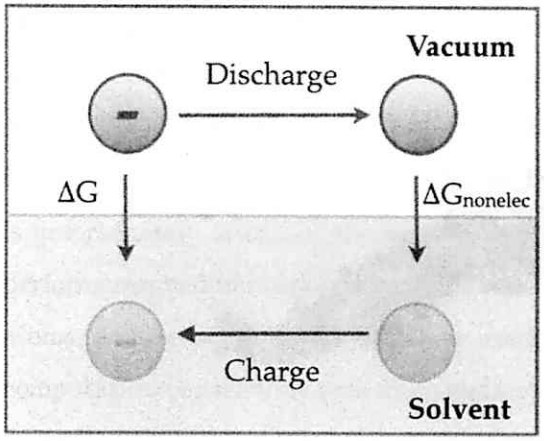


Figure 3.2: **Free energy of solvating a charge.** Thermodynamic cycle showing how the free energy of transferring a charged molecule from vacuum to solvent can be computed. Adapted from a presentation slide by Dr Nathan Baker from the Washington University at St. Louis, USA.

abbreviated as PB/SA or GB/SA. In this thesis, GB/SA is employed for the mechanical unfolding simulations.

3.1.3 Practical aspects

Here I will highlight a few of the practical aspects of MD simulations and point the interested reader to the excellent text on molecular simulations by Andrew Leach [Leach, 2001] for more information.

Bonds involving H-atoms are often constrained to their equilibrium values to remove the fastest motions and allow for a larger simulation time step. Use of SHAKE method [Ryckaert et al., 1977] in the AMBER simulation package [Case et al., 2004] that we have used allows a 2-femtosecond time-step to be used (except for high-temperature simulations where the 1-femtosecond time-step is preferred because of increased motion of atoms; see below).

Another important aspect of molecular simulations is the use of a simulation box. Imagine placing a protein inside a rectangular box of water molecules, such as that shown in Fig 3.3. Water molecules that migrated beyond the edge of the box can be lost or has to be restrained to stay within the box. To remove such edge-effects, the simulation box is often repeated infinitely in space to produce mirror images in each of the three spatial dimensions. Of course, in reality we only simulate one system and atoms which went out from one box edge would reappear on the opposite side, so called periodic boundary

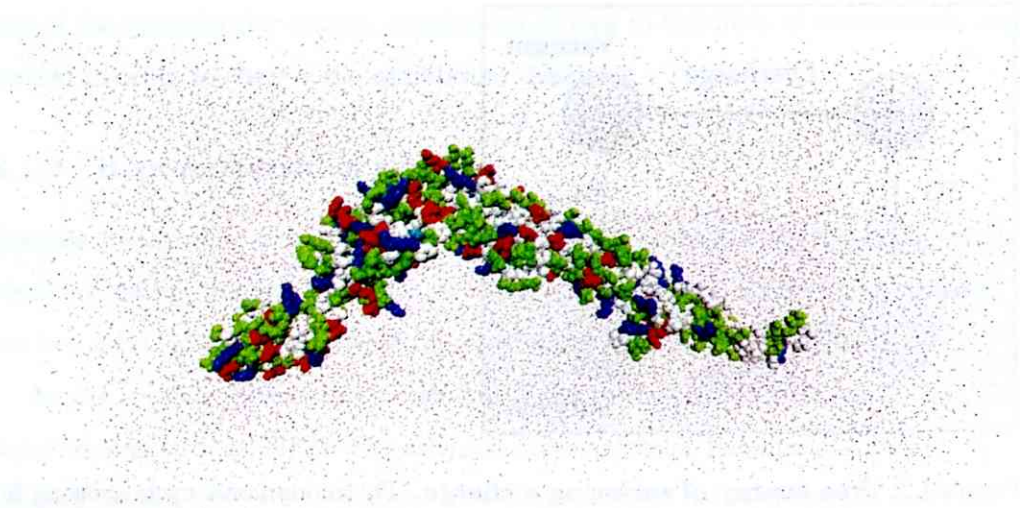


Figure 3.3: **Simulation box for flagellin.** A model of the flagellin monomer in a periodic TIP3P solvent box of dimensions 216 Å × 105 Å × 84 Å. This is the system set up for high-temperature MD simulations, with a minimum of 10 Å between protein atoms and box edges. Flagellin atoms are shown as spheres and colored according to the type of residue they belong: blue for basic, red for acidic, green for polar and white for hydrophobic. Figure rendered by molecular visualization software VMD [Humphrey et al., 1996].

conditions.

The use of periodic images leads to another issue. To save on computation time, non-bonded interaction between atoms that are far apart might be ignored. To do this, we can define a certain cutoff radius centered on each atom. In the context of periodic boxes, the value of the cutoff is often chosen such that atom i will only see a copy of atom j, that is, the cutoff radius has to be smaller than half the largest box dimension. In our explicit solvent simulations using the TIP3P water model, van der Waals interactions are subjected to such a cutoff value but not long-range electrostatic interactions. The Coulomb term was computed using the Particle Mesh Ewald (PME) method [Essmann et al., 1995] which explicitly takes periodicity into account and relies on Fast Fourier Transforms for efficiency. In contrast, both vdW and electrostatic interactions were subjected to cutoff in our implicit solvent simulations, though a large value was used.

To speed up computations, MD simulations are often performed on parallel computers. An example of such machines could be a collection of desktops with multi-core CPUs connected by Gigabit Ethernet or better network technologies. Various MD simulation systems such as AMBER, NAMD, GROMACS and CHARMM have either been designed from the beginning or eventually modified to take advantage of multiple CPUs. The two

main schemes to partition the work are atomic or spatial decomposition, with the latter more suitable for parallel computers. Information such as atomic positions and velocities have to be shared among CPUs through the network fabric connecting them. Although one might think that the more CPUs we use the faster our simulations would run, this is unfortunately often not the case. This is because of the large difference between CPU performance and network performance. Dividing a small system, say of a few thousands of atoms, among too many CPUs might result in more time spent in data transfers than in computation per CPU at each force update. We would do better to use less CPUs for such a small system to increase the computation/communication ratio. Thus, the number of CPUs in a parallel computer that can be effectively used to run a particular MD simulation has to be determined from timing short runs before long (tens of nanosecond) simulations are attempted.

3.1.4 Success and limitations of MD

Whereas experiments can tell us which parts of a protein are moving and how quickly, a key advantage of MD simulations is that they can tell us why the protein is behaving as such. This is because we know the inter-atomic interaction energies and forces. We also have a higher degree of control in simulations than experiment. We can simulate the protein of interest (either as monomer or complex; in solution or embedded in a membrane environment) under various experimental conditions like temperature, pH, salt concentration. More importantly, we can perform virtual site-directed mutagenesis and compare the behavior of various mutant proteins with the wild-type. Nevertheless, the effective use of molecular simulations calls for a closed-loop approach between simulations and experiments: simulations can help to explain experimental findings and provide mechanistic insights; experiments have to be used to validate predictions from simulations [Dodson et al., 2008].

MD uses classical mechanics to describe the motion of atoms in a biomolecule. Full quantum-mechanical treatment of biomolecules is still prohititively expensive to carry out. As a result, processes such as chemical bond formation/dissociation which involves a quantum-mechanical treatment of electrons cannot be studied with classical MD. Nevertheless, hybrid Quantum-Mechanics/Molecular-Mechanics (QM/MM) schemes have been developed to simulate the catalytic reactions of enzymes: the dynamics of the active site are handled at the QM-level whereas those of the rest of the enzyme are handled at the

MM-level. Though some success has been shown with such approaches, how to handle accurately the QM and MM interface is still an active area of research.

MD force-fields are empirical because they relied on parameters determined from higher accuracy but more computationally demanding, first-principles QM calculations or determined from best-fitting calculated structural or thermodynamic quantities of organic liquids or peptides to their experimental values [Jorgensen and Tirado-Rives, 2005]. The force-fields used in MD simulations are still under continual improvement.

Computing power available to a typical researcher would only allow he/she to simulate the atomistic dynamics of a moderate-sized biomolecular system in solvent (~100,000 atoms) for tens of nanoseconds. This is a widely recognized limitation because functionally relevant conformational changes often occurs in the millisecond time-frame. Coarse-Grained MD, an version of MD which uses a simplified representation of the biomolecular system (using just protein backbone atoms, for instance) with a correspondingly simplified force-field, has shown success in reaching tens of microseconds to study ligand-binding [Trylska et al., 2007] and protein-lipid self-assembly processes [Bond and Sansom, 2006,Shih et al., 2007]. To learn more about CG-MD methods, please see [Chng and Yang, 2008] for a review that I co-authored with a postdoc in our lab.

3.2 AFM and force-probe MD studies protein mechanics

3.2.1 Single-molecule force spectroscopy by AFM

Several high-resolution single-molecule techniques have been developed to track and manipulate the biomolecular motion, reviewed in [Greenleaf et al., 2007]. For the purpose of this thesis, only Atomic Force Microscopy (AFM) is introduced here.

AFM was developed as a high resolution imaging tool for the study of material surfaces before it was turned to the manipulation of biomolecules by using it in "force mode" [Parot et al., 2007]: the AFM cantilever tip does not scan across a surface but is repetitively pushed and retracted from it. A typical AFM setup is shown in Fig. 3.4 *a*. A drop of protein suspension in physiological salt buffer is deposited on a flat surface (cleaned cover glass or mica) mounted on a movable platform. The tip of the flexible cantilever (taken to be a spring with a known spring-constant) is then lowered onto the surface to allow for binding of protein molecules to cantilever tip atoms. By fixing one end of the protein chain to the surface, the chain can be stretched when we increase the separation between cantilever and

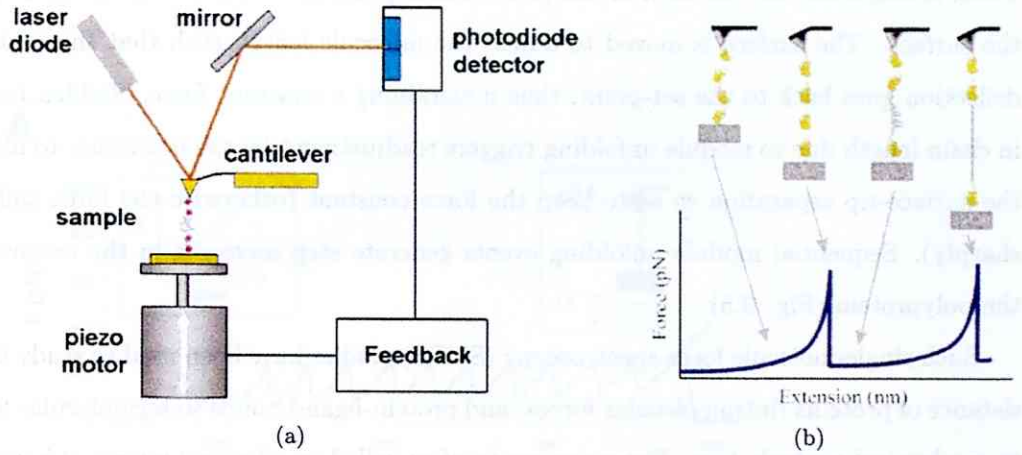


Figure 3.4: **Atomic force spectroscopy.** (*a*) Schematic of AFM setup for force spectroscopy. With kind permission from Springer Science+Business Media: Plfugers Arch. - *Eur. J. Physiol.*, "Pulling single molecules of titin by AFM – recent advances and physiological implications", **456**, 2008, 101–115, W. A. Linke and A. Grützner (*b*) Cartoon showing the sequential unfolding of domains upon stretching a polyprotein chain in a velocity-clamp AFM experiment. The restraint force produced by the protein chain increases as the chain became taut and drops sharply when one of the domains unfolded. The cycle repeats for the unfolding of another domain. Figure taken from [Oberhauser and Carrión-Vázquez, 2008] and is copyright of the American Society for Biochemistry and Molecular Biology, Inc.

surface with Angstrom precision via the piezoelectric motor. Forces in the range of a few tens to hundreds of piconewtons generated by chain resistance to extension can be measured from the cantilever deflection from its equilibrium position via a laser beam [Linke and Grützner, 2008]. Genetically engineered polyproteins (the expression of multiple copies of the same protein in a single DNA sequence) are often used in such experiments rather than single molecules. This is because the unfolding of identical modules in the polyprotein produces a series of sawtooth force peaks that are equally spaced (Fig. 3.4 *b*) which could serve as a signal that the cantilever tip has successfully attached and pulled on a protein molecule.

AFM can be operated in two modes: velocity-clamp and force-clamp. In velocity-clamp mode, the surface is moved at a constant speed away from the cantilever. The resulting degree of cantilever deflection is then measured to obtain a force-extension curve as in Fig. 3.4 *b*. On the other hand, the force applied to the polyprotein is constant under force-clamp mode, giving the extension-time curve instead. How the force-clamp AFM was developed by the Fernandez group and works is as follows [Oberhauser et al., 2001]: any deviation of the cantilever from a computer-controlled set point results in an error

which is amplified and fed back to the piezoelectric positioner which controls the height of the surface. The surface is moved to adjust the molecule length such that the cantilever deflection goes back to the set-point, thus maintaining a constant force. Sudden increase in chain length due to module unfolding triggers readjustment by the positioner to increase the surface-tip separation so as to keep the force constant (otherwise the force will drop sharply). Sequential module unfolding events generate step increases in the extension of the polyprotein (Fig. 3.5).

Such single-molecule force-spectroscopy (SMFS) studies have been used to study the resistance of proteins (intramolecular forces) and protein-ligand bonds (intermolecular forces) to mechanical perturbation. Extensive study of so-called 'mechanical proteins' have been made with SMFS, such as the giant muscle multi-domain protein titin as reviewed in [Linke and Grützner, 2008] which contains Immunoglobulin(Ig)-like and Fibronectin-III domains. Domains in these proteins have evolved resistance to repeated cycles of stretching and force release for their physiological function. Lastly, SMFS techniques (AFM and Optical Tweezers) have enabled studies of protein unfolding and refolding by altering the protein folding free-energy landscape by force [Samorì et al., 2005].

3.2.2 Force-probe MD mimics AFM *in silico*

The correspondence of force-probe (FP) MD with AFM experiments is illustrated with one example. The first FP-MD simulation was set up to study the origin of the streptavidin-biotin binding force that has been measured by AFM experiments [Grubmüller et al., 1996]. In AFM experiment (Fig. 3.6 *A*) biotin molecules (the ligand) are fixed to the cantilever tip via linker molecules and also to the agarose bead. Free tetrameric streptavidin molecules (a protein receptor with extraordinary affinity for biotin) bind to most of the biotin molecules attached to the bead. Streptavidin tetramers were also bound to biotin molecules attached to the cantilever tip. Upon contact of the tip with the bead, a few streptavidin-biotin complexes are formed between streptavidin on the tip and remaining biotins on the bead. As the cantilever is retracted, biotin molecules are pulled out of the binding pockets one at a time until eventually one ligand-receptor pair remains for some time, whose binding force is measured from the cantilever deflection. The force-probe (FP) MD mimic is shown in Fig. 3.6 *B*, where only a streptavidin monomer and one biotin molecule was included in the model. The biotin was pulled out of the binding pocket with a harmonic potential (probe "spring") acting on the same biotin atom (O2) to which the linker used in the

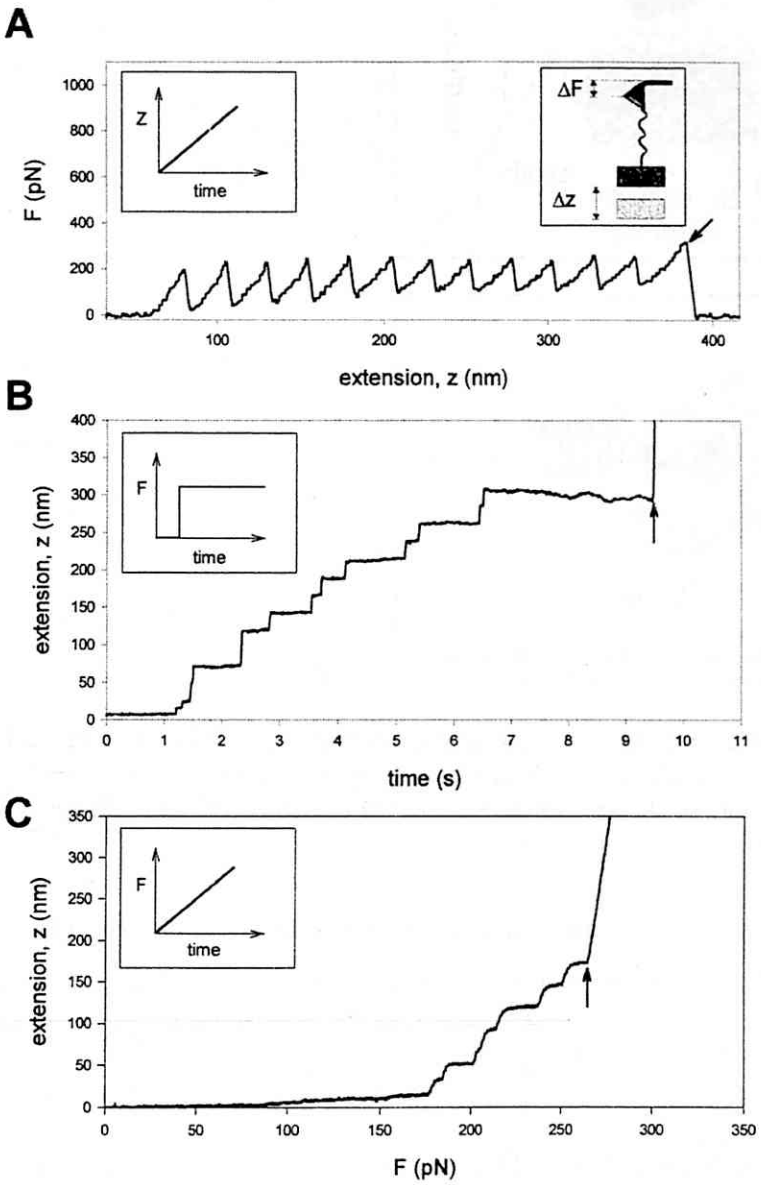


Figure 3.5: **AFM measurements.** Comparison of single protein unfolding events captured with AFM in length-clamp (*A*) and force-clamp (*B* and *C*) modes. In *A*, the surface on which a molecule with 12 repeats of the titin Ig-like domain I27 is moved at constant speed. The last peak corresponds to the detachment of the protein from the cantilever. The force applied is maintained at a constant value in *B* whereas it is increased linearly with time in *C*. Figure taken from [Oberhauser et al., 2001]. Copyright of The National Academy of Sciences of the United States of America.

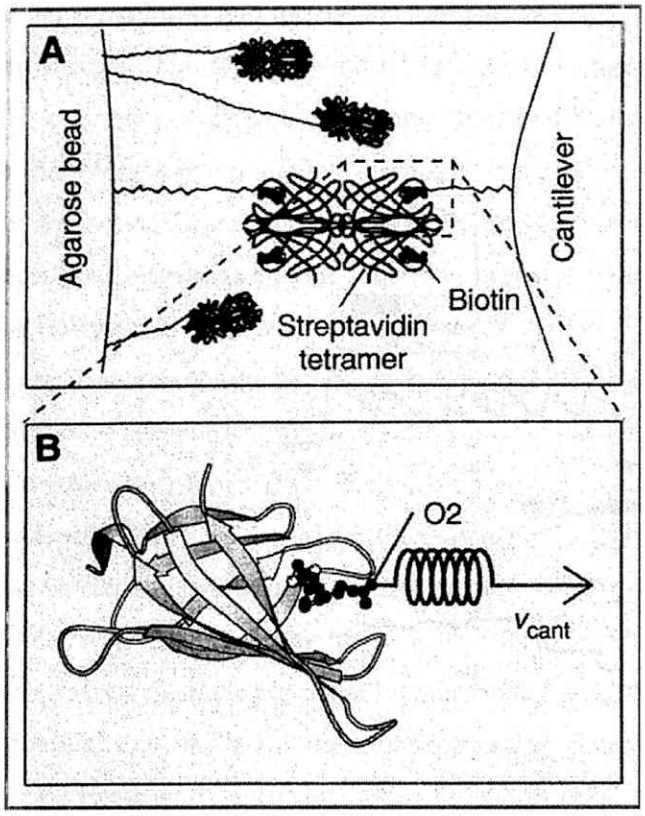


Figure 3.6: **Simulating the AFM.** Comparison between the AFM (A) and simulation (B) setups for measuring single streptavidin-biotin complex rupture forces. From H. Grubmuller, B. Heymann and P. Tavan, 1996, *Science* **271** (5251) 997-999. Reprinted with permission from AAAS.

AFM experiment is attached, with the streptavidin monomer fixed in space. The external harmonic potential E_{spring} acts only on the z-coordinate of the atom O2, z_{O2}:

$$E_{spring} = \frac{k}{2}\left[z_{O2}(t) - z_{cant}(t)\right]^2$$

where k is the spring-constant (set to 2.8 N/m or 4 $kcal/mol/\text{Å}^2$ for a "soft" spring) and $z_{cant}(t) = z_{cant}(0) + v_{cant}t$ is the displacement of the cantilever's z-coordinate at constant velocity v_{cant}. At the start of the simulation, the spring was in a relaxed state by setting $z_{cant}(0) = z_{O2}$. The position of the atom O2 was monitored as it was acted on by the moving harmonic potential and restraint force computed via $k|z(t) - z_{cant}(t)|$ as a function of z_{cant}. The resultant force-extension profile showed a series of peaks corresponding to forces required to break H-bonds formed between biotin and streptavidin binding-site residues, including those bridged by water molecules which enhanced the stability of the complex [Grubmüller et al., 1996]. Hence, FP-MD simulations could tell us the origin of

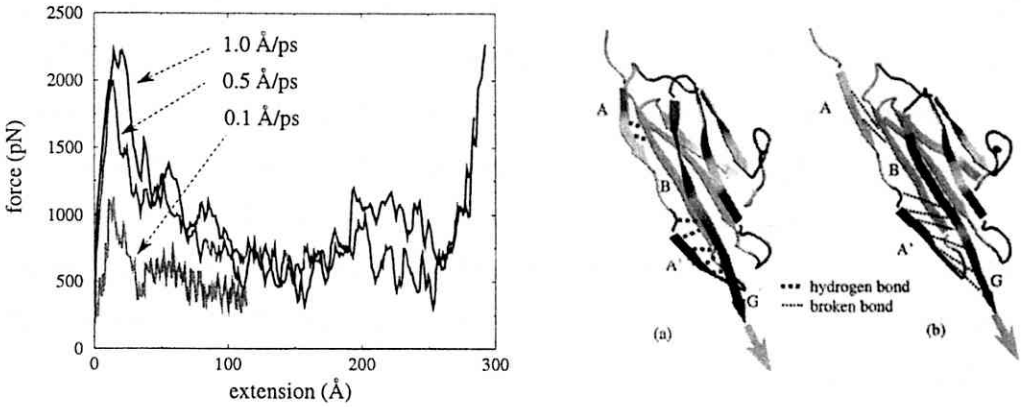


Figure 3.7: **Unfolding of titin I27.** (*Left*) Force exension curves of I27 stretching under various pulling speeds. (*Right*) Snapshots during I27 unfolding simulation: (*a*) is at 10 Å extension before the major force peak and (*b*) is afterwards (at 17 Å extension). Both figures are reprinted from *Chemical Physics*, **247**, H. Lu and K. Schulten, "Steered molecular dynamics simulation of conformational changes of immunoglobulin domain I27 interprete atomic force microscopy observations", 141–153, copyright 1999, with permission from Elsevier.

the measured binding force between ligand and binding pocket.

Force-probe (FP) MD has also been used to mimic AFM for studying protein mechanics in the computer. Selected positions on the biomolecule can be pulled apart via spring-like forces (see below) at a constant velocity or maintained under a constant force. Also known as Steered-MD (SMD), the simulation technique has produced force-extension profiles for the unfolding of Titin domains in agreement with AFM experiments despite the use of pulling speeds a millionfold larger [Lu et al., 1998]. An illustration of the constant-velocity SMD unfolding of titin I27 domain is given in Fig. 3.7, with the computed restraint forces encountered at different pulling speeds. Force-probe or Steered MD has been a great complement to AFM experiments by interpreting the origins of the force-peaks observed and revealing the unfolding pathway(s) at the atomic level [Ohta et al., 2004, Ng et al., 2005]. We now know that the force peak appearing in the force-extension curve (Fig. 3.7 left) is due to the need to break several hydrogen-bonds (H-bonds) at the same time in order to slide β-strands against each other (Fig. 3.7 right).

3.2.3 Peak forces from FP-MD are 10× larger than AFM

Events in AFM experiments occur over milliseconds whereas those in simulations have to do so within nanoseconds due to limitation in computational resources. Hence, the pulling speeds used in FP-MD simulations are at least a million times larger [Lu et al., 1998]. The

faster pulling resulted in force peaks which are 10× larger in simulations: unfolding of I27 in experiments required ~200-pN (Biomolecule Stretching Database: http://info.ifpan.edu.pl/BSDB) wherease simulations required ~2000-pN (Fig. 3.7 right). A study which performed both techniques on Ubiquitin along the same pulling directions also observed peak forces of ~2000-pN in simulations and ~200-pN for the average peak force from multiple AFM stretching of poly-Ubiquitin chains [Carrion-Vazquez et al., 2003]. Hence, often we can only get a force-extension profile that can be compared qualitatively but not quantitatively to those from AFM.

The simulated pulling process is not at thermodynamic equilibrium (due to presence of the external force and the related work performed on the simulated system) though use of slower pulling speeds would allow an approach to a quasi-equilibrium state [Pabón and Amzel, 2006]. Various ways to discount the irreversible work done on the system (which led to a measured temperature increase [Lu et al., 1998]) in order to compute the free energy barriers involved in the process investigated are discussed in [Israelewitz et al., 2001]. In this thesis, such estimations were not carried out because a large number of simulations are required which is not feasible for a large protein such as flagellin given our limited computing resources.

3.3 High-temperature MD: protein folding in reverse

3.3.1 Mechanisms of protein folding

The protein folding problem is really a set of three problems: (i) which inter-atomic interactions are key to the folding process; (ii) how can we predict the native tertiary structure of a protein from the amino-acid sequence and (iii) how can a protein fold so quickly [Dill et al., 2007]. The third question is related to the so-called Levinthal's paradox: a protein in the denatured state can adopt an astronomical number of possible conformations yet it can quickly find the one which is the native state within seconds or less time.

The three 'classic' folding models are listed below:

- The earliest model is the 'framework' model, which start with the formation of secondary structural elements according to the sequence in the absence of any tertiary structure. The local structural elements would collide and assembly into the tertiary structure via 'diffusion-collision' processes. Fig. 3.8 (*b*) depicts graphically the folding process under this model using simplified free-energy profiles.

- The classical 'nucleation' model proposes that neighboring residues in sequence form native secondary structures (β-turn or single turn of α-helices) which then act as a nucleus or seed to attract the formation of the rest of the protein structure. Thus, tertiary structure formation depends on secondary structure formation as in the model above. The simplified free-energy profile is shown in Fig. 3.8 (a).

- The 'hydrophobic collapse' model, in contrast, suggests that secondary structure forms after a rapid collapse from an extended to a compact state via long-range tertiary contacts. Secondary structures then form as rearrangements of the collapsed state. Folding intermediates are involved in the 'framework' and 'hydrophobic collapse' models but not the 'nucleation' model.

- Finally, 'nucleation-condensation' model unifies the more extreme framework and hydrophobic collapse models: the large extended folding nucleus is only weakly local and stabilized by longer-ranged tertiary contacts. Formation of nucleus and rest of the protein is coupled, making folding more efficient [Fersht, 1997]. Nucleus residues are those making the strongest interactions in the high energy folding transition state, as determined in an experiment which established that the two-state folder chymotrypsin inhibitor 2 folds via the nucleation-condensation mechanism [Itzhaki et al., 1995]. It has also been suggested that larger proteins might fold via the merging or docking of smaller folding units that individually fold via nucleation-condensation [Fersht, 1997].

Hence, a possible way to overcome Levinthal's paradox is that weak native-like interactions (folding nuclei) may remain in the denatured state which is often not fully structureless like a random-coil under physiological conditions [Daggett, 2006]. Studies have shown that such residual native-like conformations are present in regions found to fold early [Fersht, 1997].

3.3.2 Studying protein folding by simulation

Although experimental techniques such as Φ-value analysis, which can characterize which residues are in the folding transition state through introducing mutations in the protein and measuring changes to the thermodynamics, detailed information on the structural states encountered from the native to the denatured process can only be obtained from

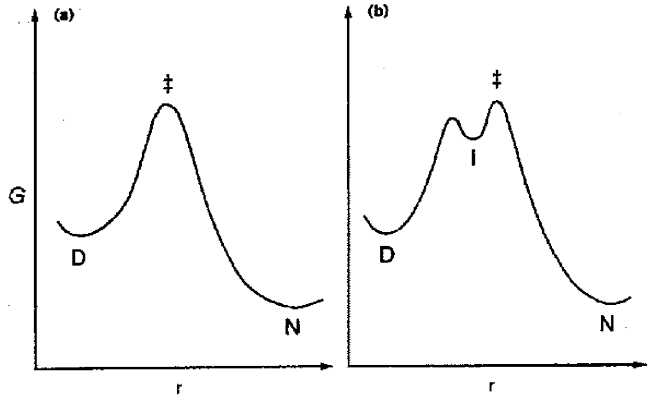


Figure 3.8: **Folding free energies.** Simplified free-energy profiles for folding from a denatured (D) state to a native (N) state. G stands for the Gibbs free energy and r is some "reaction coordinate". In (a), folding proceeds by the 'nucleation-condensation' mechanism with concerted formation of secondary and tertiary structures. In (b), Formation of secondary and tertiary structures are stepwise as in the 'framework' model: an unstable intermediate state I containing secondary structures is formed before the transition state. Reprinted from *Current Opinion in Structural Biology*, vol 7, A. R Fersht, "Nucleation mechanisms in protein folding", pages 3–9, copyright 1997, with permission from Elsevier.

atomistic simulations such as MD. Together, these techniques complement each other to reveal a more complete picture of the protein folding process [Schaeffer et al., 2008].

The fastest folding proteins have a measured folding time of a few microseconds [Kubelka et al., 2004], which implies that simulation times have to reach a microsecond or longer. Notable successes are the folding of the 36-residue ultra-fast folder villin using explicit solvent MD starting from an extended conformation carried out by Duan and Kollman in 1998 [Duan and Kollman, 1998]. A collapsed state at 4.5 Å root-mean-squared-deviation (RMSD) from the native NMR structure was achieved, after 1 microsecond of simulation which is still very long by today's standards. Direct simulations of folding from extended states often get trapped in local free-energy wells, unable to reach the true native state with the lowest free energy. Recent progress made in studying protein folding and the related structure prediction problem has been extensively reviewed in [Dill et al., 2007, Chen et al., 2008].

Studying protein folding by simulating the unfolding process can thus be viewed as a convenient alternative, with the assumption of reversibility in mind (see later). Use of high temperatures helps to accelerate the unfolding process by overcoming energy barriers to unfolding. HT-MD simulations, when used in combination with experiments, affords a method to more fully describe the unfolding pathway [Oroguchi et al., 2005, Akanuma and Yamagishi, 2005, Scott et al., 2006]. Relative stabilities of domains during unfolding have

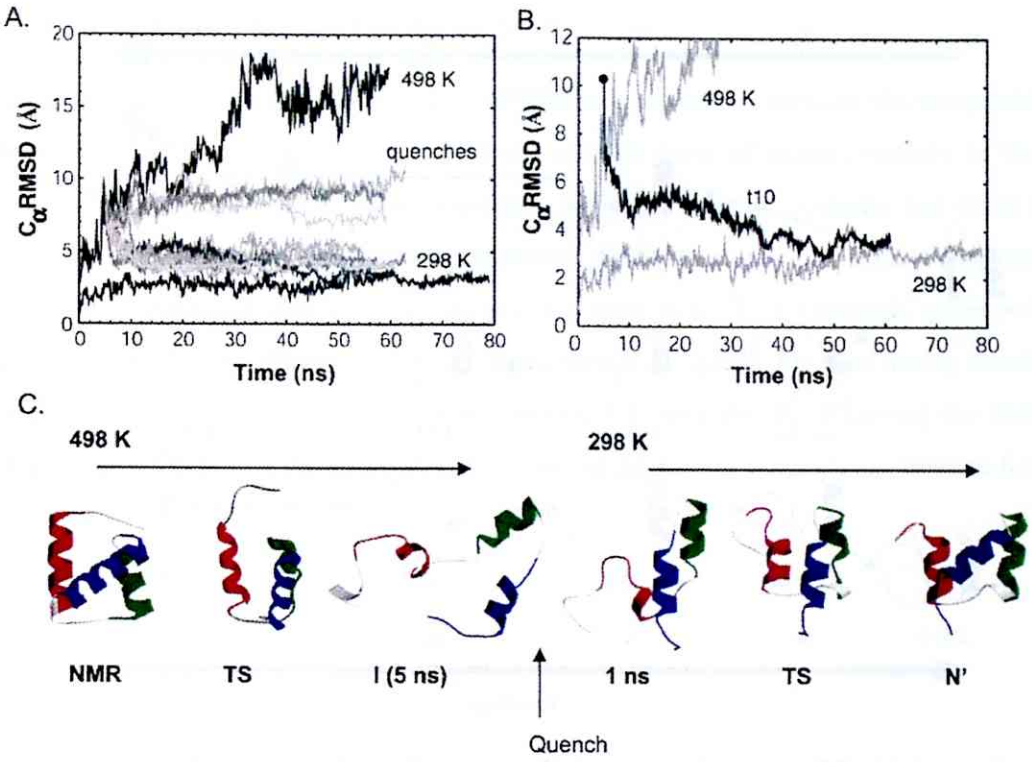


Figure 3.9: **Heating and cooling of a protein.** Unfolding and refolding of the engrailed homeodomain [Beck and Daggett, 2004]. (A) Temperature quenched simulations of the protein from 498 to 298 K show that the protein is approaching the native state in some simulations. (B) Blow-up of the y-axis in panel A for one particular target simulation, t10. (C) The thermal denaturation pathway and structures after the thermal quench of t10 show the refolding and docking of the helices, as well as the similarity between the TS ensembles for unfolding and refolding. Helices are colored differently. Figure and legend reprinted with permission from Daggett, V. *Chem. Rev.* **106**:1898-1916. Copyright 2006 American Chemical Society.

also been studied by such simulations [Sham et al., 2002].

The use of elevated temperatures in simulations was shown not to grossly affect the unfolding pathway, as thermal denaturation can be viewed as an activated process where lower energy barriers are overcome first. The overall order of events are conserved across temperatures but their timescales do differ [Day et al., 2002]. Use of lower temperatures merely meant longer simulation times are needed to reach the transition and denatured states [Daggett, 2006]. Thus, thermal unfolding at high temperatures affords a way to study the stability and kinetics involved in the folding process within reasonable computational effort despite possible bias/distortion to the pathway. Such bias can be reduced by running multiple simulations at a series of temperatures and taking an "ensemble" view of the process [Day and Daggett, 2005] as I have also done in this thesis.

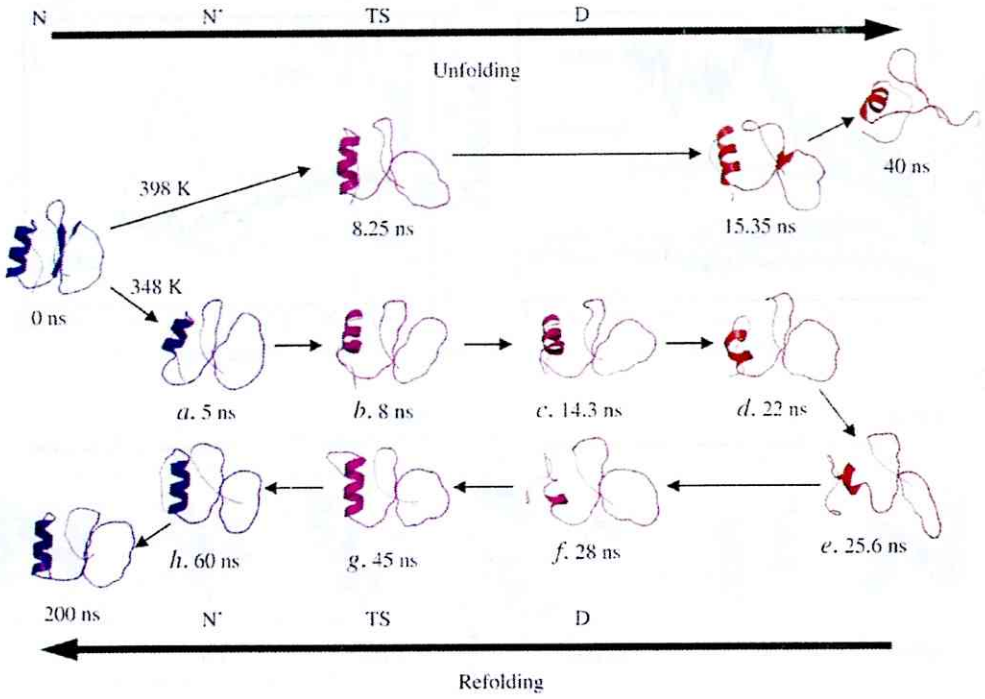


Figure 3.10: **Microscopic reversibility in action.** Representative structures from the simulation conducted by Day and Daggett [Day and Daggett, 2007]. Structures are colored from blue to red according to their degree of nativeness. Structures in the same column are judged to be similar by C_α RMSD and the position along the 'reaction coordinate' (fraction of native contacts? not defined in the paper). Figure reprinted from *Journal of Molecular Biology*, **366**, Ryan Day and Valerie Daggett, "Direct observation of microscopic reversibility in single-molecule protein folding", 677-86, copyright 2007, with permission from Elsevier.

Is unfolding truly the reverse of folding? One way to test this is to perform 'temperature-quench' (T-quench) simulations from unfolding protein conformations obtained from HT-MD. In Fig. 3.9, several of the 12 T-quench/refolding simulations started from a folding intermediate state managed to reach a native-like state after about 50-ns of MD simulation. This result suggests that refolding at the room temperature of 298 K is the reverse of unfolding at the artificially high temperature of 498 K [Beck and Daggett, 2004]. Further, unfolding/refolding has been observed in a single, long MD trajectory of chymotrypsin inhibitor 2 at a temperature close to its melting temperature where folded and unfolded states are equally populated. The refolded protein at 200-ns is again not identical to the crystal native state but can be interpreted as the native state (N') at the slightly elevated temperature (Fig. 3.10). The authors suggested that folding/unfolding from a structural nucleus do obey the principle of microscopic reversibility to a large extent, when performing simulations close to the melting temperature [Day and Daggett, 2007].

3.3.3 Control of temperature in MD

The kinetic energy of atoms is related to the system temperature through the equipartition theorem: $\frac{1}{2}m\langle v_x^2\rangle = \frac{1}{2}k_BT$, where $\langle v_x^2\rangle$ is the average (over all atoms) velocity in the x-direction and k_B is the Boltzmann constant. A similar relation holds for the other two coordinate directions. An easy way to maintain the temperature of a simulation system is thus by rescaling of atomic velocities at every time step. This approach, proposed by Berendsen and known as "weak coupling" [Berendsen et al., 1984], was used during the short simulation to obtain monomeric flagellin (Section 4.2) and also for achieving the desired temperatures during the initial phase of the thermal denaturing simulations (Section 6.2.1).

Chapter 4

Model of flagellin monomer

4.1 Terminal domain D0 is partially structured in monomeric flagellin

Limited proteolysis of monomeric flagellin showed the existence of a central resistant portion with disordered terminal region [Kostyukova et al., 1988]. The terminal helices are only marginally stable as determined from far-ultraviolet circular dichroism spectra [Vonderviszt et al., 1989] and NMR measurements [Ishima et al., 1991], but became structured during filament assembly [Aizawa et al., 1990, Tamura et al., 1997]. The polymerization of monomeric flagellin into filaments in solution made it difficult to obtain crystals of monomeric flagellin for X-ray structure determination. Hence, termini-truncated flagellin structure was first obtained by X-ray (PDB code 1IO1) [Samatey et al., 2001] and subsequently used as a guide to reconstruct the complete *polymeric* flagellin structure (PDB code 1UCU) from cryo-electron microscopy electron density map of the filament [Yonekura et al., 2003]. Inter-subunit coiled-coils form between terminal α-helices in neighboring polymeric flagellin to produce a continuous and mechanically stable filament [Yonekura et al., 2003].

Flagellin terminal regions have also been identified as disordered by the Database of Protein Disorder [Sickmeier et al., 2007]. The disorder is a dynamical one, rather than due to a lack of secondary structures. The terminal domains of the filament cap protein HAP2 are similarly unstructured in the monomeric form of the protein but became structured (forming α-helical coiled-coils) upon binding to the filament end [Maki-Yonekura et al., 2003]. Thus, several of the flagellar proteins are *intrinsically disordered* to enable their efficient assembly.

4.2 Obtaining monomeric from polymeric flagellin

To obtain the a model of *monomeric* flagellin, I performed molecular dynamics simulation in solution starting with the *polymeric* conformation. I first aligned the molecule with its longest extent along the Z-axis and immersed it in a periodic rectangular box of TIP3P waters with at least 120 Å between the protein atoms and box edge along the Z-direction and at least 8 Å in the X- and Y-directions. Such a large simulation box was meant for mechanical unfolding in solution. The biomolecular simulation software AMBER 8 [Case et al., 2004] with ff99 force-field was used for all the simulations. Electrostatics was handled with the Particle Mesh Ewald method [Essmann et al., 1995] with a non-bonded real space cutoff of 8 Å.

After energy minimization, the charge neutralized system was heated to 300 K while keeping restraints on the non-hydrogen atoms. Next, Berendsen temperature and pressure control was imposed (at 300 K and 1 atm) with restraints reduced in stages and finally turned off when equilibration has been reached. The system density approached the bulk solvent value after we activated SHAKE [Ryckaert et al., 1977] to constrain motions of chemical bonds involving hydrogen. Simulation continued until the RMSD of the new conformations from the starting conformation reached steady values for domains D1, D2 and D3 (after 1.4-ns simulation), wherease the value for D0 kept increasing which is a sign of unstructured-ness. This conformation was used to initiate mechanical unfolding simulations in explicit and implicit solvent (see Chapter 5). It has also been re-solvated in a wider simulation box and simulated for 8-ns without temperature or pressure controls (see Chapter 6). This extended 8-ns simulation served as the control for the high-temperature unfolding simulations.

4.3 Structural comparison of polymeric and monomeric flagellin

The C-terminal helix of D1 (CD1) was observed to be more straight in the polymeric flagellin in filament (1UCU) as compared to the terminal-truncated X-ray structure of monomeric flagellin (1IO1) [Yonekura et al., 2003]. Fig. 4.1 (*b*) shows the result of performing a RMS fit of terminal-truncated polymeric and monomeric flagellins using residues in the shorter ND1 helix. The RMS deviation between terminal-truncated 1UCU structure and 1IO1 structure is ~0.02 Å whereas that between 1IO1 and terminal-truncated

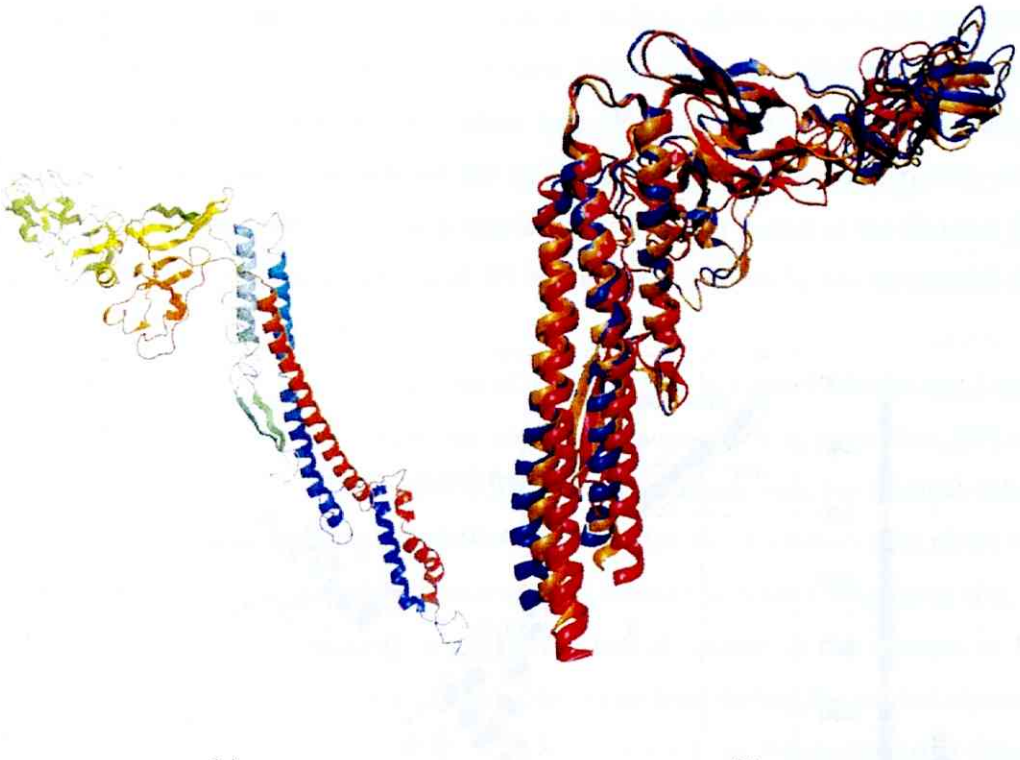


Figure 4.1: **Monomeric flagellin obtained from MD simulation.** (*a*) PyMOL rendered cartoon of the 1.4-ns equilibrium MD snapshot, to be compared to the polymeric form from 1UCU in Fig. 1.5. (*b*) Comparison of the terminal-truncated structure of monomeric flagellin obtained by simulation (red) starting from the cryo-EM polymeric structure (blue) with that from X-ray (orange), after least-squares-fitting onto the ND1 helix (residues 105 to 126). Figure created using VMD.

monomer from simulation is ~1.1 Å. Not only is the CD1 even more curved in the simulated monomer compared to the polymeric form as observed by Namba and co-workers [Yonekura et al., 2003], the longer ND1 helix also deviated from both experimental structures signifying increased structural disorder in the terminal region. Domains D2 and D3 in the simulated conformation also showed significant deviations compared to the experimental conformations, implying greater flexibility (due to inter-domain motions) exhibited by the simulated structure in solution.

4.4 Definition of persistent native contacts

The contact map of 1UCU is shown in Fig. 4.2. A symbol appears at grid location (i, j) if the residue pair $\{i,j\}$ is in *contact*, i.e. they have a minimum heavy atom separation of

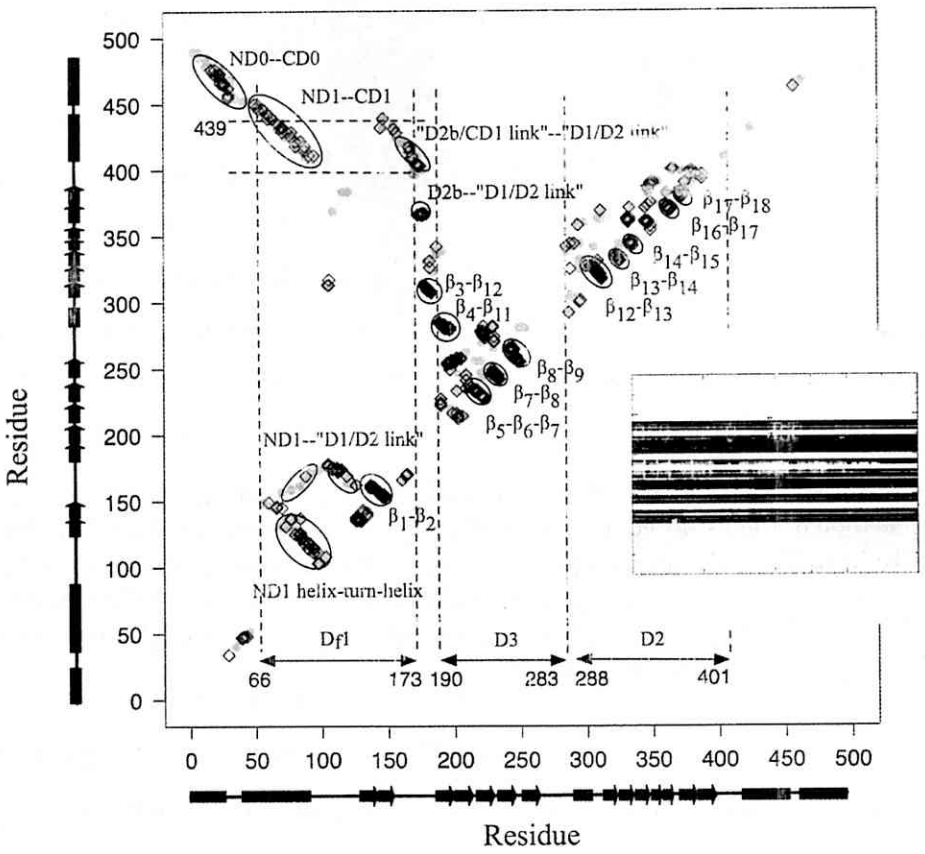


Figure 4.2: **Residue contact map of native flagellin structure.** Gray filled-circles indicate contacts in 1UCU. Black diamonds indicate persistent native contacts: contacts found in more than 70% of the snapshots taken from the last 1 ns of the extended 8-ns simulation of the monomer. Dotted lines mark location of D_f1 contact clusters (see Text).

less than 4.2 Å as determined by a Perl script in the MMTSB Toolkit [Feig et al., 2004]. The proteolytic resistant portion of D1, denoted as D_f1, which includes not just residues from the N-terminal side as originally defined [Yonekura et al., 2000] but also from the C-terminal side, as marked on the contact map (Fig. 4.2). D_f1 contains an elongated hydrophobic core that could account for its proteolytic resistance. The rigidity of D_f1 hydrophobic core has been noted in a simulation of a 44-mer model of the filament [Kitao et al., 2006]. The remaining portion of D1 was indeed found to be less structured during my simulations (see Chapter 6).

Using snapshots from the final 1-ns of the 8-ns 300 K control simulation, I defined persistent native contacts as contacting residues that appeared in more than 70% of the snapshots. The fraction of such contacts will be used for monitoring the thermal unfolding process. I have overlapped these persistent clusters on the 1UCU contact map where we can see a loss of intra- and inter-helical contacts in D0 during the control simulation (Fig. 4.2). This observation is also reflected as a loss of α-helical content in the changes to DSSP [Kabsch and Sander, 1983] assigned secondary structures during the control simulation, shown as an inset in the contact map. Contact clusters from β-stranded pairs dominate the contact map, though those between D0 and D1 helices and between unstructured domain linkers are also present.

4.5 H-bond network in D1-D2a interface

A gap exists between the $\beta_{12}\beta_{13}$ hairpin in D2a and α_2-turn-α_3 in N-terminal of D1 (ND1) (Fig. 6.1). This *interface* is deviod of solvent molecules in the cryo-EM based structure as the atomic resolution is too low. In my simulation, solvent molecules could not penetrate deep enough into this D2a-ND1 interfacial space for most part of the 300 K 8-ns control simulation, resulting in strong bridging hydrogen bonds between side-chains of D313 (in D2a) and S106/S104 (in ND1). Some water molecules finally got close enough to exchange residue-residue H-bonds with residue-solvent ones only after 7.5-ns of the 8-ns control simulation. Such solvent penetration occurred faster under higher temperatures (results not shown). In the higher resolution X-ray structure (PDB code 1IO1), interfacial solvent molecules (rather, just the oxygen positions) are present in the PDB file. Hence, I think any H-bonds across the D2a-ND1 interface appearing during my simulations should be much weaker in reality.

An alternative interpretation is that in the filament, the interface might be more closely-

packed. H-bonds were found to bridge this interface in the 44-mer model of the filament (unpublished results). This network might act to strengthen the polymeric flagellin in the filament.

4.6 Inter-domain motions of monomeric flagellin

4.6.1 Normal modes from Elastic Network Model

In the (coarse-grained) elastic network representation of a molecule, the backbone C_α atoms of each residue is a node in the network and the nodes are connected to each other via "springs" or harmonic potentials. Each node thus can only exhibit harmonic vibrations about its equilibrium position. Only neighboring atoms within a certain cutoff distance of each atom are considered in the interaction network. In physical jargon, this is a collection of coupled harmonic oscillators. The so-called Force constant matrix is constructed based on the inter-atomic separations and then diagonalized to obtain the eigenvalues (squared values of the normal mode frequencies) and eigenvectors (defining a new coordinate system for collective motions). These collective motions involving movement of one portion of the protein relative to another could have functional significance, such as in controlling access to binding sites. For a comprehensive review on ENM models, see [Yang and Chng, 2008]. Because of low computational cost, ENMs offer a quick estimation of the (though harmonic) dynamics about some equilibrium structure of a biomolecule. To characterize non-harmonic motions which may involve larger conformational transitions, long molecular dynamics simulations coupled with Principal Component Analysis would be necessary [Kitao and Go, 1999].

Here I have used the webserver oANM [Eyal et al., 2006], which implements a class of ENM called Anisotropic Network Model, to characterize the harmonic motions of flagellin molecule. Basically, "anisotropic" here means there is no simplifying assumption about atomic fluctuations in x, y, and z-directions. In constructing the atomic network, atom pairs separated by larger than the default cutoff distance of 15 Å are not included. Figure 4.3 shows the first normal mode from terminal-truncated structures (presence of D0 produced "tail-wagging" motions which dominated the normal modes). The first mode describes a "hinge-like" motion about the D1-D2 domain interface. Regions on the flagellin cartoon are color-coded according to the amount of fluctuation. Arrows indicate the size and direction of motion. The residue correlation maps (symmetric about the diagonal) indicate which

regions of the protein moves collectively: correlated regions in red and anti-correlated ones in blue. The smaller (in the center of the map) and larger red squares include residues in domains D3 and D2 respectively. These two regions are anti-correlated (dots representing inter-domain residue pairs are colored blue). D0 is weakly correlated with D3 but strongly anti-correlated with D2. All three structures show highly similar motions.

For the second normal mode, it seems to involve a sort of "hinge" motion about the D2-D3 junction this time, with a "twist" in D3 (Fig. 4.4). The same collective motions are also found from polymeric and monomeric flagellin conformations. But for the third mode, while 1UCU and 1IO1 shows a twisting motion of D3 about to the longest molecular axis, the simulated monomer (Sim) shows a motion which is more collective and involves "rocking" of D3 coupled to a "contraction" of D1 towards D2: essentially motions about both "hinges". Differences are also clear from the correlation maps[1]. The modes visualized from the 8-ns snapshot of the control simulation as well as the corresponding correlation map are similar to that from Sim (result not shown). Nevertheless, because the contributions of the normal modes to the overall collective motions are decreases from the first to the third, we can say that all three structures exhibit rather similar harmonic dynamics. This also serves as a check on the solution MD equilibration procedure, in that it had not significantly distorted the experimental structure which is located in a local harmonic energy minimum. The more flexible third normal mode from solvent simulations could simply reflect greater conformational freedom for the monomer in solution as compared to the crystal environment.

[1]I have also checked that the fourth mode from Sim also differs from the third mode from 1UCU or 1IO1, so no exchange of modes has occurred between third and fourth modes of Sim.

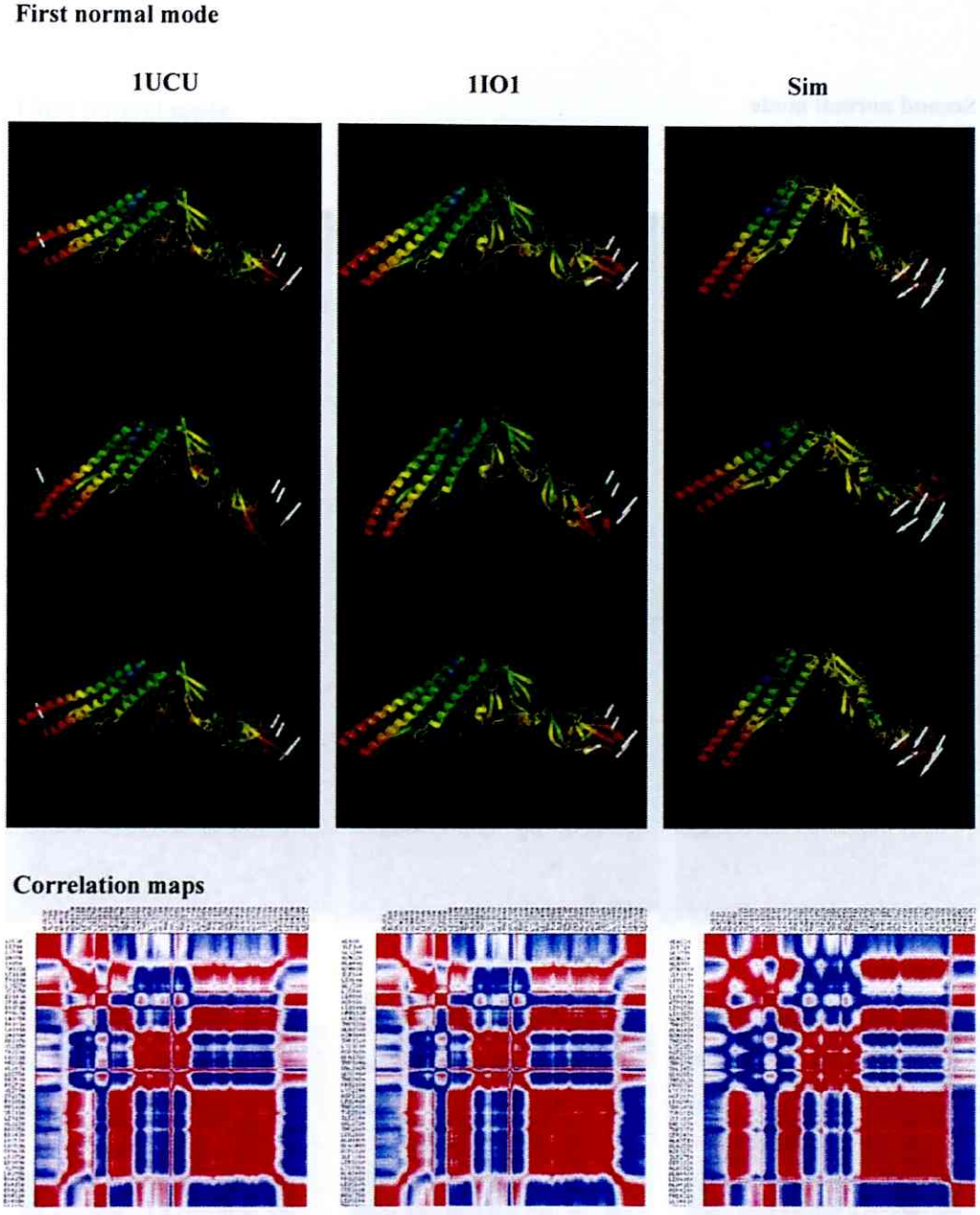


Figure 4.3: **Normal modes of flagellin.** The first normal modes from polymeric and monomeric flagellin structures, computed by oANM. Top panel shows a visualization of the collective motions from each structure. Bottom panel shows the residue correlation maps. "Sim" refers to the monomer conformation after 1.4-ns of equilibration under MD. Note that though the "open-close" sequence is reversed in Sim, similar collective motion is depicted.

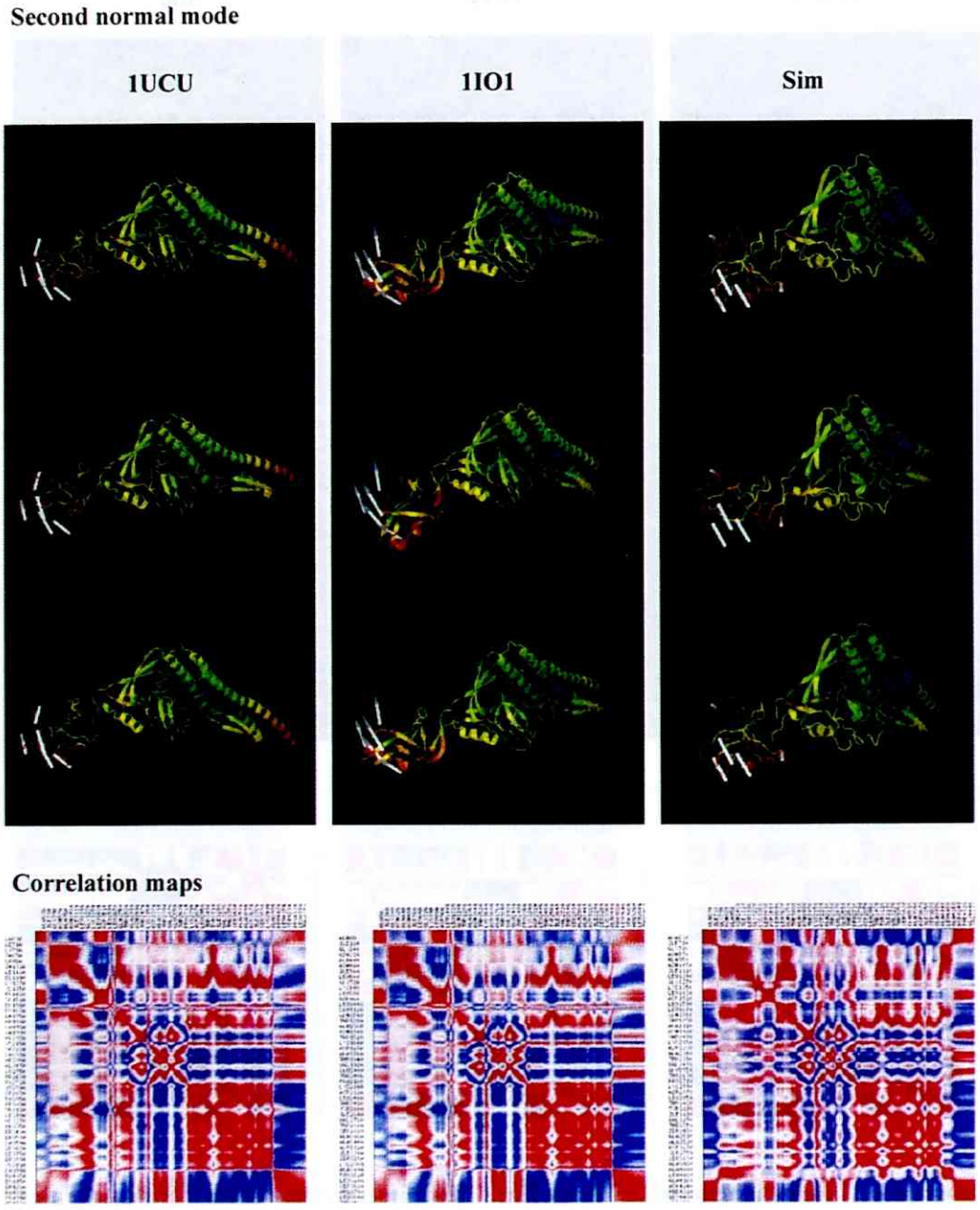


Figure 4.4: Similar to Fig. 4.3 but showing the second normal modes.

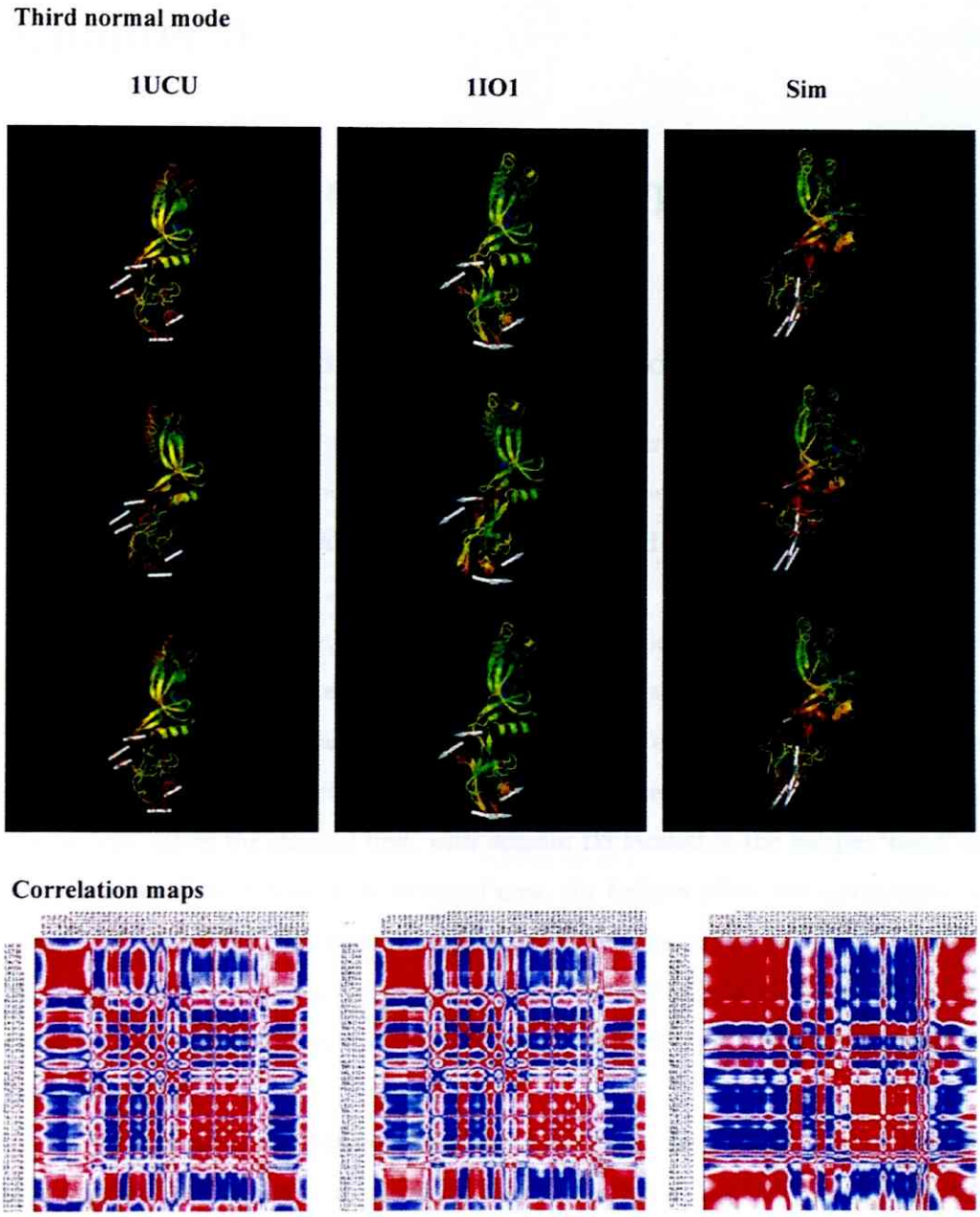


Figure 4.5: Similar to Fig. 4.3 but showing the third normal modes.

Chapter 5

Flagellin mechanical-unfolding

5.1 Two models proposed for transport form

What unfolded configuration might flagellin adopt during translocation? No indication from experiment has yet been made due to the inherent difficulty of observing translocating flagellin inside the filament channel. I proposed two models in this thesis, shown schematically in Figure 5.1.

The first, dubbed *Wire*, suggests flagellin is transported as a straight chain with N- and C-terminus on opposite ends. Because the export signal is located on the N-terminal [Kuwajima et al., 1989], I assume that the wire would enter the channel N-terminal first. The second, dubbed *Hairpin*, suggests flagellin becomes a U-shaped chain (U-loop) and the termini enters the channel first, with domain D3 located in the hairpin 'bend' entering last. In spite of the larger cross sectional area, the *hairpin* offers two advantages over the *wire*: (i) it maintains the radial arrangement of domains in the native state which would speed up refolding and; (ii) *hairpins* would only be half as long as *wires*, giving twice the transport rate. The next question is: how can the export apparatus produce either of these transport forms?

As mentioned in the Introduction, an ATPase or the Proton Motive Force (PMF) might power the mechanical unfolding and threading of flagellin proteins through the channel. The observation that ATPase IncV (homolog of FliI in the flagellar system) in the needle complex export system cannot export a chimeric protein with a mechanically strong GFP domain linked to a natural export substrate led the authors to suggest that proteins to be secreted should be made easy for the 'unfoldase' [Akeda and Galán, 2005]. It is then natural to ask if flagellin and other flagellar export proteins be easily unfolded

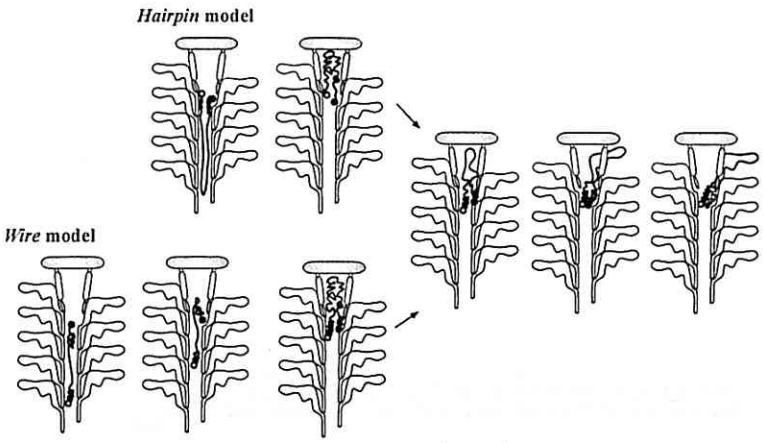


Figure 5.1: **Two models for transport forms of flagellin monomer.** In the *Wire* model, flagellin moves through the channel as a string with N-terminal (filled circle) leading. In the *Hairpin* mode, flagellin moves as a U-loop with both termini leading. Also shown are speculative binding and refolding processes in the filamet cap chamber.

along a certain mechanical pathway? To answer this, we need to investigate the mechanical unfolding pathways of flagellin. AFM and its in silico analog, force-probe (FP) MD, are useful tools to investigate the mechanical properties of proteins (see Section 3.2 for details). Experiments found that titin I27 mutants that are less resistant to mechanical unfolding are more easily imported into mitochondria through the mitochondrial import system [Sato et al., 2005]. Hence, AFM or FP-MD might give a preliminary verdict on which transport form is more likely.

Using FP-MD, I have determined the mechanical effort to obtain *wire* and *hairpin* conformers starting from a model of isolated flagellin (see Methods). A *wire* flagellin can be obtained by pulling apart or unzipping flagellin from its adjacent termini. A *hairpin* flagellin can be obtained by elongating or stretching flagellin along its molecular axis (with a radial arrangement of domains starting from termini in D0). I found that under identical simulation conditions, it takes less mechanical effort to realize a *wire*.

5.2 Methods

5.2.1 Starting structure for force-probe(FP) MD

Monomeric flagellin structure obtained from the short MD simulation in solvent (see Chapter 4) was used for most of the mechanical unfolding simulations presented here, denoted as structure S_1. The effect of using different conformations on the unfolding pathway and forces has not been extensively explored in this thesis due to limited computational

resources.

5.2.2 Use of implicit solvent model to reduce computation cost

Initial unfolding simulations for *Stretch* pathways were performed in the presence of water molecules in a very long simulation box. However, after much computational effort and repeated enlargement of the simulation box, only a straightening of flagellin along the pulling direction was obtained. This involved domain-domain movements without any unfolding. Taking into consideration that *Unzip* simulations would be far more costly in terms of computation time and hence impractical, I decided to switch over to an implicit or continuum representation of the solvent (see Section 3.1.2 for more details on implicit solvent).

In particular, the OBC model II variant [Onufriev et al., 2004] of the Generalized-Born (GB) model (setting "IGB=5" in AMBER8 software) was used to carry out GB/SA simulations with a physiological salt-concentration of 0.2 molar. GB/SA method involves computing the polar component of the solvation free energy by the GB method and the non-polar component taken to be proportional to solvent accessible Surface Area of the molecule. Default internal (protein) and external (solvent) dielectric constants were used. A large non-bonded cutoff of 25 Å was chosen for electrostatic, vdW and GB interactions. Structure S_1 (see above) were subjected to energy minimization and heating phases with positions of non-H atoms restrained using harmonic potentials. Restraints were then released in two steps of 6-ps each. Langevin dynamics with a low collision frequency of 1 ps^{-1} that also represent solvent friction and stochastic effects to some extent was used to maintain the temperature at 300 K. The resultant structures served as starting points for force-probe studies. The change to the backbone RMSD value was 0.60 Å after this preparation procedure. For the record, I have also tried to equilibrate polymeric flagellin from 1UCU under GB/SA. The structured D0 helices in 1UCU became partially structured as in TIP3P equilibration but became elongated after just 0.6-ns of simulation. The resultant conformation is hence not used for mechanical unfolding.

To give an idea of the computational effort, the complete *unzipping* of flagellin at the slowest pulling speed of 0.05 Å/ps (at each termini) required more than 350 hours (about 2-weeks) of processing time on 32 Intel Xeon 2.8 GHz processors in a PC Linux cluster connected by a Gigabit Ethernet network.

5.2.3 Implementation of constant-velocity FPMD

The natural "reaction coordinate" to monitor under constant-velocity FPMD (cv-FPMD) is the end-to-end extension along the direction of force or pulling direction. Extension is defined as the separation between two pulled-groups or between a pulled-group and a fixed-group. The difference between the desired position $x_{REF}(t)$ (reference coordinates or positions of the pulling 'spring') and the actual position $x(t)$ of the pulled-group gives rise to a harmonic restraint force $F(t) = k|x(t) - x_{REF}(t)|$ where k is the force-constant of the harmonic positional restraint (see Table 5.1 for the values used in this study). In SMD (see Chapter 3), $x_{REF}(t) = x_0 + vt$ where x_0 is the equilibrium position of the spring at the start of the simulation. But instead of continuous pulling (actually the equilibrium position is moved a distance of $v\Delta t$ every 0.1-ps interval in [Lu et al., 1998]), I incremented the restraint equilibrium position by 1 Å along the pulling direction at each step and then allow the system to equilibrate during an interval ranging from 4 to 20-ps. A set of reference coordinates representing the successive equilibrium positions were prepared before starting the pulling simulations. Although this is a "coarser" approximation to true constant-velocity pulling, the scheme offers a simple way of reaching quasi-equilibrium conditions by simply extending this time interval to say 40-ps as in the so-called "pull-and-wait" scheme [Pabón and Amzel, 2006].

Despite the difference in implementation, I could obtain a similar force-extension curve of disulphide-bond-reduced Titin I1 as reported by the Schulten group using the true SMD method in explicit solvent (compare black line in (*a*) to gray curve in (*b*) in Fig. 5.2). Snapshots during unfolding were also found to be similar (compare TIP3P 300-ps snapshot to snapshot '(f)', for instance) but detailed analysis was not performed. I also conducted unfolding of I1 under the implicit solvent model used in this work. Similar force-extension profiles between both solvent models were obtained (though lower force peak under implicit solvent, see Fig. 5.2 *a*) in addition to similar initial unfolding pathways (Fig. 5.2 *c*).

5.3 Results

5.3.1 Implicit solvent equilibrium simulation of monomeric flagellin

To assess the stability of flagellin structure under implicit solvent for nanosecond-long simulation times, I conducted a 4-ns simulation starting from structure S_1 under the same conditions as used in the force-probe MD runs. Changes to domain-level residue contacts

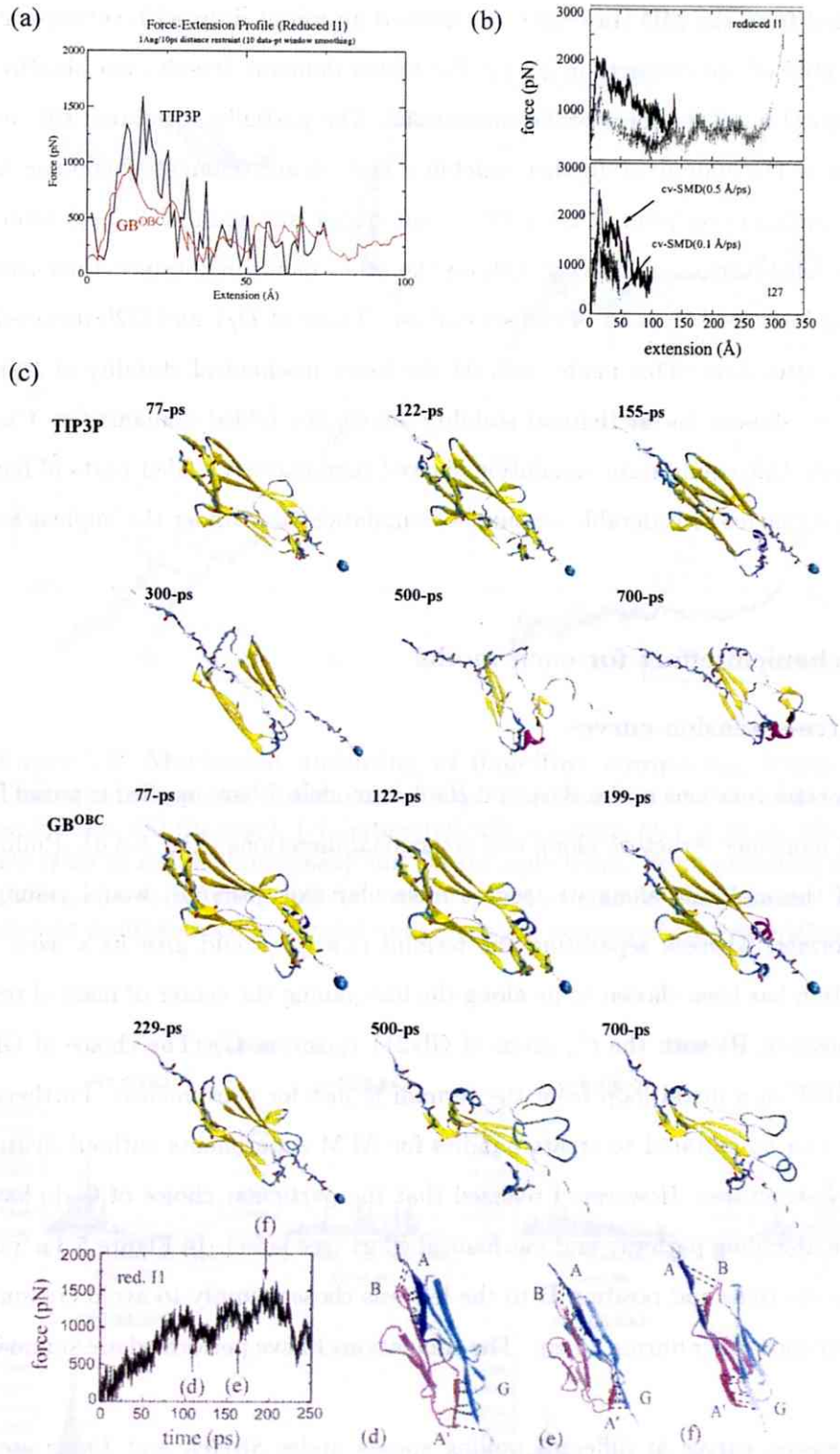


Figure 5.2: **Comparison of Titin I1 mechanical unfolding** under explicit and implicit solvent models under my cv-FPMD scheme presented in (*a*) and upper portion of (*c*) with those reported by Schulten and co-workers [Gao et al., 2002] shown in (*b*) and bottom panels in (*c*). Figures in [Gao et al., 2002] reproduced with permission from the Biophysical Society.

were monitored from the MD trajectory. I observed an initial drop with subsequent stabilization at 60% of the contacts in S_1 for the folded domains (results not shown) that was also observed in all the force-probe simulations. The partially structured D0 and the lower portion of D1 underwent further unfolding and became elongated (similar to the equilibrated conformation from 1UCU) with a non-native anti-parallel β-sheet formed in D0. Natively folded domains D2a and D3, on the other hand, maintained their contacts at ~60% from 1-ns until the end of the simulation. Those of D_f1 and D2b dropped further to ~40% after 3-ns. This might indicate the lower mechanical stability of D_f1 and D2b, which also showed lowest thermal stability among the folded domains (see Chapter 6). In summary, this equilibrium simulation showed that natively folded parts of flagellin could remain stable for considerable amounts of simulation time under the implicit solvent model.

5.3.2 Mechanical effort for each model

5.3.2.1 Force-extension curves

To realize the conformations in the *Wire* and *Hairpin* models, I have applied external forces on a flagellin monomer structure along two orthogonal directions (Fig. 5.3 *a*). Pulling on both ends of the molecule along its longest molecular axis (*Stretch*) would result in a *hairpin* conformer, whereas separating the termini (*Unzip*) would give us a *wire*. The *Stretch* direction has been chosen to be along the line joining the center of mass of termini C_α atoms (position **P**) with the C_α atom of Gly211 (position **C**). The choice of Gly211 which is located on a distal loop from the termini is just for convenience. Furthermore, loop residues can be mutated to create handles for AFM experiments without disrupting the secondary structures. However, I realized that the particular choice of **C** do have an impact on the unfolding pathway and mechanical effort (see later). In Figure 5.3 *a*, pulling position **A** to the right and position **B** to the left was chosen simply to avoid crossing the D1 helices over each other during *Unzip*. The simulations I have performed are summarized in Table 5.1.

Force-extension curves at different pulling speeds under *Stretch* and *Unzip* are presented in Fig. 5.3 (*b*) and (*c*) respectively. Under *Stretch*, the initial increase in force is independent of the pulling speed. This is during extension of the domain-domain linkers to align flagellin along the pulling direction and also includes the extension of D0 partial helices. After an initial D0-D3 extension of around 150 Å, forces increase with pulling

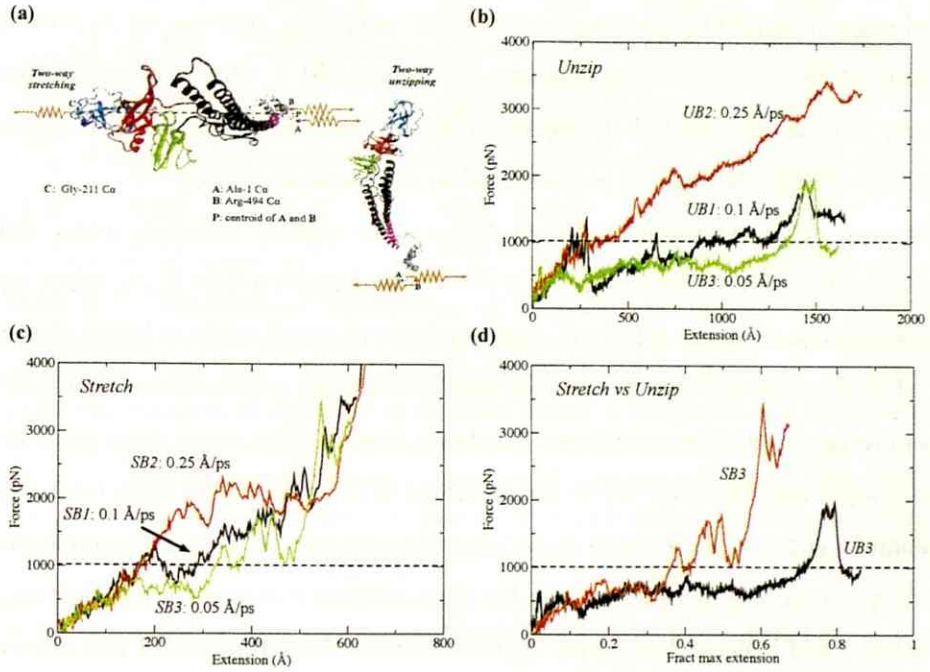


Figure 5.3: **Mechnical unfolding of flagellin: comparing *Unzip* to *Stretch*.** (*a*) Pulling directions for *Stretch* and *Unzip*. Flagellin domains and subdomains are colored as follows: D0 (in gray), D1 (proteolytically resistant D_f1 in black; the rest in magenta), D2 (D2a in red; D2b in green) and D3 (in aqua blue). Force-extension curves for different pulling speeds (at each end) for *Unzip* (*b*) and *Stretch* (*c*) simulations. (*d*) Forces from the slowest pulling speed of *Stretch* and *Unzip* are compared on a normalized extension scale.

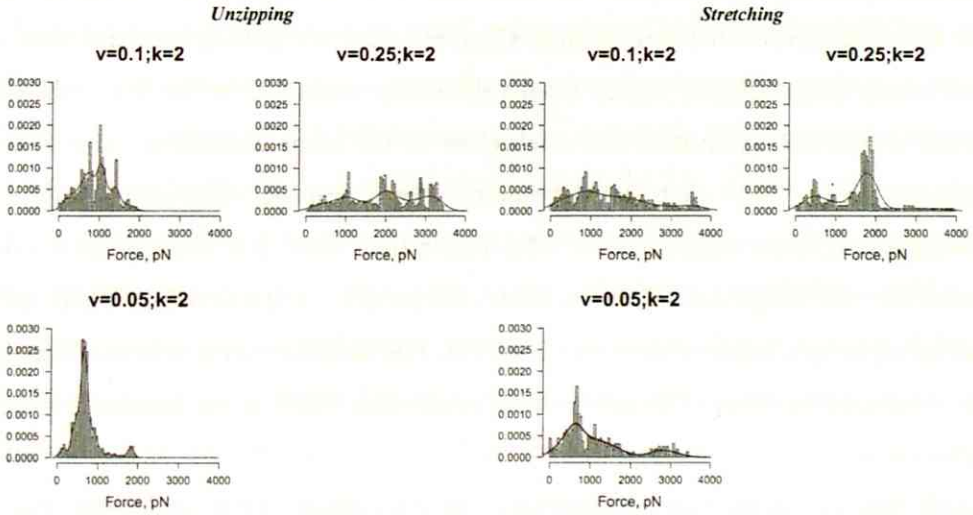


Figure 5.4: **Histograms of restraint forces** measured during each mechanical unfolding simulation.

speed though differences are slight in view of the 2 to 2.5× difference in speeds (at each end). Subsequent unfolding involves (sub)domain unfolding, showing up as force peaks. As the molecule is stretched to an extension of around 500 Å, further extension becomes increasingly difficult as reflected by the sharp increase in forces. The detailed pathway of a selected *Stretch* simulation will be presented in the next subsection.

In contrast, force-extension curves in *Unzip* show greater variation under different pulling speeds (Fig. 5.3 *c*). Although my slowest pulling speed is 0.05 Å/ps, which extends the end-to-end separation by 0.1 Å/ps, this speed is still a million times larger than values used in AFM experiments that occur over milliseconds. By extrapolation, I suspect that the forces involved would be even lower for slower speeds. The major force-peak at large extension is a pre-requisite step before the unfolding of D3. The pulled chain has to become tight enough (a climb in force) such as to rotate the domain to a more favorable position for unfolding. In AFM experiments or in the physiological system such rotation might be easy to achieve and hence a force peak might not occur. Unfolding of D3 was observed to be easy after the rotation (no further force peaks). More details will be given below.

For a better comparison of the "roughness" of the force curves, the fraction of maximum extension in *Unzip* and *Stretch* was used in place of absolute extension in Fig. 5.3 (*d*). The maximum extension in each pulling direction is determined as follows. The average C_α separation has a value of 3.8 Å in a typical 'relaxed' protein structure, often quoted in simplified models of proteins [Sulkowska and Cieplak, 2008]. Using this, the maximum end-to-end separation for *Unzip* is $(494-1)\times 3.8 - 0.75 \simeq 1873$ Å, where 0.75 is the projection of the initial separation between termini $C_\alpha$atoms onto the pulling direction (set to be approximately perpendicular to the *Stretch* direction; explained in the next subsection). Observed end-to-end extensions will reach close to the above maximum value only in a straightened chain. Note that this maximum is for a 'relaxed' state and a fully tensed and straightened chain might have a value larger than this. For *Stretch*, the maximum extension (for the longer half-chain) is $(494-211)\times 3.8 - 172.6 \simeq 902$ Å where 172.6 is the initial separation between positions **C** and **P**. The overlaid curves indicates that initial efforts are similar but that of *Stretch* became larger after ~30% of the maximum extension has been reached.

Lastly, histograms were constructed based on the restraint force time series (Fig. 5.4). These give an indication of the range of forces required as well as the average force. Forces larger than 4000-pN are excluded from each distribution. For *Unzip* under a pulling speed

Sim	Pulling speed, v (Å/ps)	Force const., k (kcal/molÅ^2)	Max. ext. (Å)	Denat. time (ns)	Denat. force (pN)	Transport work (kcal/mol)	Transport force (pN)
SB1	0.10	2	822	4.0	11500	7200	2500
SB2	0.25	2	784	1.1	2000	10230	2000
SB3	0.05	2	607	6.3	3440	9730	3440
UB1	0.10	2	1652	7.5	1950	17560	1950
UB2	0.25	2	1738	3.2	3420	40430	3420
UB3	0.05	2	1614	17.0	1960	16950	1960

Table 5.1: **Mechanical unfolding of flagellin monomer via two mechanisms: *U*nzipping or *S*tretching.** Labels *'SB'* and *'UB'* denote two-way *Stretch* or *Unzip* respectively (see Text). Denaturation time is the simulation time needed to allow native fractional contacts of domains to decrease below a threshold of 0.2. Denaturation force is the maximum restraint force encountered during the denaturation time. Transport work is the area under the force-extension curve up till extension at which the width of D3 β-sheet in C_α representation became less than 20 Å, which makes the molecule transport-capable. "Transport force" is the maximum force needed to obtain the transport form.

of 0.1 Å/ps and a "soft" spring of 2 kcal/molÅ^2, the force distribution has a mean(std-dev) of 909(404)-pN. At the slower speed of 0.05 Å/ps, the distribution is more sharply peaked with a mean(std-dev) of 721(338)-pN. The corresponding values for *Stretch* are 1508(965) and 1142(815) respectively. The results suggest that *Unzip* trajectories should incur smaller forces than their *Stretch* counterparts as we move towards more physically realistic pulling speeds.

5.3.2.2 Denaturation time and force

Table 5.1 lists the "denaturation-time" and associated "denaturation-force" for each simulation. The denaturation-time is defined as the duration for native contacts to drop below an arbitrary threshold of 20% (all domains at the end of the 500-K thermal-denaturation simulation have fractional contacts below 0.2; see Chapter 6). I then define the denaturation-force as the maximum force encountered before the denaturation time. Denaturation-force for *Stretch* are larger than *Unzip* due to the difficulty in completely unfolding D3 which requires slippage of the half-chains, possible only by breaking of multiple H-bonds in non-native anti-parallel β-sheets formed across unfolded half-chains (see region to the right of domain D2 in the 1.6-ns snapshot in Fig. 5.5 *a* lower). An exception is when I pull at the very high speed of 0.25 Å/ps, whereby D3 denatured very rapidly during the start of the simulation due to the strong local forces.

5.3.2.3 Mechanical work required to get transportable form

Besides the maximum forces encountered during unfolding, we may also wish to determine the mechanical work required. In particular, I wish to determine the effort required to obtain a "thin-enough" flagellin. Based on the C_α-only representation, the times when D3 residual β-sheet in *Stretch* trajectory snapshots has width smaller than 20 Å are determined (using measurements in molecular viewer VMD and from ellipsoidal approximation based on minor axes computed as for volume estimates in Section 6.2). The corresponding restraint forces are shown in Table 5.1 under the "Transport force" column and the "Transport work" computed as the area under the force-extension curve up to the extension with the value of "Transport force". For example, in the *SB1* entry, the time-stamp is 2.5-ns which has a restraint force of 2500-pN (Fig. 5.3 *c*). The mechanical work (from extension of 0 Å up to 484 Å) was then calculated as 7200 kcal/mol. For *Unzip* trajectories, the complete trajectories are used because D3 is thin enough only when completely unfolded. Hence, "Transport force" and "Denaturation force" are the same.

Although I have presented values for three pulling speeds, only the lower ones (0.05 or 0.1 Å/ps) are closer to experimental and physiological conditions. Hence, I decide to only use those sets in comparing *Unzip* and *Stretch*. The picture that emerges is one in which lower maximum forces are required for *Unzip* but more mechanical work required (higher ATP consumption if unfolding is powered by an ATPase). For *Stretch*, less mechanical work is needed since the maximum extension for *hairpin* is half that of *wire* but larger maximum forces may be needed.

5.3.3 Detailed mechanical unfolding pathways

In Fig. 5.5 (*a*) I present snapshots during *UB1* and *SB1* trajectories. The unfolding of each domain is monitored via the fraction of native contacts, shown in the lower panel of Fig. 5.5 (*b*). I will discuss the unfolding pathways in some detail below. The unfolding pathways under the slowest speed (*UB3* and *SB3*) are very similar.

5.3.3.1 *Stretch* pathway

Under *Stretch*, flagellin is extended along its molecular axis by pulling position **C** in domain D3 along the line joining it with position **P** (Fig. 5.3 *a*). In practice, terminal C_α atoms are separately subjected to the same pulling force. The initial response was an alignment of the domains along the pulling direction, via bending at the D1-D2 and D2-D3

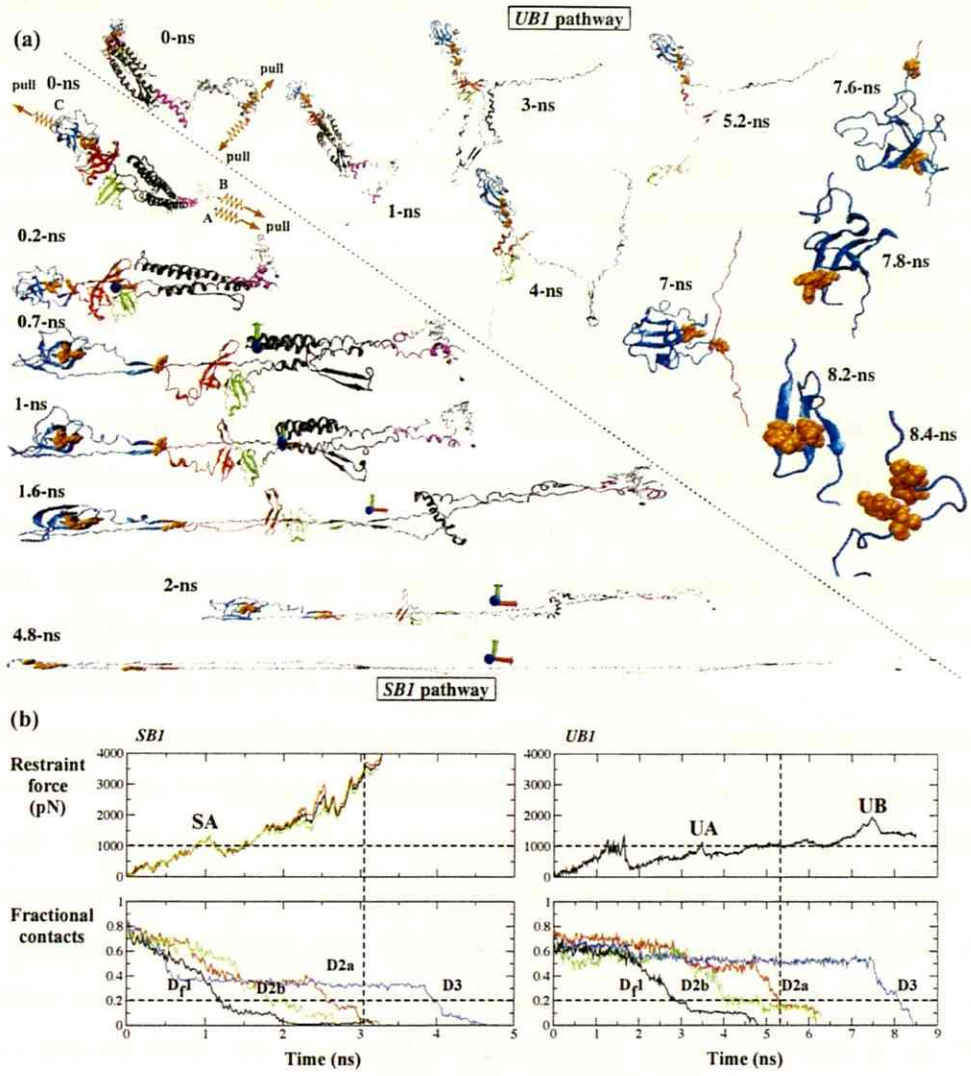


Figure 5.5: **Mechanical unfolding of flagellin: snapshots, forces and fractional contacts.** (*a*) Snapshots from *UB1* (upper portion) and *SB1* (lower portion) simulations which resulted in *wire* and *hairpin* flagellin, respectively. Flagellin domains are colored as in Fig. 5.3. Blue and red spheres represent N- and C-terminal C_α atoms. Orange spheres represent atoms in the D3 surface aromatic cluster. Colored coordinate arrows are shown as position markers to compare the amounts of stretching at either end. (*b*) Time variation of restraint forces and domain-level fractional contacts. Fractional contacts colored in the same way as their corresponding domains in (*a*). Curves in *SB1* force-time plot represent forces computed based on different end-to-end extensions (see Text). Vertical lines indicate the forces when fractional contacts in D2a dropped below the 0.2 threshold. Features 'SA', 'UA' and 'UB' are due to mechanical resistance by structural elements in flagellin shown in Fig. 5.6.

domain junctions. This is the so-called 'tertiary structural elasticity' exhibited by modular extracellular matrix proteins under weak forces [Gao et al., 2006].

Continued application of force broke up the aromatic cluster formed by Tyr190-Phe222-Tyr229 in D3 (orange spheres in Fig. 5.5) shortly after 0.2-ns. The D3 terminal β-sheet subsequently became extended along the pulling axis. The D2a-ND1 H-bond interface separated between 0.7 to 0.8-ns, with an extension and solvent exposure of the D1-D2 unstructured linker that was packed under the hydrophobic tri-helical core of D_f1 in the native state. Non-native H-bonds formed across residues with long side-chains on either side of the interface resisted its separation. However, no significant force peak results from this event. The force peak around 1-ns ('SA' in Fig. 5.5 *b*) is due to sliding of the backbone from N-terminal towards D3. Two sets of backbone H-bonds have to be broken simultaneously: (i) those between β_3 and β_{12} and (ii) those holding ND1b β-hairpin together. These H-bonds are oriented perpendicular to the force direction and are known to form 'mechanical clamps' (see Discussion). These are shown as insets to Figure 5.6. Domain D_f1 further unfolded as it is pulled from either end (manifested by a decrease in fractional contacts in the lower panel of Fig. 5.5 *b*).

Non-native anti-parallel β-sheets formed across the unfolded backbone segments at various locations need to be broken in order for the Nterm-to-D3 and D3-to-Cterm backbone segments ("half-chains") to slide past each other during further unfolding. This accounts for the steady increase in restraint forces (Fig. 5.5 *b*). In fact, for D3 unfolding at very large extensions the restraint force 'sky-rocketed' to very large values (not shown in the force-time curve). Because Gly211 chosen for position C is not located in the middle of the polypeptide chain, the Nterm-to-D3 "half-chain" is under high strain as the D3-to-Cterm "half-chain" is further extended. The same amount of force are applied to N- and C-terminus but only the C-terminus can be further extended. In AFM experiments, the N-terminus might detach from the AFM tip after the Nterm-to-D3 "half-chain" has reached a critical tension. The restraint force computed based on the longer Gly211-Arg494 separation (green line) will drop to a lower value compared to Gly211-Ala1 (red line) or the "average" based on Gly211 and Ala1-Arg494 center-of-mass (black line) after a certain time (force capped at 4000-pN in Fig. 5.5 *b*).

5.3.3.2 *Unzip* pathway

During *Unzip*, termini $C_\alpha$atoms are pulled in a direction approximately perpendicular to the flagellin's longest molecular axis (Fig. 5.3 *a*). The pulling direction was defined using the cross product of two vectors starting from the N-terminal end of helix α_2 (Fig. 5.6) but pointing towards Gly211 C_α in D3 and the Asp313 C_α at the ND1-D2a interface respectively. However, the N- and C-terminal C_α atoms in structure S_1 would collide into each other if we move them along the pulling vector. A short preparatory simulation (200-ps @ 0.05 Å/ps followed by 50-ps @ 0.1 Å/ps) was hence used to displace the atoms along the perpendicular direction to the pulling vector. After checking that the fractional native domain-level contacts have not changed significantly, I pulled the termini atoms apart along the pulling vector.

Due to the closely associated nature of the partial helices in the terminal region, a non-native 'helical-bundle' was formed when I tried to separate the N- and C-terminal backbones. Resistance from this 'helical-bundle' produced a series of force peaks (Fig. 5.5 *b*). The separation of the 'bundle' by 1.8-ns lead to a drop in the restraint force. Next, during the unfolding of D_f1, the sliding of the C-terminal (CD1) helix against the hairpin in the N-terminal (ND1) was resisted by the non-native salt-bridge Arg431-Glu153 from 2.5 to 3.5-ns, accounting for the sharp force peak around 3.5-ns ('UA' in Fig. 5.5 *b*). See Fig. 5.6 for the location of the salt-bridges. After breaking this salt-bridge, the unfolding of the rest of D1 and of D2 was without resistance or 'barrier-less'. By 6-ns, only D3 remains to be unfolded. A tightening of the unfolded chain was required in order to rotate the domain such as to resolve the "cross-over" in its terminal β-sheet (see lower-left inset to Fig. 5.6). This produced a force ramp between 6 to 7.5-ns ('UB' in Fig. 5.5 *b*). The unfolding of D3 β-sheets was again easy, as seen from the sharp drop in restraint forces. Under a larger pulling speed of 0.25 Å/ps, no force peak resulted in this unfolding step because the pulling speed is large enough to resolve the "cross-over" quickly. Put another way, the force is large enough to overcome this energy barrier easily.

In summary, there are only three major force events during *Unzip*: (i) breaking up D0 'helical-bundle', (ii) breaking of native/non-native salt-bridge in D1 and (iii) rotation of D3. I suspect that in the millisecond time-scale of AFM experiments events (i) and (iii) may not incur any mechanical resistance. Domain D3 would have time to sample different orientations relative to D2. Hence, only disruption of salt-bridges could be "rate-limiting" in the unfolding pathway.

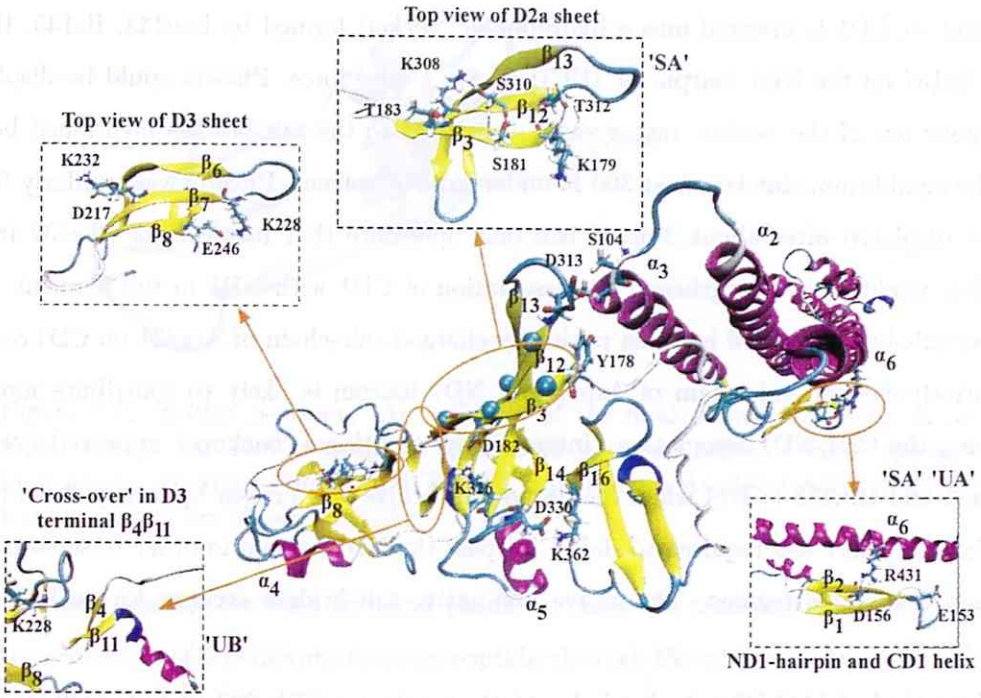


Figure 5.6: **Key load-bearing elements** in flagellin structure, indicated on the structure S_1. Secondary structures (α-helices and β-strands) assigned by STRIDE [Frishman and Argos, 1995] are numbered starting from the N-terminal. Left inset: the two salt-bridges bracketing the β_{6-8}-sheet in D3. Top-center inset: the three pairs of residues across β_3 and β_{12} in D2a that could form backbone and side-chain H-bonds are components of the 'mechanical clamp', resisting sliding of β_3 against β_{12} during all *Stretch* and some *Unzip* simulations. Bottom-right inset: the native (D156-R431) and non-native (E153-R431) salt-bridges holding ND1 and CD1 together during *Unzip* runs. The $\beta_1\beta_2$ hairpin is also a 'mechanical clamp' during *Stretch*. Labeled residues on the main figure are: D182-K326 salt-bridge spanning β_3 and β_{13}; D330-K362 salt-bridge across the D2a-D2b interface; D313, S104 and Y178 that lies across the D2a-ND1 interface (S106 hidden from view). Labels 'SA', 'UA' and 'UB' correspond to force-peaks observed in *SB1* and *UB1* unfolding.

5.3.4 Surface hydrophobic clusters and H-bond groups as load-bearing elements

Surface H-bond networks (salt-bridges or polar residue side-chains) and hydrophobic contacts might also contribute to flagellin's mechanical resistance, in addition to "longitudinal shear" of β-strands mentioned above. These side-chain interactions might help to strengthen flagellin for its role as the filament building block, though they incur a cost during mechanical unfolding.

In natively folded D1, a hydrophobic cluster and salt-bridges helped to hold C-terminal (CD1) helix α_6 to the ND1 hairpin (Fig. 5.6, lower-right Inset). The aromatic ring of

Phe432 on CD1 is inserted into a hydrophobic 'socket' formed by Leu143, Ile145, Ile155 and Ile157 on the ND1 hairpin in 1UCU or S_1. Under force, Phe432 could be displaced out gone out of the 'socket' rather easily compared to the salt-bridges mentioned below. In the equilibrium simulation at 300 K under explicit solvent, Phe432 was similarly found to be displaced after about 3-ns. I can only speculate that insertion of Phe432 in the 'socket' might have strengthened the association of CD1 with ND1 in the filament. The native salt-bridge formed between positively-charged side-chain of Arg431 on CD1 α_6 and negatively-charged side-chain of Asp156 on ND1 hairpin is likely to contribute more to keeping the CD1-ND1 association. Interestingly, additional "backups" appeared between Arg431 and Glu153 (*UB1*) when the native salt-bridge was broken by force. Thus, much mechanical effort was required to slide CD1 past the ND1 hairpin in order to separate the N- and C-terminal regions. The native/non-native salt-bridges account for peak 'UA' in Fig. 5.5 (*b*).

In natively folded D3, a hydrophobic cluster made up of Phe222 (β_6), Tyr229 (β_7) and Tyr190 (β_4) lies at the D2-D3 domain interface. The role of this cluster is still unclear, though its disruption is needed for extending the D3 termini β-sheet and increasing the separation of domains D2 and D3 as observed in the *Stretch* simulations. Aromatic interaction between the Phe222-Tyr229 pair might have contributed to $\beta_6\beta_7$ being the last structure to unfold under *Unzip*. Lastly, the salt-bridges Asp217-Lys232 and Glu246-Lys228 which lies across the edges of $\beta_6\beta_7$ and $\beta_7\beta_8$ respectively might have provided some minor resistance of the β_{6-8}-sheet to mechanical unfolding via *Unzip* and *Stretch*. Interestingly, the β_{6-8}-sheet (Fig. 5.7) was found to be a potential folding core for D3 from the thermal unfolding study (Chapter 6).

5.4 Discussion

5.4.1 Is *hairpin* small enough for channel?

In the *Hairpin* model, flagellin is transported in the form of a U-shaped chain with domain D3 located at the 'bend' in the middle. Both termini of flagellin would have to be threaded into the channel in this model. Is the channel large enough to accommodate two partially structured polypeptide chains?

From my simulations, elongated hairpin flagellin molecules show a cross sectional diameter of 20 Å or less at the terminal region that has no secondary structures. Hence,

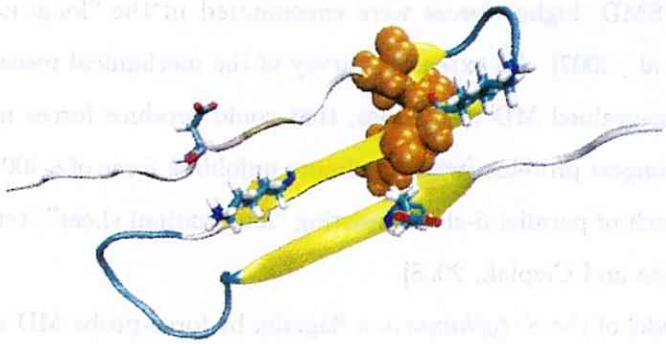


Figure 5.7: **Z-like β-sheet.** An example of the final β-sheet ($\beta_6\beta_7\beta_8$) to be unfolded under *Unzip* (the 16.6-ns snapshot from *UB3* is shown here). Orange spheres represent hydrophobic residues Phe222 and Tyr229. Salt-bridges Asp217-Lys232 (left) and Glu246-Lys228 (right) are shown in sticks.

it seems possible for both terminal chains to enter the transport channel. Furthermore, translocation of a β-hairpin-forming peptide through the ribosomal tunnel has been studied via simulation. The peptide remains folded as it moves through the tunnel if the tunnel diameter is larger than ~13.7 Å [Kirmizialtin et al., 2004]. This suggests that a hairpin conformer with β-sheets along its length could similarly pass through the flagellar channel. Setting aside the possibility that a *hairpin* might be more difficult to realize in practice, I compared the two models on an equal footing.

5.4.2 Flagellin *softness* depends on pulling geometry

How strongly a protein resist mechanical tension depends on how it is pulled: like humans, a protein has 'soft spots'. In a theoretical study published soon after single-molecule manipulation of bio-molecules became successful, Lavery and co-workers reported that it is easiest to unfold a globular protein by unzipping β-strands from the edges of β-sheets or longitudinally shearing apart α-helix bundles [Rohs et al., 1999]. The unzipping or "lateral shearing" as it is called of β-strands involves the breaking of individual H-bonds which is independent of strand length. "Longitudinal shearing" of β-strands, in contrast, involves breaking multiple H-bonds at once and is dependent on strand length [Rohs et al., 1999]. Such considerations helped to explain why E2lip3, a β-sheet protein, exhibits different mechanical resistance when pulled in two orthogonal directions in both experiments and simulations [Brockwell et al., 2003]. Also, use of SMD found that both "longitudinal shearing" and hydrophobic interactions contributed significantly to the mechanical resistance of bovine carbonic anhydrase II [Ohta et al., 2004]. Similarly, in the unbinding of edge pep-

tides from amyloid fibrils by SMD, higher forces were encountered in the "longitudinal shearing" direction [Raman et al., 2007]. An extensive survey of the mechanical resistance of PDB structures using coarse-grained MD techniques, that could produce forces in the AFM range, found that the strongest proteins (with maximum unfolding force of ~400-pN) are all β-rich and contain a patch of parallel β-sheet resisting "longitudinal shear", termed a 'mechanical clamp' [Sulkowska and Cieplak, 2008].

In this study, I pulled a model of the *S. typhimurium* flagellin by force-probe MD along directions both parallel and perpendicular to its longest molecular axis (Fig. 5.3 *a*). The pulling directions were chosen to produce the *wire* (straight chain) and *hairpin* (U-shaped chain) conformers mentioned above. Though the force responses were complicated by the multi-domain nature of this protein, "lateral" and "longitudinal" shearing of β-strands still featured prominently at the domain level. From the unfolding trajectories, *Unzip* involved mainly "lateral" shear of β-strands (unfolding of D2b, D2a and D3). In contrast, *Stretch* involved more "longitudinal" shear (separation of D2a $\beta_3\beta_{12}$ and for the many non-native β-sheets formed across the unfolded half-chains). These non-native β-sheets have to be broken in order to unfold domains adjacent to them. In Fig. 5.8 taken from a *Stretch* trajectory at pulling speed of 0.1 Å/ps, a non-native β-sheet formed across D1 (black) and D2b (green) backbones around 2-ns have to be broken in order for D2b unfolding to proceed. New non-native sheets formed across the backbones by 2.2-ns. These could account for the steady increase in restraint force under *Stretch*.

5.4.3 Which unfolding mode is preferred?

The actual transport form, *wire* or *hairpin*, depends on which unfolding mode (*Unzip* or *Stretch*, respectively) is preferred by the type III export apparatus. From my simulations, it seems that *Unzip* mode is preferred if the unfolding mechanism cannot generate very high forces. The force-extension profile from the slowest *Unzip* simulation showed a plateau (or levelling) around 600-pN, which translates to around 60-pN in AFM experiments due to the 10× difference in forces (see subsection 3.2.3 in chapter 3). The relatively constant force required would be suitable for, say, AAA type ATPases which are known to iteratively apply a uniform unfolding force during denaturation, with more ATP molecules consumed for more stable proteins [Kenniston et al., 2003]. On the other hand, only half the mechanical work or ATP molecules is required if *Stretch* mode is used. However, larger forces are required and both termini would have to be pulled into the channel.

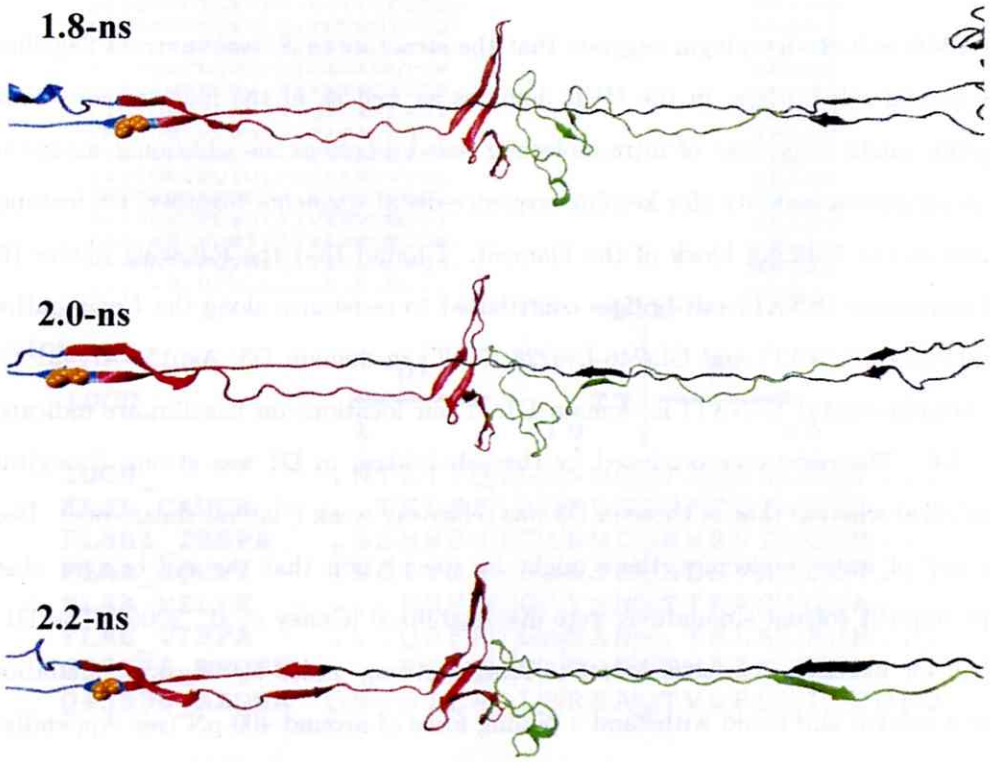


Figure 5.8: **Non-native β-sheets during *Stretch*.** Snapshots showing formation and disruption of non-native β-sheets across unfolded half-chains during *SB1* trajectory. Secondary structures assigned by program STRIDE [Frishman and Argos, 1995] in VMD [Humphrey et al., 1996].

A point to note is that although *wire* seemed to require lower peak forces (see Fig. 5.3 (*d*)), the unfolding of D3 might be delayed until the "cross-over" in its terminal β-sheet get resolved. Large increase in force (up to 2000-pN) was observed even under the slowest pulling speed. However, a change in pulling direction or simply pulling only on the N-terminal (the operational mode for unfoldases, Section 1.5) might resolve such "cross-over" with minimal effort.

5.4.4 Do salt-bridges contribute to flagellin mechanical resistance?

The VMD Salt-Bridge plugin suggests that the structure of *S. typhimurium* flagellin harbors several salt-bridges, in the HVR domains as well as in the filament-core domains. Flagellin might make use of intra-molecular salt-bridges as an additional means to increase structural stability (for keeping sequence-distal segments together, for instance) in its role as the building block of the filament. I found that the following native (NAT) and non-native (NNAT) salt-bridges contributed to resistance along the *Unzip* pathways: Asp217-Lys232 (NAT) and Glu246-Lys228 (NAT) in domain D3; Asp156-Arg431 (NAT) and Glu153-Arg431 (NNAT) in domain D1. Their locations on flagellin are indicated on Fig. 5.6. The resistance produced by the salt-bridges in D1 was strong ("longitudinal shear"-like) whereas that of those in D3 was relatively weak ("lateral shear"-like). Because of a lack of water molecules, there might be speculation that the salt-bridges observed in my implicit solvent simulations were over-stabilized [Geney et al., 2006]. The D1 salt-bridge, for instance, has been found to remain strong under equilibrium simulations in explicit solvent and could withstand a pulling force of around 400-pN (see Appendix A).

Salt-bridges have been known to play important mechanical roles in biology. The CD2-CD58 adhesion complex depends on salt-bridges, a suggestion initially made by explicit solvent SMD simulations [Bayas et al., 2003] and later validated by binding and force-measurement experiments [Bayas et al., 2007]. Salt-bridges acted as tethers during simulated unbinding of Alzheimer's β-amyloid (Aβ) peptides from amyloid fibrils [Raman et al., 2007]. The intra- and inter-molecular salt-bridges formed across windings of the infectious prion fibril may account for the β-solenoid fibril's high stability against urea at neutral pH but not at acidic or basic pH [Wasmer et al., 2008]. Inter-molecular salt-bridges have also been known to be important in the polymerization of flagellin into filament [Kitao et al., 2006].

Are these salt-bridges conserved across flagellin homologues? For D3, being a HVR

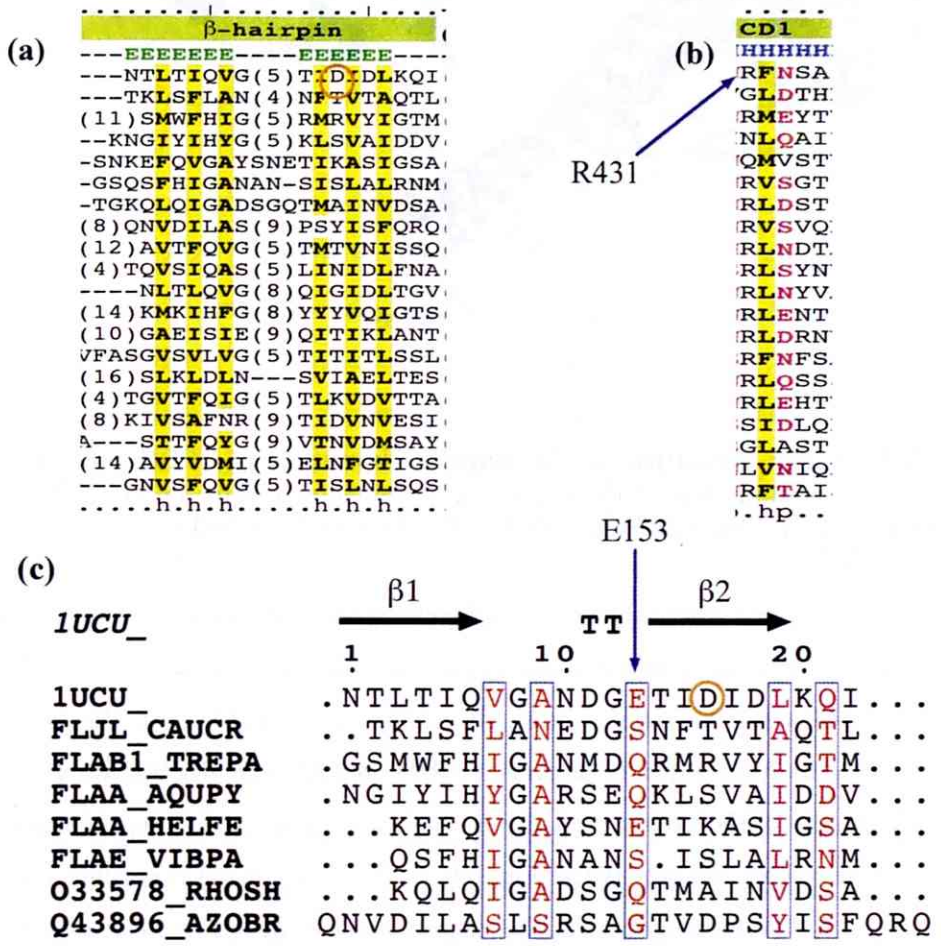


Figure 5.9: **Multiple-sequence alignment (MSA) of flagellin homologues in the ND1 and CD1 regions.** The MSA of the 20 most diverse flagellin homologs by Beatson in the ND1 hairpin region (*a*) and CD1 region (*b*), respectively. (*c*) MSA performed using ClustalW 1.83 (at http://ch.EMBnet.org, with default parameters) on the same ND1 hairpin segment for the top 8 sequences in the Beatson alignment. Post-processing by ESPript [Gouet et al., 2003] allows display of secondary structures in 1UCU over the top. Position of D156 in ND1 is highlighted by an orange circle in (*a*) and (*c*). A column is framed in blue by ESPript if more than 70% of its residues are similar according to physico-chemical properties. Sub-figures (*a*) and (*b*) are reprinted from *Trends in Microbiol.*, **14(4)**, S. A. Beatson, T. Minamino and M. J. Pallen, "Variation in bacterial flagellins: from sequence to structure", pages 151-155, copyright (2006), with permission from Elsevier.

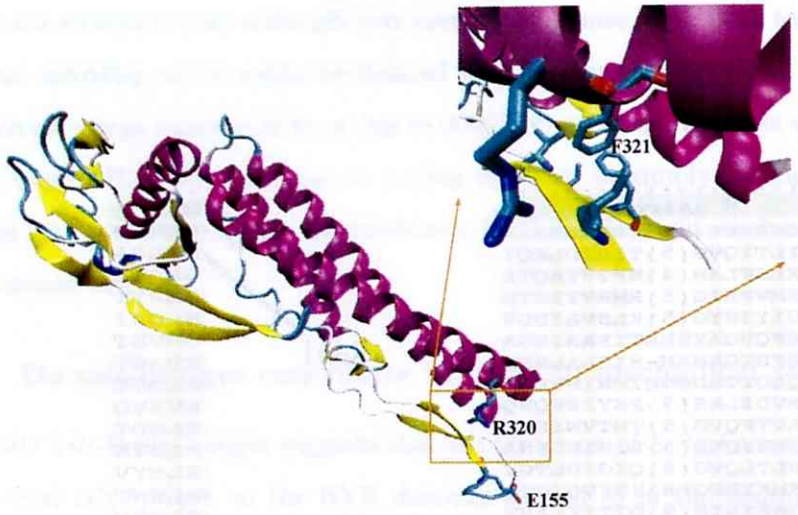


Figure 5.10: **X-ray structure of the flagellin homolog p5** (PDB code 2ZBI). Inset shows a zoomed-in view of the conserved hydrophobic cluster on the ND1 hairpin. R320(F321) is the homolog of R431(F432) in *S. typhimurium* flagellin.

domain, great diversity shown by the homologues makes any assessment of salt-bridges without tertiary structure difficult if not impossible. Domain D1, on the other hand, is highly conserved across bacterial species, especially the ND1 β-hairpin [Beatson et al., 2006]. From the multiple-sequence alignment (MSA) by Beatson, Arg appeared at residue 431 in fourteen out of the twenty most diverse sequences (Fig. 5.9 *b*). On the other hand, Asp appears at residue 156 only twice (Fig. 5.9 *a*). Because Glu153 is not included in the MSA by Beatson, I performed my own MSA using ClustalW webserver hosted by the Swiss Institute of Bioinformatics on a segment of the ND1 sequence which includes the ND1b hairpin ($\beta_1$and β_2). The first eight sequences in the Beatson alignment was used. Similarly good alignment of hydrophobic residues in β_2 region as obtained by Beatson was found (Fig. 5.9 *c*). Except exact conservation in the HELFE sequence, residue position 153 in 1UCU that is a Glu has been substituted by either Gln or Ser which also contains a hydroxyl group in the side-chains. However, TREPA (short for *Treponema pallidum*) and AQUPY (short for *Aquifex pyrophilus*) sequences contain Asp/Glu one residue before the Glu153 column in the MSA. From this observation, a non-native salt-bridge like Glu153-Arg431 in *S. typhimurium* may form in some flagellin homologs. Otherwise, a hydrogen-bond might still form across the highly conserved Arg431 and a polar residue at position 153 during *Unzipping* when ND1 and CD1 has to be separated.

The high conservation of D1 among flagellin homologs mentioned above is reaffirmed with the (terminal-truncated) structure of the homolog p5 from the cell surface of *Sphingomonas* strain A1 was determined by Murata and co-workers via X-ray diffraction in Feb 2008 [Maruyama et al., 2008]. Figure 5.10 shows the solved structure with a domain D1 highly resembling that of *S. typhimurium* flagellin (Arg431 homolog in CD1, and ND1 $\beta_1\beta_2$ hairpin are present) but with a much reduced HVR. The hydrophobic 'socket' on the surface of the hairpin as well as a Phe on CD1 that is inserted into the 'socket' are both conserved. Although no negatively-charged residues are in the direct vicinity of the conserved Arg320 (homolog of Arg431), Glu155 is located close enough to form a weak salt-bridge if its long side-chain can be oriented more towards Arg320.

If salt-bridges really do contribute to mechanical resistance, they might pose a challenge for the unfoldase. But if a proton flux through the export channel as postulated by Hughes and co-workers [Paul et al., 2008] do exist, acidic residues of proteins at the export gate might be transiently protonated and weakening the salt-bridges momentarily. However, we need a characterization of this proton flux and the resultant pH changes at the export gate to support or refute the existence of such a salt-bridge-mediated mechano-stability switch that is turned off during mechanical unfolding but turned back on during refolding.

5.4.5 Limitations of this study

I wish to mention three limitations in my study.

Firstly, the use of implicit solvent models in simulated mechanical unfolding remains controversial. On one hand, the replacement of backbone H-bonds by those made to solvent molecules have been found to be important in the unfolding process probed by explicit solvent force-probe/SMD simulations [Pabón and Amzel, 2006, Gao et al., 2006]. A force-spectroscopy study confirmed that solvent molecules are an integral part of the unfolding transition state [Dougan et al., 2008]. However, the use of implicit solvent avoided the slow solvent response under the high pulling rates used in most force-probe simulations [Ng et al., 2005]. More research is needed to assess and improve implicit solvent models, such as the commonly used Generalized-Born variants, for use in mechanical unfolding studies.

Secondly, as recognized by the pioneers of the SMD method, mechanical unfolding in silico with atomistic models requires at least a millionfold larger pulling speed than in single-molecule force experiments [Lu et al., 1998] to observe substantial unfolding events during nanosecond simulations. This may distort the unfolding pathway. Buehler and

co-workers suggested that to unfold β-sheets in simulations under the biologically-relevant slow-deformation mode, pulling speeds should be less than 0.1 Å/ps [Ackbarow et al., 2007]. I have used pulling speeds of 0.1 Å/ps or higher to reach completely denatured states within nanoseconds (Table 5.1). Our findings are highly in need of validation by AFM experiments.

Thirdly, even if my simulations could give force-extension profiles matching AFM experiments, there remains major differences between continuous AFM end-to-end pulling and in vivo pulling against a pore [Prakash and Matouschek, 2004]. In a simulation study, Tian and Andricioaei suggested that barnase import into the mitochondrion could follow the pathway with low unfolding energy barriers if pulling forces against the pore are switched off periodically to relax partially unfolded intermediates, allowing them to search out alternative pathways [Tian and Andricioaei, 2005]. Our AFM-like simulations did not allow for such pauses. Hence, future simulations should explore repetitive pulling of flagellin or other flagellar proteins against the export gate when it has been structually resolved at atomic detail.

5.5 Conclusion: flagellin transported as a *wire*?

In this study, I have used force-probe MD to conduct AFM-like two-way pulling simulations to determine the mechanical efforts as well as the detailed unfolding pathways to *Stretch* flagellin to create a U-shaped *hairpin* and to *Unzip* it to create a string-like *wire*. Although the mechanical work needed (in terms of ATP or PMF) for the *hairpin* was found to be less than half that of a *wire*, it might involve higher mechanical forces as suggested by our simulations. Furthermore, the *hairpin* runs a higher risk of being stuck during transport because it has a wider diameter than a *wire* (double versus single chain). On the other hand, a *wire* requires lower peak forces but needs more ATP (or a sustained but lower PMF). Hence, unless high mechanical forces can be generated by the export apparatus, a *wire* might be the transport form of flagellin. Through *wire* flagellin might be an almost linear chain, small secondary structures such as β-hairpins (such as those making up HVR domain folding cores, Chapter 6) might also be present and they could be accommodated inside the channel.