SIMON MITTERNACHT

PROTEIN DYNAMICS: AGGREGATION AND MECHANICAL UNFOLDING

PROTEIN DYNAMICS: AGGREGATION AND MECHANICAL UNFOLDING

Simon Mitternacht

LUNDS UNIVERSITET

2009

Thesis for the degree of Doctor of Philosophy

Computational Biology and Biological Physics
Department of Theoretical Physics
Lund University

Thesis advisor: *Anders Irbäck*
Faculty opponent: *Guido Tiana*

To be presented, with the permission of the Faculty of Science of Lund University, for public critiscism in lecture hall F of the Department of Theoretical Physics, on Friday the 17th of April 2009 at 13:30.

Organization	Document name
LUND UNIVERSITY Department of Theoretical Physics Sölvegatan 14A SE–223 62 LUND Sweden	**DOCTORAL DISSERTATION** Date of issue March 2009 Sponsoring organization
Author(s) Simon Mitternacht	

Title and subtitle
Protein dynamics: aggregation and mechanical unfolding

Abstract

The subject of this thesis is protein dynamics. Papers I–IV and VI study either of two different processes: mechanical unfolding and aggregation. Paper V presents a computationally efficient all-atom model for proteins, variants of which are used to perform Monte Carlo simulations in the other papers.

Mechanical unfolding experiments probe properties of proteins at the single molecule level. The only information obtained the experiments is the extension and resisting force of the molecule. We perform all-atom simulations to generate a detailed description of the unfolding process. Papers I and II discuss the mechanical and thermal unfolding of the protein ubiquitin. The principal finding of Paper I is that ubiquitin unfolds through a well-defined pathway and that the experimentally observed non-obligatory unfolding intermediate lies on this pathway. Paper II compares mechanical unfolding pathways with thermal unfolding pathways. In Paper IV we study the mechanical unfolding of the protein FNIII-10 and find that it has three important, mutually exclusive, unfolding pathways and that the balance between the three can be shifted by changing the pulling strength.

Paper III describes oligomerization of six-chain systems of the disease-related peptide Aβ(16-22). We find that disordered oligomers of different sizes dominate at high temperatures and as temperature is lowered, larger, more structured, oligomers form. In particular a very stable β-barrel structure forms. Paper VI is an investigation of the effect of mutations on the folding properties of the peptide Aβ42 from Alzheimer's disease. Small aggregates of this peptide are believed to be important toxic agents. We find that a disease-related mutant peptide, with an elevated aggregation propensity, has a larger conformational diversity than the wild-type peptide, whereas a mutation that is known to inhibit aggregation has the opposite effect.

Key words:
Monte Carlo simulations, Protein aggregation, Mechanical unfolding, All-atom protein models

Classification system and/or index terms (if any):

Supplementary bibliographical information:		Language English
ISSN and key title:		ISBN ISBN 978-91-628-7714-9
Recipient's notes	Number of pages 161	Price
	Security classification	

Distributor
Simon Mitternacht, Department of Theoretical Physics, Sölvegatan 14A, SE–223 62 Lund, Sweden

Sammanfattning

Proteiner är på något sätt inblandade i alla biologiska processer. Därför är det intressant att lära sig mer om dem, både av ren nyfikenhet – för att förstå naturen bättre – men också för att förstå olika sjukdomar.

Proteinmolekyler består av långa kedjor av sammanlänkade aminosyror. Essentiellt för ett proteins biologiska funktion är dess tredimensionella struktur. Denna bestäms av proteinets aminosyrasekvens som i sin tur ges av genen för proteinet. Proteinstrukturer är inte statiska: i många biologiska processer är proteiner dynamiska aktörer som utför t.ex. mekaniskt arbete. Forskningen som presenteras i den här avhandlingen undersöker genom datorsimuleringar proteiners dynamiska egenskaper i två olika processer: mekanisk uppveckning och aggregation. Dessutom beskrivs i en artikel en revision av den modell vi använder i våra datorsimuleringar.

Mekanisk uppveckning kallas den process då ett proteins struktur veckas upp genom att man fysiskt drar i det. Detta är relativt vanligt förekommande i biologiska system. Man tror till exempel att proteinerna som ger muskler deras elasticitet veckas upp när muskler dras ut för att sedan hjälpa till att dra ihop dem genom att veckas tillbaka när spänningen försvinner. En mer generell process där mekanisk uppveckning förekommer är nedbrytning av uttjänta proteiner. Mekanisk uppveckning kan studeras genom experiment, men det går bara att mäta molekylens motståndskraft och hur den förlängs. Simuleringar kan komplementera experimenten med en mer detaljerad beskrivning. Artikel I, II och IV beskriver simuleringar av mekanisk uppveckning av två proteiner. För det ena proteinet finner vi att uppveckningen nästan alltid följer samma väg. I det andra fallet finns tre olika uppveckningsvägar och vi ser att sannolikheten för att en viss väg följs beror på hur hårt man drar.

Aggregation av proteiner är ett framträdande symtom vid ett antal sjukdomar, t.ex. Alzheimers och Parkinsons sjukdomar och typ 2-diabetes. Var och en av sjukdomarna är kopplad till ett visst protein som av någon anledning börjar bilda stabila så kallade amyloida aggregat. Experiment har visat att de stora aggregat man sedan länge känner till antagligen bara är en restprodukt. Istället tror man att mindre aggregat bestående av endast ett fåtal molekyler är de som orsakar problem genom att döda celler. Artikel III beskriver hur sådana aggregat bildas för ett system bestående av sex små molekyler. I artikel VI studerar vi ett av proteinerna som aggregerar i Alzheimers sjukdom. Vissa ärftliga former av sjukdomen är kopplade till mutationer i aminosyrasekvensen för detta protein. Vi undersöker huruvida mutationerna förändrar proteinets veckningsegenskaper och spekulerar i hur detta kan påverka aggregationen.

Publications

The following publications are included in the thesis.

I Anders Irbäck, Simon Mitternacht and Sandipan Mohanty. *Dissecting the mechanical unfolding of ubiquitin.* Proc. Natl. Acad. Sci. USA 102: 13427–13432. (2005)

II Anders Irbäck and Simon Mitternacht. *Thermal versus mechanical unfolding of ubiquitin.* Protiens 65: 759–766. (2006)

III Anders Irbäck and Simon Mitternacht. *Spontaneous β-barrel formation: an all-atom Monte Carlo study of Aβ (16-22) oligomerization.* Proteins 71: 207–214. (2008)

IV Simon Mitternacht, Stefano Luccioli, Alessandro Torcini, Alberto Imparato and Anders Irbäck. *Changing the mechanical unfolding pathway of FnIII-10 by tuning the pulling strength.* Biobhys. J 96: 429–441. (2009)

V Anders Irbäck, Simon Mitternacht and Sandipan Mohanty. *Effective all-atom potential for protein studies.* LU TP 09-01 (Submitted). (2009)

VI Simon Mitternacht, Iskra Staneva, Torleif Härd and Anders Irbäck. *Effects of mutations on the folding of the Alzheimer's Aβ42 peptide.* LU TP 09-04. (2009)

I have also co-authored one publication that is not included in the thesis, during my time as PhD-student.

- Simon Mitternacht, Stefan Schnabel, Michael Bachmann, Wolfhard Janke and Anders Irbäck. *Differences in solution behavior among four semiconductor-binding peptides.* J. Phys. Chem. B 111: 4355–4360. (2007)

Contents

The first question I ask myself when something doesn't seem to be beautiful is why do I think it's not beautiful. And very shortly you discover that there is no reason.

– John Cage

Introduction

The work presented in this thesis investigates protein dynamics in different situations by computer simulations. The two main areas of interest are mechanical unfolding and aggregation. The first refers to the process of disrupting protein structures by mechanical manipulation, usually by pulling at the ends of the protein chain. Mechanical unfolding experiments are interesting for any protein because they probe macromolecules at the single-molecule level. They also allow an unusually straight-forward comparison between experiment and simulations. In a more direct manner, mechanical unfolding studies are relevant for proteins that are subject to mechanical stress in different biological processes.

The second area of interest is protein aggregation, which is a prominent symptom of a number of neuro-degenerative disorders like Alzheimer's, Huntington's and Parkinson's diseases. Each disease is associated with one or a few specific proteins that accumulate in the form of amyloid aggregates in the brains of patients. Aggregation is a difficult process to study because it is non-reversible and out-of-equilibrium. The problem is further complicated by the fact that small transient aggregates – on or off the pathway from free protein monomers to plaques – are believed to be the most important neurotoxic agents in several of the diseases.

The main body of the thesis consists of six articles. In addition to five studies of the above-mentioned topics, one of the articles presents an improvement of the model used in the simulations in the other papers. This introductory chapter gives the uninitiated reader an orientation in both the basics of protein physics and the models and methods used in the included articles. The second half of the introduction gives a background to, and summary of, the research in the different papers.

1.1 *The basics*

In modern science, proteins are studied at many levels of detail. At the most abstract level they are nodes in reaction networks between proteins, DNA, RNA and other smaller molecules involved in cell biology. Proteins can be studied in terms of their three-dimensional structure and how it relates through evolution to the structure of other proteins. The dynamics of proteins are investigated in terms of interactions with other molecules, and the internal dynamics involved in different processes like protein folding or unfolding. At the most detailed level the chemistry of specific groups of atoms are discussed, for example when binding different medical compounds. The research in this thesis focuses on the dynamics of forced unfolding and aggregation. To begin with I will briefly mention some basic properties and concepts.

The basic building blocks of proteins, the amino acids, can interlink to form a polypeptide chain as depicted in Figure 1.1. The amino acids contain a backbone group and a side-chain that has varying properties. The properties of a polypeptide chain are determined by which amino acids it contains. In biological systems inherited sequences of DNA are translated into sequences of amino acids to produce the polypeptide chains we call proteins. The difference between proteins and random polypeptides is that the amino acid sequence of the former has evolved to encode a specific biological function.

The majority of proteins are only biologically active when the chain has folded to a unique three-dimensional structure, the native state. For most proteins this is the thermal equilibrium state, which indicates that all information needed for biological function can be found in the amino acid sequence [1]. In some cases large parts of the protein structure provide a static scaffold to make sure that a few amino acids on the surface are arranged in a given way in order to interact with some specific molecule. In many cases however, proteins are highly dynamic systems. There is for example a number of proteins that are capable of performing mechanical work (see chapters 10 and 11 of reference [2]). There are also many proteins that are natively unfolded [3].

Protein structure is often discussed in terms of primary, secondary, tertiary and quaternary structure. The term primary structure simply refers to the sequence of amino acids. Local structure motifs defined by just a few amino acids are called secondary structure, and are usually associated with specific values of the backbone Ramachandran angles φ and ψ (see Figure 1.1). The most important are α-helices and β-sheets which are stabilized by hydrogen bonds between backbone dipoles of different amino acids as indicated in Figure 1.2. It turns out that not only do these structures allow favorable hydrogen bonds, they also correspond to the values of φ and ψ that best avoid

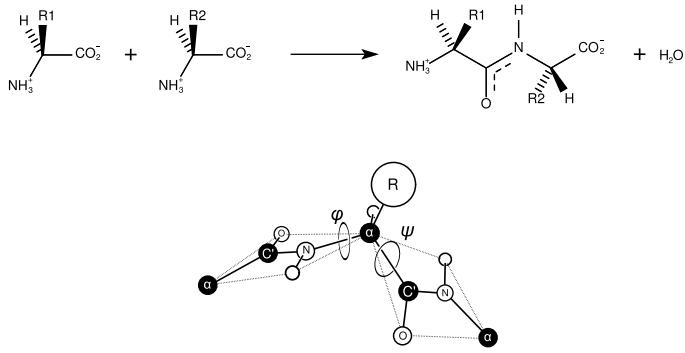

Figure 1.1: The chemical reaction that leads to formation of a polypeptide chain (top) and the polypeptide backbone with its principal degrees of freedom, the Ramachandran angles φ and ψ (bottom). Side chains are represented by the letter R. The planar peptide units (CONH) are indicated by thinner lines in the figure. This particular configuration has $\varphi = -139°$, $\psi = 169°$.

steric clashes [4]. The term tertiary structure is used to describe the relative arrangement of secondary structure elements; a few examples are given in Figure 1.3. Large proteins often consist of several folded domains and the relation between these is called quaternary structure.

The process by which a protein finds its native state, the folding process, can be described in several ways and is different for different proteins. Many proteins display cooperative transitions between the unfolded and the folded state [5, 6], but there are also proteins that have no significant free energy barrier between the two states [7]. A discussion of how a protein can find its native state among the astronomical number of possible conformations in a limited time started in the 1960's and 70's (see for example references [8–10]) and much progress has been made to this date (for review articles see references [7, 11–13]).

1.2 *Protein models*

All protein simulations require a model to represent the protein and its interactions. A common approach is to model every atom as a partially charged particle interacting with other atoms via Coulomb and van der Waals-forces [14]. Chemical bonds are represented as stiff springs that enforce correct bond lengths, bond angles and torsion angles. The potential energy for interactions between atoms can then be written as a sum of simple functions of the

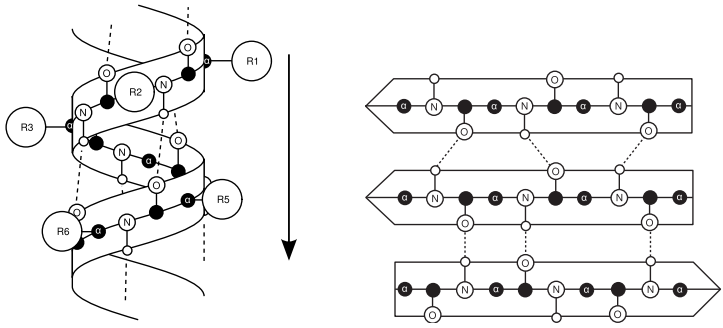

Figure 1.2: Sketch of α-helix (left) and β-sheet (right). The arrows indicate the direction of the chain and the dotted lines hydrogen bonds.

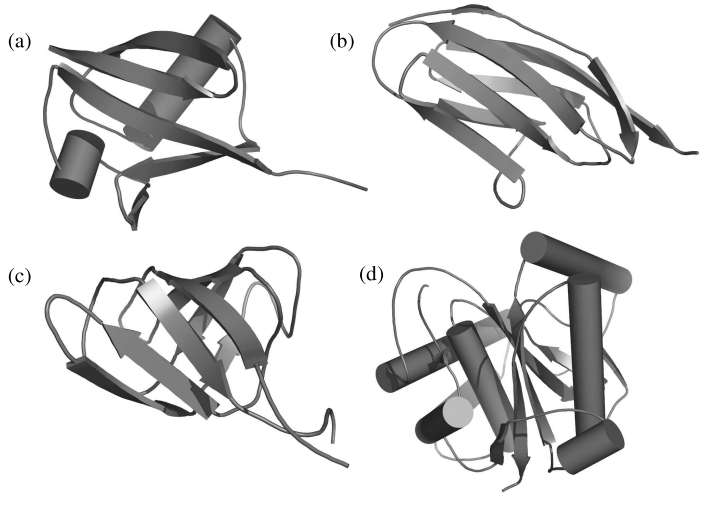

Figure 1.3: Examples of common folds. Helices are represented by cylinders and β-strands by flat arrows. Loops have been smoothed for clarity. (a) $\alpha\beta$-fold (pdb-code 1ubq). (b) β-sandwich (1tit). (c) β-barrel (1d1n). (d) A rare and complicated knotted fold (chain A from 1xd3).

inter-atomic distances. The angle terms will depend on the position of three or four atoms. The molecules of the solvent are often represented explicitly. This approach is intuitively appealing because it gives a simple mathematical form for the potential and all the parameters can in theory be estimated from first principles. On the other hand the parameters are so many that it is very difficult to determine a balanced set [15, 16]. Neither is it clear that a static representation with fixed point charges and harmonic bond potentials is a good approximation. Furthermore, computations with explicit solvent are very demanding because a realistic simulation requires a large number of water molecules. The long range of the Coulomb potential introduces additional difficulties. The high computational cost of all-atom, explicit solvent simulations limits the applicability to studies of very small systems, or, for larger molecules, perturbations around local free energy minima.

At the other end of the spectrum there are the very coarse-grained models. A classic example is the HP model where proteins are modeled as self-avoiding chains on a lattice [17]. Every amino acid is modeled as a single bead on the chain. A bead can either be hydrophobic (H) or polar (P). The model favors configurations with a hydrophobic core by giving H–H contacts a negative energy. There are several models at this level of simplicity that can be used to investigate general properties of proteins [18]. Their simplicity makes it relatively easy to compute global properties like different thermodynamic quantities. More detailed models are needed, however, to describe the properties of specific proteins.

Another class of simplified models is Gō-type models [19]. Here the protein can be modeled either atomistically or in some simplified representation with a reduced number of atoms/beads. The potential consists of some general terms for self-avoidance and covalent bonds, and a term which favors the native state. The last term is constructed such that contacts between atoms present in the native state are given a negative energy if present in a given conformation. The native state is thus, by construction, the ground state of the model. Gō-models can be used to study different questions related to protein folding and unfolding and can provide a good description in cases where the free energy landscape is funnel-shaped, i.e. dominated by the native state. Examples of commonly studied topics are questions regarding the effect of fold topology for the dynamics of folding and unfolding. There are cases where it is necessary to extend the Gō-potential with non-native interactions for folding to occur at all [20].

In view of the computational inefficiency of the available all-atom models and the lack of detail in the coarse-grained models we have developed a new,

computationally efficient all-atom model. It is detailed enough to give what we believe is a realistic description of a number of small peptides, but simple enough to facilitate thermodynamic simulations. It includes all protein atoms, but the solvent is treated implicitly and there are only torsional degrees of freedom. Bond lengths and bond angles are kept fixed. All interactions are short-ranged to allow fast evaluation of the energy function. The model is sequence-based, meaning that no experimental information about a peptide is used, the only input to the model is the amino acid sequence of the molecule under study. Different versions of our model are used in the included publications. The latest version is presented in Paper V.

The parameters of the model are calibrated by a manual time-consuming procedure. The key concepts are backwards compatibility and thermodynamics. The first version of the model was calibrated to give a reasonable description of a set of small peptides. This means that the native structures and thermodynamics were correct for all peptides in the set. A subsequent model revision is only accepted if it still gives a correct description of the initial set of peptides and expands the set with more peptides. By this procedure we aim to develop a model that can describe any peptide given only the sequence. We hope that the model will also be capable of describing larger proteins for which calibration simulations are very costly. Our approach to calibration is time-consuming because every change needs to be tested on a large number of systems to be accepted, and the simulations require extensive sampling of configurational space. As we see it, there really is no better option if one wants to create a general, realistic model.

The two most important attractive interactions in proteins are hydrogen bonds and hydrophobicity. The strength and nature of these two interactions can only be understood when taking solvent entropy into account. In the model presented here the solvent is treated implicitly, which means that any entropic contribution of the solvent to an interaction is modeled as part of a potential energy. In a sense we have averaged out the behavior of the solvent to provide a potential of mean force. A possible complication comes from the fact that the entropic and energetic contributions to the free energy of a state have different temperature dependencies. In principle our implicit solvent model should have temperature dependent energy terms. Experiments have indeed found a non-negligible temperature dependence of the free energy of hydrophobic association in the range 0–100 °C [21, 22]. Omitting these effects, by having constant parameters, is however a convenient approximation that appears to give a reasonable temperature dependence of a number of observables. The thermodynamic properties of many proteins are dominated

by a sharp cooperative folding transition, the temperature range of which is typically much narrower than that of significant variations in the strength of for example hydrophobicity.

The energy function in the model consists of four main terms

$$E = E_{\text{ev}} + E_{\text{loc}} + E_{\text{sc}} + E_{\text{hb}}\,. \tag{1.1}$$

The first term, E_{ev}, represents excluded volume effects, i. e. the fact that two atoms can not be in the same place. The term E_{loc} models local interactions along the protein backbone and within the side-chains. Attraction between hydrophobic side-chains and attraction/repulsion between charged side-chains are described by the term E_{sc}. The last term, E_{hb}, describes hydrogen bonds. In the following discussion of these terms, focus will lie on their physical motivation, rather than the mathematical details (which can be found in Paper V).

Excluded volume. A simple but important component of any atomistic molecular model is that no two atoms can be in the same place. We assign a radius σ_i to each atom based on its element. The excluded volume energy for a pair of atoms is proportional to $(\sigma_i + \sigma_j)^{12}/r_{ij}^{12}$. The term E_{ev} is a sum over all pairs of atoms. It has the simplest functional form of the terms in the potential but is actually the most demanding to calculate because all atoms are involved. In its naïve form the computational cost scales as the number of atoms squared. By introducing a cutoff distance, beyond which the repulsion is zero, computation time instead scales linearly with the size of the system.

Local interactions. The repulsion between atoms makes sure that no parts of the chain overlap, but it also affects which values are allowed for a given torsion angle. Ramachandran et al. showed that steric clashes between atoms limit the allowed values of the backbone torsion angles to only a fraction of the (φ, ψ)-plane [4]. The experimentally measured distributions of these angles agrees relatively well with this picture, but Ramachandran's model, based purely on hard sphere repulsion, does not capture modulations of the distribution due to for example electrostatic interactions between the partially charged atoms of the backbone. In particular the balance between the propensity for α and β secondary structures is shifted. To alleviate this we introduce the term E_{loc} that models the electrostatic interaction between neighboring peptide units along the backbone. Similarly, local steric constraints do not fully explain the observed distributions of side-chain torsion angles, and to get these correct we include an additional term in E_{loc} in which each sub-term is a simple

trigonometric function of the respective angles that suppresses disallowed values.

Hydrogen bonds. A special property of water and ice is the network of hydrogen bonds that binds the molecules together. In general, hydrogen bonds form between a hydrogen atom covalently bound to an electronegative atom like oxygen or nitrogen (donor), and another electronegative atom (acceptor). The two electronegative atoms "compete for the same hydrogen atom" [21]. In water, hydrogen bonds have the structure O–H$\cdots$O. The bonds are directional, and strongest when the three atoms are aligned. Acceptors can bind two donors, but it rarely occurs that a donor binds two acceptors. Each water molecule can thus participate in four hydrogen bonds, which is what happens in ice crystals. Liquid water has the same basic network of hydrogen bonds as ice, but in a less rigid form.

Some polar atom groups in proteins are capable of forming hydrogen bonds, both between themselves and with water. The protein-protein hydrogen bonds are energetically comparable to the corresponding bonds with water molecules and so strong that no dipoles can afford to be left free without binding either water or another protein dipole at physiological temperatures [22]. When a hydrogen bond between different parts of the chain forms, the two water molecules – which were previously bound to the respective protein dipoles – are freed and will each bind to other water molecules instead. In total, the number of hydrogen bonds is thus preserved, but the water molecules gain translational and rotational entropy, which makes the reaction favorable [22]. Hydrogen bonds between the NH and CO groups of the backbone are particularly important for example for stabilizing secondary structure (see Figure 1.2).

In our implicit solvent model, intra-chain hydrogen bonds are represented by a potential well. To generate a negative contribution to E_{hb} the involved dipoles need to be aligned and the distance between donor and acceptor atoms in a given range. The depth of the potential does not correspond to the electrostatic energy of a hydrogen bond in vacuum, the strength instead reflects the change in free energy of the combined protein-solvent system.

Side-chain interactions. As we have seen molecules other than water that have acceptor or donor groups can take part in the hydrogen bond network of water without disrupting it too much. Non-polar molecules, on the other hand, will not interact favorably with water and the water molecules need to rearrange and be fixated to accommodate a non-polar molecule. At least for

small non-polar molecules, this rearrangement usually has no energetic cost, but the loss of entropy from forming a rigid cage of water is large [22]. The cost of inserting a non-polar molecule into water increases with the surface area because the number of locked water molecules increases. As a result, non-polar molecules will stick together to minimize their water-interface, which is what happens when for example water and oil are mixed. Because non-polar molecules avoid water as much as possible they are called hydrophobic. Hydrophobic amino acid side-chains are typically found in the center of a folded protein, shielded from water, and polar amino acids at the surface.

The first part of the term E_{sc} in our model models the hydrophobic effect in the form of a pairwise attraction between hydrophobic side-chains. The interaction energy for each pair is a product of two factors: the first measures the number of side-chain atoms in contact, the second is a constant parameter that sets the overall strength of the interaction for that particular pair of amino acid types.

Although it is a totally different interaction, it turns out that we can model the interaction between charged atom groups using the same functional form as for the hydrophobic attraction and this constitutes the second half of the term E_{sc}. In reality this interaction probably has very specific requirements on geometry, including the orientation of the surrounding water molecules. Since we do not know these details, and we have no explicit water molecules, a reasonable approximation is to use the more "fuzzy" geometry of our hydrophobicity term. To model the high dielectric constant of water the strength of this interaction is set relatively weak.

1.3 *Monte Carlo methods*

Once a model has been defined it needs to be evaluated by some integration method. There are basically two approaches, either one solves Newton's equations of motion for the system (probably modified with a thermostat and/or barostat), or the Boltzmann distribution is sampled by some Monte Carlo scheme. As the title of this section implies all the work presented here uses different Monte Carlo methods, based on the Metropolis-Hastings algorithm [23, 24]. The algorithm was first proposed by Metropolis et al. in 1953 [23] to sample the Boltzmann distribution for a relatively simple system. Hastings provided a more general formulation for sampling any probability distribution in 1970 [24]. The basic idea is to generate a discrete Markov chain based on a transition probability p_{ij}. In each step of the chain, an update of the system from one state to another is drawn from a distribution q_{ij} and

then accepted with a probability α_{ij}, giving $p_{ij} = q_{ij}\alpha_{ij}$. If the update isn't accepted the original state is counted twice. Hastings showed how to construct a q_{ij}-dependent acceptance probability α_{ij} such that a desired distribution π_i is the stationary distribution of p_{ij} [24].

A common application of the algorithm is to sample the Boltzmann distribution, or some variant thereof. Typically, a model in statistical mechanics is formulated in terms of an energy function. Hence, the energy and Boltzmann weight of any given state in the model is known. The Metropolis algorithm is used to integrate over the microstates of the model with proper weights to measure macroscopic quantities like averages and distributions of different observables. An important feature of the Metropolis algorithm is that unknown normalizing factors of π_i cancel in p_{ij}. This property is of great advantage in statistical mechanics because simulations can be done without knowing the partition function.

The original form of the algorithm works well for systems with smooth free energy landscapes without large barriers, or for problems where one is only interested in the surroundings of a certain free energy minimum. The performance is however relatively poor for systems with high free energy barriers. Several methods have been devised that expand the Metropolis algorithm in different ways to address this and other issues [25–28]. Here I will present the method simulated tempering [29, 30], which is used in Papers III, V and VI. The standard Metropolis algorithm applied on the Boltzmann distribution is used in Papers I, II and IV to investigate non-equilibrium properties of two different proteins.

Simulated tempering. In simulated tempering the system is allowed to dynamically jump between the temperatures of a predefined set $\{T_m\}$. Because free energy barriers are generally lower at high temperatures, one can avoid getting stuck in local minima by changing temperature. As a bonus, measurements can be done at several temperatures in the same simulation and, as will be shown, the difference in free energy between temperatures can also be measured. To ensure mobility in temperature space a set of tunable parameters, $\{g_m\}$, is introduced. Using the Metropolis algorithm one explores the joint distribution $P(x,m) \propto \exp(-E(x)/k_\mathrm{B}T_m + g_m)$, for states x and temperatures T_m. The conditional distribution $P(x|m)$ is equivalent to the Boltzmann distribution at the temperature T_m. The marginal probability $P(m)$ can be written in terms of the canonical partition function,

$$Z_m = \sum_x \exp(-E(x)/k_\mathrm{B}T_m), \tag{1.2}$$

as $P(m) \propto Z_m \exp(g_m)$. Using the free energy $F_m = -k_B T_m \ln(Z_m)$ we find

$$P(m) \propto \exp(-F_m / k_B T_m + g_m). \tag{1.3}$$

This means that choosing $g_m = F_m / k_B T_m$ will give a uniform $P(m)$. Sampling all temperatures equally is desirable: we will stay long enough in a free energy minimum to explore some of its details, but jump in temperature often enough to explore many minima. The free energies F_m are not known beforehand (if we could calculate them analytically we probably would not need to do any simulations), but the g_m can be tuned by an iterative procedure: Start from an initial guess and measure $P(m)$ in a short simulation. The free energies can then be estimated from Equation 1.3 and the parameters g_m adjusted accordingly. Reiterate this procedure until a satisfactory $P(m)$ is obtained.

Monte Carlo updates. A well-balanced choice of the proposal distribution q_{ij}, for generating updates in the Monte Carlo step, is important for efficiency. If the updates are too small, too many steps are required to move between distant configurations. If they are too large very few updates will be accepted because most will likely lead to low-probability states. Obviously the updates must also provide ergodicity, that is, allow the system to visit all parts of conformational space. In our protein simulations we use five different updates: The two simplest updates are rigid body rotations and translations, which are important when there is more than one protein in the simulation. Another simple update is rotation of single side-chain torsion angles, which only affects a few atoms in the corresponding side-chain. Updates of single backbone torsion angles are also used. This update can give rise to dramatic conformational changes because it modifies the relative orientation of two large segments of the chain. To allow fine-tuning of local backbone structure without disrupting the rest of the structure a fifth update called Biased Gaussian Steps (BGS) is introduced [31]. This update rotates up to seven or eight consecutive backbone angles in a manner that keeps the end points of the corresponding backbone segment approximately fixed. The probability for suggesting these five updates is proportional to the total number of degrees of freedom affected by the update. At high temperatures, only the single angle update is used to update the backbone, at low temperatures only BGS, and at intermediate temperatures both. In simulated tempering simulations temperature is updated in single steps up or down. The probability of suggesting a temperature update is relatively low.

In Papers I, II and IV the Metropolis algorithm is used to simulate non-equilibrium unfolding. The generated Markov chains are interpreted as time series and compared with experimentally measured unfolding trajectories.

Comparisons between Langevin and Monte Carlo dynamics have shown that large scale movements that are determined by the general shape of the free energy landscape can be described well by Monte Carlo dynamics [32], provided that each update is small [33]. To ensure realistic dynamics in the unfolding simulations, the backbone is only updated using the local BGS move.

1.4 *Mechanical unfolding*

In mechanical unfolding experiments a single molecule is stretched by pulling at its ends. Two different experimental techniques have been used to study proteins in this respect, atomic force microscopy (AFM) and optical tweezers [34]. Many of the proteins studied are multidomain chains consisting of several independently folding units. When single-domain proteins are studied the sequence is often multiplied to produce a larger molecule consisting of several identical domains. The force response of such a tandem protein is periodic, providing a clear identification signal of unfolding events stemming from the correct molecule. The most common protocol in mechanical unfolding experiments is to pull at a constant velocity. If the domains have any mechanical resistance, force will increase as the molecule is stretched, until one of the domains breaks. This is often a sudden event that unfolds the domain fully or partially. A plot of force versus time for multidomain molecules thus typically displays a saw-tooth pattern. The extension of the molecule in each unfolding event is measured to determine whether there was full unfolding, or unfolding to an intermediate. It is also possible to perform experiments at constant force (see for example reference [35]).

Studying the mechanical response of a protein is particularly interesting because it can be done at the single molecule level. This gives unambiguous and independent data on the dynamical properties of the molecule. It is a way to measure the strength of the interactions that stabilize proteins in a very direct manner and can give information about structural and folding properties [36]. Of more direct biological relevance is the relation of these experiments to processes that involve mechanically induced unfolding. Certain proteins have functions directly linked to their force-response, like fibronectin of the extracellular membrane, or the muscle protein titin. Translocation through membranes and protein degradation are examples of two generic processes that involve mechanical unfolding by pulling at one end of the protein [37]. Interestingly, the protein UCH-L3 utilizes a tightly knotted fold to provide mechanical resistance against degradation [38].

Mechanical unfolding is a process that lends itself particularly well to simulation studies for several reasons. The process is essentially non-equilibrium with given start and end points, and therefore does not require complete sampling of conformational space to achieve relevant results. This means that one is also able to study relatively large proteins. The experiments only measure end-to-end distance and force as function of time – it has not been possible to obtain any conformational data from these experiments. Simulations, on the other hand, generate a fully atomistic description of the process, revealing details not accessible in experiments. In addition, the simulated force-extension profiles can be compared with experimental data in a straight-forward manner. Simulation studies have so far focused on identifying important stabilizing structures and characterizing unfolding pathways and intermediates. In the majority of previous studies very large forces and pulling velocities were used to speed up the all-atom explicit solvent simulations. Furthermore, only a few trajectories were generated, prohibiting a quantitative analysis. In contrast, the computational efficiency of our protein model allows us to produce a large number of unfolding trajectories at experimental values of the force.

1.5 *Protein aggregation*

Anyone who has boiled an egg has experienced the irreversible aggregation of denatured proteins. As temperature increases proteins unfold and expose their hydrophobic amino acids to the solvent. If the concentration of proteins is high enough the proteins will form amorphous hydrophobic clusters to reduce this exposure. When the denaturing conditions are neutralized the proteins are all tangled up in aggregates and unable to fold back to their native structures. Proteins that are slow to fold or unstable may aggregate also under physiological conditions or under small variations of their environment [39, 40]. To prevent this, there are cellular mechanisms both to guide slow-folding proteins and to dissolve already formed aggregates [40].

Particularly interesting are the amyloid aggregates [41] which, in addition to hydrophobicity, are stabilized by hydrogen bonds between the chains. In amyloid aggregates each molecule forms one or several strands in a multiprotein β-sheet. As the sheet extends it can eventually form long amyloid fibrils [42]. Amyloid aggregation is normally associated with disease but there are also functional amyloid structures in nature [39], one example being spider web. Amyloid diseases include Type-II diabetes, and Alzheimer's, Huntington's and Parkinson's diseases. Each syndrome is associated with one or a few specific proteins that aggregate to form amyloid fibrils [39, 43]. Initially, the fibrils

were thought to be the primary toxic agents in these diseases [44], a notion supported by the finding that many familial forms of amyloidosis are carried by mutations in either the genes coding the aggregating protein, or genes involved in the production or clearance thereof [45–48].

Recent research indicates that small oligomers, containing only a few copies of the protein, are the toxic species in several of the amyloid diseases [39, 49, 50]. A proposed, simple explanation for the elevated toxicity of oligomers compared to fibrils is the larger surface to volume ratio, which implies an increased exposure of groups normally buried in folded proteins [39]. A more specific proposal is that they form pores in the cell membrane, disrupting the chemical balance between the cell and its environment [50]. A solution of amyloid-competent proteins will be a mixture of oligomers, fibrils and free monomers at the same time. Because of their transient nature and heterogeneity it is difficult to isolate and study a given class of oligomer experimentally. Several studies have identified specific disease-related oligomers [51, 52], but a detailed characterization of them has not been possible yet.

1.6 *The articles*

The research articles presented in the main part of the thesis can be divided into three groups. Three of them treat the mechanical unfolding of ubiquitin and a fibronectin domain (Papers I, II and IV). Paper III discusses aggregation of a small peptide and Paper VI the folding properties of a peptide that aggregates in Alzheimer's disease. Variants of the same model are used in all papers, and Paper V presents the newest revision of this model. Here I will discuss the motivation for the research in the papers, its relation to the work of others, and give a brief summary of the results.

Model revision. The physical motivation for the model presented in Paper V was discussed in section 1.2. The paper presents a revision of an older version of the model which was used in Papers I–IV. The biggest changes are the addition of an interaction between charged side-chains and a major modification of the local interaction represented by E_{loc} in Equation 1.1. The model was developed by comparing results from simulations with experimental thermodynamical data for a number of peptides. In the end, the energy scale of the model is defined by comparing the simulated thermodynamics of the Trp-cage peptide with experimental data: we define the temperature at the peak in specific heat to be equal to the experimentally determined melting temperature, which fixates the last free parameter of the model.

The article presents results for folding simulations of 17 small peptides and 3 larger systems. We find that the model predicts native structures that are in good agreement with those determined experimentally for all 20 systems. Also, in the cases where there is data to compare with, the thermodynamics in our simulations agree well with experiment. At the end, we give examples of two short sequences that our model fails to describe correctly.

As mentioned in section 1.2, an important motivation for developing this model was the absence of a detailed, realistic *and* computationally efficient model. As an illustration of the speed of our simulations it is worth mentioning the largest system studied in Paper V: the 67 amino acid three-helix bundle protein GS-α_3W. On average, using a single, standard CPU we generate one independent folding event per day for this system.

Mechanical unfolding. In Paper I the mechanical unfolding of ubiquitin is investigated. Ubiquitin is, among other things, involved in labeling other proteins for degradation by the proteasome [53]. It is therefore perhaps not surprising that it is relatively stable against pulling at the ends [54]. Constant force AFM unfolding experiments found that ubiquitin can unfold in two ways, either directly in one large step or via an intermediate in two smaller steps [55]. In our paper we describe mechanical unfolding simulations for the three different force magnitudes studied in the experiment. We find, in agreement with experiments, that there are two major types of unfolding events in our simulations: direct, or via an intermediate. The end-to-end length of the intermediate state in the simulations agrees with that found in the experiment. The structure of the intermediate is well-defined. We find that the importance of the intermediate increases as force is lowered. Furthermore, we identify a single unfolding pathway that is followed in a majority of both one- and two-step unfolding events – the difference between the two is only whether the system halts during the process. The pathway we see can be understood by simple geometrical considerations: the sections of the chain that are first exposed to force are the first to break. Our predictions have later been supported by simulations by other groups [56–58].

The predicted mechanical unfolding pathway for ubiquitin does not agree with the thermal unfolding pathway that can be deduced from experiments measuring the thermal stability of different structure elements [59, 60] and folding pathways [61, 62]. To verify that our model gives a correct description of thermal unfolding, we performed the simulations presented in Paper II. We find that thermal unfolding displays a larger variation in unfolding pathways than mechanical unfolding. The dominating pathways are distinct from our

predicted mechanical unfolding pathway and in agreement with those proposed based on experimental data. Our predictions regarding mechanical unfolding are strengthened by the fact that our model gives a different but correct description of an unrelated unfolding process.

In Paper IV we study the mechanical unfolding of the tenth type III module of fibronectin (FnIII-10). Fibronectin molecules can associate to form stabilizing fibrillar structure in the extracellular matrix in response to mechanical tension [63]. Like ubiquitin, fibronectin unfolds via intermediate states. Both experimental [64] and computational studies [65, 66] have indicated that there are several intermediates of practically indistinguishable end-to-end distance. Our simulations support this view. We find three different intermediates that are associated with specific, mutually exclusive unfolding pathways. All three intermediates have similar end-to-end distance. The observed unfolding pathways can be rationalized in the same way as for ubiquitin in Paper I by simply studying the topology of the folded structure. We perform both constant force and constant velocity simulation and find that at low force/velocity the three pathways have comparable probabilities. When pulling strength is increased one pathway comes to dominate and unfolding is nearly deterministic. This variability in response suggests a potential mechanism for the cell to translate different force magnitudes into distinct chemical signals. In addition, we show that given enough resolution it would be possible to distinguish between the three intermediates in an experiment. The different states actually display slight but significant differences in their end-to-end distance. If possible, such an experiment would provide a valuable test of our predictions.

The unfolding forces we apply in our constant force simulations, and observe in constant velocity simulations, are equal to or smaller than those seen in experiments but are potentially much larger than physiological values. We use a version of Jarzynski's equality [67] to estimate the global free energy as function of extension based on the work performed in constant-velocity unfolding. Having determined the free energy landscape, we are able to investigate what happens at forces smaller than those simulated. The reduction of the multidimensional configurational space to a one-dimensional free energy curve obscures some important properties, like the height of the unfolding barrier, but gives a rough estimate of the force magnitudes at which unfolding becomes important.

A general conclusion from these three papers is the fact that a strong pulling force dictates one or a few well-defined unfolding pathways which can be deduced by simple topological considerations. Earlier research has also addressed this question, but computational cost has prohibited quantitative

analysis using transferable all-atom models. The non-trivial dependence of folding pathway on pulling strength, investigated in Paper IV, is a novel finding made possible by the large number of trajectories analyzed.

Aggregation. One of the proteins aggregating in Alzheimer's disease is called amyloid β (Aβ) and has 39–43 amino acids. It is produced by cleavage of the larger membrane protein APP [44]. Certain mutations in APP and also in the genes of the cleaving complex, are known to increase the risk of developing the disease [46]. In vitro experiments have shown that disease-related mutations of Aβ aggregate more readily (see for example reference [68]). The central hydrophobic core of Aβ at positions 17–21 is crucial for aggregation. A mutation of the hydrophobic phenylalanine at position 20 to negatively charged glutamic acid (F20E), for example, reduces the aggregation propensity of Aβ dramatically [68]. In addition, the small peptide formed by the amino acids at positions 16–22 of Aβ (Aβ_{16-22}) is capable of forming amyloid fibrils [69]. This peptide is a convenient model system because the molecule is very small but still has many of the features of full-length Aβ. It has been studied intensely both experimentally and by computer simulations.

The aim of Paper III is to investigate dynamics and structural preferences of Aβ_{16-22} oligomer formation. We perform equilibrium simulations of six peptides in a box with periodic boundary conditions. All simulations start from random configurations. At high temperatures the monomers are essentially free, at intermediate temperatures disordered oligomers of all sizes are found, and at low temperature ordered β-sheet aggregates containing four to six chains dominate, including a particularly stable β-barrel. This barrel is interesting not only because it is the most stable structure seen, but also because it is closed and cannot grow further by monomer addition which means it can be of particular relevance also in systems containing more than six chains. Barrels have been proposed as candidates for the pore-forming oligomers that are believed to be toxic in various amyloid diseases [70, 71]. The barrel seen in our simulations is too small to be able to form a membrane-spanning pore, but our findings indicate that the β-barrel is a stable and accessible oligomeric state for an amyloid-forming peptide.

The work in Paper VI is inspired by the connection between disease-related mutations of Aβ and its aggregation properties. We investigate how the folding properties of monomeric Aβ are affected by three different mutations. The principal finding of the paper is that the conformational diversity and stability of a bend centered at residues 23–26 has a correlation with experimentally measured aggregation propensity. We find that the mutation F20E, which

has low aggregation propensity [68], has a more stable and well-defined conformation in this region than the wild type peptide. The arctic mutant E22G on the other hand, which increases aggregation propensity – and is associated with early-onset Alzheimer's disease – displays a larger conformational diversity in the bend than both wild type and F20E. The fourth peptide, the double-mutant E22G/I31E, forms fibrils as readily as E22G *in vitro*, but seems to do so without passing through potentially toxic oligomeric states [68]. In our simulations this peptide has a reduced probability of forming the bend around residues 23–26, and displays a more flexible folding landscape than E22G. This behavior might indicate a qualitatively different aggregation mechanism for this peptide. The specific role of the observed folding properties for aggregation will be the subject of future simulation studies of systems containing more than one peptide.

List of contributions. The following lists my contributions to the included papers. The research was co-designed by me and the collaborators in all projects. In all projects I took part in writing the text of the articles.

I I performed and analyzed an initial set of simulations in the context of my master's project. When we wrote the paper the simulations were redone by Sandipan Mohanty, using a slightly different protocol. The paper is not simply a reproduction of my master's thesis but contains an extended analysis with new results and a clearer presentation.

II I ran all simulations and did the data analysis.

III I ran all simulations and did the data analysis.

IV I ran all of the all-atom simulations. I performed all analyses except the Jarzynski-part and the worm-like chain fits.

V The revision of the model is the result of three years of continuous work, done primarily by Anders Irbäck, and discussions between me and him. In the end I performed and analyzed all the production simulations for the small peptides.

VI The simulations were done by me. I performed the analysis of accessible surface areas and secondary structure content.

Acknowledgments

A PhD thesis only has one name on the cover, but would of course be impossible to produce without the contributions of a number of people.

First and foremost I would like to thank my supervisor Anders Irbäck for intensive guidance, support and inspiration throughout the more than four years we have worked together. You have never been too busy to listen to any of my (not always well thought out) questions and ponderings. It has been an honor to work closely with such a devoted yet down-to-earth scientist.

I also wish to express my utmost gratitude to Hanna, my family and my friends for love and support.

The open atmosphere in the Computational Biology and Biological Physics group means that help is never far away and I am grateful for all the enlightening discussions throughout the years. Working in this environment has taught me to focus on the big picture and not get lost in the theoretical details that could easily lure people of our disposition into scientific dead-ends. A big thank you goes to Sandipan Mohanty for inspiring discussions and lots of tech support.

Work is not only work, and a number of people have made my period at the department pleasant. Thank you Pontus for the indecent amount of time in the coffee room, and innumerous lunches, discussing everything and nothing. I have appreciated it greatly and I already miss it. Thank you Carl, Michael and Patrik for having made day to day life at the department a little brighter, and Caroline and Marianne for all the non-geeky lunch conversations. Stefano, your short but intense visits to Lund and Malmö have been a true pleasure – the green room has a standing reservation for you.

Lastly, I want to thank Iskra Staneva and Anders for their critical reading of this introduction.

References

1. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230.
2. Nelson P (2004) Biological physics: energy, information, life. WH Freeman and Company, New York.
3. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6:197–208.
4. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain. J Mol Biol 7:95–99.
5. Privalov PL (1989) Thermodynamic problems of protein structure. Annu Rev Biophys Biophys Chem 18:47–69.
6. Jackson SE, Fersht AR (1991) Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. Biochemistry 30:10428–10435.
7. Muñoz V (2007) Conformational dynamics and ensembles in protein folding. Annu Rev Biophys Biomol Struct 36:395–412.
8. Tsong TY, Baldwin RL, McPhie P, Elson EL (1972) A sequential model of nucleation-dependent protein folding: Kinetic studies of ribonuclease a. J Mol Biol 63:453–469.
9. Wetlaufer DB (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci USA 70:697–701.
10. Karplus M, Weaver DL (1976) Protein-folding dynamics. Nature 260:404–406.
11. Mirny L, Shakhnovich E (2001) Protein folding theory: from lattice to all-atom models. Annu Rev Biophys Biomol Struct 30:361–396.
12. Oliveberg M, Wolynes PG (2005) The experimental survey of protein-folding energy landscapes. Quarterly Reviews of Biophysics 38:245–288.
13. Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The protein folding problem. Annu Rev Biophys 37:289–316.
14. Ponder JW, Case DA (2003) Force fields for protein simulation. Adv Protein Chem 66:27–85.
15. Yoda T, Sugitab Y, Okamotoa Y (2004) Comparisons of force fields for proteins by generalized-ensemble simulations. Chem Phys Lett 386:460–467.
16. Best RB, Buchete N, Hummer G (2008) Are current molecular dynamics force fields too helical? Biophys J 95:L07–L09.
17. Lau KF, Dill KA (1989) A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules 22:3986–3997.
18. Tozzini V (2005) Coarse-grained models for proteins. Curr Opin Struct Biol 15:144–50.
19. Gō N (1983) Theoretical studies of protein folding. Annu Rev Bioph Bioeng 12:183–210.
20. Wallin S, Zeldovich KB, Shakhnovich EI (2007) The folding mechanics of a knotted protein. J Mol Biol 368:884–893.
21. Creighton TE (1993) Proteins: structures and molecular properties. WH Freeman and Company, New York.
22. Finkelstein AV, Ptitsyn OG (2002) Protein physics: a course of lectures. Academic Press.
23. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092.

24. Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.
25. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680.
26. Swendsen RH, Wang JS (1986) Replica Monte Carlo simulation of spin glasses. Phys Rev Lett 57:2607–2609.
27. Berg BA, Neuhaus T (1991) Multicanonical algorithms for first order phase transitions. Phys Lett B 267:249–253.
28. Wang F, Landau DP (2001) Efficient multiple range random walk algorithm to calculate density of states. Phys Rev Lett 86:2050.
29. Lyubartsev AP, Martsinovski AA, Shevkunov SV, Vorontsov-Velyaminov PN (1992) New approch to Monte Carlo calculation of the free energy: Method of expanded ensembles. J Chem Phys 96:1776–1783.
30. Marinari E, Parisi G (1992) Simulated tempering: a new Monte Carlo scheme. Europhys Lett 19:451–458.
31. Favrin G, Irbäck A, Sjunnesson F (2001) Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. J Chem Phys 114:8154–8158.
32. Rey A, Skolnick J (1991) Comparison of lattice Monte Carlo dynamics and Brownian dynamics folding pathways of α-helical hairpins. Chem Phys 158:199–219.
33. Tiana G, Sutto L, Broglia RA (2007) Use of the Metropolis algorithm to simulate the dynamics of protein chains. Physica A 380:241–249.
34. Neuman KC, Nagy A (2008) Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. Nat Methods 5:491–505.
35. Fernandez JM, Li H (2004) Force-clamp spectroscopy monitors the folding trajectory of a single protein. Science 303:1674–1678.
36. Forman JR, Clarke J (2007) Mechanical unfolding of proteins: insights into biology, structure and folding. Curr Opin Struct Biol 17:58–66.
37. Prakash S, Matoushek A (2004) Protein unfolding in the cell. Trends Biochem Sci 29:593–600.
38. Virnau P, Mirny LA, Kardar M (2006) Intricate knots in proteins: function and evolution. PLOSCB 2:1074–1079.
39. Chiti F, Dobson CM (2006) Protein misfolding, functional amyloid and human disease. Annu Rev Biochem 75:333–366.
40. Liberek K, Lewandowska A, Ziętkiewicz S (2008) Chaperones in control of protein disaggregation. EMBO J 27:328–335.
41. Sipe JD, Cohen AS (2000) Review: History of the amyloid fibril. J Struct Biol 130:88–98.
42. Nelson R, Eisenberg D (2006) Structural models of amyloid-like fibrils. In: Kajava A, Squire JM, Parry DAD, editors, Fibrous Proteins: Amyloids, Prions and Beta Proteins, Academic Press, volume 73 of *Advances in Protein Chemistry*. pp. 235 – 282.
43. Selkoe DJ (2003) Folding proteins in fatal ways. Nature 426:900–904.
44. Hardy J, Selkoe DJ (2002) The amyloid hypothesis of Alzheimer's disease. Science 297:353–356.
45. Martin JB (1999) Molecular basis of the neurodegenerative disorders. N Engl J Med 340:1970–1980.
46. Chai CK (2007) The genetics of Alzheimer's disease. Am J Alzheimers Dis Other Demen 22:37–41.

47. Biskup S, Gerlach M, Kupsch A, Reichmann H, Riederer P, et al. (2008) Genes associated with Parkinson syndrome. J Neurol 255:8–17.
48. Imarisio S, Carmichael J, Korolchuk V, Chen CW, Saiki S, et al. (2008) Huntington's disease: from pathology and genetics to potential therapies. Biochem J 412:191–209.
49. Kirkitadze MD, Bitan G, Teplow DB (2002) Paradigm shifts in Alzheimer's disease and other neurodegenerative disorders: the emerging role of oligomeric assemblies. J Neurosci Res 69:567–577.
50. Lashuel HA, Lansbury Jr PT (2006) Are amyloid diseases caused by protein aggregates that mimic bacterial pore-forming toxins? Q Rev Biophys 39:167–201.
51. Lesné S, Koh MT, Kotilinek L, Kayed R, Glabe CG, et al. (2006) A specific amyloid-β protein assembly in the brain impairs memory. Nature 440:352–356.
52. Shankar GM, Li S, Mehta TH, Garcia-Munoz A, Shepardson NE, et al. (2008) Amyloid-β protein dimers isolated directly from Alzheimer's brains impair synaptic plasticity and memory. Nat Med 14:837–842.
53. Chau V, Tobias JW, Bachmair A, Marriott D, Ecker DJ, et al. (1989) A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein. Science 243:1576–1583.
54. Carrion-Vazquez M, Li H, Lu H, Marszalek PE, Oberhauser AF, et al. (2003) The mechanical stability of ubiquitin is linkage dependent. Nat Struct Biol 10:738–743.
55. Schlierf M, Li H, Fernandez JM (2004) The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques. Proc Natl Acad Sci USA 101:7299–7304.
56. Kleiner A, Shakhnovich E (2007) The mechanical unfolding of ubiquitin through all-atom Monte Carlo simulation with a Gō-type potential. Biophys J 92:2054–2061.
57. Li MS, Kouza M, Hu CK (2007) Refolding upon force quench and pathways of mechanical and thermal unfolding of ubiquitin. Biophys J 92:547–561.
58. Imparato A, Pelizzola A (2008) Mechanical unfolding and refolding pathways of ubiquitin. Phys Rev Lett 100:158104.
59. Cordier F, Grzesiek S (2002) Temperature-dependence of protein hydrogen bond properties as studied by high-resolution NMR. J Mol Biol 317:739–752.
60. Chung HS, Khalil M, Smith AW, Ganim Z, Tokmakoff A (2005) Conformational changes during the nanosecond-to-millisecond unfolding of ubiquitin. Proc Natl Acad Sci USA 102:612–617.
61. Went HM, Benitez-Cardoza CG, Jackson SE (2004) Is an intermediate state populated on the folding pathway of ubiquitin? FEBS Lett 567:333–338.
62. Krantz BA, Dothager RS, Sosnick TR (2004) Discerning the structure and energy of multiple transition states in protein folding using ψ-analysis. J Mol Biol 337:463–475.
63. Vogel V (2006) Mechanotransduction involving multimodular proteins: converting force into biochemical signals. Annu Rev Biophys Biomol Struct 35:459–488.
64. Li L, Huang HHL, Badilla CL, Fernandez JM (2005) Mechanical unfolding intermediates observed by single-molecule force spectroscopy in a fibronectin type III module. J Mol Biol 345:817–826.
65. Paci E, Karplus M (1999) Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulation. J Mol Biol 288:441–459.
66. Gao M, Craig D, Vogel V, Schulten K (2002) Identifying unfolding intermediates of FnIII-10 by steered molecular dynamics. J Mol Biol 323:939–950.

67. Jarzynski C (1997) Nonequilibrium equality for free energy differences. Phys Rev Lett 78:2690–2693.
68. Luheshi LM, Tartaglia GG, Brorsson A, Pawar AP, Watson IE, et al. (2007) Systematic in vivo analysis of the intrinsic determinants of amyloid β pathogenicity. PLoS Biol 5:e290.
69. Balbach JJ, Ishii Y, Antzutkin ON, Leapman RD, Rizzo NW, et al. (2000) Amyloid fibril formation by Aβ(16-22), a seven-residue fragment of the Alzheimer's β-amyloid peptide, and structural characterization by solid state NMR. Biochemistry 39:13748–13759.
70. Durrel SR, Guy HR, Arispe N, Rojas E, Pollard HB (1994) Theoretical models of the ion channel structure of amyloid β-protein. Biophys J 67:2137–2145.
71. Jang H, Zheng J, Nussinov R (2007) Models of β-amyloid ion channels in the membrane suggest that channel formation in the bilayer is a dynamic process. Biophys J 93:1938–1949.

PAPER I

Dissecting the mechanical unfolding of ubiquitin

Anders Irbäck[1], Simon Mitternacht[1] and Sandipan Mohanty[1,2]

[1]Computational Biology and Biological Physics, Department of Theoretical Physics, Lund University, Sweden. [2]Currently at Institute of Advanced Simulations, Jülich Supercomputing Center, Forschungszentrum Jülich, Germany.

Proc. Natl. Acad. Sci. USA 102: 13427–13432. (2005)

The unfolding behavior of ubiquitin under the influence of a stretching force was recently investigated experimentally by single-molecule constant-force methods. Many observed unfolding traces had a simple two-state character, whereas others showed clear evidence of intermediate states. Here we use Monte Carlo (MC) simulations to investigate the force-induced unfolding of ubiquitin at the atomic level. In agreement with experimental data, we find that the unfolding process can occur either in a single step or via intermediate states. In addition to this randomness, we find that many quantities, such as the frequency of occurrence of intermediates, show a clear systematic dependence on the strength of the applied force. Despite this diversity, one common feature can be identified in the simulated unfolding events, which is the order in which the secondary-structure elements break. This order is the same in two- and three-state events, and at the different forces studied. The observed order remains to be verified experimentally but appears physically reasonable.

Introduction

The 76-residue protein ubiquitin fulfills many important regulatory functions in eukaryotic cells through its covalent attachment to other proteins [1, 2]. In many cases, the ubiquitin tag consists of a chain of ubiquitin domains (polyubiquitin), which is formed by linkages between an exposed lysine side chain (Lys11, Lys29, Lys48 or Lys63) of the last ubiquitin of a growing chain and the C terminus of a new ubiquitin. The fate of a polyubiquitin-tagged protein depends on the linkage. For example, Lys48-C linked polyubiquitin marks the protein substrate for proteasomal degradation [3].

Recently, Fernandez and coworkers [4–7] and Chyan et al. [8] investigated the mechanical properties of polyubiquitin by single-molecule force spectroscopy. It was shown that Lys48-C linked as well as end-to-end (N-C) linked polyubiquitin can withstand a stretching force; the average unfolding force was 85 pN for Lys48-C linkage and about 200 pN for N-C linkage [4]. In these experiments, the polyubiquitin chains were pulled with a constant velocity. In another experiment on N-C linked polyubiquitin, the stretching force was kept constant [6]. At constant force, the fraction of unfolded ubiquitin domains was found to show an approximately single-exponential time dependence, as expected if the unfolding of individual domains is a simple Markovian two-state process. Nevertheless, the unfolding of individual domains sometimes occurred via intermediate states. The precise nature of these different unfolding pathways, including the structure of the intermediate states, remains to be determined. Let us stress that these intermediates are states along forced unfolding trajectories. To what extent there are significant folding intermediates for small proteins is a debated [9, 10] but different issue.

Here we use MC simulations to examine the unfolding of ubiquitin under a constant stretching force in atomic detail. Our calculations are based on a simplified force field, which was developed through folding studies of several α-helical and β-sheet peptides with about 20 residues [11–13]. The same model was also used to study the oligomerization properties of a fibril-forming fragment of the Alzheimer's Aβ peptide, with very promising results [14]. The model is computationally efficient and allows for the collection of large amounts of unfolding events for ubiquitin, which is important because of the existence of multiple unfolding pathways.

It turns out that the model is able to reproduce key features observed in the above-mentioned experiments. In particular, we find that both one-step unfolding and unfolding via intermediate states occur in our simulations. Having verified this, we turn to more detailed measurements aimed, in particular, at characterizing the typical intermediate state. This is a non-trivial challenge

because this state is non-obligatory and is located far away from the native state, so that a few representative unfolding events would be inadequate to characterize it. In a previous study, Li and Makarov [15] used the CHARMM force field and a continuum solvation model to calculate the force required to initiate the unfolding of ubiquitin. The intermediate appears at a later stage of the unfolding process, in a region not explored in their study. An intermediate was, by contrast, observed in simulations of force-induced unfolding of the I27 immunoglobulin domain from titin [16, 17]. This domain was found to unfold via an obligatory intermediate located close to the native state. Atomic-level simulations of force-induced unfolding have also been performed for other proteins [18, 19], and several groups have used simplified protein representations to study the mechanisms of force-induced unfolding [20–25].

Model and definitions

The model we use [11–13] contains all atoms of the protein chain, including hydrogen atoms, but no explicit water molecules. It assumes fixed bond lengths, bond angles and peptide torsion angles (180°), so that each amino acid only has the Ramachandran torsion angles φ, ψ and a number of side-chain torsion angles as its degrees of freedom. In absence of the stretching force, the interaction potential

$$E_0 = E_{\mathrm{loc}} + E_{\mathrm{ev}} + E_{\mathrm{hb}} + E_{\mathrm{hp}} \tag{I.1}$$

is composed of four terms. The term E_{loc} is local in sequence and represents an electrostatic interaction between adjacent peptide units along the chain. The other three terms are non-local in sequence. The excluded volume term E_{ev} is a $1/r^{12}$ repulsion between pairs of atoms. E_{hb} represents two kinds of hydrogen bonds: backbone-backbone bonds and bonds between charged side chains and the backbone. The last term E_{hp} represents an effective hydrophobic attraction between non-polar side chains. It is a simple pairwise additive potential based on the degree of contact between two non-polar side chains. The precise form of the different interaction terms and the numerical values of all the geometry parameters held constant can be found elsewhere [11, 13]. All our simulations were carried out at a fixed temperature of 288 K.

This potential does not require prior knowledge of the native structure, making it a sequence-based rather than Gō-type [26] potential. Computationally, it is almost as fast as a Gō potential, because the most expensive term, E_{ev}, is essentially the same. The terms E_{hb} and E_{hp} contain non-native interactions which are ignored in a simple Gō-type potential. Such interactions may play

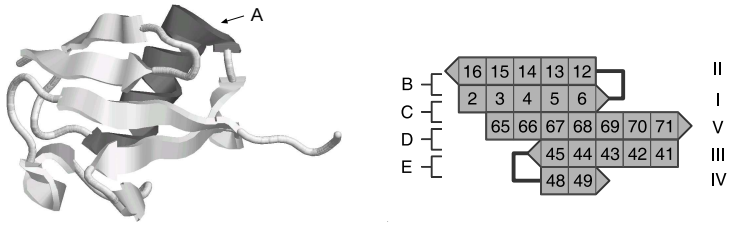

Figure I.1: Left: Schematic illustration of the native structure of ubiquitin (Protein Data Bank code 1d3z, first model). Drawn with RasMol [28]. Right: The organization of the β-sheet. Residue numbers, strand labels (I-V) and two β-hairpin turns are indicated.

a role even in force-induced unfolding; in fact, non-native hydrogen bonds do form, and break, in our ubiquitin simulations (see below). It has been shown [13] that the potential in Equation I.1, despite its simplicity, is able to provide a good description of the structure and folding thermodynamics of several peptides with different native geometries, for one and the same choice of parameters. Ubiquitin is significantly larger than these peptides. However, our present study focuses on unfolding, which is easier to simulate than folding to a unique native state.

Specifically, we here investigate the response of ubiquitin to constant stretching forces $\vec{F}$ and $-\vec{F}$ acting on the C and N termini, respectively. In the presence of these forces, the energy function becomes

$$E = E_0 - \vec{F} \cdot \vec{R}, \tag{I.2}$$

where E_0 is given by Equation I.1 and $\vec{R}$ denotes the vector from the N to the C terminus.

Figure I.1 shows the NMR-derived [27] native structure for ubiquitin, which contains a five-stranded β-sheet, an α-helix (residues 23–34) and two short 3_{10}-helices (residues 38–40 and 57–59). The organization of the β-sheet is illustrated to the right.

Despite its limited degrees of freedom, the model offers an accurate representation of the experimental structure. A model approximation of this structure was derived by using the auxiliary energy function $\tilde{E} = E_0 + \kappa\Delta^2$, where κ is a parameter and Δ denotes the root-mean-square deviation from the NMR structure (calculated over all non-hydrogen atoms). By simulated-annealing based minimization of $\tilde{E}$, a structure with $\Delta < 0.5$ Å was found. This optimized model structure served as the starting point for all our unfolding simulations.

To get a precise picture of the unfolding process, native backbone hydrogen bonds were identified and monitored in the simulations. A hydrogen bond is considered formed if the energy is lower than a cutoff [13].

To identify unfolding intermediates, we constructed a histogram of $r = |\vec{R}|$ for each unfolding event, based on measurements taken at regular time intervals. During an unfolding event, r increases essentially monotonically, and significant peaks in the histogram of r are candidates for intermediate states. To reduce noise, a moving average smoothing procedure [29] was used, in which data in each bin was replaced by the average over the three nearest bins. After this smoothing, a cutoff in the height and area of the peaks was used to filter out all but the most significant peaks. Pairs of peaks for which the level between the two maxima does not fall below half the average height of the peaks were combined into single peaks. In most events, all the bins at intermediate r were so sparsely populated that the above procedure left no peak which would count as an intermediate, corresponding to two-state unfolding. However, some events were qualitatively different with one or two very prominent peaks at intermediate r.

Our simulations were performed using MC methods. To avoid large unphysical deformations of the chain, a semi-local update was used for the backbone degrees of freedom. This update, Biased Gaussian Steps [30], works with up to eight backbone torsion angles which are turned in a coordinated manner. Side-chain torsion angles were updated one by one. In addition to these updates of internal coordinates, we also included small rigid-body rotations of the whole molecule. The fractions of attempted backbone moves, side-chain moves and rigid-body rotations were 24 %, 75 % and 1 %, respectively.

The system was restarted from the native state as soon as an unfolding event had occurred. The maximum length of a run was 10^8 elementary MC steps. If this limit was reached, the run was stopped even if the chain was still folded, and a new run was started. The fraction of runs in which the system remained folded after 10^8 MC steps varied from 65 % at 100 pN to 1% at 200 pN.

Results

We study the unfolding of ubiquitin for three different strengths of the applied force, namely 100 pN, 140 pN and 200 pN, a choice that enables us to directly compare with the experiments at constant force by Schlierf et al. [6]. For each force, a set of more than 500 MC runs was performed, all of which were started from the native structure but with different random number seeds.

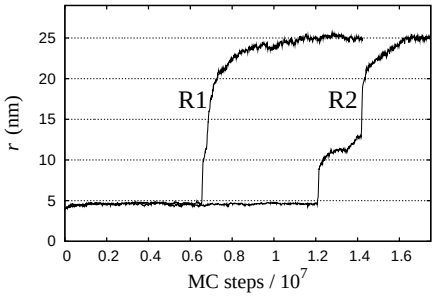

Figure I.2: MC evolution of the end-to-end distance r in two representative runs. The protein chain unfolds in a single step in run R1, and via an intermediate state in run R2. The pulling force is 100 pN in both runs.

Let us first briefly describe the different phases encountered in the simulations. Figure I.2 shows the time evolution of the end-to-end distance $r = |\vec{R}|$ in two typical runs. All runs start with a rapid relaxation of r from its native value of 3.9 nm to a value r_f, which is $r_\mathrm{f} = 4.6$ nm for 100 pN and $r_\mathrm{f} = 4.7$ nm for 200 pN. Here, the chain ends get stretched out, while the rest of the chain remains largely unaffected. This initial adjustment of r is followed by an extended "waiting time", a phase where r is confined to a small range around r_f. In this phase, structural fluctuations occur, but the protein remains native-like with essentially intact secondary structure. Actually, in some runs, the β-sheet temporarily increases in length along the applied force, through the formation of non-native hydrogen bonds. This phase is terminated by a sudden jump in r. In many runs, the unfolding from the native-like state occurs in a single step, as in run R1 in Figure I.2. However, in agreement with the experiments [6], we also observe several examples of unfolding via intermediate states, as in run R2 in Figure I.2. In all our runs, r shows an essentially monotonic increase with time; no transition from an intermediate state back to the native state was observed. In the final phase of the runs, r fluctuates around a force-dependent mean r_u, which increases from $r_\mathrm{u} = 25.2$ nm to $r_\mathrm{u} = 25.9$ nm as the force increases from 100 pN to 200 pN. The unfolded value r_u may be compared to the value $76 \times 0.36 = 27.4$ nm for a fully extended 76-residue protein, and to the average value 23.0 nm for a worm-like chain [31] with contour length 27.4 nm and persistence length 4 Å [32], at 100 pN and 288 K.

For the total step size $\Delta r_\mathrm{tot} = r_\mathrm{u} - r_\mathrm{f}$ a value of $\Delta r_\mathrm{tot} = 20.3 \pm 0.9$ nm was reported from the experiments [6], which was an average over data at different forces. In our model, we find a weak but steady increase in Δr_tot with force:

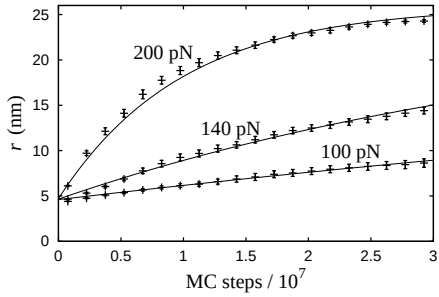

Figure I.3: End-to-end distance r against MC time for the three different forces studied. Each data point represents an average over all the runs for a fixed force. The curves are fitted single exponentials.

$\Delta r_{\mathrm{tot}} = 20.6$ nm for 100 pN, $\Delta r_{\mathrm{tot}} = 21.0$ nm for 140 pN, and $\Delta r_{\mathrm{tot}} = 21.2$ nm for 200 pN. These results are consistent with the experimental value. However, further experimental data are required to verify the force dependence of Δr_{tot}.

In the experiments, the average unfolding curve showed an approximately single-exponential time dependence, with a rate constant that increased exponentially with force [6]. Figure I.3 shows the average of r against time from our simulations. A single exponential provides a reasonable description of the data at 200 pN, whereas higher statistics would be required to study the functional form at 100 pN and 140 pN. Our fitted time constant at 200 pN is $\tau = 1.0 \cdot 10^7$ MC steps. At 100 pN, the best way to estimate τ is from the folded population after 10^8 MC steps (65 %) which, assuming a single exponential, gives $\tau = 2.3 \cdot 10^8$ MC steps. The experiments obtained $\tau = 0.05$ s at 200 pN and $\tau = 2.77$ s at 100 pN [6], so unfolding was 55 times faster at 200 pN than at 100 pN. This ratio is a factor 2 smaller in our simulations.

The time behavior of the experimental data is consistent with a simple two-state picture. Nevertheless, unfolding intermediates were observed in the experiments. Schlierf et al. [6] saw intermediates in about 5% of their 800 unfolding events, recorded at different forces. The most common distance between the initial and intermediate states was $\Delta r = 8.1 \pm 0.7$ nm.

Although the steps are not as sharp as in the experiments, unfolding proceeds in a clear step-wise fashion in the simulations as well (see Figure I.2), and a useful operational definition of intermediate states can be easily devised (see Model and Definitions). To avoid the initial and final states, we restrict our analysis of intermediate states to the range $6.5\,\mathrm{nm} \leq r \leq 18.5\,\mathrm{nm}$. Here the upper limit is rather conservatively chosen. The reason for this is that the

Table I.1: The total number of observed unfolding events and the respective fractions of two-, three- and four-state events, for the three different forces studied.

force	no. of events	two-state	three-state	four-state
100 pN	183	0.74	0.24	0.02
140 pN	357	0.96	0.04	0
200 pN	583	0.99	0.01	0

last part of the unfolding process is somewhat slow in the simulations (see Figure I.2), which prevents us from unambiguously identifying intermediates with $r > 18.5$ nm.

All our unfolding events contain either zero, one or two intermediate states. The relative frequencies of two-, three- and four-state events are shown in Table I.1. For all three force magnitudes, two-state events are more common than three-state events, which in turn are more common than four-state events. The fraction of events with intermediate states is comparable to the above-mentioned value of 5 %, which was an average over experiments at different forces. From Table I.1 we also note that intermediates occur more frequently as the force gets lower. This force dependence was not investigated experimentally.

Figure I.4 shows the distribution of r for the intermediate states we observed at 100 pN. The distribution is sharply peaked around a typical unfolding step of $\Delta r \approx 7$ nm, which is slightly lower than the most common step size in the experiments (8.1 ± 0.7 nm). This deviation could indicate that the stability of the intermediate state is somewhat low in the model, so that we loose this state before it reaches its optimal orientation. However, the deviation is small, and it should be kept in mind that the experimental value represents an average over experiments under varying conditions. Our step-size distribution at 140 pN (not shown) is similar to that in Figure I.4 but noisier, due to fewer events. For scarcity of events, the step-size analysis is not meaningful for 200 pN. Neither is it possible to draw any statistical conclusions specifically about four-state events. We note, however, that the intermediate state with $\Delta r \approx 7$ nm occurs in our four-state events as well (see Figure I.4). Further, this intermediate state has the longest life-time among the observed intermediates (data not shown).

To check whether the character of the unfolding event depends on the waiting time, we divided the 183 unfolding trajectories for 100 pN into two groups of 93 and 90 events. Those in the first group unfolded in the first third of the maximal simulation time. The second group had waiting times greater than that. Both the frequency of occurrence of intermediate states and their

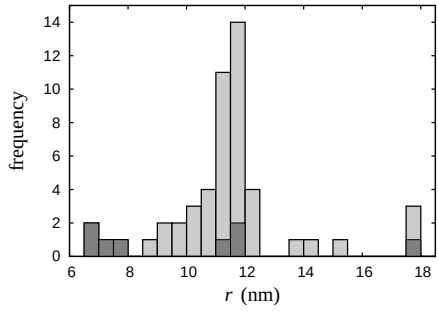

Figure I.4: Histogram of r for intermediate states at 100 pN. For this force, we observed 43 three-state events and 4 four-state events. The light- and dark-grey parts of the bars correspond to intermediates from three- and four-state events, respectively.

typical location in r are very similar between these groups, indicating that the unfolding behavior is largely independent of the waiting time.

We now turn to a more detailed description of the unfolding process. To delineate the unfolding process, we follow five key elements of the native structure, labeled A to E and indicated in Figure I.1. The structure A is the α-helix, whereas B, C, D and E are the four different pairs of adjacent strands in the β-sheet. The experiments mainly focused on the end-to-end distance, and provide therefore only limited information about the structure of the intermediate states. However, Schlierf et al. [6] proposed a three-state scenario, based on the observation that the ubiquitin sequence can be split into two halves that correspond to well defined clusters packing against each other in the native structure. The first cluster includes the structures A and B, whereas the second cluster includes the structures D and E (C has one strand in each cluster). The unraveling of the second cluster would give an intermediate unfolding step of about 8 nm, which agrees with the most common step size in the experiments [6]. In this scenario, the intermediate is composed of the structures A and B.

To examine the order in which the structures A to E break, we study the presence of native backbone hydrogen bonds as a function of r. To this end, simulated conformations were divided into different intervals in r. For each interval, the frequency of occurrence of the different hydrogen bonds was computed. Figure I.5 shows the result of this calculation for all the native backbone hydrogen bonds in the structures A to E, at 100 pN. From this figure it is immediately clear that the structures do not break in a random order, but

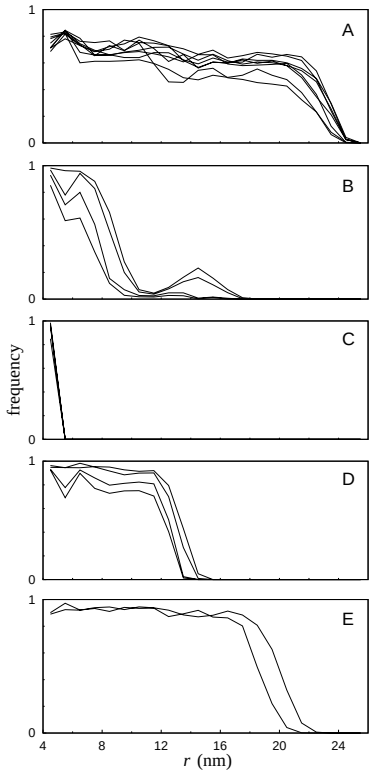

Figure I.5: The frequency of occurrence as a function of the end-to-end distance r for all native backbone hydrogen bonds in the respective structures A to E, at 100 pN. Each curve represents one hydrogen bond. The numbers of native backbone hydrogen bonds are 8, 4, 4, 4 and 2 for A, B, C, D and E, respectively.

instead in a statistically preferred order, namely CBDEA. That C breaks first is expected, because the rest of the structure is shielded from the action of the pulling force as long as C is in place (see Figure I.1). The structures C and B tend to break below the typical r for intermediate states, $r_\mathrm{i} \approx 12\,\mathrm{nm}$ (see Figure I.4), whereas D, E and A tend to break above r_i. The data in Figure I.5 thus suggest that the typical intermediate is composed of A, D and E rather than A and B. An analogous analysis using native contacts was also performed, with very similar results.

Figure I.6 shows the results obtained when performing the same analysis as in Figure I.5 for two- and three-state events separately. For clarity, Figure I.6

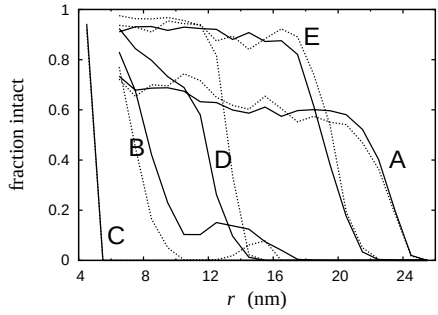

Figure I.6: The fraction of formed hydrogen bonds as a function of r for the structures A to E, for two-state (full lines) and three-state (dashed lines) events, at 100 pN. Each curve represents an average over all the native backbone hydrogen bonds of a given structure. For the structure C, the two curves coincide.

does not show data for individual bonds, but only averages for the different structures. Overall the curves for two- and three-state events are similar in shape, which in particular suggests that the structures A to E break in the same order in both cases. The biggest difference we see is for the structure D, which tends to break just above r_i. At r_i, D is essentially intact in three-state events, whereas one or two of its hydrogen bonds typically are missing in two-state events. This difference strongly indicates that the structure D plays a crucial stabilizing role in the typical intermediate state. A small difference between two- and three-state events can also be seen in the results for the structure B; remnants of B are present near r_i in two-state events but not in three-state events. The results for the three structures A, C and E show, by contrast, no significant differences between two- and three-state events.

The analysis of Figs. 5 and 6 does not tell how strong the statistical preference is for the unfolding order CBDEA. To check this, we directly analyzed the time of breaking of the structures A to E and determined a path (a permutation of ABCDE) for each individual event. For this purpose, we regard a structure to be intact if one third or more of its native backbone hydrogen bonds are present. To filter out short-term thermal fluctuations and thereby focus on genuine long-term changes because of unfolding, we define the time of breaking of a structure as the last point in time after which the structure is never found intact.

In the 100 pN case, we find that 61 % of the events follow the path CBDEA unambiguously. Another 23 % of the events have the order of B and D apparently interchanged, the path being CDBEA. In these events B does unfold

before D, but then partially reforms after D is gone, so that the definition above assigns a later breaking time for B. This partial refolding can actually be seen in Figure I.5; the β-hairpin B is almost completely dissolved at r_i, but two of its four hydrogen bonds (those closest to the turn) occur again around $r = 14$ nm with a non-negligible frequency. The partial refolding of B takes place just after D breaks, and is possible because a large chain segment is released as D breaks. As this segment gets stretched out, B dissolves again. The reformation of B is not a step back toward the native state, because when B reforms, D is gone and the system has a larger end-to-end distance. Hence, this class of events can be regarded as a minor variation of the path CBDEA, which means that 84 % of our unfolding events follow the same basic pathway. All other of the 120 possible paths have probabilities below 5 %. Given the uncertainties due to the intrinsically somewhat arbitrary definition of a path, these low-probability paths were not analyzed any further. As expected from Figure I.6, separate analyses of two- and three-state events showed that the fraction of events following the main pathway is similar in both cases.

To check the force dependence of our findings, the analysis of Figs. 5 and 6 as well as the path analysis of individual events were repeated for 140 pN and 200 pN, with results similar to those at 100 pN. While intermediate states occur more frequently at low force, we thus find that the order in which the structures break is the same at the different forces.

That the α-helix A survives up to very large r hints at the possibility that its intrinsic stability could be very high. Therefore, we performed equilibrium simulations of the α-helix segment in isolation. We found that this excised segment makes an α-helix in the model, which unfolds at about 15 pN. This force is much smaller than those studied above, which implies that the effective force felt by the α-helix in the ubiquitin simulations must be small compared to the applied force.

Finally, Figure I.7 shows a snapshot of the typical intermediate state, taken from the run R2 in Figure I.2. The structures A, D and E are still present, whereas B and C are missing. The structure B, which together with A forms the intermediate in the above-mentioned three-state unfolding scenario of Schlierf et al. [6], breaks early in our simulations.

Summary and discussion

We have performed a MC study to elucidate at the atomic level the unfolding behavior of ubiquitin under a constant stretching force. The study consists of two parts. Following the experiments, we first investigated the behavior

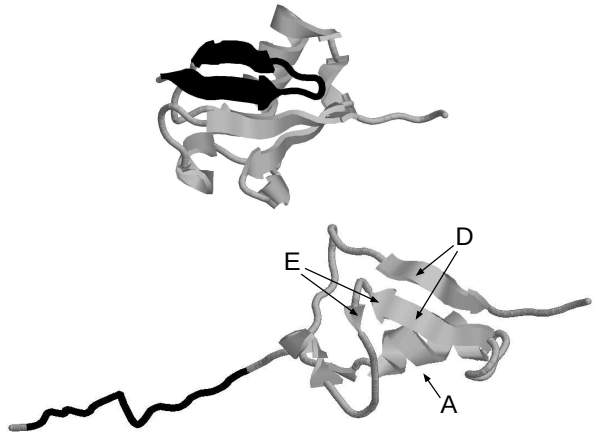

Figure I.7: The native structure and a snapshot illustrating the shape of the typical intermediate state in our simulations. The chain segment that makes the structure B, a β-hairpin, is marked in black.

of the end-to-end distance. Properties such as the size of the unfolding step, the frequency of occurrence of intermediate states, and the position of the typical intermediate state were all found to be in reasonable agreement with experimental data. For the range of forces studied here, we further found that the size of the unfolding step increased with force, whereas intermediate states occurred more frequently at lower force. These dependencies on the strength of the applied force remain to be tested experimentally.

In the second part of our study, we analyzed the order of breaking of five major secondary-structure elements. This analysis revealed that these structures, A to E (see Figure I.1), tend to break in a definite order. Disregarding interchanges of B and D due to the partial refolding of the β-hairpin B, we found that more than 80 % of the events followed the same unfolding path, CBDEA. The order was the same in events with and without intermediate states, and neither did it change between the three different forces studied.

To what extent this predicted unfolding order is correct is not obvious, given the lack of detailed experimental data to examine it. In fact, as mentioned in Results, a markedly different unfolding scenario has been proposed [6]. Therefore, it should be stressed that several aspects of the calculated unfolding behavior can be understood in terms of topology and pulling geometry. That C breaks first is inevitable; the other parts cannot sense the force until C is broken (see Figure I.1). The native state is mechanically resistant because C is pulled longitudinally, so that several bonds must break simultaneously. Once

C is gone, nothing keeps the β-hairpin B from unzipping, one bond at a time; unzipping requires less force than separation by longitudinal pulling [33, 34]. Similarly, for the three structures A, D and E, which form the typical intermediate state, it is clear that the β-hairpin E is protected by D (see Figure I.7), because the two strands of D are on different sides of E along the sequence. Visual inspection of snapshots from the simulations suggests that D is pulled in a direction neither parallel nor perpendicular to its β-strands making it semi-resistant, which is consistent with our conclusion that D is important for the stability of the intermediate (see Figure I.6). When D is gone, E is free to unzip. Compared to unwinding the α-helix A, the unzipping of E requires less force, which makes E break before A.

This picture of the unfolding process highlights the importance of topology and pulling geometry. These two factors make the four β-sheet structures B to E behave very differently. C and D play important stabilizing roles in the native and typical intermediate states, respectively. B and E are, by contrast, easy to pull apart, and survive only as long as they are protected; they do not cause any traps on the unfolding pathway. Previous studies [4, 34] have shown the importance of pulling geometry, in relation with native topology, as a determinant of a protein's mechanical stability. We extend that picture and show that geometrical and topological factors play important stabilizing roles along the entire unfolding pathway.

Our study was based on nonequilibrium simulations, but contains a sufficient number of unfolding events to, nevertheless, speculate on the underlying free-energy landscape. Of particular interest is the typical intermediate state with $\Delta r \approx 7\,\mathrm{nm}$. Our step-size and secondary-structure analyses strongly suggest that this unfolding intermediate corresponds to a local free-energy minimum with a rather well-defined three-dimensional structure. It is worth stressing that in the limit of zero force, this minimum might very well be irrelevant, because the smallest force studied, 100 pN, is still large enough to strongly tilt the energy landscape ($100\,\mathrm{pN} \times 7\,\mathrm{nm} \approx 100\,\mathrm{kcal/mol}$). The folding and unfolding of ubiquitin at zero force has been extensively studied by using both chemical denaturants [35] and temperature denaturation [36, 37], and the flexibility of the native state has also been examined [38]. In a study of thermal unfolding based on IR spectroscopy, Chung et al. [37] found that the β-strands I and II are more stable than the β-strands III-V, which contrasts sharply with our results. It thus appears that the pathways followed in thermal and force-induced unfolding of ubiquitin indeed are different.

The physical forces governing the force-induced unfolding of proteins are the same as those governing protein folding and aggregation, which makes it

tempting to search for a unified computational approach to these phenomena. The model used in the present calculations has previously been used to study folding [13] as well as aggregation [14]. All the parameters of the model could be kept unchanged between these three different studies. This fact does not, of course, imply that the model is perfect, but is nevertheless encouraging. In particular, it strongly suggests that computational studies of force-induced unfolding could complement detailed experimental studies to provide a better physical picture of the mechanical unfolding of different proteins. Such studies could also provide useful information for refining the force field, which in turn would improve our understanding of protein folding and aggregation.

Acknowledgments. We thank Michael Schlierf for useful discussions. This work was in part supported by the Swedish Research Council. The computer simulations were performed at the LUNARC facility at the Lund University.

References

1. Pickart CM (2001) Mechanisms underlying ubiquitination. Annu Rev Biochem 70:503–533.
2. Weissman AM (2001) Themes and variations on ubiquitylation. Nat Rev Mol Cell Biol 2:169–178.
3. Chau V, Tobias JW, Bachmair A, Marriott D, Ecker DJ, et al. (1989) A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein. Science 243:1576–1583.
4. Carrion-Vazquez M, Li H, Lu H, Marszalek PE, Oberhauser AF, et al. (2003) The mechanical stability of ubiquitin is linkage dependent. Nat Struct Biol 10:738–743.
5. Fernandez JM, Li H (2004) Force-clamp spectroscopy monitors the folding trajectory of a single protein. Science 303:1674–1678.
6. Schlierf M, Li H, Fernandez JM (2004) The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques. Proc Natl Acad Sci USA 101:7299–7304.
7. Sarkar A, Robertson RB, Fernandez JM (2004) Simultaneous atomic force microscope and fluorescence measurements of protein unfolding using a calibrated evanescent wave. Proc Natl Acad Sci USA 101:12882–12886.
8. Chyan CL, Lin FC, Peng H, Yuan JM, Chang CH, et al. (2004) Reversible mechanical unfolding of single ubiquitin molecules. Biophys J 87:3995–4006.
9. Jackson SE, Fersht AR (1991) Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. Biochemistry 30:10428–10435.
10. Sanchéz IE, Kiefhaber T (2003) Evidence for sequential barriers and obligatory intermediates in apparent two-state protein folding. J Mol Biol 325:367–376.
11. Irbäck A, Samuelsson B, Sjunnesson F, Wallin S (2003) Thermodynamics of α- and β-structure formation in proteins. Biophys J 85:1466–1473.
12. Irbäck A, Sjunnesson F (2004) Folding thermodynamics of three β-sheet peptides: a model study. Proteins 56:110–116.

13. Irbäck A, Mohanty S (2005) Folding thermodynamics of peptides. Biophys J 88:1560–1569.
14. Favrin G, Irbäck A, Mohanty S (2004) Oligomerization of amyloid Aβ(16-22) peptides using hydrogen bonds and hydrophobicity forces. Biophys J 87:3657–3664.
15. Li PC, Makarov DE (2004) Simulation of the mechanical unfolding of ubiquitin: probing different unfolding reaction coordinates by changing the pulling geometry. J Chem Phys 121:4826–4832.
16. Lu H, Schulten K (2000) The key event in force-induced unfolding of titin's immunoglobulin domains. Biophys J 79:51–65.
17. Fowler SB, Best RB, Herrera JLT, Rutherford TJ, Steward A, et al. (2002) Mechanical unfolding of a titin Ig domain: structure of unfolding intermediate revealed by combining AFM, molecular dynamics simulations, NMR and protein engineering. J Mol Biol 322:841–849.
18. Lu H, Schulten K (1999) Steered molecular dynamics simulations of force-induced protein domain unfolding. Proteins 35:453–463.
19. Bryant Z, Pande VS, Rokhsar DS (2000) Mechanical unfolding of a β-hairpin using molecular dynamics. Biophys J 78:584–589.
20. Socci ND, Onuchic JN, Wolynes PG (1999) Stretching lattice models of protein folding. Proc Natl Acad Sci USA 96:2031–2035.
21. Klimov DK, Thirumalai D (2000) Native topology determines force-induced unfolding pathways in globular proteins. Proc Natl Acad Sci USA 97:7254–7259.
22. Cieplak M, Hoang TX, Robbins MO (2002) Thermal unfolding and mechanical unfolding pathways of protein secondary structures. Proteins 49:104–113.
23. Geissler PL, Shakhnovich EI (2002) Reversible stretching of random heteropolymers. Phys Rev E 65:056110.
24. Shen T, Canino LS, McCammon JA (2002) Unfolding proteins under external force: a solvable model under the self-consistent pair contact probability approximation. Phys Rev Lett 89:068103.
25. Marenduzzo D, Maritan A, Rosa A, Seno F (2003) Stretching of a polymer below the theta point. Phys Rev Lett 90:088301.
26. Gō N (1983) Theoretical studies of protein folding. Annu Rev Bioph Bioeng 12:183–210.
27. Cornilescu G, Marquardt JL, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute crystalline phase. J Am Chem Soc 120:6836–6837.
28. Sayle RA, Milner-White EJ (1995) Rasmol: biomolecular graphics for all. Trends Biochem Sci 20:374–376.
29. Box GEP, Jenkins GM, Reinsel GC (1994) Time Series Analysis: Forecasting & Control. Prentice Hall, Englewood Clifs, 3rd edition.
30. Favrin G, Irbäck A, Sjunnesson F (2001) Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. J Chem Phys 114:8154–8158.
31. Marko JF, Siggia ED (1995) Stretching DNA. Macromolecules 28:8759–8770.
32. Oesterhelt F, Oesterhelt D, Pfeiffer M, Engel A, Gaub HE, et al. (2000) Unfolding pathways of individual bacteriorhodopsins. Science 288:143–146.
33. Rohs R, Etchebest C, Lavery R (1999) Unraveling proteins: a molecular mechanics study. Biophys J 76:2760–2768.
34. Brockwell DJ, Paci E, Zinober RC, Beddard GS, Olmsted PD, et al. (2003) Pulling geometry defines the mechanical resistance of a β-sheet protein. Nat Struct Biol

10:731–737.
35. Khorasanizadeh S, Peters ID, Roder H (1996) Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. Nat Struct Biol 3:193–205.
36. Sabelko J, Ervin J, Gruebele M (1999) Observation of strange kinetics in protein folding. Proc Natl Acad Sci USA 96:6031–6036.
37. Chung HS, Khalil M, Smith AW, Ganim Z, Tokmakoff A (2005) Conformational changes during the nanosecond-to-millisecond unfolding of ubiquitin. Proc Natl Acad Sci USA 102:612–617.
38. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. Nature 433:128–132.

PAPER II

Thermal versus mechanical unfolding of ubiquitin

Anders Irbäck and Simon Mitternacht

Computational Biology and Biological Physics, Department of Theoretical Physics, Lund University, Sweden.

Proteins 65: 759–766. (2006)

We study the temperature-induced unfolding of ubiquitin by all-atom Monte Carlo (MC) simulations. The unfolding behavior is compared to that seen in previous simulations of the mechanical unfolding of this protein, based on the same model. In mechanical unfolding, secondary-structure elements were found to break in a quite well-defined order. In thermal unfolding, we see somewhat larger event-to-event fluctuations, but the unfolding pathway is still far from random. Two long-lived secondary-structure elements can be identified in the simulations. These two elements have been found experimentally to be the thermally most stable ones. Interestingly, one of these long-lived elements, the first β-hairpin, was found to break early in the mechanical unfolding simulations. Our combined simulation results thus enable us to predict in detail important differences between the thermal and mechanical unfolding behaviors of ubiquitin.

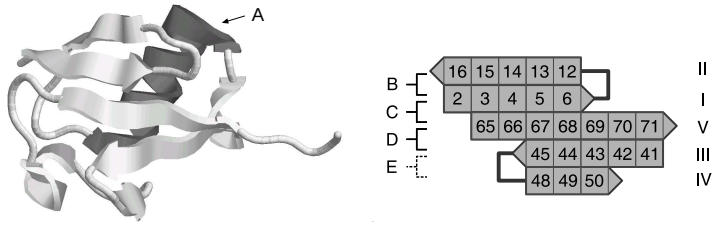

Figure II.1: Schematic illustrations of the native structure of ubiquitin with our labels for secondary-structure elements, A–E. (Left) A 3D model (Protein Data Bank code 1d3z, first model) drawn with RasMol [7]. (Right) The organization of the β-sheet. Residue numbers, strand labels (I-V) and two β-hairpin turns are indicated.

Introduction

Ubiquitin is a 76-residue α/β protein, whose unfolding and refolding properties have been extensively studied experimentally. It was first thought that an intermediate state is populated during the folding of this protein [1]. More recent studies suggest, however, that ubiquitin folds in a two-state manner in the absence of stabilizing salt [2, 3].

In its native form, ubiquitin contains an α-helix packed against a five-stranded β-sheet (see Figure II.1). The thermally most stable parts of the native structure seem to be the α-helix and the first (N-terminal) β-hairpin, which is part of the β-sheet. An NMR-based study of the temperature dependence of hydrogen bonds found that these secondary-structure elements are more resistant to temperature increases than the remaining three β-strands [4]. That the first β-hairpin is thermally more stable than the rest of the β-sheet has also been concluded from IR spectroscopy [5]. Interestingly, the α-helix and the first β-hairpin have been found to be relatively resistant to cold denaturation as well [6].

There is evidence that the same two secondary-structure elements, the α-helix and the first β-hairpin, form early as ubiquitin folds to its native conformation. An extensive φ-value analysis, based on 27 mutations throughout the protein, found that both these structures are present in the transition state, but that the rest of the molecule remains largely unstructured at this stage of folding [8]. It has also been suggested, based on a ψ-value analysis, that two additional β-strands are present in the transition state [9]. The reason for these somewhat different conclusions has been discussed [10, 11].

In addition to studies of full-length ubiquitin, there have also been experiments on various excised fragments of the protein, including the first β-hairpin [12] (residues 1–17) and the α-helix [13] (residues 21–35). Both these peptides showed a tendency, although weak, to make nativelike structure, whereas the 36–76-residue fragment showed little or no such tendency [13].

Together, the experiments strongly suggest that the α-helix and the first β-hairpin are relatively resistant to several different types of perturbations. Recently, however, we suggested, based on all-atom MC simulations, that the situation is different in mechanical unfolding [14]. The mechanical unfolding of ubiquitin has been studied in several single-molecule experiments [15–19]. One of these studies investigated the unfolding behavior under a constant stretching force, using end-to-end linked polyubiquitin [17]. In this study, ubiquitin was seen to unfold in a two-state manner in most events, but several examples of unfolding through intermediate states were also observed. The difference in end-to-end distance between the native and typical intermediate states was consistent with what one would expect if the α-helix and the first β-hairpin survived whereas the rest of the β-sheet unfolded [17], which would be in line with the above-mentioned findings at zero force. However, in the presence of a stretching force of ≥ 100 pN, the energy landscape is strongly tilted, and the unfolding behavior might very well be different from that at zero force. Important differences between mechanical and thermal unfolding have in fact been seen in simulations of other proteins, such as the I27 domain of titin [20, 21].

The results from our previous study of the mechanical unfolding of ubiquitin [14] are indeed different from the experimental results on the thermal unfolding of this protein. In particular, we found that the first β-hairpin broke early in the simulations, so that the typical unfolding intermediate contained the α-helix and the other three β-strands, but not the first β-hairpin. This behavior contrasts sharply with the observed stability of this β-hairpin in zero-force experiments. The mechanical unfolding experiments [17] monitored only the end-to-end distance. Our results for this quantity were in good agreement with the experimental results. Both one-step unfolding and unfolding through intermediate states occurred in our simulations, and properties such as the size of the unfolding step, the frequency of occurrence of intermediate states and the position of the typical intermediate state were all found to be in reasonable agreement with the experimental data.

Here we perform an MC study of the thermal unfolding of ubiquitin, using the same model and methods as in our study of the force-induced unfolding of this protein [14]. In this way, we can directly compare thermal and force-

induced unfolding. To ensure representative sampling, we generated a set of 800 thermal unfolding events for our analysis.

Computational studies of ubiquitin folding and unfolding at zero force have been reported by several groups. Perhaps the work most relevant to ours is an all-atom molecular dynamics study with explicit water [22], which found that the native contacts between the β-strands I and V (see Figure II.1) as well as those in the first β-hairpin were lost early in thermal unfolding. These findings are not in perfect agreement with experimental data. However, the statistics were limited to two unfolding trajectories, and it might be worth noting that contacts in the first β-hairpin did not disappear completely but fluctuated with time. In addition to this all-atom molecular dynamics study, there have been studies using coarse-grained approaches [23–25], threading algorithms [26] and an MC design strategy [27]. The force-induced unfolding of ubiquitin has also been studied using both all-atom [14, 15, 28] and coarse-grained [29, 30] models. Atomic-level simulations of force-induced unfolding have been reported for several other proteins as well, such as immunoglobulin and fibronectin type III domains of titin [20, 31–33].

Materials and methods

The model we use combines an atomistically detailed chain representation with a simplified force field, which was developed by folding studies of a set of well characterized peptides [34, 35]. Both α-helical and β-sheet peptides were studied. In addition to peptide folding, this model has also been used to study peptide aggregation [36] and the force-induced unfolding of ubiquitin [14], without changing any model parameters.

The model contains all atoms of the protein chain, including hydrogen atoms, but no explicit water molecules. It assumes fixed bond lengths, bond angles and peptide torsion angles (180°), so that each amino acid only has the Ramachandran torsion angles φ, ψ and a number of side-chain torsion angles as its degrees of freedom. The interaction potential

$$E = E_{\mathrm{loc}} + E_{\mathrm{ev}} + E_{\mathrm{hb}} + E_{\mathrm{hp}} \tag{II.1}$$

is composed of four terms. The term E_{loc} is local in sequence and represents an electrostatic interaction between adjacent peptide units along the chain. The other three terms are non-local in sequence. The excluded volume term E_{ev} is a $1/r^{12}$ repulsion between pairs of atoms. E_{hb} represents two kinds of hydrogen bonds: backbone-backbone bonds and bonds between charged side chains and the backbone. The last term E_{hp} represents an effective hydrophobic attraction between non-polar side chains. It is a simple pairwise

additive potential based on the degree of contact between two non-polar side chains. The precise form of the different interaction terms and the numerical values of all geometry parameters can be found elsewhere [34, 35].

The potential in Equation II.1 provides, despite its simplicity, a good description of the structure and folding thermodynamics of several peptides with different native geometries [35]. For ubiquitin, which is significantly larger than these peptides, it is unclear whether the native state corresponds to the global free-energy minimum of the model at physiological temperatures. In our thermal unfolding simulations, the temperature was set to 368 K, which is higher than, but close to, the melting temperature of ubiquitin ($\sim 90^{\circ}$C) [37]. At this temperature, refolding should not be extremely rare, but it did not occur in our simulations. This is not unexpected; as indicated above, it is possible that the model underestimates the weight of the native state relative to the rest of conformational space. To be able to study the unfolding process, it is, however, sufficient that the model reproduces the free-energy landscape in a limited neighborhood of the native state, which is a much more modest requirement. From our previous ubiquitin study [14], it is known that the native state is a pronounced local free-energy minimum of the model, and also that the model reproduces key observations in the single-molecule constant-force experiments [17].

Our simulations were carried out using the program package PROFASI [38], which is a C++ implementation of this model. A model structure with a root-mean-square deviation (RMSD) of < 0.5 Å from the NMR-derived native structure [39] was determined by simulated-annealing-based minimization. This optimized model structure served as the starting point for our simulations. The unfolding process was simulated using MC dynamics, with two move types. For the backbone, we used a semi-local method, Biased Gaussian Steps [40], rather than single-variable updates, in order to avoid large unphysical deformations of the chain. This method turns up to eight adjacent angles simultaneously, with a bias toward local deformations of the chain. Side-chain angles, on the other hand, were updated one by one. The fractions of attempted backbone and side-chain moves were 25 % and 75 %, respectively. A total of 800 unfolding simulations were performed, each comprising $2.5 \cdot 10^7$ elementary MC steps.

A study of folding or unfolding kinetics based on MC dynamics must be interpreted with care, as discussed, for example, by Shakhnovich and coworkers [41]. Many detailed questions regarding the kinetics cannot be addressed this way; the short-time evolution of an individual MC trajectory may have very little to do with real dynamics. However, the sequence of major events

will be dictated by the free-energy landscape rather than the precise choice of dynamics, provided that a "small-step", detailed-balance-obeying algorithm is used. Here it is essential that the events are separated by many elementary MC steps, and that they are observed in an ensemble of trajectories. A comparison of MC and Brownian dynamics was made in a study of α-helical hairpins [42]. It was found that the two methods gave the same folding pathway.

To delineate the unfolding process, we examine the order of breaking of four secondary-structure elements, A–D, which are indicated in Figure II.1. The structure A is the α-helix, whereas B–D are three pairs of adjacent strands in the β-sheet. In our analysis of mechanical unfolding [14], we also followed the fourth pair of adjacent strands in the five-stranded β-sheet, labeled E in Figure II.1. The structure E is a small β-hairpin with one pair of properly hydrogen bonded amino acids, namely 45 and 48, and one additional backbone hydrogen bond between amino acids 43 and 50. This structure, with residues that are close in sequence, broke and partly reappeared many times in the simulations, although complete restoration was very rare. As a result, the time of breaking becomes difficult to define for E. Therefore, and because of its small size, this structure is omitted from our analysis.

To what degree the structures A–D are present at a given stage in unfolding is determined by monitoring native backbone hydrogen bonds. A hydrogen bond is considered formed if the energy is lower than a cutoff [35]. The structures A–D contain a total of 20 backbone hydrogen bonds. The fraction of these 20 bonds that are present in a given conformation is denoted by n_{hb}.

The times at which the structures A–D break in the simulations show large event-to-event fluctuations (see Results); both the "waiting time" before unfolding starts and the actual unfolding time vary substantially from event to event. To filter these variations out of the analysis and focus on the order of events, we study the unfolding process as a function of some unfolding coordinate x rather than MC time.

To investigate the dependence of the individual hydrogen bonds on the unfolding coordinate x, we divide the x-axis into bins. For each unfolding event, we then compute the fraction of configurations in the different bins that contain a given hydrogen bond. In this way, we obtain an x-profile for each hydrogen bond and each event. Finally, we compute an average x-profile for each hydrogen bond, by averaging over all events. This procedure assigns equal weight to all the events, irrespective of the time spent in the different bins. In our study of mechanical unfolding [14], we averaged over all data without first computing profiles for the individual events, which effectively means that the events were weighted by the time spent in the different bins.

The definition adopted here is somewhat more robust. The change of definition alters the precise shape of the curves, but does not affect any of our previous conclusions regarding the mechanical unfolding simulations.

In the case of mechanical unfolding, the end-to-end distance r is an obvious choice for the unfolding coordinate x. For thermal unfolding, there are several conceivable choices of x, including the above-mentioned fraction of native hydrogen bonds, n_{hb}, the RMSD from the native state, and the fraction of native contacts. Below, we use n_{hb} as unfolding coordinate. A difference between this coordinate and the fraction of native contacts is that the former does not take into account contacts between the α-helix and the β-sheet. However, in our simulations, these two coordinates were strongly correlated and therefore essentially equivalent. The correlation coefficient was 0.96. The correlation between n_{hb} and the RMSD from the native state was slightly weaker, with a coefficient of -0.78. We found n_{hb} to be a more suitable choice than the RMSD when studying the order of breaking of the structures A–D. The separation between these structures was clearer in the profiles with n_{hb} as unfolding, or x, coordinate than in those with the RMSD as x coordinate.

In addition to this event-averaged analysis of the unfolding order of the structures A–D, we also perform an analysis of all individual events. Here we directly analyze the time of breaking of these structures, rather than using n_{hb}.

Results

Using the model and methods described in Material and Methods, we performed 800 high-temperature simulations of ubiquitin at 368 K, all starting from the native structure but with different random number seeds. Figures II.2a and II.2b illustrate three representative runs, by showing the time evolution of the RMSD from the native structure (calculated over all non-hydrogen atoms) and the native hydrogen bond fraction n_{hb} (see Materials and Methods), respectively. The protein unfolded within the simulated MC time interval in virtually all runs, and often early on. However, many runs begin with a 'waiting period', in which the protein molecule remains nativelike, typically with $n_{\mathrm{hb}} \sim 0.7$ and an RMSD of ~ 5 Å. This period, whose extent varies from run to run, is followed by rapid unfolding. The existence of a waiting period shows that the native state is a local free-energy minimum of the model at the temperature studied, although thermal fluctuations are relatively large, as can be seen from the RMSD and n_{hb} data.

Figure II.2c shows a control simulation at low temperature, 288 K. At this temperature, the molecule remained nativelike throughout the simulated time

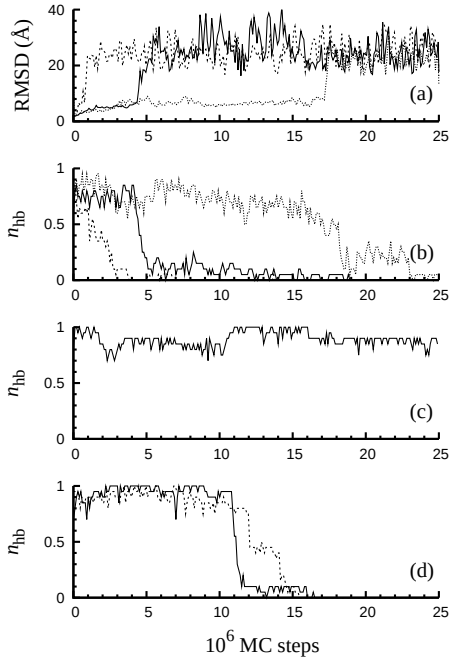

Figure II.2: MC evolution in representative runs with and without applied force. (a) RMSD from the native structure in three runs at 368 K and zero force. (b) The fraction of native backbone hydrogen bonds in the structures A–D, n_{hb}, in the same three runs. (c) n_{hb} in a control simulation at 288 K and zero force. (d) n_{hb} in two simulations of mechanical unfolding at 288 K and a force strength of 100 pN.

interval, with n_{hb} typically above 0.8, which is higher than in the waiting phase at 368 K.

Finally, Figure II.2d shows two runs from our study of force-induced unfolding [14], for a force of 100 pN and a temperature of 288 K. In one of these runs, unfolding is seen to occur in a single step. The other trajectory halts for a while at $n_{\mathrm{hb}} \sim 0.4$ before unfolding continues. The plateau at $n_{\mathrm{hb}} \sim 0.4$ signals a significantly populated intermediate state. Comparing our temperature- and force-induced events, we find that the unfolding process tends to be somewhat less abrupt in the thermal case, although event-to-event variations are large. Another difference is that we do not find any evidence of well-defined reoccurring intermediate states in thermal unfolding.

To get a more detailed picture of the unfolding process, we now turn to the order of breaking of the four secondary-structure elements labeled A–D in Figure II.1, which is investigated by measuring native backbone hydrogen bonds (see Materials and Methods). Figure II.3 illustrates how the presence of these hydrogen bonds varies with the degree of unfolding in our simulations of thermal as well as mechanical unfolding. The unfolding coordinate is $1 - n_{\mathrm{hb}}$ in the thermal case and the end-to-end distance r in the mechanical case. In the mechanical unfolding simulations (see Figure II.3b), it is immediately clear that the structures A–D tend to break in a certain order, namely CBDA. The points at which the different structures break are less well defined and less separated in the thermal case (see Figure II.3a). Nevertheless, it is evident that the unfolding order is not entirely random in the thermal unfolding simulations either. Specifically, it can be seen that C again has a tendency to break first. However, in the thermal case, C is followed by D rather than B. The order in which the remaining two structures A and B break is harder to decide; they are both partially present for essentially all values of the unfolding parameter $1 - n_{\mathrm{hb}}$. Inevitably, all structures disappear as $1 - n_{\mathrm{hb}}$ approaches one.

Figure II.3 provides an idea of what the preferred unfolding orders are in the thermal and mechanical unfolding simulations, respectively. In order to quantify how strong these statistical preferences are, it is necessary to refine the analysis of Figure II.3, which deals solely with averages over all events and hence does not tell how large event-to-event fluctuations are. An extended event-based analysis was carried out for the mechanical unfolding simulations [14]. Here we basically follow the same procedure, although some details are different. For each individual event, we determine an unfolding path (a permutation of ABCD), by directly analyzing the time of breaking of the structures A–D. A structure is deemed intact, at a given time, if one-half

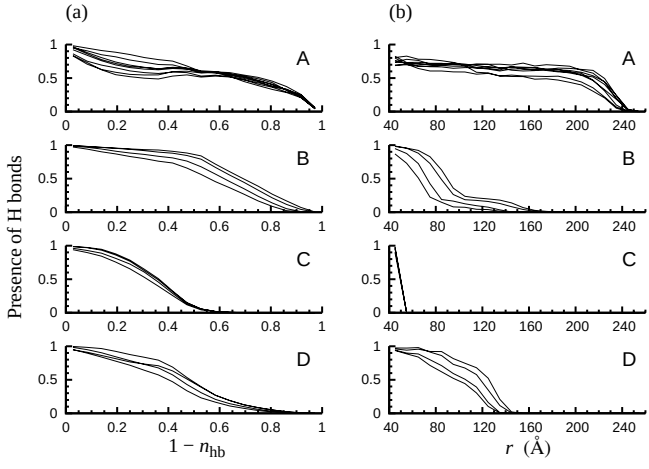

Figure II.3: Presence of native backbone hydrogen bonds during unfolding, for the structures A–D (see Figure II.1). Each curve represents one hydrogen bond, and is an average over all runs. The structure A has eight native backbone hydrogen bonds, whereas the other three structures have four bonds each. (a) Temperature-induced unfolding at 368 K (and zero force), with $1 - n_{\mathrm{hb}}$ as unfolding coordinate, where n_{hb} is the fraction of native backbone hydrogen bonds in A–D. (b) Force-induced unfolding at 288 K and 100 pN, with the end-to-end distance r as unfolding coordinate.

or more of its native backbone hydrogen bonds are present, and the time of breaking is defined as the last point in time after which the structure is never found intact. With these criteria, we find that 66 % of our mechanical unfolding events follow the path CBDA (at 100 pN and 288 K). Another 21 % of the events have the order of B and D apparently interchanged, the path being CDBA. In these events, B does unfold before D but then partially reforms after D is gone. This behavior can be regarded as a minor variation of the path CBDA, which means that 87 % of the events follow the same basic pathway. The partial refolding of B in the apparent CDBA events is responsible for the tail of the curves for B in Figure II.3b.

Repeating the same analysis for our thermal unfolding simulations leads to a different picture. As expected from Figure II.3a, there is no similarly dominant path in this case; the most common path is CDBA with a frequency of occurrence of 44 %. Nevertheless, some definite trends can be seen in the thermal unfolding data as well. In particular, there is a clear tendency for C and D to be the first structures to break. Among the six possible pairs of the structures A–D, this pair breaks first in 76 % of the events. The event-based analysis thus confirms that both C and D tend to break before B, as suggested by Figure II.3a. Compared to the preferred mechanical unfolding path, this result implies that the order of B and D is interchanged in a majority of the thermal unfolding events.

Especially for thermal unfolding, where event-to-event fluctuations are larger, it is instructive to directly compare the times of breaking of the structures A–D by means of scatter plots. Figure II.4 shows t_i, t_j scatter plots, where t_i and t_j are the times of breaking of structures i and j, for all the six possible pairs of the structures A–D. In all six cases, there is a clear tendency for one of the structures to break before the other. Particularly clear are the relations $t_C < t_A$, $t_C < t_B$ and $t_D < t_A$, which are fulfilled in 796, 786 and 747 events, respectively (out of 800). These strong statistical preferences clearly reveal that the unfolding order is less random than what one might expect from the event-averaged results shown in Figure II.3a. For the remaining three pairs of structures, we find that $t_C < t_D$, $t_D < t_B$ and $t_B < t_A$ hold in 644, 640 and 633 of the events, respectively.

Discussion

Using an all-atom model with a simplified interaction potential, we have examined the thermal unfolding of ubiquitin. In particular, we analyzed the unfolding order of four major secondary-structure elements, labeled A–D,

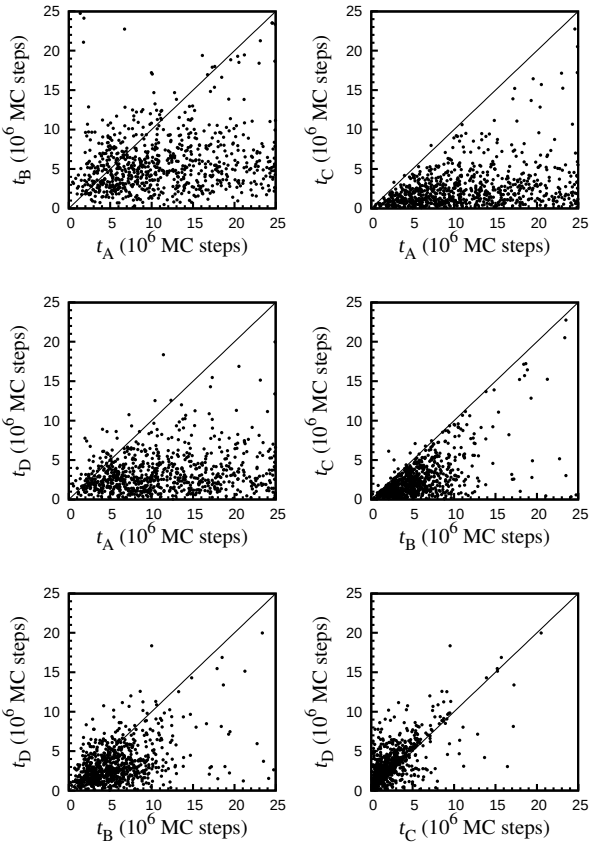

Figure II.4: t_i, t_j scatter plots, where t_i and t_j denote the times of breaking of structures i and j, for all the six possible pairs of the structures A–D. Each data point represents one unfolding simulation.

based on a set of 800 unfolding events. This analysis showed that there was no single statistically dominant unfolding pathway, but the unfolding order of A–D was nevertheless far from random. In particular, we found that the structures A and B, the α-helix and the first β-hairpin, typically survived longer than C and D in the simulations. Among six possible pairs of structures, the pair A and B was the last to break in ~ 75 % of the events. Experimentally, these two structures have been found to be the thermally most stable ones [4, 5]. Our results support this conclusion.

There are both similarities and differences between our thermal and mechanical [14] unfolding results for this protein. One difference is that the unfolding behavior was more deterministic in the mechanical case, and a statistically preferred unfolding order could be identified (CBDA). Another difference is that while B typically was the second structure to break in those simulations, B along with A are the structures that tend to survive longest in the thermal unfolding simulations. A similarity between the two sets of unfolding events is that the structure C tends to break first in both cases.

The results from the mechanical unfolding simulations can be compared with data from single-molecule constant-force experiments [17]. However, these experiments monitored only the end-to-end distance, and therefore do not provide any detailed information about the unfolding pathway. There is, by contrast, experimental information available about the thermal stability of different parts of the ubiquitin structure [4, 5], and our thermal unfolding results are, as mentioned above, consistent with these experiments. This agreement in the thermal case provides support for our calculated but experimentally unverified unfolding order for the mechanical case, since the model and methods were the same in both our studies.

The fact that we observe different pathways in thermal and mechanical unfolding is not surprising, because our study of the mechanical unfolding was, like the experiments [17], carried out for stretching forces of 100 pN or more. Such a large force shifts the energy balance between the native and typical intermediate states by as much as $\sim 100\,\text{pN}\cdot 7\,\text{nm} \sim 100$ kcal/mol. Hence, there is no reason to expect the unfolding behavior to be the same as at zero force.

Nevertheless, it is interesting to try to identify and compare the major factors dictating the thermal and mechanical unfolding behaviors. It is well-known that pulling geometry, in relation with native topology, is a major determinant of a protein's mechanical resistance [15, 43]. Pulling geometry and topology also explain why the ubiquitin elements C and B break early in our mechanical unfolding events. That C breaks first is inevitable, because

the stretching forces act on C and cannot be sensed by the other parts until C is broken. The native state is mechanically resistant because C is pulled longitudinally, so that several hydrogen bonds must break simultaneously. Once C is gone, the β-hairpin B is free to unzip, one bond at a time. Unzipping requires less force than separation by longitudinal pulling, which makes B likely to break quickly when no longer protected by C.

As mentioned earlier, differences between thermal and mechanical unfolding have been seen in previous simulations of, for example, the I27 domain of titin [20, 21]. For this β-sandwich protein, it was found that the breaking of contacts between its two β-sheets was an early event when stretching the protein, but occurred late in thermal unfolding. On the other hand, it was found that the same three β-strands are unfolded last in both thermal and mechanical unfolding. For ubiquitin, our results suggest that the most long-lived β-strands are different in thermal and mechanical unfolding, respectively.

In pulling the ends of a protein some elements of the structure are exposed to force while others are not. In thermal unfolding, there is no external force acting selectively on certain parts of the protein. Instead, unfolding is driven by thermal fluctuations. As a result, intrinsic stabilities should become more important. The parts of ubiquitin that are thermally most stable, A and B, both consist of connected stretches of residues along the sequence. The local character makes it possible for these structures to reform spontaneously. Experimentally, it has been found that the excised peptides A and B both have a tendency to make nativelike structure [12, 13], which shows that these structures possess an intrinsic stability. The structure element that tends to break first in the simulations, C, is, on the other hand, the only parallel β-sheet structure in the native state, and might have a relatively low intrinsic stability due to a different hydrogen bond geometry and other factors. These observations show that differences in intrinsic stability might, indeed, partly explain the behavior seen in our thermal unfolding simulations. However, there are also other factors that should be important. In particular, it is worth noting that the structure C has a key role in the native topology; it connects the two ends of the chain and occupies a central position in the β-sheet. As a result, C is likely to play an important protecting role in thermal unfolding as well, and not only in the mechanical case where the external forces acted on this particular structure.

Finally, we wish to stress that our studies of the thermal and the mechanical unfolding of ubiquitin both were performed using the same model, without adjusting any parameters. There are peptides that this model, developed by studies of peptide folding, fails to fold [35]. Nevertheless, it is encouraging

that this model, with its relatively simple potential, is able to capture relevant features seen in thermal as well as mechanical unfolding experiments on ubiquitin. To what extent the applicability of the model can be extended to a wider spectrum of sequences, by refinement of the potential, remains to be seen.

Acknowledgments. We thank Sandipan Mohanty for discussions and computational assistance. This work was in part supported by the Swedish Research Council.

References

1. Khorasanizadeh S, Peters ID, Roder H (1996) Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. Nat Struct Biol 3:193–205.
2. Krantz BA, Sosnick TR (2000) Distinguishing between two-state and three-state models for ubiquitin folding. Biochemistry 39:11696–11701.
3. Went HM, Benitez-Cardoza CG, Jackson SE (2004) Is an intermediate state populated on the folding pathway of ubiquitin? FEBS Lett 567:333–338.
4. Cordier F, Grzesiek S (2002) Temperature-dependence of protein hydrogen bond properties as studied by high-resolution NMR. J Mol Biol 317:739–752.
5. Chung HS, Khalil M, Smith AW, Ganim Z, Tokmakoff A (2005) Conformational changes during the nanosecond-to-millisecond unfolding of ubiquitin. Proc Natl Acad Sci USA 102:612–617.
6. Babu CR, Hilser VJ, Wand AJ (2004) Direct access to the cooperative substructure of proteins and the protein ensemble via cold denaturation. Nat Struct Mol Biol 11:352–357.
7. Sayle RA, Milner-White EJ (1995) Rasmol: biomolecular graphics for all. Trends Biochem Sci 20:374–376.
8. Went HM, Jackson SE (2005) Ubiquitin folds through a highly polarized transition state. Protein Eng Des Sel 18:229–237.
9. Krantz BA, Dothager RS, Sosnick TR (2004) Discerning the structure and energy of multiple transition states in protein folding using ψ-analysis. J Mol Biol 337:463–475.
10. Fersht AR (2004) φ value versus ψ analysis. Proc Natl Acad Sci USA 101:17327–17328.
11. Sosnick TR, Dothager RS, Krantz BA (2004) Differences in the folding transition state of ubiquitin indicated by φ and ψ analyses. Proc Natl Acad Sci USA 101:17377–17382.
12. Zerella R, Evans PA, Ionides JMC, Packman LC, Trotter BW, et al. (1999) Autonomous folding of a peptide corresponding to the β-hairpin from ubiquitin. Protein Sci 8:1320–1331.
13. Jourdan M, Searle MS (2000) Cooperative assembly of a nativelike ubiquitin structure through peptide fragment complexation: energetics of peptide association and folding. Biochemistry 39:12355–12364.

14. Irbäck A, Mitternacht S, Mohanty S (2005) Dissecting the mechanical unfolding of ubiquitin. Proc Natl Acad Sci USA 102:13427–13432.
15. Carrion-Vazquez M, Li H, Lu H, Marszalek PE, Oberhauser AF, et al. (2003) The mechanical stability of ubiquitin is linkage dependent. Nat Struct Biol 10:738–743.
16. Fernandez JM, Li H (2004) Force-clamp spectroscopy monitors the folding trajectory of a single protein. Science 303:1674–1678.
17. Schlierf M, Li H, Fernandez JM (2004) The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques. Proc Natl Acad Sci USA 101:7299–7304.
18. Sarkar A, Robertson RB, Fernandez JM (2004) Simultaneous atomic force microscope and fluorescence measurements of protein unfolding using a calibrated evanescent wave. Proc Natl Acad Sci USA 101:12882–12886.
19. Chyan CL, Lin FC, Peng H, Yuan JM, Chang CH, et al. (2004) Reversible mechanical unfolding of single ubiquitin molecules. Biophys J 87:3995–4006.
20. Paci E, Karplus M (2000) Unfolding proteins by external forces and temperature: the importance of topology and energetics. Proc Natl Acad Sci USA 97:6521–6526.
21. Cieplak M, Sułkowska J (2005) Thermal unfolding of proteins. J Chem Phys 123:194908.
22. Alonso DOV, Daggett V (1998) Molecular dynamics simulations of hydrophobic collapse of ubiquitin. Protein Sci 7:860–874.
23. Sorenson JM, Head-Gordon T (2002) Toward minimalist models of larger proteins: a ubiquitin-like protein. Proteins 46:368–379.
24. Sosnick TR, Berry RS, Colubri A, Fernández A (2002) Distinguishing foldable proteins from nonfolders: when and how do they differ? Proteins 49:15–23.
25. Zhang J, Qin M, Wang W (2005) Multiple folding mechanisms of protein ubiquitin. Proteins 59:565–579.
26. Gilis D, Rooman M (2001) Identification and ab initio simulations of early folding units in proteins. Proteins 42:164–176.
27. Michnick SW, Shakhnovich E (1998) A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. Fold Des 3:239–251.
28. Li PC, Makarov DE (2004) Simulation of the mechanical unfolding of ubiquitin: probing different unfolding reaction coordinates by changing the pulling geometry. J Chem Phys 121:4826–4832.
29. Kirmizialtin S, Huang L, Makarov DE (2005) Topography of the free-energy landscape probed via mechanical unfolding of proteins. J Chem Phys 122:234915.
30. Szymczak P, Cieplak M (2006) Stretching of proteins in a force-clamp. J Phys: Condens Matter 18:L21–L28.
31. Lu H, Schulten K (1999) Steered molecular dynamics simulations of force-induced protein domain unfolding. Proteins 35:453–463.
32. Lu H, Schulten K (2000) The key event in force-induced unfolding of titin's immunoglobulin domains. Biophys J 79:51–65.
33. Fowler SB, Best RB, Herrera JLT, Rutherford TJ, Steward A, et al. (2002) Mechanical unfolding of a titin Ig domain: structure of unfolding intermediate revealed by combining AFM, molecular dynamics simulations, NMR and protein engineering. J Mol Biol 322:841–849.
34. Irbäck A, Samuelsson B, Sjunnesson F, Wallin S (2003) Thermodynamics of α- and β-structure formation in proteins. Biophys J 85:1466–1473.

35. Irbäck A, Mohanty S (2005) Folding thermodynamics of peptides. Biophys J 88:1560–1569.
36. Favrin G, Irbäck A, Mohanty S (2004) Oligomerization of amyloid Aβ(16-22) peptides using hydrogen bonds and hydrophobicity forces. Biophys J 87:3657–3664.
37. Wintrode PL, Makhatadze GI, Privalov PL (1994) Thermodynamics of ubiquitin unfolding. Proteins 18:246–253.
38. Irbäck A, Mohanty S (2006) PROFASI: a Monte Carlo simulation package for protein folding and aggregation. J Comput Chem 27:1548–1555.
39. Cornilescu G, Marquardt JL, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute crystalline phase. J Am Chem Soc 120:6836–6837.
40. Favrin G, Irbäck A, Sjunnesson F (2001) Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. J Chem Phys 114:8154–8158.
41. Shimada J, Kussell EL, Shakhnovich EI (2001) The folding thermodynamics and kinetics of crambin using an all-atom monte carlo simulation. J Mol Biol 308:79–95.
42. Rey A, Skolnick J (1991) Comparison of lattice Monte Carlo dynamics and Brownian dynamics folding pathways of α-helical hairpins. Chem Phys 158:199–219.
43. Brockwell DJ, Paci E, Zinober RC, Beddard GS, Olmsted PD, et al. (2003) Pulling geometry defines the mechanical resistance of a β-sheet protein. Nat Struct Biol 10:731–737.

PAPER III

Spontaneous β-barrel formation: an all-atom Monte Carlo study of Aβ(16–22) oligomerization

Anders Irbäck and Simon Mitternacht

Computational Biology and Biological Physics, Department of Theoretical Physics, Lund University, Sweden.

Proteins 71: 207–214. (2008)

Using all-atom Monte Carlo (MC) simulations with implicit water combined with a cluster size analysis, we study the aggregation of $A\beta_{16-22}$, a peptide capable of forming amyloid fibrils. We consider a system of six initially randomly oriented $A\beta_{16-22}$ peptides, and investigate the thermodynamics and structural properties of aggregates formed by this system. The system is unaggregated without ordered secondary structure at high temperature, and forms β-sheet rich aggregates at low temperature. At the crossover between these two regimes, we find that clusters of all sizes occur, whereas the β-strand content is low. In one of several runs, we observe the spontaneous formation of a β-barrel with six antiparallel strands. The β-barrel stands out as the by far most long-lived aggregate seen in our simulations.

Introduction

Amyloid fibril formation is a symptom of several human neurodegenerative diseases, but increasing evidence suggests that the neurotoxic agent is not the fibrils themselves [1]. Much current research is directed at characterizing small soluble oligomers of amyloid proteins, in order to identify the major toxic species [2]. For example, in a study of Alzheimer's amyloid-β protein, Aβ, it was found that 56-kDa Aβ assemblies could be linked to loss of memory function in mouse and rat [3].

While atomic-level structural models have now emerged for some amyloid fibrils [4–7], less is known about the detailed structure of the oligomeric states. These states are difficult to characterize due to their transient nature — they can transform into other classes of oligomers, break up into monomers or move onto fibril formation (which is probably an irreversible process, so that there is no equilibrium distribution). On the other hand, there are some common classes of oligomeric states that have been observed for several amyloid proteins, like spherical, chain-like and annular species [8]. One possible explanation of neurotoxicity is that annular pore-like aggregates cause membrane permeabilization [8, 9].

Here we explore the structure and stability of small Aβ_{16-22} oligomers, by all-atom MC simulations for a system of six Aβ_{16-22} peptides. This seven-residue fragment of Aβ is known to be able to make amyloid fibrils [10]. Furthermore, it has been demonstrated, by solid-state NMR, that the β-strand organization is antiparallel in Aβ_{16-22} fibrils [10, 11]. A recent study of Aβ_{16-22} by infrared spectroscopy (IR), found evidence for antiparallel β-sheet structure also in solution [12].

The availability of experimental data, its small size, and the fact that it spans an aggregation prone region of Aβ [13], make the Aβ_{16-22} peptide a suitable model system for computational studies, and simulations of Aβ_{16-22} aggregation have been reported by several groups [14–23]. Here we study Aβ_{16-22} oligomerization by unbiased thermodynamic simulations started from random initial conformations. We use a simple but novel procedure to identify which clusters are formed in a given multichain conformation. Another novelty is that we observe the spontaneous formation of a β-barrel. The formation of annular, β-barrel-like structures [24, 25] and open high-curvature β-sheets [26, 27] has previously been seen in simulations based on coarse-grained models, but as far as we know, not in atomic-level simulations for Aβ_{16-22} or any other sequence. Reviews of computational aggregation studies of amyloid peptides can be found in two recent articles [28, 29].

Calculations similar to those presented here, but without any cluster size analysis, have been reported earlier [16]. In that study no β-barrel was observed. The fact that we do so here could, in part, simply be due to improved statistics (by a factor 6). Another difference is that we here use a slightly modified energy function that incorporates attractions and repulsions between side-chain charges (see Materials and methods), which increase the stability of the β-barrel (see Results and Discussion).

Materials and methods

Model. The system we study consists of six Aβ_{16-22} peptides (acetyl-Lys-Leu-Val-Phe-Phe-Ala-Glu-NH_2) contained in a periodic box of size $(50.4\,\text{Å})^3$, with implicit water. All atoms of the peptide chains are included in our calculations, but we assume fixed bond lengths, bond angles and peptide torsion angles (180°), so that each residue only has the Ramachandran torsion angles φ, ψ and a number of side-chain torsion angles as its degrees of freedom. Numerical values of the geometrical parameters held constant can be found elsewhere [30].

The energy function we use is a close variant of an energy function [30, 31] that has been used to study the folding of several peptides with about 20 residues [31], the aggregation of Aβ_{16-22} [16], and the mechanical and thermal unfolding of ubiquitin [32, 33]. It is composed of four terms,

$$E = E_{\text{loc}} + E_{\text{ev}} + E_{\text{hb}} + E_{\text{sc}}\,. \tag{III.1}$$

The term E_{loc} is an intrachain potential and local in sequence. It represents an electrostatic interaction between adjacent peptide units along the chain. The other three terms are both intra- and interchain potentials, and non-local in sequence. The excluded volume term E_{ev} is a $1/r^{12}$ repulsion between pairs of atoms. E_{hb} represents two kinds of H bonds: backbone-backbone bonds and bonds between charged side chains and the backbone. The last term E_{sc} represents interactions between pairs of side chains. It is a simple pairwise additive potential based on the degree of contact between two side chains.

There is one major difference between the energy function used here and that used in the previous studies of this model [16, 31–33]. The difference is in the term E_{sc}, which in the previous studies represented an effective hydrophobic attraction between pairs of non-polar side chains. In the present study, we have incorporated attraction and repulsion between charged side chains into this term, while leaving the hydrophobicity part unchanged. The interaction between two charged side chains is assumed to be short range, due

to screening by water, and is, for simplicity, taken to have the same functional form as the hydrophobic interaction between two non-polar side chains [31]. Specifically, the new part is of the form

$$E_{\mathrm{sc}}^{(\mathrm{q})} = \epsilon_{\mathrm{q}} \sum_{I<J} q_I q_J C_{IJ}\,, \tag{III.2}$$

where I and J denote charged residues, q_I and q_J are charges (± 1), and ϵ_{q} sets the strength of the interaction (≈ 2.0 kcal/mol). C_{IJ} is a measure of the degree of contact between two amino acids

$$C_{IJ} = \frac{1}{N_I + N_J}\left[\sum_{i\in A_I} f(\min_{j\in A_J} r_{ij}^2) + \sum_{j\in A_J} f(\min_{i\in A_I} r_{ij}^2)\right]. \tag{III.3}$$

A_I denotes a predefined set of atoms: for Glu it is the two side-chain oxygens and for Lys the hydrogens in the NH_3-group (there are no Asp or Arg residues in Aβ_{16-22}). N_I and N_J are the number of atoms in the sets A_I and A_J. The function $f(x)$ is given by $f(x) = 1$ if $x < A$, $f(x) = 0$ if $x > B$, and $f(x) = (B-x)/(B-A)$ otherwise [$A = (3.5\text{Å})^2$, $B = (4.5\text{Å})^2$]. The form of the other energy terms has been described elsewhere [31].

Having modified the energy function, we also recalibrated the energy scale of the model, by folding simulations for the Trp cage peptide. The energy scale of the model is determined by using the model prediction for the melting temperature of Trp cage and the experimental value [34] for the same (315 K). On the internal scale of the model, the melting temperature changed by 2 % (from 0.470 to 0.479) with the new energy function.

Simulation methods. The thermodynamics of aggregation for this system is investigated by using simulated tempering [35–37], in which the temperature is a dynamical variable. This method is closely related to the replica exchange, or parallel tempering, method [38–40]. The main difference is that simulated tempering works with only one copy of the system, whereas the replica exchange method simulates several copies of the system in parallel (which exchange temperatures with each other). Our simulations are carried out using the software package PROFASI [41]. A total of 30 independent simulated-tempering runs is collected. 10 of the runs span six temperatures from 293 K to 362 K, whereas the other 20 runs span five temperatures from 306 K to 362 K. The six-temperature runs each comprise 10^{10} elementary MC steps. The length of each five-temperature run is $6 \cdot 10^9$ elementary MC steps. All the runs are started from random conformations.

For the backbone degrees of freedom, we use two different elementary moves: single-variable updates of individual torsion angles, which is a non-

local method, and Biased Gaussian Steps [42], a semi-local move that simultaneously updates up to eight angles. Side-chain angles are updated one by one. In addition to these updates, for computational efficiency, we also include rigid-body translations and rotations of whole chains. Every update involves a Metropolis accept/reject step, thus ensuring detailed balance.

Measurements. To monitor the aggregation state of the system, we use a cluster size analysis. Two chains I and J are said to be in the same cluster if the sum of their interchain side-chain interactions, $E_{\rm sc}(I,J)$, and interchain backbone-backbone H bond energy, $E_{\rm hb}^{\rm bb}(I,J)$, is lower than a cutoff, $E_{\rm sc}(I,J)+E_{\rm hb}^{\rm bb}(I,J)<-1.5\epsilon_{\rm hb}^{(1)}$, corresponding to 2–3 H bonds ($\epsilon_{\rm hb}^{(1)}$ sets the strength of backbone-backbone H bonds [31]). The cutoff is chosen to exclude brief random contacts, without being too restrictive. The size of the largest cluster in a given conformation is denoted by Λ. The clusters obtained by this definition are referred to as *general* clusters, or simply clusters, and may completely lack ordered secondary structure. We also use a stricter cluster definition. These clusters are referred to as *ordered* clusters. An ordered cluster is formed by pairs I, J of chains with $E_{\rm hb}^{\rm bb}(I,J)<-1.5\epsilon_{\rm hb}^{(1)}$ and a β-strand content, as defined below, higher than 0.3 for both chains. An ordered cluster is thus always part of a general cluster. The size of the largest ordered cluster is denoted by $\Lambda_{\rm o}$.

For a chain with N amino acids, we define the α-helix and β-strand contents as the fractions of the $N-2$ inner amino acids with their (φ,ψ) pair in the α-helix and β-strand regions of the Ramachandran space. We assume that α-helix corresponds to $-90^\circ<\varphi<-30^\circ$, $-77^\circ<\psi<-17^\circ$ and that β-strand corresponds to $-150^\circ<\varphi<-90^\circ$, $90^\circ<\psi<150^\circ$. The average α-helix and β-strand contents, over all the chains of the system, are denoted by n_α and n_β, respectively.

To determine the amounts of parallel and antiparallel β-sheet structure in a given multichain conformation, we consider all possible pairs of chains. We first identify all chain pairs I, J such that (i) their interchain backbone-backbone H bond energy satisfies $E_{\rm hb}^{\rm bb}(I,J)<-1.5\epsilon_{\rm hb}^{(1)}$, and (ii) both chains I and J have a β-strand content higher than 0.5. For each such pair of chains, we then calculate the scalar product of their normalized end-to-end unit vectors. If this scalar product is greater than 0.7 (less than -0.7), we say that the two chains are parallel (antiparallel). The numbers of parallel and antiparallel pairs of chains are denoted by n_+ and n_-, respectively.

The simulation data are analyzed using multi-histogram techniques [43]. All statistical uncertainties quoted are 1σ errors obtained by the jackknife method [44].

Figures of 3D structures were prepared using PyMOL [45].

β-barrel geometry. The geometrical features of regular β-barrels are determined by the number of strands, n, and the shear number, S [46–50]. The shear number can be obtained by drawing a curve around the barrel perpendicular to the strands, from some reference strand until this strand is reached again. Let i and j be the residues on this strand at which the curve starts and ends, respectively. The shear number is $|i-j|$. In this definition, the curve can be traversed in either of two possible directions. If the direction is specified, the shear number can be given a sign. Because β-barrels have a right-handed twist, the sign is, however, not needed.

The tilt angle of the strands relative to the barrel axis, α, is given by $\tan\alpha = Sa/nb$, where the constants $a = 3.3\,\text{Å}$ and $b = 4.4\,\text{Å}$ are the C_α-C_α distance along the strands and the interstrand distance, respectively.

Results and discussion

Thermodynamics. Using the methods described above, we study the system of six Aβ_{16-22} peptides in the temperature range $293\,\text{K} < T < 362\,\text{K}$. Figure III.1a shows the calculated specific heat curve, which exhibits a sharp peak centered at $T = T_{\max} \approx 335\,\text{K}$. Our cluster size analysis (see Materials and methods) reveals that large aggregated structures start to form around this temperature. Figure III.1b shows the probability of having a cluster with five or six chains, $P(\Lambda \geq 5)$, against temperature. This probability increases from close to 0 at high temperature to more than 0.9 at low temperature, through a sigmoid-like transition centered in the vicinity of $T_{\max}$. There is also a clear increase in β-strand content, $\langle n_\beta \rangle$, as the temperature decreases (see Figure III.1c). The α-helix content, $\langle n_\alpha \rangle$, is, by contrast, small throughout the temperature range studied (data not shown). We thus find that the chains lack ordered secondary structure at high temperature, but form β-sheet structure at low temperature.

An interesting detail in Figure III.1 is that $P(\Lambda \geq 5)$ starts to increase slightly before $\langle n_\beta \rangle$, as the temperature is decreased. At $T = T_{\max}$, clusters with 5 or 6 chains occur with a significant frequency, whereas the β-strand content is small.

Figure III.2a shows the probability that the largest cluster is of size n, $P(\Lambda = n)$, against temperature, for different n. The maximum of $P(\Lambda = 2)$ is at 357 K, which is well above the specific-heat maximum $T_{\max}$. As n is increased, the maximum of $P(\Lambda = n)$ shifts toward lower temperature. Near $T_{\max}$, all Λ are roughly equally probable, showing that clusters of all sizes occur. At low

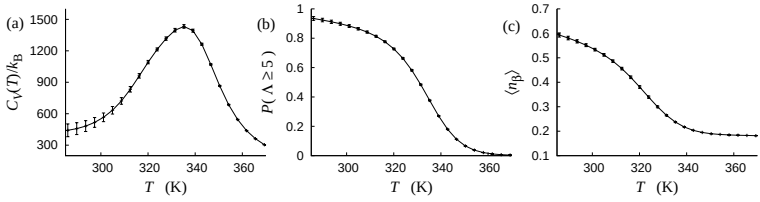

Figure III.1: Temperature dependence of (a) the specific heat $C_V = (\langle E^2 \rangle - \langle E \rangle^2)/k_B T^2$, (b) the probability that the size of the largest (general) cluster is 5 or 6, $P(\Lambda \geq 5)$, and (c) the β-strand content $\langle n_\beta \rangle$.

temperature, $\Lambda = 6$ is the by far most common value, so all the chains tend to form a single cluster.

Figure 2b shows an analysis similar to that in Figure III.2a but for ordered clusters. Since an ordered cluster is always part of a general cluster, the size of the largest ordered cluster, Λ_o, cannot exceed Λ. Large ordered clusters are, in contrast to large general clusters, very rare at $T = T_{max}$, where $P(\Lambda_o = 5)$ and $P(\Lambda_o = 6)$ both are close to 0. The absence of large ordered clusters is consistent with the finding that the β-strand content is small at $T = T_{max}$ (see Figure III.1c). Large ordered clusters are, by contrast, common at the lowest temperatures studied, where $P(\Lambda_o \geq 5) > 0.5$.

The ordered aggregates seen at low temperature contain β-sheets, which can be either parallel, antiparallel or mixed parallel/antiparallel. All these three kinds of β-sheet structure occur in our simulations. To find out whether there is a preference for parallel or antiparallel organization, we examine the probability distribution of the variables n_+ and n_- (see Materials and methods), $P(n_+, n_-)$. Figure III.3 shows $P(n, 0)$ and $P(0, n)$ as functions of n at a fixed temperature of 306 K. $P(n, 0)$ is markedly smaller than $P(0, n)$ for all n, implying that antiparallel structures are more common than parallel ones. For comparison, Figure III.3 also shows the corresponding results from simulations with the interactions between side-chain charges switched off. As previously reported [16], we find a clear preference for the antiparallel organization in this case, too, although slightly

weaker. This finding suggests that the interactions between side-chain charges are not alone responsible for the antiparallel organization seen in Aβ_{16-22} experiments [10–12].

Figure III.4 shows the MC evolution of the energy E in two of our runs, along with some snapshots of long-lived aggregated structures. The simulation time is long enough for aggregated structures to both form and dissolve in the

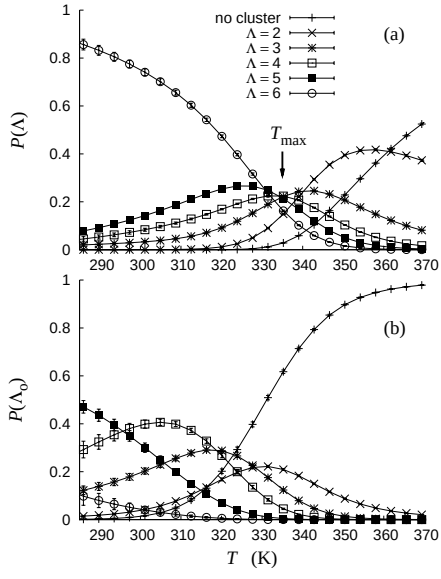

Figure III.2: (a) The probability that the largest general cluster is of size n, $P(\Lambda = n)$, against temperature, for $n = 2, \ldots, 6$. The curve labeled 'no cluster' is the probability that there is no cluster of size ≥ 2. The specific-heat maximum, $T_{\max}$, is marked with an arrow. (b) The corresponding plot for ordered clusters.

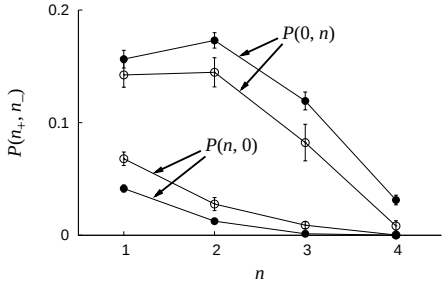

Figure III.3: The probability distribution $P(n_+, n_-)$ for $(n_+, n_-) = (n, 0)$ and $(n_+, n_-) = (0, n)$, against n, at 306 K, as obtained from simulations with (•) and without (∘) interactions between charged side chains. The variables n_+ and n_- count parallel and antiparallel pairs of interacting β-strands (see Materials and methods).

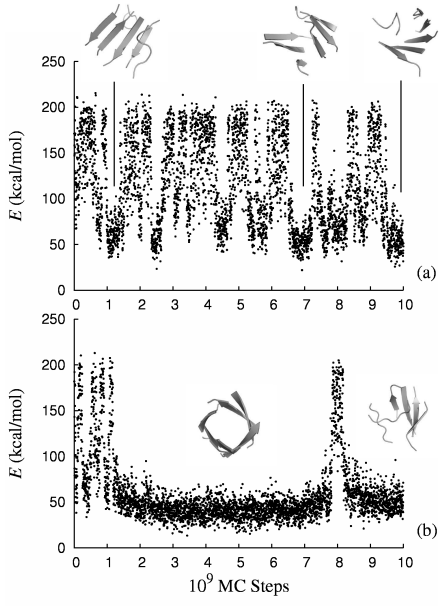

Figure III.4: MC evolution of the energy in two simulated-tempering runs for the system of six Aβ_{16-22} peptides. Also shown are some snapshots of long-lived aggregates that formed in the course of the runs.

course of the runs, but simulating this system with >900 atoms is nevertheless a challenge. To test the convergence of our results, we computed the specific heat with statistical errors using two different data sets, the five- and six-temperature runs, respectively (see Materials and Methods). The results of the two analyses were in perfect agreement. This agreement suggests that the relative weights of high- and low-energy states are properly sampled, so that quantities like the specific heat can be reliably estimated.

Longer simulations would, by contrast, be needed in order to determine the relative weights of different low-energy states. Nevertheless, we next take a closer look at one particular low-energy state, the β-barrel. The β-barrel only occurs in one of our runs, the one shown in Figure III.4b, although several runs contain curved, almost closed β-sheets. The β-barrel has lower energy than any other observed state and stays intact over a very long period, about $5 \cdot 10^9$ MC steps (see Figure III.4b). It is by far the most long-lived state seen in any of our simulations.

Another caveat of the analysis presented above is the small number of chains used. In order to make testable predictions for thermodynamic quanti-

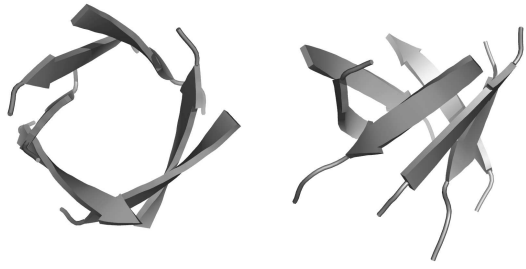

Figure III.5: Schematic snapshot of the β-barrel that formed in one of our runs. The two pictures are different views of the same structure.

ties like the temperature at which aggregation sets in, it would be necessary to study larger systems. New techniques for encapsulating peptides are, however, being developed [51], which have the potential to facilitate future comparisons of experimental and computational studies.

Structure and stability of the β-barrel. The geometry of regular β-barrel structures can be classified by the number of strands, n, and the shear number, S (see Materials and methods) [46–50]. It has been argued that regular β-barrels with good β-sheet geometries and well-packed interiors can be obtained only for a limited set of 10 different (n,S) pairs, namely $(n,8)$ with $4 \leq n \leq 8$, $(n,10)$ with $5 \leq n \leq 8$, and $(n,S) = (6,12)$ [48, 49]. For $n = 6$, the preferred values of S are 8, 10 and 12. The corresponding tilt angles are in the range 45° to 56°.

Figure III.5 shows a schematic snapshot of the β-barrel observed in our simulations. It is (right-)twisted, as it should, and composed of six antiparallel strands that are tilted relative to the barrel axis. A closer inspection reveals that among the six pairs of adjacent strands, the alignment is in register for one pair and out of register by two residue units for the other five pairs, leading to a shear number of $S = 10$. The observed barrel thus has one of the three preferred S values for regular six-stranded barrels. In fact, $S = 10$ is expected to be optimal with respect to β-sheet geometry, given $n = 6$ [48]. The (n,S) classification of this barrel is thus perfectly consistent with its high apparent stability (see Figure III.4b).

At first glance, a β-barrel may seem inconsistent with IR experiments on $A\beta_{16-22}$ in solution [12], which found evidence for an antiparallel in-register β-sheet structure, corresponding to $S = 0$. However, this behavior was observed

after a relaxation period, during which the spectra changed with time. A non-negligible β-barrel population might have been present in the early stages.

Our calculations aim at exploring thermodynamically relevant states of the system, rather than the kinetics of the aggregation process. Nevertheless, they can elucidate possible pathways for the formation of Aβ_{16-22} oligomers. Figure III.6 illustrates how the configuration of ordered clusters evolves with MC time in two runs. The first run (Figure III.6a) is the one containing the β-barrel. The barrel is present over a period extending roughly from $2 \cdot 10^9$ MC steps to $7 \cdot 10^9$ MC steps, during which the chains tend to form an ordered cluster of size 6. Immediately before the formation of the barrel, there is a brief phase in which the largest ordered cluster is typically of size 5. This period is, in turn, preceded by a stage dominated by '3+3' conformations, which have two ordered clusters of size 3. During the whole duration of the intermediate peak for ordered clusters of size 5, it turns out that the probability of having a *general* cluster of size 6 is above 0.8 (data not shown). So, all chains are in contact most of this time, but one chain forms less ordered contacts with the other chains. The emerging picture is that of two three-stranded β-sheets merging to form the barrel, although it is not a simple docking of two rigid structures.

An ordered cluster of size 6 occurs in the second run as well (Figure III.6b). It is present approximately between $1.0 \cdot 10^9$ MC steps and $1.8 \cdot 10^9$ MC steps. Inspection of snapshots from the run shows that this cluster corresponds to a six-stranded β-sheet that is curved but not closed. The formation of this state occurs through addition of chains one by one to a growing β-sheet, rather than through fusion of two smaller β-sheets. Growth by monomer addition has been experimentally verified as a viable mechanism of fibril formation [52].

The Aβ_{16-22} peptide has two charged side chains, those of the end residues Lys16 and Glu22. The formation of the β-barrel, with antiparallel strands, brings oppositely charged side chains close to each other. The interactions between these charges might be crucial for the stability of the β-barrel, especially since the number of chains is even, so that a closed structure can be created without getting any conflicting pair of nearby like charges. To test the importance of the interactions between side-chain charges, we study the stability of the β-barrel both with and without these interactions in the model. For each of the two cases, 100 runs are performed at a constant temperature of 334 K, which is near the specific-heat maximum $T_{\max}$. In these stability tests, we do not use the non-local single-variable update of backbone angles (see Materials and methods). The runs are started with the β-barrel as the initial conformation. Figure III.7 shows the MC evolution of the probability

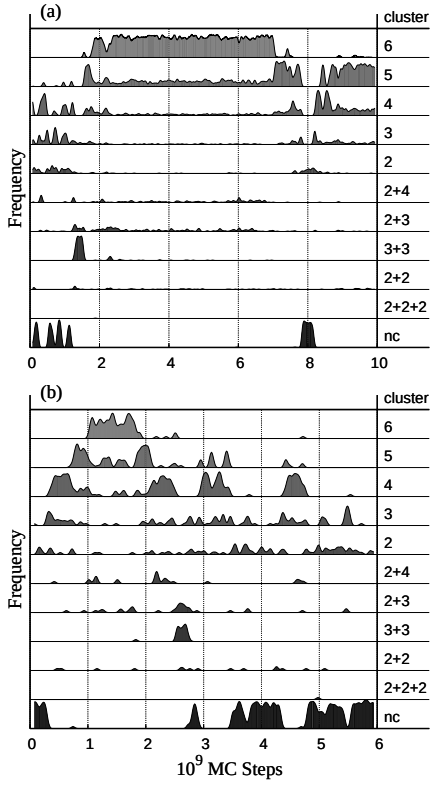

Figure III.6: The probability of different configurations of ordered clusters against MC time in two simulated-tempering runs. A label 'n' stands for conformations with one ordered cluster of size n, '$n+m$' for conformations with two ordered clusters of sizes n and m, etc., whereas 'nc' denotes conformations without any ordered cluster. The probabilities are calculated as sliding averages over the respective time series, using a Gaussian weight factor with a variance of $3 \cdot 10^7$ MC steps. (a) The run in which a β-barrel formed (same as in Figure III.4b). (b) A run in which a curved but not closed six-stranded β-sheet occurs.

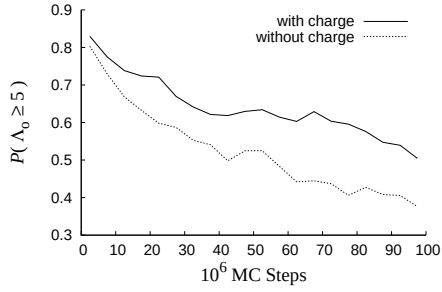

Figure III.7: MC evolution of the probability of having an ordered cluster of size $\Lambda_o \geq 5$ at 334 K, in runs with the β-barrel in Figure III.5 as the initial conformation. The two curves represent simulations with (solid) and without (dashed) interactions between charged side chains, respectively. To reduce noise, raw data are binned into subintervals of size $5 \cdot 10^6$ MC steps. Each curve is an average over 100 runs.

of having an ordered cluster of size 5 or 6, $P(\Lambda_o \geq 5)$, as obtained from these two sets of runs. Initially, this probability is 1, since $\Lambda_o = 6$ for the β-barrel conformation. The subsequent decay of $P(\Lambda_o \geq 5)$ is seen to be faster if the interactions between side-chain charges are removed from the model, which shows that the β-barrel indeed is less stable in this case. The difference is, however, quantitative rather than qualitative. The β-barrel remains a local free-energy minimum, although less pronounced, in the absence of these interactions.

Conclusion

To explore the nature of small oligomers of the fibril-forming Aβ_{16-22} peptide, we have performed an all-atom study of six Aβ_{16-22} peptides enclosed in a periodic box. A simple but useful cluster size analysis was devised and employed to characterize multichain conformations with respect to aggregates. The analysis distinguishes between general and ordered clusters. At the onset of aggregation, where the specific heat has a sharp peak, general clusters of all sizes were found to occur. Large ordered clusters are, by contrast, rare at this temperature. This difference indicates that hydrophobic association tends to precede secondary-structure formation in the aggregation process. The formation of ordered aggregates may involve fusion of two smaller aggregates, or occur by monomer addition to a growing β-sheet. The runs discussed in

Figure III.6, although not kinetic simulations, illustrate these two types of behavior.

At low temperature, the aggregated structures were found to be β-sheet rich, with either parallel, antiparallel or mixed parallel/antiparallel strands. While all these three possibilities occur in the simulations, a clear statistical preference was found for antiparallel over parallel alignment, which is consistent with Aβ_{16-22} experiments [10–12]. It is worth noting that the antiparallel preference persists, although slightly reduced, upon removal of the interactions between side-chain charges. This finding suggests that these interactions are not alone responsible for the antiparallel alignment, but other factors play a significant role, too.

In one of our runs, the six chains spontaneously self-assembled into a β-barrel. It occurred only once in 30 runs, but once formed, the β-barrel remained intact over an extraordinary long period, about $5 \cdot 10^9$ MC steps. This behavior suggests that the β-barrel represents a sharp but not easily accessible minimum on the free-energy landscape. Many other aggregated conformations were also seen in the simulations, indicating a rugged free-energy landscape with many distinct minima.

The observed β-barrel has six antiparallel strands and a shear number of $S = 10$. This value of S is what one expects for a six-stranded barrel with optimal β-sheet geometry [48], which in part explains the high apparent stability of the state.

For peptides in the diverse class of amyloid-forming sequences, the β-barrel motif is a natural but not obvious candidate for a relatively stable oligomeric state. A β-barrel, indeed, is the most long-lived species found in our Aβ_{16-22} simulations. To address the question of whether the ability to form β-barrels is a common property among amyloid-forming peptides, it would be interesting to extend these calculations to other sequences.

Acknowledgment. We thank Sandipan Mohanty for valuable discussions. This work was in part supported by the Swedish Research Council. The simulations were performed at the LUNARC facility at Lund University.

References

1. Lansbury Jr PT, Lashuel HA (2006) A century-old debate on protein aggregation and neurodegeneration enters the clinic. Nature 443:774–779.
2. Teplow DB, Lazo ND, Bitan G, Bernstein S, Wyttenbach T, et al. (2006) Elucidating amyloid β-protein folding and asembly: a multidisciplinary approach. Acc Chem Res 39:635–645.

3. Lesné S, Koh MT, Kotilinek L, Kayed R, Glabe CG, et al. (2006) A specific amyloid-β protein assembly in the brain impairs memory. Nature 440:352–356.
4. Jaroniec CP, MacPhee CE, Bajaj VS, McMahon MT, Dobson CM, et al. (2004) High-resolution molecular structure of a peptide in an amyloid fibril determined by magic angle spinning NMR spectroscopy. Proc Natl Acad Sci USA 101:711–716.
5. Luehrs T, Ritter C, Adrian M, Riek-Loher D, Bohrmann B, et al. (2005) 3D structure of Alzheimers amyloid-(1-42) fibrils. Proc Natl Acad Sci USA 102:17342–17347.
6. Nelson R, Sawaya MR, Balbirnie M, Madsen AØ, Riekel C, et al. (2005) Structure of the cross-β spine of amyloid-like fibrils. Nature 435:773–778.
7. Makin OS, Atkins E, Sikorski P, Johansson J, Serpell LC (2005) Molecular basis for amyloid fibril formation and stability. Proc Natl Acad Sci USA 102:315–320.
8. Lashuel HA, Lansbury Jr PT (2006) Are amyloid diseases caused by protein aggregates that mimic bacterial pore-forming toxins? Q Rev Biophys 39:167–201.
9. Arispe N, Pollard HB, Rojas E (1994) Beta-amyloid Ca^{2+}-channel hypothesis for neuronal death in Alzheimer disease. Mol Cell Biochem 140:119–125.
10. Balbach JJ, Ishii Y, Antzutkin ON, Leapman RD, Rizzo NW, et al. (2000) Amyloid fibril formation by Aβ(16-22), a seven-residue fragment of the Alzheimer's β-amyloid peptide, and structural characterization by solid state NMR. Biochemistry 39:13748–13759.
11. Gordon DJ, Balbach JJ, Tycko R, Meredith SC (2004) Increasing the amphiphilicity of an amyloidogenic peptide changes the β-sheet structure in the fibrils from antiparallel to parallel. Biophys J 86:428–434.
12. Petty SA, Decatur SM (2005) Experimental evidence for the reorganization of β-strands within aggregates of the Aβ(16-22) peptide. J Am Chem Soc 127:13488–13489.
13. Tjernberg LO, Näslund J, Lindqvist F, Johansson J, Karlström AR, et al. (1996) Arrest of β-amyloid fibril formation by pentapeptide ligand. J Biol Chem 271:8545–8548.
14. Ma B, Nussinov R (2002) Stabilities and conformations of Alzheimer's β-amyloid peptide oligomers (Aβ(16-22), Aβ(16-35), and Aβ(10-35)): sequence effects. Proc Natl Acad Sci USA 99:14126–14131.
15. Klimov DK, Thirumalai D (2003) Dissecting the assembly of Aβ(16-22) amyloid peptides into antiparallel β sheets. Structure 11:295–307.
16. Favrin G, Irbäck A, Mohanty S (2004) Oligomerization of amyloid Aβ(16-22) peptides using hydrogen bonds and hydrophobicity forces. Biophys J 87:3657–3664.
17. Santini S, Mousseau N, Derreumaux P (2004) In silico assembly of Alzheimer's Aβ(16-22) peptide into β-sheets. J Am Chem Soc 126:11509–11516.
18. Hwang W, Zhang S, Kamm RD, Karplus M (2004) Kinetic control of dimer structure formation in amyloid fibrillogenesis. Proc Natl Acad Sci USA 101:12916–12921.
19. Klimov DK, Straub JE, Thirumalai D (2004) Aqeous urea solution destabilizes Aβ(16-22) oligomers. Proc Natl Acad Sci USA 101:14760–14765.
20. Röhrig UF, Laio A, Tantalo N, Parrinello M, Petronzio R (2006) Stability and structure of oligomers of the Alzheimer peptide Aβ(16-22): from the dimer to the 32-mer. Biophys J 91:3217–3229.
21. Gnanakaran S, Nussinov R, García AE (2006) Atomic-level description of amyloid β-dimer formation. J Am Chem Soc 128:2158–2159.
22. Meinke JH, Hansmann UHE (2007) Aggregation of β-amyloid fragments. J Chem Phys 126:014706.

23. Nguyen PH, Li MS, Stock G, Straub JE, Thirumalai D (2007) Monomer adds to preformed structured oligomers of Aβ-peptides by a two-stage dock-lock mechanism. Proc Natl Acad Sci USA 104:111–116.
24. Friedel M, Shea JE (2004) Self-assembly of peptides into a β-barrel motif. J Chem Phys 120:5809–5823.
25. Marchut AJ, Hall CK (2006) Spontaneous formation of annular structures observed in molecular dynamics simulations of polyglutamine peptides. Comput Biol Chem 30:215–218.
26. Wei G, Mousseau N, Derreumaux P (2004) Sampling the self-assembly pathways of KFFE hexamers. Biophys J 87:3648–3656.
27. Melquiond A, Mousseau N, Derreumaux P (2006) Structures of soluble amyloid oligomers from computer simulations. Proteins 65:180–191.
28. Ma B, Nussinov R (2006) Simulations as analytical tools to understand protein aggregation and predict amyloid formation. Curr Opin Chem Biol 10:1–8.
29. Gsponer J, Vendruscolo M (2006) Theoretical approaches to protein aggregation. Protein Pept Lett 13:287–293.
30. Irbäck A, Samuelsson B, Sjunnesson F, Wallin S (2003) Thermodynamics of α- and β-structure formation in proteins. Biophys J 85:1466–1473.
31. Irbäck A, Mohanty S (2005) Folding thermodynamics of peptides. Biophys J 88:1560–1569.
32. Irbäck A, Mitternacht S, Mohanty S (2005) Dissecting the mechanical unfolding of ubiquitin. Proc Natl Acad Sci USA 102:13427–13432.
33. Irbäck A, Mitternacht S (2006) Thermal versus mechanical unfolding of ubiquitin. Proteins 65:759–766.
34. Neidigh JW, Fesinmeyer RM, Andersen NH (2002) Desiging a 20-residue protein. Nat Struct Biol 9:425–430.
35. Lyubartsev AP, Martsinovski AA, Shevkunov SV, Vorontsov-Velyaminov PN (1992) New approch to Monte Carlo calculation of the free energy: Method of expanded ensembles. J Chem Phys 96:1776–1783.
36. Marinari E, Parisi G (1992) Simulated tempering: a new Monte Carlo scheme. Europhys Lett 19:451–458.
37. Irbäck A, Potthast F (1995) Studies of an off-lattice model for protein folding: sequence dependence and improved sampling at finite temperature. J Chem Phys 103:10298–10305.
38. Tesi MC, van Rensburg EJJ, Orlandini E, Whittington SG (1996) Monte Carlo study of the interacting self-avoiding walk model in three dimensions. J Stat Phys 82:155–181.
39. Hukushima K, Nemoto K (1996) Exchange Monte Carlo method and application to spin glass simulations. J Phys Soc (Jap) 65:1604–1608.
40. Hansmann UHE (1997) Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett 281:140–150.
41. Irbäck A, Mohanty S (2006) PROFASI: a Monte Carlo simulation package for protein folding and aggregation. J Comput Chem 27:1548–1555.
42. Favrin G, Irbäck A, Sjunnesson F (2001) Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. J Chem Phys 114:8154–8158.
43. Ferrenberg AM, Swendsen RH (1989) Optimized Monte Carlo data analysis. Phys Rev Lett 63:1195–1198.
44. Miller RG (1974) The jackknife – a review. Biometrika 61:1–15.

45. DeLano WL (2002). The PyMOL molecular graphics system. San Carlos, CA: DeLano Scientific.
46. McLachlan AD (1979) Gene duplication in the structural evolution of chymotrypsin. J Mol Biol 128:49–79.
47. Chou KC, Carlacci L, Maggiora GG (1990) Conformational and geometrical properties of idealized β-barrels in proteins. J Mol Biol 168:389–407.
48. Murzin AG, Lesk AM, Chothia C (1994) Principles determining the structure of β-sheet barrels in proteins. I. A theoretical analysis. J Mol Biol 236:1369–1381.
49. Murzin AG, Lesk AM, Chothia C (1994) Principles determining the structure of β-sheet barrels in proteins. II. The observed structures. J Mol Biol 236:1382–1400.
50. Liu WM (1998) Shear numbers of protein β-barrels: definition refinements and statistics. J Mol Biol 275:541–545.
51. Lazar KL, Kurutz JW, Tycko R, Meredith SC (2006) Encapsulation and NMR on an aggregating peptide before fibrillogenesis. J Am Chem Soc 128:16460–16461.
52. Collins SR, Douglas A, Vale RD, Weissman JS (2004) Mechanism of prion propagation: amyloid growth occurs by monomer addition. PLoS Biol 2:1582–1590.

PAPER IV

Changing the mechanical unfolding pathway of FnIII-10 by tuning the pulling strength

Simon Mitternacht[1], Stefano Luccioli[2], Alessandro Torcini[2], Alberto Imparato[3] and Anders Irbäck[1]

[1]Computational Biology and Biological Physics, Department of Theoretical Physics, Lund University, Sweden. [2]Istituto dei Sistemi Complessi, CNR, Sesto Fiorentino, Italy; INFN Sezione di Firenze, Sesto Fiorentino, Italy. [3]ISI Foundation, Torino, Italy. Current address: Department of Physics and Astronomy, University of Aarhus, Denmark.

Biophys. J. 96: 429–441. (2009)

We investigate the mechanical unfolding of the tenth type III domain from fibronectin, FnIII-10, both at constant force and at constant pulling velocity, by all-atom Monte Carlo simulations. We observe both apparent two-state unfolding and several unfolding pathways involving one of three major, mutually exclusive intermediate states. All the three major intermediates lack two of seven native β-strands, and share a quite similar extension. The unfolding behavior is found to depend strongly on the pulling conditions. In particular, we observe large variations in the relative frequencies of occurrence for the intermediates. At low constant force or low constant velocity, all the three major intermediates occur with a significant frequency. At high constant force or high constant velocity, one of them, with the N- and C-terminal β-strands detached, dominates over the other two. Using the extended Jarzynski equality, we also estimate the equilibrium free-energy landscape, calculated as a function of chain extension. The application of a constant pulling force leads to a free-energy profile with three major local minima. Two of these correspond to the native and fully unfolded states, respectively, whereas the third one can be associated with the major unfolding intermediates.

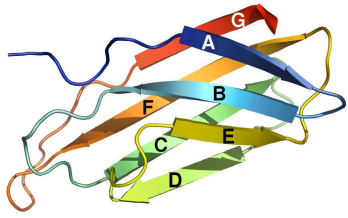

Figure IV.1: Schematic illustration of the NMR-derived [11] native structure of FnIII-10 (Protein Data Bank ID 1ttf). Its seven β-strands are labeled A–G in sequence order.

Introduction

Fibronectin is a giant multimodular protein that exists in both soluble (dimeric) and fibrillar forms. In its fibrillar form, it plays a central role in cell adhesion to the extracellular matrix. Increasing evidence indicates that mechanical forces exerted by cells are a key player in initiation of fibronectin fibrillogenesis as well as in modulation of cell-fibronectin adhesion, and thus may regulate the form and function of fibronectin [1, 2].

Each fibronectin monomer contains more than 20 modules of three types, called FnI–III. The most common type is FnIII, with ∼90 amino acids and a β-sandwich fold. Two critical sites for the interaction between cells and fibronectin are the RGD motif Arg78-Gly79-Asp80 [3] on the tenth FnIII module, FnIII-10, and a synergistic site [4] on the ninth FnIII module, which bind to cell-surface integrins. In the native structure of FnIII-10, shown in Figure IV.1, the RGD motif is found on the loop connecting the C-terminal β-strands F and G. It has been suggested that a stretching force can change the distance between these two binding sites sufficiently to affect the cell-adhesion properties, without deforming the sites themselves [2]. Force could also influence the adhesion properties by causing full or partial unfolding of the FnIII-10 module, and thereby deformation of the RGD motif [5]. Whether or not mechanical unfolding of fibronectin modules occurs in vivo is controversial. It is known that cell-generated force can extend fibronectin fibrils to several times their unstretched length [6]. There are experiments indicating that this extensibility is due to changes in quaternary structure rather than unfolding [7], while other experiments indicate that the extensibility originates from force-induced unfolding of FnIII modules [8, 9]. Also worth noting is that the FnIII-10 module is capable of fast refolding [10].

Atomic force microscopy (AFM) experiments have provided important insights into the mechanical properties of FnIII modules [12–14]. Interestingly, it

was found that, although thermodynamically very stable [15], the cell-binding module FnIII-10 is mechanically one of the least stable FnIII modules [12]. Further, it was shown that the force-induced unfolding of FnIII-10 often occurs through intermediate states [13]. While apparent one-step events were seen as well, a majority of the unfolding events had a clear two-step character [13]. A recent AFM study of pH dependence [14] suggests that electrostatic contributions are less important for the mechanical stability of FnIII-10 than previously thought.

Several groups have used computer simulations to investigate the force-induced unfolding of FnIII-10 [5, 16–21]. An early study predicted the occurrence of intermediate states [16]. In these simulations, two unfolding pathways were seen, both proceeding through partially unfolded intermediate states. Both intermediates lacked two of the seven native β-strands. The missing strands were A and B in one case, and A and G in the other (for strand labels, see Figure IV.1). A more recent study reached somewhat different conclusions [18]. This study found three different pathways, only one of which involved a partially unfolded intermediate state, with strands A and B detached. The experiments [13] are consistent with the existence of the two different intermediates seen in the early simulations [16], but do not permit an unambiguous identification of the states. When comparing the experiments with these simulations, it should be kept in mind that the forces studied in the simulations were larger than those studied experimentally.

Here we use an implicit-water all-atom model with a simple and computationally convenient energy function [22, 23] to investigate how the response of FnIII-10 to a stretching force depends on the pulling strength. We study the unfolding behavior both at constant force and at constant pulling velocity. Some previous studies were carried out using explicit-solvent models [5, 18, 19]. These models might capture important details that our implicit-solvent model ignores, like weakening of specific hydrogen bonds through interactions between water molecules and the protein backbone [24]. The advantage of our model is computational convenience. The relative simplicity of the model makes it possible for us to generate a large set of unfolding events, which is important when studying a system with multiple unfolding pathways.

Our analysis of the generated unfolding trajectories consists of two parts. The first part aims at characterizing the major unfolding pathways and unfolding intermediates. In the second part, we use the extended Jarzynski equality (EJE) [25–27] to estimate the equilibrium free-energy landscape, calculated as a function of end-to-end distance. This analysis extends previous work on simplified protein models [28–31] to an atomic-level model. This level of

detail may be needed to facilitate comparisons with future EJE reconstructions based on experimental data. Indeed, two applications of this method to experimental protein data were recently reported [32, 33].

Model and methods

Model. We use an all-atom model with implicit water, torsional degrees of freedom, and a simplified energy function [22, 23]. The energy function

$$E = E_{\mathrm{loc}} + E_{\mathrm{ev}} + E_{\mathrm{hb}} + E_{\mathrm{hp}} \tag{IV.1}$$

is composed of four terms. The term E_{loc} is local in sequence and represents an electrostatic interaction between adjacent peptide units along the chain. The other three terms are non-local in sequence. The excluded volume term E_{ev} is a $1/r^{12}$ repulsion between pairs of atoms. E_{hb} represents two kinds of hydrogen bonds: backbone-backbone bonds and bonds between charged side chains and the backbone. The last term E_{hp} represents an effective hydrophobic attraction between nonpolar side chains. It is a simple pairwise additive potential based on the degree of contact between two nonpolar side chains. The precise form of the different interaction terms and the numerical values of all geometry parameters can be found elsewhere [22, 23].

It has been shown that this model, despite its simplicity, provides a good description of the structure and folding thermodynamics of several peptides with different native geometries [23]. For the significantly larger protein FnIII-10, it is computationally infeasible to verify that the native structure is the global free-energy minimum. However, in order to study unfolding, it is sufficient that the native state is a local free-energy minimum. In our model, with unchanged parameters [22, 23], the native state of FnIII-10, indeed, is a long-lived state corresponding to a free-energy minimum, as will be seen below.

The same model has previously been used to study both mechanical and thermal unfolding of ubiquitin [34, 35]. In agreement with AFM experiments [36], it was found that ubiquitin, like FnIII-10, displays a mechanical unfolding intermediate far from the native state, and this intermediate was characterized [34]. The picture emerging from this study [34] was subsequently supported by ubiquitin simulations based on completely different models [37–39].

The energy function E of Equation IV.1 describes an unstretched protein. In our calculations, the protein is pulled either by a constant force or with a

constant velocity. In the first case, constant forces $-\vec{F}$ and $\vec{F}$ act on the N and C termini, respectively. The full energy function is then given by

$$E_{\text{tot}} = E - \vec{F} \cdot \vec{R} \tag{IV.2}$$

where $\vec{R}$ is the vector from the N to the C terminus. In the constant-velocity simulations, the pulling of the protein is modeled using a harmonic potential in the end-to-end distance $L = |\vec{R}|$ whose minimum $L_v(t)$ varies linearly with Monte Carlo (MC) time t. With this external potential, the full, time-dependent energy function becomes

$$E_{\text{tot}}(t) = E + \frac{k}{2}\,[L_v(t) - L]^2 = E + \frac{k}{2}\,[L_0 + v\,t - L]^2 \tag{IV.3}$$

where k is a spring constant, v is the pulling velocity, and L_0 is the initial equilibrium position of the spring. The spring constant, corresponding to the cantilever stiffness in AFM experiments, is set to $k = 37\,\text{pN/nm}$. The experimental FnIII-10 study of [13] reported a typical spring constant of $k \sim 50\,\text{pN/nm}$.

Simulation methods. Using MC dynamics, we study six constant force magnitudes F (50 pN, 80 pN, 100 pN, 120 pN, 150 pN and 192 pN) and four constant pulling velocities v (0.03 fm/MC step, 0.05 fm/MC step, 0.10 fm/MC step and 1.0 fm/MC step), at a temperature of 288 K. Three different types of MC updates are used: (i) Biased Gaussian Steps [40], BGS, which are semi-local updates of backbone angles; (ii) single-variable Metropolis updates of side-chain angles; and (iii) small rigid-body rotations of the whole chain. The BGS move simultaneously updates up to eight consecutive backbone angles, in a manner that keeps the chain ends approximately fixed. In the constant-velocity simulations, the time-dependent parameter $L_v(t)$ is changed after every attempted MC step.

As a starting point for our simulations, we use a model approximation of the experimental FnIII-10 structure (backbone root-mean-square deviation $\approx 0.2\,\text{nm}$), obtained by simulated annealing. All simulations are started from this initial structure, with different random number seeds. However, in the constant-velocity runs, the system is first thermalized in the potential $E + k(L_0 - L)^2/2$ for 10^7 MC steps ($L_0 = 3.8\,\text{nm}$), before the actual simulation is started at $t = 0$. The thermalization is a prerequisite for the Jarzynski analysis (see below).

The constant-force simulations are run for a fixed time, which depends on the force magnitude. There are runs in which the protein remains folded over the whole time interval studied. The constant-velocity simulations are run

until the spring has been pulled a distance of $v\,t = 35\,\mathrm{nm}$. At this point, the protein is always unfolded.

Our simulations are carried out using the program package PROFASI [41], which is a C++ implementation of this model. 3D structures are drawn with PyMOL [42].

Analysis of pathways and intermediates. To characterize pathways and intermediates, we study the evolution of the native secondary-structure elements along the unfolding trajectories. For this purpose, during the course of the simulations, all native hydrogen bonds connecting two β-strands (see Figure IV.1) are monitored. A bond is defined as present if the energy of that bond is lower than a cutoff ($-2.4\,k_\mathrm{B}\,T$). Using this data, we can describe a configuration by which pairs of β-strands are formed. A β-strand pair is said to be formed if more than a fraction 0.3 of its native hydrogen bonds are present. Whether individual β-strands are present or absent is determined based on which β-strand pairs the conformation contains.

The characterization of intermediate states requires slightly different procedures in the respective cases of constant force and constant velocity. For constant force, a histogram of the end-to-end distance L, covering the interval $3\,\mathrm{nm} < L < 27\,\mathrm{nm}$, is made for each unfolding trajectory. Each peak in the histogram corresponds to a metastable state along the unfolding pathway. To reduce noise the histogram is smoothed with a sliding L window of 0.3 nm. Peaks higher than a given cutoff are identified. Two peaks that are close to each other are only considered separate states if the values between them drop below half the height of the smallest peak. The position of an intermediate, L_I, is calculated as a weighted mean over the corresponding peak. The area under the peak provides, in principle, a measure, τ_I, of the life time of the state. However, due to statistical difficulties, we do not measure average life times of intermediate states.

In the constant-velocity runs, the unraveling of the native state or an intermediate state is associated with a rupture event, at which a large drop in force occurs. To ascertain that we register actual rupture events and not fluctuations due to thermal noise, the force versus time curves are smoothed with a sliding time window of $T_\mathrm{w} = 0.3\,\mathrm{nm}/v$, where v is the pulling velocity. Rupture events are identified as drops in force that are larger than 25 pN within a time less than T_w. The point of highest force just before the drop defines the rupture force, F_I, and the end-to-end distance, L_I, of the corresponding state. Only rupture events with a time separation of at least $2\,T_\mathrm{w}$ are considered

separate events. The rupture force F_I is a stability measure statistically easier to estimate than the life time τ_I at constant force.

For a peak with a given L_I, to decide which β-strands the corresponding state contains, we consider all stored configurations with $|L - L_\mathrm{I}| < 0.1$ nm. All β-strand pairs occurring at least once in these configurations are considered formed in the state. With this prescription, it happens that separate peaks from a single run exhibit the same set of β-strand pairs. Distinguishing between different substates with the same secondary-structure elements is beyond the scope of the present work. Such peaks are counted as a single state, with L_I set to the weighted average position of the merged peaks.

Jarzynski analysis. From the constant-velocity trajectories, we estimate the equilibrium free-energy landscape $G_0(L)$, as a function of the end-to-end distance L, for the unstretched protein by using EJE [25–27, 43]. For our system, this identity takes the form

$$\mathrm{e}^{-G_0(L)/k_\mathrm{B}T} = \text{constant} \cdot \mathrm{e}^{k[L-L_v(t)]^2/2k_\mathrm{B}T} \langle \delta(L - L(C_t)) \mathrm{e}^{-W_t/k_\mathrm{B}T} \rangle_t \qquad \text{(IV.4)}$$

where k_B is Boltzmann's constant, T the temperature, and C_t stands for the configuration of the system at time t. In this equation, $\langle \ldots \rangle_t$ denotes an average over trajectories C_τ, $0 < \tau < t$, with starting points C_0 drawn from the Boltzmann distribution corresponding to $E_\mathrm{tot}(0)$ (see Equation IV.3). The quantity W_t is the work done on the system along a trajectory and is given by

$$W_t = \int_0^t k\, v\, [L_v(\tau) - L(C_\tau)]\, \mathrm{d}\tau = \int F\, \mathrm{d}L_v \qquad \text{(IV.5)}$$

As discussed in [27, 43], combining Equation IV.4 with the weighted histogram method [44], one finds that the optimal estimate of the target function $G_0(L)$ is given by

$$G_0(L) = -k_\mathrm{B}T \ln \left[\frac{\sum_t \langle \delta(L - L(C_t)) \mathrm{e}^{-W_t/k_\mathrm{B}T} \rangle_t / \langle \mathrm{e}^{-W_t/k_\mathrm{B}T} \rangle}{\sum_t \mathrm{e}^{-k[L-L_v(t)]^2/2k_\mathrm{B}T} / \langle \mathrm{e}^{-W_t/k_\mathrm{B}T} \rangle} \right], \qquad \text{(IV.6)}$$

up to an additive constant. As in an experimental situation, for each unfolding trajectory, we sample the end-to-end distance $L(C_t)$ and the work W_t at discrete time intervals $k\Delta\tau$, with $k = 0, \ldots, n$ and $n\Delta\tau = t$. The sums appearing in Equation IV.6 thus run over these discrete times.

Let L_min and L_max be the minimal and maximal end-to-end distances, respectively, observed in the unfolding trajectories. We divide the interval $[L_\mathrm{min}, L_\mathrm{max}]$ into sub-intervals of length ΔL and evaluate $G_0(L_i)$ for each $L_i = L_\mathrm{min} + (i+1/2)\Delta L$ by exploiting Equation IV.6. The two averages appearing in this equation are estimated as $\overline{\theta_i(L(C_t)) \exp(-W_t/k_\mathrm{B}T)}$ and $\overline{\exp(-W_t/k_\mathrm{B}T)}$,

Table IV.1: Number of runs and the length of each run, in number of elementary MC steps, at the different pulling conditions studied.

pulling force or velocity	runs	MC steps/10^6
50 pN	98	1 000
80 pN	100	1 000
100 pN	100	250
120 pN	200	100
150 pN	340	50
192 pN	600	30
0.03 fm/MC step	100	1 167
0.05 fm/MC step	99	700
0.10 fm/MC step	99	350
1.0 fm/MC step	200	35

where the bar indicates an average over trajectories and the function $\theta_i(x)$ is defined as $\theta_i(x)=1$ if $|x-L_i|<\Delta L/2$ and $\theta_i(x)=0$ otherwise. Further details on the scheme used can be found in [43].

Results

Description of the calculated unfolding traces. We study the mechanical unfolding of FnIII-10 for six constant forces and four constant velocities. Table IV.1 shows the number of runs and the length of each run in these ten cases. At low force or low velocity, it takes longer for the protein to unfold, which makes it necessary to use longer and computationally more expensive trajectories.

Figure IV.2 shows the time evolution of the end-to-end distance L in a representative set of runs at constant force (100 pN). Typically each trajectory starts with a long waiting phase with $L\sim 5$ nm, where the molecule stays close to the native conformation. In this phase, the relative orientation of the two β-sheets (see Figure IV.1) might change, but all native β-strands remain unbroken. The waiting phase is followed by a sudden increase in L. This step typically leads either directly to the completely unfolded state with $L\sim 30$ nm or, more commonly, to an intermediate state at $L\sim 12$–16 nm. The intermediate is in turn unfolded in another abrupt step that leads to the completely stretched state. In a small fraction of the trajectories, depending on force, the protein is still in the native state or an intermediate state when the simulation stops. Intermediates outside the range 12–16 nm are unusual

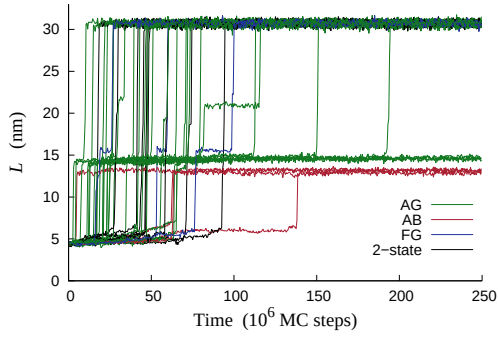

Figure IV.2: MC time evolution of the end-to-end distance in 42 independent simulations with a constant pulling force of 100 pN. The three most frequent intermediates lack different pairs of native β-strands: AG, FG, or AB. Trajectories in which these states occur are labeled green, blue and red, respectively. Apparent two-state events are colored black.

but occur in some runs. For example, a relatively long-lived intermediate at 21 nm can be seen in one of the runs in Figure IV.2.

Figure IV.3 shows samples of unfolding traces at constant velocity (0.05 fm/MC step). Here force is plotted against end-to-end distance. As in the constant-force runs, there are two main events in most trajectories. First, the native state is pulled until it ruptures at $L \sim 5$ nm. The chain is then elongated without much resistance until it, in most cases, reaches an intermediate at $L \sim 12$–16 nm. Here the force increases until there is a second rupture event. After that, the molecule is free to elongate towards the fully unfolded state with $L \sim 30$ nm. Some trajectories have force peaks at other L. An unusually large peak of this kind can be seen at 22 nm in Figure IV.3. Inspection of the corresponding structure reveals that it contains a three-stranded β-sheet composed of the native CD hairpin and a non-native strand. This sheet is pulled longitudinally, which explains why the stability is high. Another feature worth noting in Figure IV.3 is that the pulling velocity is sufficiently small to permit the force to drop to small values between the peaks.

There are several similarities between the unfolding events seen at constant force and at constant velocity. In most trajectories, there are stable intermediates, and the unfolding from both the native and intermediate states is abrupt. Also, the vast majority of the observed intermediates have a similar end-to-end distance, in the range 12–16 nm. It should be noticed that experiments typically measure contour-length differences rather than end-to-end distances. Below we analyze contour-length differences between the native state and

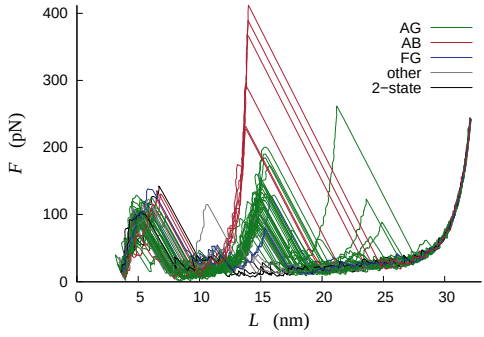

Figure IV.3: Force versus end-to-end distance in 55 independent simulations with a constant pulling velocity of 0.05 fm/MC step. Noise has been filtered out using a sliding time window of $6 \cdot 10^6$ MC steps. The color coding is the same as in Figure IV.2, with the addition of a new category for a few trajectories not belonging to any of the four categories in that figure. These trajectories are colored grey.

our calculated intermediates, which turn out to be in good agreement with experimental data.

The trajectories can be divided into three categories: apparent two-state unfolding, unfolding through intermediate states, and trajectories in which no unfolding takes place. Table IV.2 shows the relative frequencies of these groups at the different pulling conditions. The number of trajectories in which the protein remains folded throughout the run obviously depends on the trajectory length. More interesting to analyze is the ratio between the two kinds of unfolding, with or without intermediate states. In the constant-force runs, this ratio depends strongly on the magnitude of the applied force; unfolding through intermediates dominates at the lowest force, but is less common than apparent two-state unfolding at the highest force. In the constant-velocity runs, unfolding through intermediates is much more probable than apparent two-state unfolding at all the velocities studied.

Identifying pathways and intermediates. The fact that most observed intermediates fall in the relatively narrow L interval of 12–16 nm does not mean that they are structurally similar. Actually, the data in Figures IV.2 and IV.3 clearly indicate that these intermediates can be divided into three groups with similar but not identical end-to-end distances. The β-strand analysis (see Model and methods) reveals that these three groups correspond to the detachment of different pairs of β-strands, namely A and G, A and B, or F and G. The

Table IV.2: The fractions of trajectories in which unfolding occurs either in an apparent two-state manner (labeled $n = 2$) or through intermediate states (labeled $n \geq 3$). "No unfolding" refers to the fraction of trajectories in which the protein remains folded throughout the run (with $L < 8$ nm).

pulling force or velocity	$n = 2$	$n \geq 3$	no unfolding
50 pN	0.01	0.79	0.20
80 pN	0.21	0.79	0
100 pN	0.23	0.77	0
120 pN	0.24	0.76	0
150 pN	0.29	0.72	<0.01
192 pN	0.54	0.46	0
0.03 fm/MC step	0.04	0.96	0
0.05 fm/MC step	0.07	0.93	0
0.10 fm/MC step	0.03	0.97	0
1.0 fm/MC step	0	1.0	0

prevalence of these particular intermediate states is not surprising, given the native topology. When pulling the native structure of FnIII-10, the interior of the molecule is shielded from force by the N- and C-terminal β-strands, A and G. Consequently, in 95 % or more of our runs, either strand A or G is the first to detach, for all the pulling conditions studied. Most commonly, this detachment is followed by a release of the other strand of the two. But, when A (G) is detached, B (F) is also exposed to force. We thus have three main options for detaching two strands, AG, AB or FG, which actually correspond to the three major intermediates we observe.

Intermediates outside the interval 12–16 nm also occur in our simulations. When applied to the intermediates with $L < 12$ nm, the β-strand analysis identifies two states with one strand detached, A or G. The intermediates with $L > 16$ nm are scattered in L and correspond to rare states with more than two strands detached. The intermediate at 21 nm seen in one of the runs in Figure IV.2 lacks, for example, four strands (A, B, F and G). However, in these relatively unstructured states with more than two strands detached, the remaining strands are often disrupted, which makes the binary classification of strands as either present or absent somewhat ambiguous. Moreover, it is not uncommon that these large-L intermediates contain some non-native secondary structure. In what follows, we therefore focus on the five states seen with only one or two strands detached.

For convenience, the intermediates will be referred to by which strands are detached. The intermediate with strands A and B unfolded will thus be labeled AB, etc. Tables IV.3 and IV.4 show basic properties of the A, G, AB, AG

Table IV.3: Frequency f and average extension $\bar{L}_\mathrm{I}$ (in nm) of intermediate states in the constant-force simulations. The label of a state indicates which β-strands are detached, that is the state AG lacks strands A and G, etc. The frequency f is the number of runs in which a given state was seen, divided by the total number of runs in which unfolding occurred. The statistical uncertainties on $\bar{L}_\mathrm{I}$ are about 0.1 nm or smaller. "—" indicates not applicable.

	50 pN		80 pN		100 pN		120 pN		150 pN		192 pN	
state	f	$\bar{L}_\mathrm{I}$	f	$\bar{L}_\mathrm{I}$	f	$\bar{L}_\mathrm{I}$	f	$\bar{L}_\mathrm{I}$	f	$\bar{L}_\mathrm{I}$	f	$\bar{L}_\mathrm{I}$
AG	0.46	13.9	0.49	14.3	0.65	14.3	0.69	14.5	0.69	14.6	0.45	14.7
AB	0.35	12.4	0.14	12.9	0.09	13.1	0.03	13.2	<0.01	—	<0.01	—
FG	0.15	14.8	0.13	15.2	0.03	15.5	0.03	15.7	<0.01	—	<0.01	—
G	0.19	11.1	0.04	11.8	0	—	0	—	0	—	0	—
A	0.13	6.7	0	—	0	—	0	—	0	—	0	—

Table IV.4: Frequency f, average rupture force $\bar{F}_\mathrm{I}$ (in pN) and average extension $\bar{L}_\mathrm{I}$ (in nm) of intermediate states in the constant-velocity simulations. The statistical uncertainties are 10–20 % on F_I, about 0.1 nm or smaller on L_I for AG and AB, and about 0.5 nm on L_I for FG, G and A.

	0.03 fm/MC step			0.05 fm/MC step			0.10 fm/MC step			1.0 fm/MC step		
state	f	$\bar{F}_\mathrm{I}$	$\bar{L}_\mathrm{I}$	f	$\bar{F}_\mathrm{I}$	$\bar{L}_\mathrm{I}$	f	$\bar{F}_\mathrm{I}$	$\bar{L}_\mathrm{I}$	f	$\bar{F}_\mathrm{I}$	$\bar{L}_\mathrm{I}$
AG	0.60	115	14.9	0.69	121	14.9	0.78	131	14.8	0.81	198	15.0
AB	0.14	283	13.7	0.09	289	13.8	0.08	333	13.9	0.04	318	13.9
FG	0.15	119	15.6	0.08	107	15.3	0.08	162	16.0	0.04	216	15.7
G	0.05	54	10.5	0.08	73	10.8	0.20	46	9.9	0.06	67	10.3
A	0.06	43	6.2	0.07	53	7.2	0.09	57	6.9	0.03	81	7.2

and FG intermediates, as observed at constant force and constant velocity, respectively.

From Tables IV.3 and IV.4, several observations can be made. A first one is that the average end-to-end distance, $\bar{L}_\mathrm{I}$, of a given state increases slightly with increasing force. More importantly, it can be seen that the relative frequencies with which the different intermediates occur depend strongly on the pulling conditions. At high force or high velocity, the AG intermediate stands out as the by far most common one. By contrast, at low force or low velocity, there is no single dominant state. In fact, at $F = 50$ pN as well as at $v = 0.03$ fm/MC step, all the five states occur with a significant frequency.

Table IV.4 also shows the average rupture force, $\bar{F}_\mathrm{I}$, of the different states, at the different pulling velocities. Although the data are somewhat noisy, there is a clear tendency that $\bar{F}_\mathrm{I}$, for a given state, slowly increases with increasing pulling velocity, which is in line with the expected logarithmic v

dependence [45]. Comparing the different states, we find that those with only one strand detached (A and G) are markedly weaker than those with two strands detached (AG, AB and FG), as will be further discussed below. Most force-resistant is the AB intermediate. This state occurs much less frequently than the AG intermediate, especially at high velocity, but is harder to break once formed. Compared to experimental data, our $\bar{F}_{\mathrm{I}}$ values for the intermediates are somewhat large. The experiments found a relatively wide distribution of unfolding forces centered at 40–50 pN [13], which is a factor two or more lower than what we find for the AG, AB and FG intermediates. Our results for the unfolding force of the native state are consistent with experimental data. For the native state, the experiments found unfolding forces of 75 ± 20 pN [12] and 90 ± 20 pN [13]. Our corresponding results are 88 ± 2 pN, 99 ± 2 pN and 114 ± 3 pN at $v = 0.03$ fm/MC step, $v = 0.05$ fm/MC step and $v = 0.10$ fm/MC step, respectively.

The AG, AB and FG intermediates do not only require a significant rupture force in our constant-velocity runs, but are also long-lived in our constant-force simulations. In fact, in many runs, the system is still in one of these states when the simulation ends, which means that their average life times, unfortunately, are too long to be determined from the present set of simulations. Nevertheless, there is a clear trend that the AB intermediate is more long-lived than the other two, which in turn have similar life times. The relative life times of these states in the constant-force runs are thus fully consistent with their force-resistance in the constant-velocity runs.

At high constant force, we see a single dominant intermediate, the AG state, but also a large fraction of events without any detectable intermediate. Interestingly, it turns out that the same two strands, A and G, are almost always the first to break in the apparent two-state events as well. Table IV.5 shows the fraction of all trajectories, with or without intermediates, in which A and G are the first two strands to break, at the different forces studied. At 192 pN, this fraction is as large as 98 %. Although the time spent in the state with strands A and G detached varies from run to run, there is thus an essentially deterministic component in the simulated events at high force.

The unfolding behavior at low force or velocity is, by contrast, complex, with several possible pathways. Figure IV.4 illustrates the relations between observed pathways at the lowest pulling velocity, 0.03 fm/MC step. The main unfolding path begins with the detachment of strand G, followed by the formation of the AG intermediate, through the detachment of A. There are also runs in which the same intermediate occurs but A and G detach in the opposite order. Note that for the majority of the trajectories the boxes A and G in

Table IV.5: The fractions of all unfolding events in which the first two strands to break are A & G, F & G, and A & B, respectively, at different constant forces. The first pair to break was always one of these three.

first pair	50 pN	80 pN	100 pN	120 pN	150 pN	192 pN
A & G	0.50	0.69	0.87	0.935	0.973	0.980
A & B	0.35	0.15	0.09	0.025	0.006	0.007
F & G	0.15	0.16	0.04	0.040	0.021	0.013

Figure IV.4 only indicate passage through these states, not the formation of an intermediate state. In a few events, it is impossible to say which strand breaks first. In these events, the initial step is either that the hairpin AB detaches as one unit, or that strands A and G are unzipped simultaneously. Detachment of the FG hairpin in one chunk does not occur in the set of trajectories analyzed for Figure IV.4. Finally, we note that in the few trajectories where G occurs as an intermediate, the FG intermediate is always visited as well, but never AG. Similarly, the few trajectories where the A intermediate occurs also contain the AG intermediate, but not AB. We find no example where the AB intermediate is preceded by another intermediate.

The unfolding pattern illustrated in Figure IV.4 can be partly understood by counting native hydrogen bonds. The numbers of hydrogen bonds connecting the strand pairs AB, BE, CF and FG are $n_{\mathrm{AB}} = 7$, $n_{\mathrm{BE}} = 5$, $n_{\mathrm{CF}} = 8$ and $n_{\mathrm{FG}} = 6$, respectively. In our as well as in a previous study [18], two hydrogen bonds near the C terminus break early in some cases, which reduces the number of FG bonds to $n_{\mathrm{FG}} = 4$. The transition frequencies seen in Figure IV.4 match well with the ordering $n_{\mathrm{BE}} \sim n_{\mathrm{FG}} < n_{\mathrm{AB}} < n_{\mathrm{CF}}$. The first branch point in Figure IV.4 is the native state. Transitions from this state to the G state, N→G, are more common than N→A transitions, in line with the relation $n_{\mathrm{FG}} < n_{\mathrm{AB}}$. The second layer of branch points is the A and G states. That transitions G→AG are more common than G→GF and that A→AG and A→AB have similar frequencies, match well with the relations $n_{\mathrm{AB}} < n_{\mathrm{CF}}$ and $n_{\mathrm{FG}} \sim n_{\mathrm{BE}}$, respectively. Finally, there are fewer hydrogen bonds connecting the AB hairpin to the rest of the native structure than what is the case for the FG hairpin, $n_{\mathrm{BE}} < n_{\mathrm{CF}}$, which may explain why the AB hairpin, unlike the FG hairpin, detaches as one unit in some runs.

Another feature seen from Figure IV.4 is that the remaining native-like core rotates during the course of the unfolding process. The orientation of the core is crucial, because a strand is much more easily released if it can be unzipped one hydrogen bond at a time, rather than by longitudinal pulling. The detachment of the first strand leads, irrespective of whether it is A or G, to

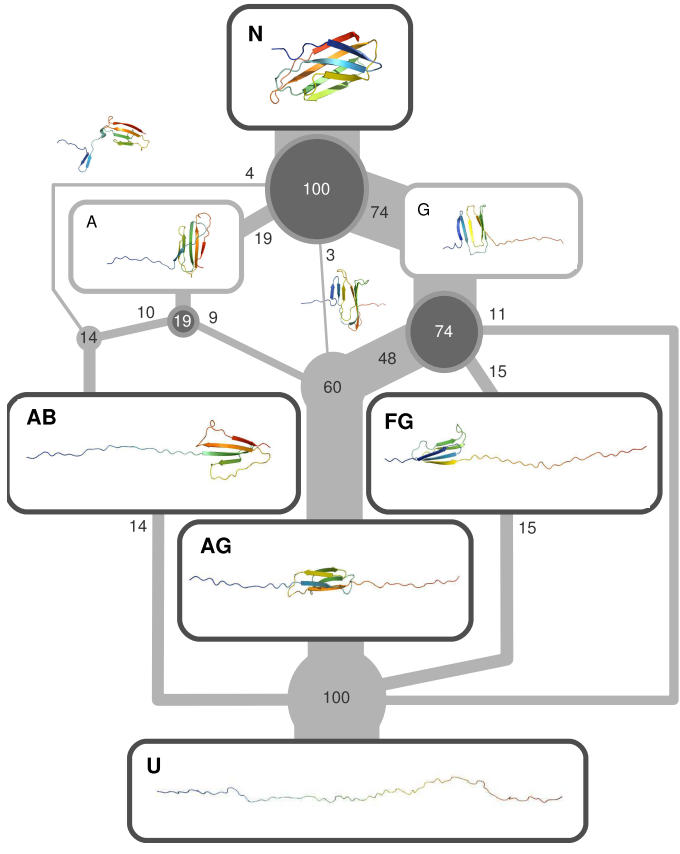

Figure IV.4: Illustration of the diversity of unfolding pathways in the 100 constant-velocity unfolding simulations at $v = 0.03\,\mathrm{fm/MC\ step}$. The numbers indicate how many of the trajectories follow a certain path. The boxes illustrate important structures along the pathways and boxes with dark rims correspond to the most long-lived states. Dark circles mark branch points. Most trajectories pass through G or A, but only a fraction spend a significant amount of time there (see Table IV.4). The line directly from G to U corresponds to events that either have no intermediate at all or only have intermediates other than the main three. The direct lines N→AB and N→AG describe events that do not clearly pass through A or G and examples of structures seen in those events are illustrated by the unboxed cartoons next to the lines.

an arrangement such that two strands are favorably positioned for unzipping, which explains why the intermediates with only A or G detached have a low force-resistance (see Tables IV.3 and IV.4). The AG, AB and FG intermediates, on the other hand, have cores that are pulled longitudinally, which makes them more resistant. Also worth noting is that the core of the AG intermediate is flipped 180°, which is not the case for the AB and FG intermediates.

The end-to-end distance of the intermediates cannot be directly compared with experimental data. The experiments measured [13] contour-length differences rather than L, through worm-like chain (WLC) [46] fits to constant-velocity data. Using data at our lowest pulling velocity (0.03 fm/MC step), we now mimic this procedure. For each force peak, we determine a contour length L_c by fitting the WLC expression

$$F = \frac{k_B T}{\xi}\left[\frac{1}{4(1-z/L_c)^2} - \frac{1}{4} + \frac{z}{L_c}\right] \tag{IV.7}$$

to data. Here ξ denotes the persistence length and z is the elongation, defined as $z = L - L_N$, where L_N is the end-to-end distance of the native state. Following [13], we use a fixed persistence length of $\xi = 0.4$ nm.

After each rupture peak follows a region where the force is relatively low. Here it sometimes happens that the newly released chain segment forms α-helical structures, indicating that our system is not perfectly described by the simple WLC model. Nevertheless, the WLC model provides a quite good description of our unfolding traces, as illustrated by Figure IV.5. The figure shows a typical unfolding trajectory with three force peaks, corresponding to the native (N), intermediate (I) and unfolded (U) states, respectively. From the fitted L_c values, the contour-length differences $\Delta L_c(\mathrm{N}\to\mathrm{I})$, $\Delta L_c(\mathrm{I}\to\mathrm{U})$ and $\Delta L_c(\mathrm{N}\to\mathrm{U})$ can be calculated.

Figure IV.6 shows a histogram of $\Delta L_c(\mathrm{N}\to\mathrm{I})$, based on our 100 trajectories for $v = 0.03$ fm/MC step. For a small fraction of the force peaks, a WLC fit is not possible; e.g., the A state cannot be analyzed due to its closeness to the native state. All intermediates analyzed have a $\Delta L_c(\mathrm{N}\to\mathrm{I})$ in the range 6–27 nm. They are divided into five groups: AB, AG, FG, G and "other". Most of those in the category "other" have five strands detached (CDEFG or ABEFG) and a $\Delta L_c(\mathrm{N}\to\mathrm{I})$ larger than 21 nm. These intermediates were not identified in the experimental study [13], which did not report any $\Delta L_c(\mathrm{N}\to\mathrm{I})$ values larger than 18 nm. These high-L intermediates mainly occur as a second intermediate, following one of the main intermediates, which perhaps explains why they were not observed in the experiments. The few remaining intermediates in the category "other" are all of the same kind, ABG, but show a large variation in $\Delta L_c(\mathrm{N}\to\mathrm{I})$, from 10 to 19 nm. The small values correspond

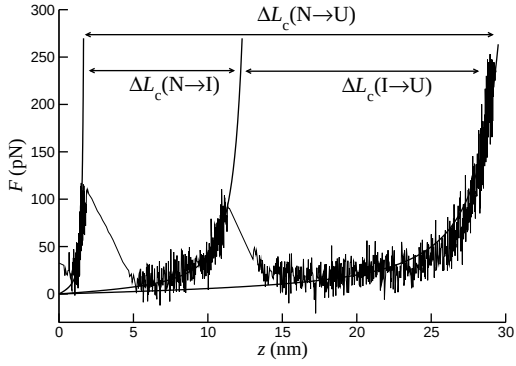

Figure IV.5: WLC fits (Equation IV.7) to a typical force-extension curve at $v = 0.03$ fm/MC step. The arrows indicate contour-length differences extracted from the fits: $\Delta L_c(\mathrm{N} \to \mathrm{I})$, $\Delta L_c(\mathrm{I} \to \mathrm{U})$ and $\Delta L_c(\mathrm{N} \to \mathrm{U})$.

Table IV.6: The average contour-length difference $\overline{\Delta L_c(\mathrm{N} \to \mathrm{I})}$ for different intermediates, as obtained by WLC fits (Equation IV.7) to our data for $v = 0.03$ fm/MC step.

state	$\overline{\Delta L_c(\mathrm{N} \to \mathrm{I})}$ (nm)
AG	12.1 ± 0.3
AB	10.1 ± 0.1
FG	13.4 ± 0.3
G	8.2 ± 0.9

to states where strand B actually is attached to the structured core, but through non-native hydrogen bonds.

The three major peaks in the $\Delta L_c(\mathrm{N} \to \mathrm{I})$ histogram (Figure IV.6) correspond to the AG, AB and FG intermediates. Although similar in size, these states give rise to well separated peaks, the means of which differ in a statistically significant way (see Table IV.6). For comparison, Figure IV.6 also shows the experimental $\Delta L_c(\mathrm{N} \to \mathrm{I})$ distribution [13]. The statistical uncertainties appear to be larger in the experiments, because the distribution has a single broad peak extending from 6 to 18 nm. All our $\Delta L_c(\mathrm{N} \to \mathrm{I})$ data for the AB, AG, FG and G intermediates fall within this region. The occurrence of these four intermediates is thus consistent with the experimental $\Delta L_c(\mathrm{N} \to \mathrm{I})$ distribution. The highest peak, corresponding to the AG intermediate, is located near the center of the experimental distribution.

Transitions from the native state directly to the unfolded state do not occur in the trajectories analyzed for Figure IV.6. For the contour-length difference

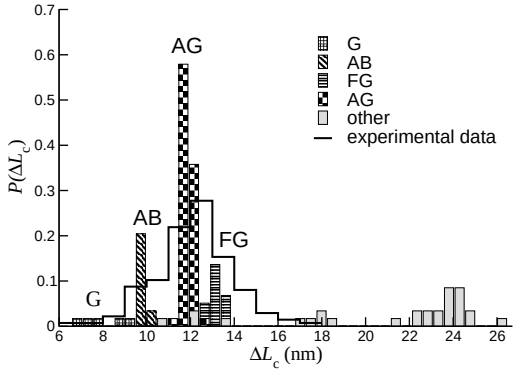

Figure IV.6: Histogram of the contour-length difference $\Delta L_c(\mathrm{N} \to \mathrm{I})$, obtained by WLC fits (Equation IV.7) to our data for $v = 0.03$ fm/MC step. A total of 121 force peaks corresponding to intermediate states are analyzed. The intermediates are divided into five groups: AB, AG, FG, G and "other". The experimental $\Delta L_c(\mathrm{N} \to \mathrm{I})$ distribution, from [13], is also indicated.

between these two states, we find a value of $\Delta L_c(\mathrm{N} \to \mathrm{U}) = 30.9 \pm 0.1$ nm, in perfect agreement with experimental data [13].

Estimating the free-energy profile. We now present the free-energy profile obtained by applying Eqs. IV.4–IV.6 to the constant-velocity trajectories. The number of trajectories analyzed can be seen in Table IV.1. Figure IV.7 shows the free-energy landscape at zero force, $G_0(L)$, against the end-to-end distance L, as obtained using different velocities v. We observe a collapse of the curves in the region of small-to-moderate L. Furthermore, the range of L where the curves superimpose, expands as v decreases. As discussed in [29–31, 43], the collapse of the reconstructed free-energy curves, as the manipulation rate is decreased, is a clear signature of the reliability of the evaluated free-energy landscape. Given our computational resources, we are not able to further decrease the velocity v, and for $L > 15$ nm there is still a difference of $\sim 40 k_B T$ between the two curves corresponding to the lowest velocities. The best estimate we currently have for $G_0(L)$ is the curve obtained with $v = 0.03$ fm/MC step. This curve will be used in the following analysis.

Let us consider the case where a constant force F is applied to the chain ends. The free energy then becomes $G(L) = G_0(L) - F \cdot L$. The tilted free-energy landscape $G(L)$ is especially interesting for small forces for which the unfolding process is too slow to be studied through direct simulation.

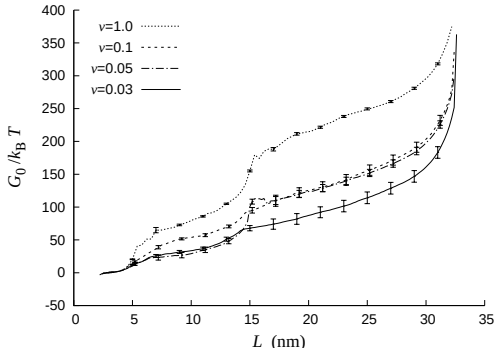

Figure IV.7: Free-energy landscape $G_0(L)$ calculated as a function of the end-to-end distance L, using data at different pulling velocities v (given in fm/MC step). In the calculations, L is discretized with a bin size of $\Delta L = 0.4$ nm for $v = 1.0$ fm/MC step and $\Delta L = 0.2$ nm for all the other velocities (see Model and methods).

Figure IV.8 shows our calculated $G(L)$ for four external forces in the range 10–50 pN. At $F = 10$ pN, the state with minimum free energy is still the native one, and no additional local minima have appeared. At $F = 25$ pN, the situation has changed. For $20 \lesssim F \lesssim 60$ pN, we find that $G(L)$ exhibits three major minima: the native minimum and two other minima, one of which corresponds to the fully unfolded state. The fully unfolded state takes over as the global minimum beyond $F = F_c \approx 22$ pN. The statistical uncertainty on the force at which this happens, F_c, is large, due to uncertainties on $G(L)$ for large L, as will be further discussed below. For $F = 25$ pN, the positions of the three major minima are 4.3 nm, 12 nm and 25 nm. As F increases, the minima move slightly toward larger L; for $F = 50$ pN, their positions are 4.6 nm, 14 nm and 29 nm. The first two minima become increasingly shallow with increasing F. For $F \gtrsim 60$ pN, the only surviving minimum is the third one, corresponding to the completely unfolded state.

These results have to be compared with the analysis above, which showed that the system, on its way from the native to the fully unfolded state, often spends a significant amount of time in some partially unfolded intermediate state with L around 12–16 nm. These intermediates should correspond to local free-energy minima along different unfolding pathways, but may or may not correspond to local minima of the global free energy $G(L)$, which is based on an average over the full conformational space. As we just saw, it turns out that $G(L)$ actually exhibits a minimum around 12–16 nm, where the most common intermediates are found. It is worth noting that above ~ 25 pN this minimum

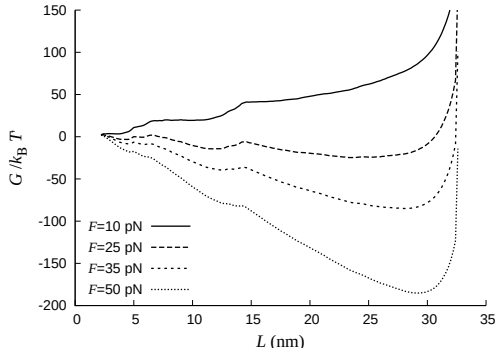

Figure IV.8: Tilted free-energy landscape $G(L) = G_0(L) - F \cdot L$ for four different forces F. The unperturbed landscape $G_0(L)$ corresponds to the curve shown in Figure IV.7 for $\nu = 0.03$ fm/MC step. The minima of $G(L)$ are discussed in the text.

gets weaker with increasing force. This trend is in agreement with the results shown in Table IV.2: the fraction of apparent two-state events, without any detectable intermediate, increases with increasing force.

For $F = 25$ pN and $F = 35$ pN, a fourth minimum can also be seen in Figure IV.8, close to the native state. Its position is ≈ 6 nm. This minimum is weak and has already disappeared for $F = 50$ pN. It corresponds to a state in which the two native β-sheets are slightly shifted relative to each other and aligned along the direction of the force, with all strands essentially intact. The appearance of this minimum is in good agreement with the results of Gao et al. [18]. In their unfolding trajectories, Gao et al. saw two early plateaus with small L, which in terms of our $G(L)$ should correspond to the native minimum and this $L \approx 6$ nm minimum. In our model, the $L \approx 6$ nm minimum represents a non-obligatory intermediate state; in many unfolding events, especially at high force, the molecule does not pass this state.

Finally, Figure IV.9 illustrates a more detailed analysis of the native minimum of $G(L)$, for $20\,\text{pN} < F < 60\,\text{pN}$. In this force range, we find that the first barrier is always located at $L = 5.0$ nm, whereas the position of the native minimum varies with force (see inset of Figure IV.9). Hence, the distance between the native minimum and the barrier, x_u, depends on the applied force, as expected [47–50]. Figure IV.9 shows the force-dependence of the barrier height, $\Delta G(F)$. The solid line is a linear fit with slope $x_u = 0.4$ nm, which describes the data quite well in the force range 25–56 pN. At lower force, the force-dependence is steeper; a linear fit to the data at low force gives a

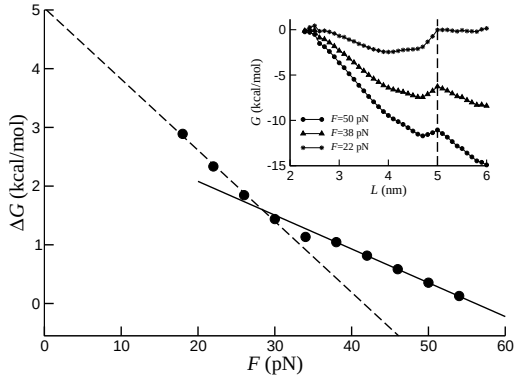

Figure IV.9: Free-energy barrier ΔG, separating the native state from extended conformations, as a function of the pulling force F. The solid line is a linear fit to the data for forces $F > 25$ pN, while the dashed line refers to a linear fit to the data in the interval $15\,\text{pN} \leq F \leq 30\,\text{pN}$. The inset shows the free energy $G(L)$ in the vicinity of the native state for three values of the force. The vertical dashed line indicates the position of the barrier.

slope of $x_u = 0.8$ nm (dashed line). Using this latter fit to extrapolate to zero force, we obtain a barrier estimate of $\Delta G(0) \approx 5$ kcal/mol. Due to the existence of the non-obligatory $L \approx 6$ nm intermediate, it is unclear how to relate this one-dimensional free-energy barrier to unfolding rates. Experimentally, barriers are indirectly probed, using unfolding kinetics. For FnIII-10, experiments found a zero-force barrier of 22.2 kcal/mol [12], using kinetics. For the unfolding length, an experimental value of $x_u = 0.38$ nm was reported [12], based on data in the force range 50–115 pN. Our result $x_u = 0.4$ nm obtained using the overlapping force range 25–56 pN, is in good agreement with this value.

Discussion

By AFM experiments, Li et al. [13] showed that FnIII-10 unfolds through intermediates when stretched by an external force. AFM data for the wild-type sequence and some engineered mutants were consistent with the existence of two distinct unfolding pathways with different intermediates, one being the AB state with strands A and B detached and the other being either the AG or the FG state [13]. This conclusion is in broad agreement with simulation results obtained by Paci and Karplus [16] and by Gao et al. [18].

Comparing our results with these previous simulations, one finds both differences and similarities. In our simulations, three major intermediates

are observed: AB, which was seen by Paci and Karplus as well as by Gao et al.; AG, also seen by Paci and Karplus; and FG, which was not observed in previous studies. The most force-resistant intermediate is AB in our as well as in previous studies. Frequencies of occurrence of the intermediates are difficult to compare because the previous studies were based on fewer trajectories. Nevertheless, one may note that the most common intermediate in our simulations, AG, is one of two intermediates seen by Paci and Karplus, and corresponds to one of three pathways observed by Gao et al. A and G often being the first two strands to break is also in agreement with the simulation results of Klimov and Thirumalai [17], who studied several different proteins using a simplified model. Unlike us, these authors found a definite unfolding order for the β-strands. The first strand to break was G, followed by A.

A key issue in our study is how the unfolding pathway depends on the pulling strength. This question was addressed by Gao et al. [18]. Based on a simple analytical model rather than simulations, it was argued that there is a single unfolding pathway at low force and multiple unfolding pathways at high force. Our results show the opposite trend. At our lowest force, 50 pN, we observe several different unfolding pathways, and all the three major intermediates occur with a significant frequency. At our highest force, 192 pN, unfolding occurs either in one step or through one particular intermediate, the AG state. Moreover, at 192 pN, the same two strands, A and G, are almost always the first to break in the apparent one-step events as well. Hence, at our highest force, we find that the unfolding behavior has an essentially deterministic component. The trend that the unfolding pathway becomes more deterministic with increasing force can probably be attributed to a reduced relative importance of random thermal fluctuations.

There is a point of disagreement between our results and experimental data, which is that the rupture forces of the three major intermediates are higher in our constant-velocity simulations than they were in the experiments [13]. Although the statistical uncertainties are non-negligible and the pulling conditions are not identical (e.g., we consider a single FnIII-10 module, while the experiments studied multimodular constructs), we do not see any plausible explanation of this discrepancy. It thus seems that our model overestimates the rupture force of these intermediates. Our calculated rupture force for the native state is consistent with experimental data (see above). To make sure that this agreement is not accidental, we also measured the rupture force of the native state for three other domains, namely FnIII-12, FnIII-13 and the titin I27 domain. AFM experiments (at 0.6 μm/s) found that these domains differ in force-resistance, following the order FnIII-13 ($\sim$ 90 pN) $<$ FnIII-12 ($\sim$ 120 pN)

< I27 ($\sim$ 200 pN) [12]. For each of these domains, we carried out a set of 60 unfolding simulations, at a constant velocity of 0.10 fm/MC step. The average rupture forces were 108 ± 4 pN for FnIII-13, 135 ± 4 pN for FnIII-12, and 159 ± 6 pN for I27, which is in reasonable agreement with experimental data. In particular, our model correctly predicts that the force-resistance of the native state decreases as follows: I27 > FnIII-12> FnIII-13$\sim$ FnIII-10. Similar findings have been reported for another model [19].

Throughout the paper, times have been given in MC steps. In order to roughly estimate what one MC step corresponds to in physical units, we use the average unfolding time of the native state, which is $\sim 4 \cdot 10^8$ MC steps at our lowest force, 50 pN. Assuming that the force-dependence of the unfolding rate is given by $k(F) = k_0 \exp(F x_u / k_B T)$ [51] with $x_u = 0.38$ nm [12], this unfolding time corresponds to a zero-force unfolding rate of $k_0 \sim 1/(4 \cdot 10^{10}$ MC steps). Setting this quantity equal to its experimental value, $k_0 = 0.02\,\mathrm{s}^{-1}$ [12], gives the relation that one MC step corresponds to $1 \cdot 10^{-9}$ s. Using this relation to translate our pulling velocities into physical units, one finds, for example, that 0.05 fm/MC step corresponds to 0.05 μm/s. This estimate suggests that the effective pulling velocities in our simulations are comparable to or lower than the typical pulling velocity in the experiments [13], which was 0.4 μm/s. That the effective pulling velocity is low in our simulations is supported by the observation made earlier that the force drops to very small values between the rupture peaks.

The force range studied in our simulations is comparable to that studied in AFM experiments [12–14]. The exact forces acting on fibronectin under physiological conditions are not known, but might be considerably smaller. For comparison, it was estimated that physiologically relevant forces for the muscle protein titin are ~ 4 pN per I-band molecule [52]. For so small forces, the unfolding of FnIII-10 occurs too slowly in the model to permit direct simulation. Therefore, we cannot characterize unfolding pathways and possible intermediates for these forces. On the other hand, we have an estimate of the free-energy profile $G(L)$ for arbitrary force, which can be used, in particular, to estimate the force F_c, beyond which the fully extended state has minimum free energy. Using our best estimate of $G(L)$, one finds an F_c of 22 pN (see above). Now, F_c depends on the behavior of $G(L)$ for large L, where the uncertainties are large and not easy to accurately estimate. As a test, we therefore repeated the same analysis using the Ising-like model of [29, 30, 39] (unpublished results), which gave us the estimate $F_c \sim 20$ pN, in quite good agreement with the value found above (22 pN). Together, we take these results to indicate that $F_c \gtrsim 15$ pN, which might be large compared to physiologically relevant forces

(see above). For stretching forces F significantly smaller than F_c, the statistical weight of the fully stretched state is small. To estimate the suppression, let L_N and L_s be the end-to-end distances of the native and stretched states. The free energies of these states at force F can be written as $G_N = G_N^c - (F - F_c)L_N$ and $G_s = G_s^c - (F - F_c)L_s$, where G_N^c and G_s^c are the free energies at F_c. Assuming $G_N^c = G_s^c$, $F_c \gtrsim 15$ pN and $L_s - L_N \sim 20$ nm, one finds, for example, that $G_s - G_N \gtrsim 25\, k_B T$ for $F \lesssim 10$ pN. Our estimate $F_c \gtrsim 15$ pN thus indicates that unfolding of FnIII-10 to its fully stretched state is a rare event for stretching forces $F \lesssim 10$ pN. The major intermediates are also suppressed compared to the native state for $F \lesssim 10$ pN (see Figure IV.8). However, our results indicate that the major intermediates are more likely to be observed than the fully stretched state for these forces.

By extrapolating from experimental data at zero force, the force at which the native and fully stretched states have equal free energy has been estimated to be 3.5–5 pN for an average FnIII domain [53]. Our results suggest that the native state remains thermodynamically dominant at so small forces.

The reconstructed free energies $G_0(L)$ and $G(L)$ are thermodynamical potentials describing the equilibrium behavior of the system in the absence and presence of an external force F, respectively. On the other hand, the long-lived intermediate states observed during the unfolding of the molecule are a clear signature of out-of-equilibrium behavior. They indicate an arrest of the unfolding kinetics, typically in the L range 12–16 nm, on the way from the old (native) equilibrium state to the new (fully unfolded) equilibrium state. The calculated (equilibrium) landscape $G(L)$ (see Figure IV.8) is to some extent able to describe this out-of-equilibrium behavior. For $20 \lesssim F \lesssim 60$ pN, this function exhibits three major minima corresponding to the folded state, the most common intermediates, and the fully unfolded state, respectively. However, since $G(L)$ describes the system in terms of a single coordinate L and "hides" the microscopic configuration, one cannot extract the full details of individual unfolding pathways from this function. For example, one cannot, based on $G(L)$, distinguish the AG, AB and FG intermediates, which have quite similar L.

The height of the first free-energy barrier, ΔG, can be related to the unfolding length x_u, a parameter typically extracted from unfolding kinetics, assuming the linear relationship $\Delta G(F) = \Delta G_0 - F \cdot x_u$. The parameter x_u measures the distance between the native state and the free-energy barrier, which generally depends on force. Our data for x_u indeed show a clear force-dependence (see inset of Figure IV.9). However, over a quite large force interval,

our x_u is almost constant and similar to its experimental value [13], which was based on an overlapping force interval.

Conclusion

We have used all-atom MC simulations to study the force-induced unfolding of the fibronectin module FnIII-10, and in particular how the unfolding pathway depends on the pulling conditions. Both at constant force and at constant pulling velocity, the same three major intermediates were seen, all with two native β-strands missing: AG, AB or FG. Contour-length differences $\Delta L_c(\mathrm{N} \rightarrow \mathrm{I})$ for these states were analyzed, through WLC fits to constant-velocity data. We found that the states, in principle, can be distinguished based on their $\Delta L_c(\mathrm{N} \rightarrow \mathrm{I})$ distributions, but the differences between the distributions are small compared to the resolution of existing experimental data.

The unfolding behavior at constant force was examined in the range 50–192 pN. The following picture emerges from this analysis:

1. At the lowest forces studied, several different unfolding pathways can be seen, and all the three major intermediates occur with a significant frequency.

2. At the highest forces studied, the AB and FG intermediates are very rare. Unfolding occurs either in an apparent single step or through the AG intermediate.

3. The unfolding behavior becomes more deterministic with increasing force. At 192 pN, the first strand pair to break is almost always A and G, also in apparent two-state events.

The dependence on pulling velocity in the constant-velocity simulations was found to be somewhat less pronounced, compared to the force-dependence in the constant-force simulations. Nevertheless, some clear trends could be seen in this case as well. In particular, with increasing velocity, we found that the AG state becomes increasingly dominant among the intermediates. Our results thus suggest that the AG state is the most important intermediate both at high constant force and at high constant velocity.

The response to weak pulling forces is expensive to simulate; our calculations, based on a relatively simple and computationally efficient model, extended down to 50 pN. The Jarzynski method for determining the free energy $G(L)$ opens up a possibility to partially circumvent this problem. Our estimated $G(L)$, which matches well with several direct observations from

the simulations, indicates, in particular, that stretching forces below 10 pN only rarely unfold FnIII-10 to its fully extended state. Although supported by calculations based on a different model, this conclusion should be verified by further studies, because accurately determining $G(L)$ for large L is a challenge.

Acknowledgments. This work has been in part supported by the Swedish Research Council and by the European Community via the STREP project EMBIO NEST (contract no. 12835).

References

1. Geiger B, Bershadsky A, Pankov R, Yamada KM (2001) Transmembrane crosstalk between the extracellular matrix and the cytoskeleton. Nat Rev Mol Cell Biol 2:793–805.
2. Vogel V (2006) Mechanotransduction involving multimodular proteins: converting force into biochemical signals. Annu Rev Biophys Biomol Struct 35:459–488.
3. Ruoslahti E, Pierschbacher MD (1987) New perspectives in cell adhesion: RGD and integrins. Science 238:491–497.
4. Aota SI, Nomizu M, Yamada KM (1994) The short amino acid sequence Pro-His-Ser-Arg-Asn in human fibronectin enhances cell-adhesive function. J Biol Chem 269:24756–24761.
5. Krammer A, Lu H, Isralewitz B, Schulten K, Vogel V (1999) Forced unfolding of the fibronectin type III module reveals a tensile molecular recognition switch. Proc Natl Acad Sci USA 96:1351–1356.
6. Ohashi T, Kiehart DP, Erickson HP (1999) Dynamics and elasticity of the fibronectin matrix in living cell culture visualized by fibronectin-green fluorescent protein. Proc Natl Acad Sci USA 96:2153–2158.
7. Abu-Lail NI, Ohashi T, Clark RL, Erickson HP, Zauscher S (2006) Understanding the elasticity of fibronectin fibrils: unfolding strengths of FN-III and GFP domains measured by single molecule force spectroscopy. Matrix Biol 25:175–184.
8. Baneyx G, Baugh L, Vogel V (2002) Fibronectin extension and unfolding within cell matrix fibrils controlled by cytoskeletal tension. Proc Natl Acad Sci USA 99:5139–5143.
9. Smith ML, Gourdon D, Little WC, Kubow KE, Andresen Eguiluz R, et al. (2007) Force-induced unfolding of fibronectin in the extracellular matrix of living cells. PLoS Biol 5:e268.
10. Plaxco KW, Spitzfaden C, Campbell ID, Dobson CM (1997) A comparison of the folding kinetics and thermodynamics of two homologous fibronection type III modules. J Mol Biol 270:763–770.
11. Main AL, Harvey TS, Baron M, Boyd J, Campbell ID (1992) The three-dimensional structure of the tenth type III module of fibronectin: an insight into RGD-mediated interactions. Cell 71:671–678.
12. Oberhauser AF, Badilla-Fernandez C, Carrion-Vazquez M, Fernandez JM (2002) The mechanical hierarchies of fibronectin observed by single-molecule AFM. J Mol Biol 319:433–447.

13. Li L, Huang HHL, Badilla CL, Fernandez JM (2005) Mechanical unfolding intermediates observed by single-molecule force spectroscopy in a fibronectin type III module. J Mol Biol 345:817–826.
14. Ng SP, Clarke J (2007) Experiments suggest that simulations may overestimate electrostatic contributions to the mechanical stability of a fibronectin type III domain. J Mol Biol 371:851–854.
15. Cota E, Clarke J (2000) Folding of beta-sandwich proteins: three-state transition of a fibronectin type III module. Protein Sci 9:112–120.
16. Paci E, Karplus M (1999) Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulation. J Mol Biol 288:441–459.
17. Klimov DK, Thirumalai D (2000) Native topology determines force-induced unfolding pathways in globular proteins. Proc Natl Acad Sci USA 97:7254–7259.
18. Gao M, Craig D, Vogel V, Schulten K (2002) Identifying unfolding intermediates of FnIII-10 by steered molecular dynamics. J Mol Biol 323:939–950.
19. Craig D, Gao M, Schulten K, Vogel V (2004) Tuning the mechanical stability of fibronectin type III modules through sequence variations. Structure 12:21–30.
20. Sułkowska JI, Cieplak M (2007) Mechanical stretching of proteins – a theoretical survey of the Protein Data Bank. J Phys: Condens Matter 19:283201.
21. Li MS (2007) Secondary structure, mechanical stability, and location of transition state of proteins. Biophys J 93:2644–2654.
22. Irbäck A, Samuelsson B, Sjunnesson F, Wallin S (2003) Thermodynamics of α- and β-structure formation in proteins. Biophys J 85:1466–1473.
23. Irbäck A, Mohanty S (2005) Folding thermodynamics of peptides. Biophys J 88:1560–1569.
24. Lu H, Schulten K (2000) The key event in force-induced unfolding of titin's immunoglobulin domains. Biophys J 79:51–65.
25. Jarzynski C (1997) Nonequilibrium equality for free energy differences. Phys Rev Lett 78:2690–2693.
26. Crooks GE (1999) Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. Phys Rev E 60:2721–2726.
27. Hummer G, Szabo A (2001) Free energy reconstruction from nonequilibrium single-molecule pulling experiments. Proc Natl Acad Sci USA 98:3658–3661.
28. West DK, Olmsted PD, Paci E (2006) Free energy for protein folding from nonequilibrium simulations using the Jarzynski equality. J Chem Phys 125:204910.
29. Imparato A, Pelizzola A, Zamparo M (2007) Ising-like model for protein mechanical unfolding. Phys Rev Lett 98:148102.
30. Imparato A, Pelizzola A, Zamparo M (2007) Protein mechanical unfolding: a model with binary variables. J Chem Phys 127:145105.
31. Imparato A, Luccioli S, Torcini A (2007) Reconstructing the free-energy landscape of a mechanically unfolded model protein. Phys Rev Lett 99:168101.
32. Harris NC, Song Y, Kiang CH (2007) Experimental free energy surface reconstruction from single-molecule force spectroscopy using Jarzynski's equality. Phys Rev Lett 99:068101.
33. Imparato A, Sbrana F, Vassalli M (2008) Reconstructing the free-energy landscape of a polyprotein by single-molecule experiments. Europhys Lett 82:58006.
34. Irbäck A, Mitternacht S, Mohanty S (2005) Dissecting the mechanical unfolding of ubiquitin. Proc Natl Acad Sci USA 102:13427–13432.

35. Irbäck A, Mitternacht S (2006) Thermal versus mechanical unfolding of ubiquitin. Proteins 65:759–766.
36. Schlierf M, Li H, Fernandez JM (2004) The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques. Proc Natl Acad Sci USA 101:7299–7304.
37. Li MS, Kouza M, Hu CK (2007) Refolding upon force quench and pathways of mechanical and thermal unfolding of ubiquitin. Biophys J 92:547–561.
38. Kleiner A, Shakhnovich E (2007) The mechanical unfolding of ubiquitin through all-atom Monte Carlo simulation with a Gō-type potential. Biophys J 92:2054–2061.
39. Imparato A, Pelizzola A (2008) Mechanical unfolding and refolding pathways of ubiquitin. Phys Rev Lett 100:158104.
40. Favrin G, Irbäck A, Sjunnesson F (2001) Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. J Chem Phys 114:8154–8158.
41. Irbäck A, Mohanty S (2006) PROFASI: a Monte Carlo simulation package for protein folding and aggregation. J Comput Chem 27:1548–1555.
42. DeLano WL (2002). The PyMOL molecular graphics system. San Carlos, CA: DeLano Scientific.
43. Imparato A, Peliti L (2006) Evaluation of free energy landscapes from manipulation experiments. J Stat Mech :P03005.
44. Ferrenberg AM, Swendsen RH (1989) Optimized Monte Carlo data analysis. Phys Rev Lett 63:1195–1198.
45. Evans E, Ritchie K (1997) Dynamic strength of molecular adhesion bonds. Biophys J 72:1541–1555.
46. Marko JF, Siggia ED (1995) Stretching DNA. Macromolecules 28:8759–8770.
47. Li PC, Makarov DE (2003) Theoretical studies of the mechanical unfolding of the muscle protein titin: bridging the time-scale gap between simulation and experiment. J Chem Phys 119:9260–9268.
48. Hyeon C, Thirumalai D (2006) Forced-unfolding and force-quench refolding of RNA hairpins. Biophys J 90:3410–3427.
49. West DK, Paci E, Olmsted PD (2006) Internal protein dynamics shifts the distance to the mechanical transition state. Phys Rev E 74:061912.
50. Dudko OK, Mathé J, Szabo A, Meller A, Hummer G (2007) Extracting kinetics from single-molecule force spectroscopy: nanopore unzipping of DNA hairpins. Biophys J 92:4186–4195.
51. Bell GI (1978) Models for specific adhesion of cells to cells. Science 200:618–627.
52. Li H, Linke WA, Oberhauser AF, Carrion-Vazquez M, Kerkvliet JG, et al. (2002) Reverse engineering of the giant muscle protein titin. Nature 418:998–1002.
53. Erickson HP (1994) Reversible unfolding of fibronectin type III and immunoglobulin domains provides the structural basis for stretch and elasticity of titin and fibronectin. Proc Natl Acad Sci USA 91:10114–10118.

PAPER V

An effective all-atom potential for proteins

Anders Irbäck[1], Simon Mitternacht[1] and Sandipan Mohanty[2]

[1]Computational Biology and Biological Physics, Department of Theoretical Physics, Lund University, Sweden. [2]Jülich Supercomputing Center, Institute for Advanced Simulation, Forschungszentrum Jülich, Germany

Submitted (LU TP 09-01)

We describe and test an implicit solvent all-atom potential for simulations of protein folding and aggregation. The potential is developed through studies of structural and thermodynamic properties of 17 peptides with diverse secondary structure. Results obtained using the final form of the potential are presented for all these peptides. The same model, with unchanged parameters, is furthermore applied to a heterodimeric coiled-coil system, a mixed α/β protein and a three-helix-bundle protein, with very good results. The computational efficiency of the potential makes it possible to investigate the free-energy landscape of these 49–67-residue systems with high statistical accuracy, using only modest computational resources by today's standards.

Introduction

A molecular understanding of living systems requires modeling of the dynamics and interactions of proteins. The relevant dynamics of a protein may amount to small fluctuations about its native structure, or reorientations of its ordered parts relative to each other. In either case, a tiny fraction of the conformational space is explored. For flexible proteins, perhaps with large intrinsically disordered parts [1, 2], the situation is different. When studying such proteins or conformational conversion processes like folding or amyloid aggregation, the competition between different minima on the free-energy landscape inevitably comes into focus. Studying these systems by computer simulation is a challenge, because proper sampling of all relevant free-energy minima must be ensured. This goal is very hard to achieve if explicit solvent molecules are included in the simulations. The use of coarse-grained models can alleviate this problem, but makes important geometric properties like secondary structure formation more difficult to describe.

Here we present an implicit solvent all-atom protein model especially aimed at problems requiring exploration of the global free-energy landscape. It is based on a computationally convenient effective potential, with parameters determined through full-scale thermodynamic simulations of a set of experimentally well characterized peptides. Central to the approach is the use of a single set of model parameters, independent of the protein studied. This constraint is a simple but efficient way to avoid unphysical biases, for example, toward either α-helical or β-sheet structure [3, 4]. Imposing this constraint is also a way to enable systematic refinement of the potential.

An earlier version [5, 6] of this potential has proven useful, for example, for studies of aggregation [7–9] and mechanical unfolding [10, 11]. Also, using a slightly modified form of the potential [12], the folding mechanisms of a 49-residue protein, Top7-CFr, were investigated [13, 14]. Here we revise this potential, through studies of an enlarged set of 17 peptides (see Table V.1 and Figure V.1). We show that the model, in its final form, folds these different sequences to structures similar to their experimental structures, using a single set of potential parameters. The description of each peptide is kept brief, to be able to discuss all systems and thereby address the issue of transferability in a direct manner. The main purpose of this study is model development rather than detailed characterization of individual systems.

Whether or not this potential, calibrated using data on peptides with typically $\sim$20 residues, will be useful for larger systems is not obvious. Therefore, we also apply our potential, with unchanged parameters, to three larger systems with different geometries. These systems are the mixed α/β protein

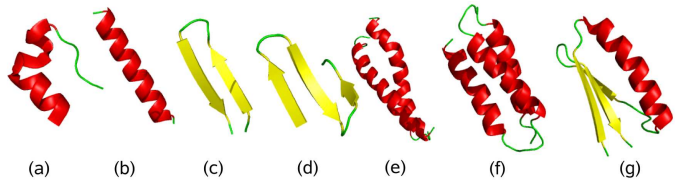

Figure V.1: Schematic illustration of native geometries studied. (a) the Trp-cage, (b) an α-helix, (c) a β-hairpin, (d) a three-stranded β-sheet, (e) an α-helix dimer (1U2U), (f) a three-helix bundle (1LQ7), and (g) a mixed α/β protein (2GJH).

Table V.1: Amino acid sequences. Suc stands for succinylic acid.

System	PDB code	Sequence
Trp-cage	1L2Y	NLYIQ WLKDG GPSSG RPPPS
E6apn1	1RIJ	Ac–ALQEL LGQWL KDGGP SSGRP PPS–NH_2
C		Ac–KETAA AKFER AHA–NH_2
EK		Ac–YAEAA KAAEA AKAF–NH_2
F_s		Suc–AAAAA AAARA AAARA AAARA A–NH_2
GCN4tp	2OVN	NYHLE NEVAR LKKLV GE
HPLC-6	1WFA	DTASD AAAAA ALTAA NAKAA AELTA ANAAA AAAAT AR–NH_2
Chignolin	1UAO	GYDPE TGTWG
MBH12	1J4M	RGKWT YNGIT YEGR
GB1p		GEWTY DDATK TFTVT E
GB1m2		GEWTY NPATG KFTVT E
GB1m3		KKWTY NPATG KFTVQ E
trpzip1	1LE0	SWTWE GNKWT WK–NH_2
trpzip2	1LE1	SWTWE NGKWT WK–NH_2
betanova		RGWSV QNGKY TNNGK TTEGR
LLM		RGWSL QNGKY TLNGK TMEGR
beta3s		TWIQN GSTKW YQNGS TKIYT
AB zipper	1U2U	Ac–EVAQL EKEVA QLEAE NYQLE QEVAQ LEHEG–NH_2
		Ac–EVQAL KKRVQ ALKAR NYALK QKVQA LRHKG–NH_2
Top7-CFR	2GJH	ERVRI SITAR TKKEA EKFAA ILIKV FAELG YNDIN
		VTWDG DTVTV EGQL
GS-α_3W	1LQ7	GSRVK ALEEK VKALE EKVKA LGGGG RIEEL KKKWE
		ELKKK IEELG GGGEV KKVEE EVKKL EEEIK KL

Top7-CFr, a three-helix-bundle protein with 67 residues, and a heterodimeric leucine zipper composed of two 30-residue chains.

Protein folding simulations are by necessity based on potentials whose terms are interdependent and dependent on the choice of geometric representation. Therefore, we choose to calibrate our potential directly against folding properties of whole chains. To make this feasible, we deliberately omit many details included in force fields like Amber, CHARMM and OPLS (for a review, see [15]). With this approach, we might lose details of a given free-energy minimum, but, by construction, we optimize the balance between competing minima.

Two potentials somewhat similar in form to ours are the μ-potential of the Shakhnovich group [16] and the PFF potential of the Wenzel group [17]. These groups also consider properties of entire chains for calibration, but use folded PDB structures or sets of decoys rather than full-scale thermodynamic simulations. Our admittedly time-consuming procedure implies that our model is trained on completely general structures, which might be an advantage when studying the dynamics of folding. Another potential with similarities to ours is that developed by the Dokholyan group for discrete molecular dynamics simulations [18].

Methods

Our model belongs to the class of implicit solvent all-atom models with torsional degrees of freedom. All geometrical parameters, like bond lengths and bond angles, are as described earlier [5].

The interaction potential is composed of four major terms:

$$E = E_{\mathrm{loc}} + E_{\mathrm{ev}} + E_{\mathrm{hb}} + E_{\mathrm{sc}}\,. \tag{V.1}$$

The first term, E_{loc}, contains local interactions between atoms separated by only a few covalent bonds. The other three terms are non-local in character: E_{ev} represents excluded-volume effects, E_{hb} is a hydrogen-bond potential, and E_{sc} contains residue-specific interactions between pairs of side-chains. Next we describe the precise form of these four terms. Energy parameters are given in a unit called eu. The factor for conversion from eu to kcal/mol will be determined in the next section, by calibration against the experimental melting temperature for one of the peptides studied, the Trp-cage.

Local potential. The local potential $E_{\mathrm{loc}} = E^{(1)}_{\mathrm{loc}} + E^{(2)}_{\mathrm{loc}} + E^{(3)}_{\mathrm{loc}}$ can be divided into two backbone terms, $E^{(1)}_{\mathrm{loc}}$ and $E^{(2)}_{\mathrm{loc}}$, and one side-chain term, $E^{(3)}_{\mathrm{loc}}$. In describing

the potential, the concept of a peptide unit is useful. A peptide unit consists of the backbone C′O group of one residue and the backbone NH group of the next residue.

- The potential $E_{\rm loc}^{(1)}$ represents interactions between partial charges of neighboring peptide units along the chain. It is given by

$$E_{\rm loc}^{(1)} = \kappa_{\rm loc}^{(1)} \sum_{\rm n.n.} \sum_i \sum_j \frac{q_i q_j}{r_{ij}/\text{Å}}, \tag{V.2}$$

  where the outer sum runs over all pairs of nearest-neighbor peptide units and each of the two inner sums runs over atoms in one peptide unit (if the N side of the peptide unit is proline the sum runs over only C′ and O). The partial charge q_i is taken as ± 0.42 for C′ and O atoms and ± 0.20 for N and H atoms. The parameter $\kappa_{\rm loc}^{(1)}$ is set to 6 eu, corresponding to a dielectric constant of $\epsilon_{\rm r} \approx 41$. Two peptide units that are not nearest neighbors along the chain interact through hydrogen bonding (see below) rather than through the potential $E_{\rm loc}^{(1)}$.

- The term $E_{\rm loc}^{(2)}$ provides an additional OO and HH repulsion for neighboring peptide units, unless the residue flanked by the two peptide units is a glycine. This repulsion is added to make doubling of hydrogen bonds less likely. Glycine has markedly different backbone energetics compared to other residues. The lack of C_β atom makes glycine more flexible. However, the observed distribution of Ramachandran φ, ψ angles for glycine in PDB structures [19] is not as broad as simple steric considerations would suggest. $E_{\rm loc}^{(2)}$ provides an energy penalty for glycine ψ values around $\pm 120°$, which are sterically allowed but relatively rare in PDB structures.

  The full expression for $E_{\rm loc}^{(2)}$ is

$$E_{\rm loc}^{(2)} = \kappa_{\rm loc}^{(2)} \sum_{\text{non-Gly}} [f(u_I) + f(v_I)] + \kappa_{\rm loc,G}^{(2)} \sum_{\rm Gly} (\cos\psi_I + 2\cos 2\psi_I), \tag{V.3}$$

  where $\kappa_{\rm loc}^{(2)} = 1.2$ eu, $\kappa_{\rm loc,G}^{(2)} = -0.15$ eu, I is a residue index, and

$$u_I = \min[d(\mathrm{H}_I, \mathrm{N}_{I+1}), d(\mathrm{N}_I, \mathrm{H}_{I+1})] - d(\mathrm{H}_I, \mathrm{H}_{I+1}) \tag{V.4}$$

$$v_I = \min[d(\mathrm{O}_I, \mathrm{C}'_{I+1}), d(\mathrm{C}'_I, \mathrm{O}_{I+1})] - d(\mathrm{O}_I, \mathrm{O}_{I+1}) \tag{V.5}$$

$$f(x) = \max(0, \tanh 3x) \tag{V.6}$$

  The function $f(u_I)$ is positive if the $\mathrm{H}_I\mathrm{H}_{I+1}$ distance, $d(\mathrm{H}_I, \mathrm{H}_{I+1})$, is smaller than both of the $\mathrm{H}_I\mathrm{N}_{I+1}$ and $\mathrm{N}_I\mathrm{H}_{I+1}$ distances, and zero otherwise. This term thus provides an energy penalty when H_I and H_{I+1}

Table V.2: Classification of side-chain angles, χ_i. The parameters of the torsion angle potential $E^{(3)}_{\rm loc}$ are $(\kappa^{(3)}_{{\rm loc},i}, n_i) = (0.6\,{\rm eu}, 3)$ for class I, $(\kappa^{(3)}_{{\rm loc},i}, n_i) = (0.3\,{\rm eu}, 3)$ for class II, $(\kappa^{(3)}_{{\rm loc},i}, n_i) = (0.4\,{\rm eu}, 2)$ for class III, and $(\kappa^{(3)}_{{\rm loc},i}, n_i) = (-0.4\,{\rm eu}, 2)$ for class IV.

Residue	χ_1	χ_2	χ_3	χ_4
Ser, Cys, Thr, Val	I			
Ile, Leu	I	I		
Asp, Asn	I	IV		
His, Phe, Tyr, Trp	I	III		
Met	I	I	II	
Glu, Gln	I	I	IV	
Lys	I	I	I	I
Arg	I	I	I	III

are exposed to each other (it is omitted if residue I or $I+1$ is a proline). Similarly, $f(v_I)$ is positive when O_I and O_{I+1} are exposed to each other.

- $E^{(3)}_{\rm loc}$ is an explicit torsion angle potential for side-chain angles, χ_i. Many side-chain angles display distributions resembling what one would expect based on simple steric considerations. The use of the torsion potential is particularly relevant for χ_2 in asparagine and aspartic acid and χ_3 in glutamine and glutamic acid. The torsion potential is defined as

$$E^{(3)}_{\rm loc} = \sum_i \kappa^{(3)}_{{\rm loc},i} \cos n_i \chi_i \,, \tag{V.7}$$

where $\kappa^{(3)}_{{\rm loc},i}$ and n_i are constants. Each side-chain angle χ_i belongs to one of four classes associated with different values of $\kappa^{(3)}_{{\rm loc},i}$ and n_i (see Table V.2).

Excluded volume. Excluded-volume effects are modeled using the potential

$$E_{\rm ev} = \kappa_{\rm ev} \sum_{i<j} \left[\frac{\lambda_{ij}(\sigma_i + \sigma_j)}{r_{ij}} \right]^{12} , \tag{V.8}$$

where the summation is over all pairs of atoms with a non-constant separation, $\kappa_{\rm ev} = 0.10\,{\rm eu}$, and $\sigma_i = 1.77$, 1.75, 1.53, 1.42 and 1.00 Å for S, C, N, O and H atoms, respectively. The parameter λ_{ij} is unity for pairs connected by three covalent bonds and $\lambda_{ij} = 0.75$ for all other pairs. To speed up the calculations, $E_{\rm ev}$ is evaluated using a cutoff of $4.3\lambda_{ij}$ Å.

Hydrogen bonding. Our potential contains an explicit hydrogen-bond term, $E_{\rm hb}$. All hydrogen bonds in the model are between NH and CO groups. They connect either two backbone groups or a charged side-chain (aspartic acid, glutamic acid, lysine, arginine) with a backbone group. Two neighboring peptide units, which interact through the local potential (see above), are not allowed to hydrogen bond with each other.

The form of the hydrogen-bond potential is

$$E_{\rm hb} = \epsilon_{\rm hb}^{(1)} \sum_{\rm bb-bb} u(r_{ij}) v(\alpha_{ij}, \beta_{ij}) + \epsilon_{\rm hb}^{(2)} \sum_{\rm sc-bb} u(r_{ij}) v(\alpha_{ij}, \beta_{ij}), \tag{V.9}$$

where $\epsilon_{\rm hb}^{(1)} = 3.0$ eu and $\epsilon_{\rm hb}^{(2)} = 2.3$ eu set the strengths of backbone-backbone and sidechain-backbone bonds, respectively, r_{ij} is the HO distance, α_{ij} is the NHO angle, and β_{ij} is the HOC angle. The functions $u(r)$ and $v(\alpha,\beta)$ are given by

$$u(r) = 5\left(\frac{\sigma_{\rm hb}}{r}\right)^{12} - 6\left(\frac{\sigma_{\rm hb}}{r}\right)^{10} \tag{V.10}$$

$$v(\alpha,\beta) = \begin{cases} (\cos\alpha\cos\beta)^{1/2} & \text{if } \alpha,\beta > 90^\circ \\ 0 & \text{otherwise} \end{cases} \tag{V.11}$$

where $\sigma_{\rm hb} = 2.0$ Å. A 4.5 Å cutoff is used for $u(r)$.

Side-chain potential. Our side-chain potential is composed of two terms, $E_{\rm sc} = E_{\rm hp} + E_{\rm ch}$. The $E_{\rm ch}$ term represents interactions among side-chain charges. The first and more important term, $E_{\rm hp}$, is meant to capture the effects of all other relevant interactions, among which effective hydrophobic attraction is assumed to be most important. For convenience, $E_{\rm hp}$ and $E_{\rm ch}$ have a similar form,

$$E_{\rm hp} = -\sum_{I<J} M_{IJ}^{\rm (hp)} C_{IJ}^{\rm (hp)} \qquad E_{\rm ch} = -\sum_{I<J} M_{IJ}^{\rm (ch)} C_{IJ}^{\rm (ch)}. \tag{V.12}$$

Here the sums run over residue pairs IJ, $C_{IJ}^{\rm (hp)}$ and $C_{IJ}^{\rm (ch)}$ are contact measures that take values between 0 and 1, and $M_{IJ}^{\rm (hp)}$ and $M_{IJ}^{\rm (ch)}$ are energy parameters.

It is assumed that ten of the twenty natural amino acids contribute to $E_{\rm hp}$, see Table V.3. Included among these ten are lysine and arginine, which are charged but have large hydrophobic parts. To reduce the number of parameters, the hydrophobic contact energies are taken to be additive, $M_{IJ}^{\rm (hp)} = m_I + m_J$. It is known that the statistically derived Miyazawa-Jernigan contact matrix [20] can be approximately decomposed this way [21]. The m_I parameters can be found in Table V.3. $M_{IJ}^{\rm (hp)}$ is set to 0 if residues I and J are

Table V.3: The parameter m_I of the hydrophobicity potential $E_{\rm hp}$.

Residue	m_I (eu)
Arg	0.3
Met, Lys	0.4
Val	0.6
Ile, Leu, Pro	0.8
Tyr	1.1
Phe, Trp	1.6

nearest neighbors along the chain, and is reduced by a factor 2 for next-nearest neighbors.

The residues taken as charged are aspartic acid, glutamic acid, lysine and arginine. The charge-charge contact energy is $-M_{IJ}^{(\rm ch)} = 1.5 s_I s_J$ eu, where s_I and s_J are the signs of the charges (± 1).

The contact measure $C_{IJ}^{(\rm hp)}$ is calculated using a predetermined set of atoms for each amino acid, denoted by $A_I^{(\rm hp)}$ (see Table V.4). Let n_I be the number of atoms in $A_I^{(\rm hp)}$ and let

$$\Gamma_{IJ}^{(\rm hp)} = \sum_{i \in A_I^{(\rm hp)}} g(\min_{j \in A_J^{(\rm hp)}} r_{ij}^2), \tag{V.13}$$

where $g(x)$ is unity for $x < (3.7\text{Å})^2$, vanishes for $x > (4.5\text{Å})^2$, and varies linearly for intermediate x. The contact measure can then be written as

$$C_{IJ}^{(\rm hp)} = \frac{\min(\gamma_{IJ}(n_I + n_J), \Gamma_{IJ}^{(\rm hp)} + \Gamma_{JI}^{(\rm hp)})}{\gamma_{IJ}(n_I + n_J)}, \tag{V.14}$$

where γ_{IJ} is either 1 or 0.75. For $\gamma_{IJ} = 1$, $C_{IJ}^{(\rm hp)}$ is, roughly speaking, the fraction of atoms in $A_I^{(\rm hp)}$ and $A_J^{(\rm hp)}$ that are in contact with some atom from the other of the two sets. A reduction to $\gamma_{IJ} = 0.75$ makes it easier to achieve a full contact ($C_{IJ}^{(\rm hp)} = 1$). The value $\gamma_{IJ} = 0.75$ is used for interactions within the group proline, phenylalanine, tyrosine and tryptophan, to make face-to-face stacking of these side-chains less likely. It is also used within the group isoleucine, leucine and valine, because a full contact is otherwise hard to achieve for these pairs. In all other cases, γ_{IJ} is unity.

The definition of $C_{IJ}^{(\rm ch)}$ is similar. The γ_{IJ} parameter is unity for charge-charge interactions, and the sets of atoms used, $A_I^{(\rm ch)}$, can be found in Table V.5.

Chain ends. Some of the sequences we study have extra groups attached at one or both ends of the chain. The groups occurring are N-terminal acetyl and succinylic acid, and C-terminal NH_2. When such a unit is present, the

Table V.4: Atoms used in the calculation of the contact measure $C_{IJ}^{(\mathrm{hp})}$.

Residue	Set of atoms (A_I)
Pro	C_β, C_γ, C_δ
Tyr	C_γ, $C_{\delta 1}$, $C_{\delta 2}$, $C_{\epsilon 1}$, $C_{\epsilon 2}$, C_ζ
Val	C_β, $C_{\gamma 1}$, $C_{\gamma 2}$
Ile	C_β, $C_{\gamma 1}$, $C_{\gamma 2}$, C_δ
Leu	C_β, C_γ, $C_{\delta 1}$, $C_{\delta 2}$
Met	C_β, C_γ, S_δ, C_ϵ
Phe	C_γ, $C_{\delta 1}$, $C_{\delta 2}$, $C_{\epsilon 1}$, $C_{\epsilon 2}$, C_ζ
Trp	C_γ, $C_{\delta 1}$, $C_{\delta 2}$, $C_{\epsilon 3}$, $C_{\zeta 3}$, $C_{\eta 2}$
Arg	C_β, C_γ
Lys	C_β, C_γ, C_δ

Table V.5: Atoms used in the calculation of the contact measure $C_{IJ}^{(\mathrm{ch})}$.

Residue	Set of atoms (A_I)
Arg	N_ϵ, C_ζ, $N_{\eta 1}$, $N_{\eta 2}$
Lys	$^1H_\zeta$, $^2H_\zeta$, $^3H_\zeta$
Asp	$O_{\delta 1}$, $O_{\delta 2}$
Glu	$O_{\epsilon 1}$, $O_{\epsilon 2}$

model assumes polar NH and CO groups beyond the last C_α atom to hydrogen bond like backbone NH/CO groups but with the strength reduced by a factor 2 (multiplicatively). The charged group of succinylic acid interacts like a charged side-chain.

In the absence of end groups, the model assumes the N and C termini to be positively and negatively charged, respectively, and to interact like charged side-chains.

Monte Carlo details. We investigate the folding thermodynamics of this model by Monte Carlo (MC) methods. The simulations are done using either simulated tempering (ST) [22, 23] or parallel tempering/replica exchange (PT) [24, 25], both with temperature as a dynamical variable. For small systems we use ST, with seven geometrically distributed temperatures in the range 279 K–367 K. For each system, ten independent ST runs are performed. For our largest systems we use PT with a set of sixteen temperatures, spanning the same interval. Using fourfold multiplexing [26], one run comprising 64 parallel trajectories is performed for each system. The PT temperature distribution is determined by an optimization procedure [26]. The length of our different simulations can be found in Table V.6.

Table V.6: Algorithm used and total number of elementary MC steps for all systems studied.

System	Method	MC steps
Trp-cage, E6apn1	ST	$10 \times 1.0 \times 10^9$
C, EK, F_s, GCN4tp	ST	$10 \times 1.0 \times 10^9$
HPLC-6	ST	$10 \times 3.0 \times 10^9$
Chignolin	ST	$10 \times 0.5 \times 10^9$
MBH12	ST	$10 \times 1.0 \times 10^9$
GB1p	ST	$10 \times 2.0 \times 10^9$
GB1m2, GB1m3	ST	$10 \times 1.0 \times 10^9$
trpzip1, trpzip2	ST	$10 \times 1.0 \times 10^9$
betanova, LLM	ST	$10 \times 1.0 \times 10^9$
beta3s	ST	$10 \times 2.0 \times 10^9$
AB zipper	PT	$64 \times 3.0 \times 10^9$
Top7-CFR	PT	$64 \times 2.4 \times 10^9$
GS-α_3W	PT	$64 \times 3.5 \times 10^9$

Three different conformational updates are used in the simulations: single variable updates of side-chain and backbone angles, respectively, and Biased Gaussian Steps (BGS) [27]. The BGS move is semi-local and updates up to eight consecutive backbone degrees of freedom in a manner that keeps the ends of the segment approximately fixed. The ratio of side-chain to backbone updates is the same at all temperatures, whereas the relative frequency of the two backbone updates depends on the temperature. At high temperatures the single variable update is the only backbone update used, and at low temperatures only BGS is used. At intermediate temperatures both updates are used.

The AB zipper, a two-chain system, is studied using a periodic box of size $(158\,\text{Å})^3$. In addition to the conformational updates described above, the simulations of this system used rigid body translations and rotations of individual chains.

Our simulations are performed using the open source C++-package PROFASI [28]. Future public releases of PROFASI will include an implementation of the force field described here. While this force field has been implemented in PROFASI in an optimized manner, this optimization does not involve a parallel evaluation of the potential on many processors. Therefore, in our simulations the number of processors used is the same as the number of MC trajectories generated. For a typical small peptide, a trajectory of the length as given in Table V.6 takes ~ 18 hours to generate on an AMD Opteron processor with ~ 2.0 GHz clock rate. For the largest system studied, GS-α_3W, the simula-

tions, with a proportionately larger number of MC updates, take ~ 10 days to complete.

Analysis. In our simulations, we monitor a variety of different properties. Three important observables are as follows.

1. α-helix content, h. A residue is defined as helical if its Ramachandran angle pair is in the region $-90^\circ < \varphi < -30^\circ$, $-77^\circ < \psi < -17^\circ$. Following [29], a stretch of $n > 2$ helical residues is said to form a helical segment of length $n-2$. For an end residue that is not followed by an extra end group, the (φ, ψ) pair is poorly defined. Thus, for a chain with N residues, the maximum length of a helical segment is $N-4$, $N-3$ or $N-2$, depending on whether there are zero, one or two end groups. The α-helix content h is defined as the total length of all helical segments divided by this maximum length.

2. Root-mean-square deviation from a folded reference structure, bRMSD/RMSD/pRMSD. bRMSD is calculated over backbone atoms, whereas RMSD is calculated over all heavy atoms. All residues except the two end residues are included in the calculation, unless otherwise stated. For the case of the dimeric AB zipper, the periodic box used for the simulations has to be taken into account. The two chains in the simulation might superficially appear to be far away when they are in fact close, because of periodicity. For this case we evaluate backbone RMSD over atoms taken from both chains in the dimer, and minimize this value with respect to periodic translations. We denote this as pRMSD.

3. Nativeness measure based on hydrogen bonds, q_{hb}. This observable has the value 1 if at most two native backbone-backbone hydrogen bonds are missing, and is 0 otherwise. A hydrogen bond is considered formed if its energy is less than -1.03 eu.

In many cases, it turns out that the temperature dependence of our results can be approximately described in terms of the simple two-state model

$$X(T) = \frac{X_1 + X_2 K(T)}{1 + K(T)} \qquad K(T) = \exp\left[\left(\frac{1}{RT} - \frac{1}{RT_{\mathrm{m}}}\right)\Delta E\right] \tag{V.15}$$

where $X(T)$ is the quantity studied, X_1 and X_2 are the values of X in the two states, and $K(T)$ is the effective equilibrium constant (R is the gas constant). In this first-order form, $K(T)$ contains two parameters: the melting temperature T_{m} and the energy difference ΔE. The parameters T_{m}, ΔE, X_1 and X_2 are determined by fitting to data.

Thermal averages and their statistical errors are calculated by using the jackknife method [30], after discarding the first 20 % of each MC trajectory for thermalization.

Figures of 3D structures were prepared using PyMOL [31].

Results

We study a total of 20 peptide/protein systems, listed in Table V.1 (amino acid sequences can be found in this table). Among these, there are 17 smaller systems with 10–37 residues and 3 larger ones with ≥ 49 residues. Many of the smaller systems have been simulated by other groups, in some cases with explicit water (for a review, see [32]). Two of the three larger systems, as far as we know, have not been studied using other force fields. A study of the 67-residue three-helix-bundle protein GS-α_3W using the ECEPP/3 force field was recently reported [33]. The simulations presented here use the same geometric representation and find about a hundred times the number of independent folding events, while consuming much smaller computing resources.

Trp-cage and E6apn1. The Trp-cage is a designed 20-residue miniprotein with a compact helical structure [34]. Its NMR-derived native structure (see Figure V.1) contains an α-helix and a single turn of 3_{10}-helix [34]. The E6apn1 peptide was designed using the Trp-cage motif as a scaffold, to inhibit the E6 protein of papillomavirus [35]. E6apn1 is three residues larger than the Trp-cage but has a similar structure, except that the α-helix is slightly longer [35].

As indicated earlier, we use melting data for the Trp-cage to set the energy scale of the model. For this peptide, several experiments found a similar melting temperature, $T_\mathrm{m} \sim 315$ K [34, 36, 37]. In our model, the heat capacity of the Trp-cage displays a maximum at $RT = 0.4722 \pm 0.0008$ eu. Our energy unit eu is converted to kcal/mol by setting this temperature equal to the experimental melting temperature (315 K). Having done that, there is no free parameter left in the model. Other systems are thus studied without tuning any model parameter. For E6apn1, the experimental melting temperature is $T_\mathrm{m} \sim 305$ K [35].

Figure V.2a shows the helix content h against temperature for the Trp-cage and E6apn1, as obtained from our simulations. In both cases, the T dependence is well described by the simple two-state model of Equation V.15. The fitted melting temperatures are $T_\mathrm{m} = 309.6 \pm 0.7$ K and $T_\mathrm{m} = 304.0 \pm 0.5$ K for the Trp-cage and E6apn1, respectively. This T_m value for the Trp-cage is slightly lower than that we obtain from heat capacity data, 315 K. A fit to our data

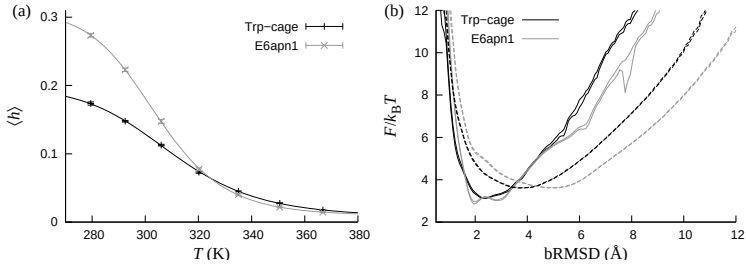

Figure V.2: The Trp-cage and E6apn1. (a) Helix content h against temperature. The lines are two-state fits ($T_m = 309.6 \pm 0.7$ K and $\Delta E = 11.3 \pm 0.3$ kcal/mol for the Trp-cage; $T_m = 304.0 \pm 0.5$ K and $\Delta E = 14.2 \pm 0.3$ kcal/mol for E6apn1). (b) Free energy F calculated as a function of bRMSD at two different temperatures, 279 K and 306 K. The double lines indicate the statistical errors.

for the hydrophobicity energy E_{hp} (not shown) gives instead a slightly larger T_m, 321.1 ± 0.8 K. This probe dependence of T_m implies an uncertainty in the determination of the energy scale. By using the Trp-cage, this uncertainty is kept small (~ 2 %). For many other peptides, the spread in T_m is much larger (see below).

Figure V.2b shows the free energy calculated as a function of bRMSD for the Trp-cage and E6apn1 at two different temperatures. The first temperature, 279 K, is well below T_m. Here native-like conformations dominate and the global free-energy minima are at 2.4 Å and 2.0 Å for the Trp-cage and E6apn1, respectively. At the second temperature, 306 K, the minima are shifted to higher bRMSD. Note that these free-energy profiles, taken near T_m, show no sign of a double-well structure. Hence, these peptides do not show a genuine two-state behavior in our simulations, even though the melting curves (Figure V.2a) are well described by a two-state model, as are many experimentally observed melting curves.

The α-helices C, EK, F_s, GCN4tp and HPLC-6. Our next five sequences form α-helices. Among these, there are large differences in helix stability, according to CD studies. The least stable are the C [38] and EK [39] peptides, which are only partially stable at $T \sim 273$ K. The original C peptide is a 13-residue fragment of ribonuclease A, but the C peptide here is an analogue with two alanine substitutions and a slightly increased helix stability [40]. The EK peptide is a designed alanine-based peptide with 14 residues.

Our third α-helix peptide is the 21-residue F_s [41], which is also alanine-based. F_s is more stable than C and EK [41, 42], with estimated T_m values of

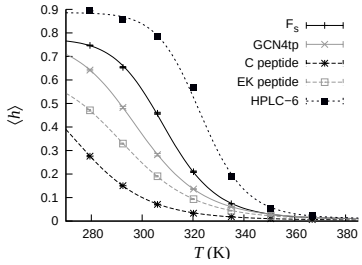

Figure V.3: The C, EK, F_s, GCN4tp and HPLC-6 peptides. Helix content h against temperature. The lines are two-state fits ($T_m = 276.3 \pm 2.4$ K and $\Delta E = 11.7 \pm 0.4$ kcal/mol for C; $T_m = 293.9 \pm 0.4$ K and $\Delta E = 12.6 \pm 0.2$ kcal/mol for EK; $T_m = 309.2 \pm 0.3$ K and $\Delta E = 18.7 \pm 0.4$ kcal/mol for F_s; $T_m = 298.9 \pm 0.1$ K and $\Delta E = 14.1 \pm 0.1$ kcal/mol for GCN4tp; $T_m = 323.3 \pm 1.2$ K and $\Delta E = 23.6 \pm 2.2$ kcal/mol for HPLC-6).

308 K [42] and 303 K [43] from CD studies and 334 K from an IR study [44]. Even more stable is HPLC-6, a winter flounder antifreeze peptide with 37 residues. CD data suggest that the helix content of HPLC-6 remains non-negligible, ~ 0.10, at temperatures as high as ~ 343 K [45]. Our fifth helix-forming sequence, which we call GCN4tp, has 17 residues and is taken from a study of GCN4 coiled-coil formation [46]. Its melting behavior has not been studied, as far as we know, but its structure was characterized by NMR [46].

These five peptides are indeed α-helical in our model. At 279 K, the calculated helix content h is 0.28 for the C peptide, 0.47 for the EK peptide, and > 0.60 for the other three peptides. Figure V.3 shows the temperature dependence of h. By fitting Equation V.15 to the data for the three stable sequences, we find melting temperatures of 298.9 ± 0.1 K, 309.2 ± 0.3 K and 323.3 ± 1.2 K for GCN4tp, F_s and HPLC-6, respectively.

For the four peptides whose melting behavior has been studied experimentally, these results are in good agreement with experimental data. In particular, we find that HPLC-6 indeed is more stable than F_s in the model, which in turn is more stable than both C and EK. The model thus captures the stability order among these peptides.

The β-hairpins chignolin and MBH12. We now turn to β-sheet peptides and begin with the β-hairpins chignolin [47] and MBH12 [48] with 10 and 14 residues, respectively. Both are designed and have been characterized by NMR. For chignolin, T_m values in the range 311–315 K were reported [47], based on CD and NMR. We are not aware of any melting data for MBH12.

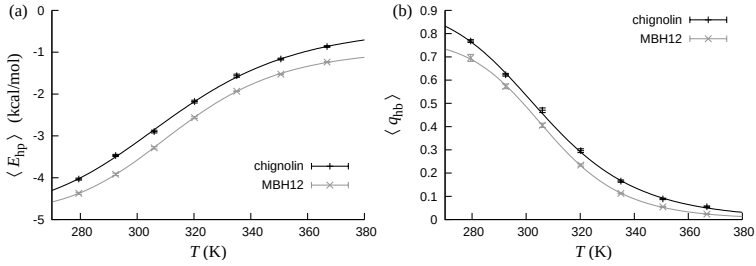

Figure V.4: Chignolin and MBH12. (a) Hydrophobicity energy E_{hp} against temperature. The lines are two-state fits ($T_m = 311.0 \pm 0.5$ K and $\Delta E = 9.6 \pm 0.2$ kcal/mol for chignolin; $T_m = 315.4 \pm 1.3$ K and $\Delta E = 9.9 \pm 0.9$ kcal/mol for MBH12). (b) Nativeness q_{hb} against temperature. The lines are two-state fits ($T_m = 305.4 \pm 0.5$ K and $\Delta E = 10.4 \pm 0.1$ kcal/mol for chignolin; $T_m = 309.2 \pm 0.7$ K and $\Delta E = 13.5 \pm 0.2$ kcal/mol for MBH12).

Figure V.4 shows the temperature dependence of the hydrophobicity energy E_{hp} and the nativeness parameter q_{hb} for these peptides. By fitting to E_{hp} data, we obtain $T_m = 311.0 \pm 0.5$ K and $T_m = 315.4 \pm 1.3$ K for chignolin and MBH12, respectively. Using q_{hb} data instead, we find $T_m = 305.4 \pm 0.5$ K for chignolin and $T_m = 309.2 \pm 0.7$ K for MBH12. These T_m values show a significant but relatively weak probe dependence. The values for chignolin can be compared with experimental data, and the agreement is good.

Because these peptides have only four native hydrogen bonds each, one may question our definition of q_{hb} (see Methods), which takes a conformation as native-like ($q_{hb} = 1$) even if two hydrogen bonds are missing. Therefore, we repeated the analysis using the stricter criterion that native-like conformations ($q_{hb} = 1$) may lack at most one hydrogen bond. The resulting decrease in native population, as measured by the average q_{hb}, was ~ 0.1 or smaller at all temperatures. Even with this stricter definition, we find native populations well above 0.5 at low temperature for both peptides.

The β-hairpins GB1p, GB1m2 and GB1m3. GB1p is the second β-hairpin of the B1 domain of protein G (residues 41–56). Its folded population has been estimated by CD/NMR to be 0.42 at 278 K [49] and ~ 0.30 at 298 K [50], whereas a Trp fluorescence study found a T_m of 297 K [51], corresponding to a somewhat higher folded population. GB1m2 and GB1m3 are two mutants of GB1p with significantly enhanced stability [50]. At 298 K, the folded population was found to be 0.74 ± 0.05 for GB1m2 and 0.86 ± 0.03 for GB1m3, based on CD

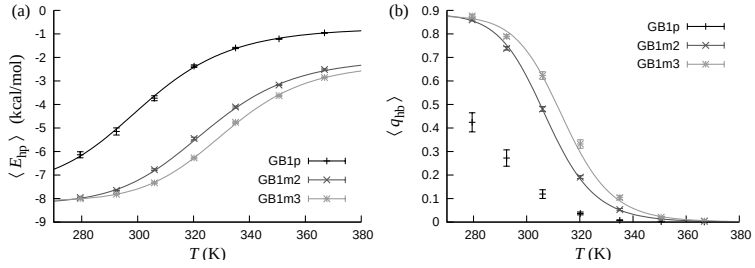

Figure V.5: GB1p, GB1m2 and GB1m3. (a) Hydrophobicity energy E_{hp} against temperature. The lines are two-state fits ($T_m = 301.7 \pm 3.3$ K and $\Delta E = 11.3 \pm 1.1$ kcal/mol for GB1p; $T_m = 324.4 \pm 1.4$ K and $\Delta E = 13.2 \pm 1.0$ kcal/mol for GB1m2; $T_m = 331.4 \pm 0.7$ K and $\Delta E = 14.8 \pm 0.5$ kcal/mol for GB1m3). (b) Nativeness q_{hb} against temperature. The lines are two-state fits ($T_m = 307.5 \pm 0.5$ K and $\Delta E = 20.7 \pm 0.5$ kcal/mol for GB1m2; $T_m = 313.9 \pm 1.4$ K and $\Delta E = 21.4 \pm 1.1$ kcal/mol for GB1m3).

and NMR measurements [50]. It was further estimated that $T_m = 320 \pm 2$ K for GB1m2 and $T_m = 333 \pm 2$ K for GB1m3 [50].

All these three peptides are believed to adopt a structure similar to that GB1p has as part of the protein G B1 domain (PDB code 1GB1). This part of the full protein contains seven backbone-backbone hydrogen bonds. These hydrogen bonds are the ones we consider when evaluating q_{hb} for these peptides.

Figure V.5 shows the observables E_{hp} and q_{hb} against temperature for these peptides. Fits to the data give E_{hp}-based T_m values of 301.7 ± 3.3 K, 324.4 ± 1.1 K and 331.4 ± 0.7 K for GB1p, GB1m2 and GB1m3, respectively, and q_{hb}-based T_m values of 307.5 ± 0.5 K and 313.9 ± 1.4 K for GB1m2 and GB1m3, respectively. The q_{hb} data do not permit a reliable fit for the less stable GB1p. At 298 K, we find q_{hb}-based folded populations of 0.20, 0.64 and 0.74 for GB1p, GB1m2 and GB1m3, respectively, which can be compared with the above-mentioned experimental results (0.30, 0.74 and 0.86).

These results show that, in the model, the apparent folded populations of these peptides depend quite strongly on the observable studied. Our E_{hp}-based results agree quite well with experimental data, especially for GB1m2 and GB1m3, whereas our q_{hb} results consistently give lower folded populations for all peptides. The stability order is the same independent of which of the two observables we study, namely GB1p < GB1m2 < GB1m3, which is the experimentally observed order.

The stability difference between GB1m2 and GB1m3 is mainly due to charge-charge interactions. In our previous model [6], these interactions were ignored,

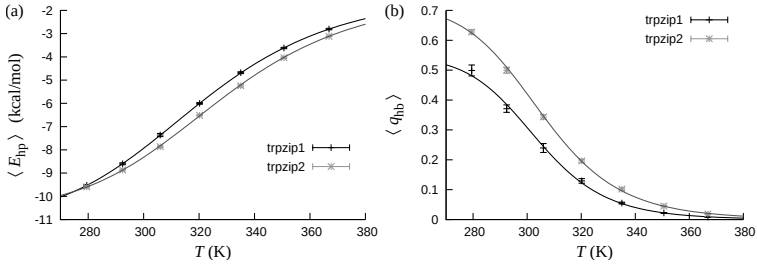

Figure V.6: Trpzip1 and trpzip2. (a) Hydrophobicity energy E_{hp} against temperature. The lines are two-state fits ($T_m = 319.7 \pm 0.2$ K and $\Delta E = 7.9 \pm 0.1$ kcal/mol for trpzip1; $T_m = 327.1 \pm 0.8$ K and $\Delta E = 8.3 \pm 0.4$ kcal/mol for trpzip2). (b) Nativeness q_{hb} against temperature. The lines are two-state fits ($T_m = 303.2 \pm 1.8$ K and $\Delta E = 14.1 \pm 0.5$ kcal/mol for trpzip1; $T_m = 305.0 \pm 1.1$ K and $\Delta E = 12.6 \pm 0.3$ kcal/mol for trpzip2).

and both peptides had similar stabilities. The present model splits this degeneracy. Moreover, the magnitude of the splitting, which sensitively depends on the strength of the charge-charge interactions, is consistent with experimental data.

The β-hairpins trpzip1 and trpzip2. The 12-residue trpzip1 and trpzip2 are designed β-hairpins, each containing two tryptophans per β-strand [52]. The only difference between the two sequences is a transposition of an asparagine and a glycine in the hairpin turn. CD measurements suggest that trpzip1 and trpzip2 are remarkably stable for their size, with T_m values of 323 K and 345 K, respectively [52]. A complementary trpzip2 study, using both experimental and computational methods, found T_m values to be strongly probe-dependent [53].

Figure V.6 shows our melting curves for these peptides, based on the observables E_{hp} and q_{hb}. The E_{hp}-based T_m values are 319.7 ± 0.2 K and 327.1 ± 0.8 K for trpzip1 and trpzip2, respectively. Using q_{hb} data instead, we find $T_m = 303.2 \pm 1.1$ K for trpzip1 and $T_m = 305.0 \pm 1.1$ K for trpzip2.

Like for the other β-hairpins discussed earlier, our q_{hb}-based folded populations are low compared to estimates based on CD data, whereas those based on E_{hp} are much closer to experimental data. For trpzip2, the agreement is not perfect but acceptable, given that T_m has been found to be strongly probe-dependent for this peptide [53].

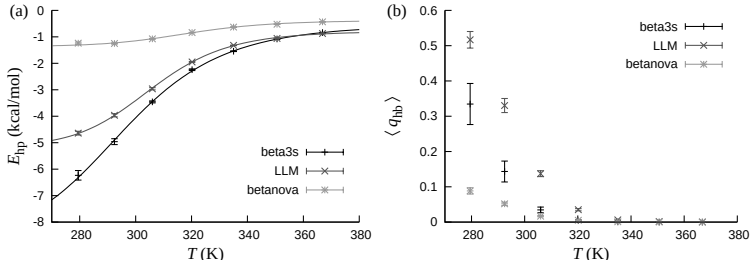

Figure V.7: Betanova, LLM and beta3s. (a) Hydrophobicity energy E_{hp} against temperature. The lines are two-state fits ($T_m = 318.8 \pm 2.5$ K and $\Delta E = 13.3 \pm 2.1$ kcal/mol for betanova; $T_m = 305.6 \pm 1.7$ K and $\Delta E = 13.4 \pm 1.0$ kcal/mol for LLM; $T_m = 295.7 \pm 3.1$ K and $\Delta E = 9.7 \pm 0.5$ kcal/mol for beta3s). (b) Nativeness q_{hb} against temperature. Two-state fits were not possible.

Three-stranded β-sheets: betanova, LLM and beta3s. Betanova [54], the betanova triple mutant LLM [55] and beta3s [56] are designed 20-residue peptides forming three-stranded β-sheets. All the three peptides are marginally stable. NMR studies suggest that the folded population at 283 K is 0.09 for betanova [55], 0.36 for LLM [55], and 0.13–0.31 for beta3s [56].

Figure V.7 shows our E_{hp} and q_{hb} data for these peptides. From the q_{hb} data, T_m values cannot be extracted, because the stability of the peptides is too low. At 283 K, the q_{hb}-based folded populations are 0.08, 0.47, 0.28 for betanova, LLM and beta3s, respectively, in good agreement with the experimental results. Fits to E_{hp} data can be performed. The obtained T_m values are 318.8 ± 2.5 K, 305.6 ± 1.7 K and 295.7 ± 3.1 K for betanova, LLM and beta3s, respectively.

These E_{hp}-based T_m values are high compared to the experimentally determined folded populations, especially for betanova. Note that betanova has a very low hydrophobicity. The correlation between E_{hp} and folding status is therefore likely to be weak for this peptide.

In contrast to the E_{hp}-based folded populations, those based on q_{hb} agree quite well with experimental data. In this respect, the situation is the opposite to what we found for the β-hairpins studied above. A possible reason for this difference is discussed below.

AB zipper. The AB zipper is a designed heterodimeric leucine zipper, composed of an acidic A chain and a basic B chain, each with 30 residues [57]. The dimer structure has been characterized by NMR, and a melting temperature of ~ 340 K was estimated by CD measurements (at neutral pH) [57].

The lowest energy state seen in our simulations is a conformation in which pRMSD calculated over backbone atoms of all residues in both chains is ~ 2.7 Å. In this structure, the bRMSD (all residues) of the individual chains A and B to their counterparts in the PDB structure are ~ 2.5 Å and ~ 2.4 Å, respectively. Unlike for the other systems described in this article, the boundary conditions have a non-trivial role for this dimeric system. A proper discussion of periodicity, concentration and temperature dependence of this system is beyond the scope of this article. In Figure V.8a, we show the energy landscape, i.e., the mean energy as a function of two order parameters for this system. The X-axis shows the measure pRMSD described earlier. The Y-axis represents the sum of the backbone RMSD of the individual chains. pRMSD can be very large even if the sum of bRMSDs is small: the two chains can be folded without making the proper inter-chain contacts. Indeed, the figure shows that the major energy gradients are along the Y-axis, showing that it is energetically favorable for both chains to fold to their respective helical states. The correct dimeric native state is energetically more favorable by ~ 20 kcal/mol compared to two folded helices without proper inter-chain contacts. This is seen more clearly in Figure V.8b, where we plot the average energy as a function of pRMSD for states with two folded chains. We also simulated the two chains A and B of the dimer in isolation. Both chains folded to their native helical conformations. The melting temperatures estimated based on helix content for chains A and B are 314 K and 313 K, respectively. As indicated above, for the dimer, thermodynamic parameters like T_{m} cannot be directly estimated from the present simulations.

Top7-CFr. Top7-CFr, the C-terminal fragment of the designed 93-residue α/β-protein Top7 [58], is the most complex of all molecules studied here. It has both α-helix and β-strand secondary structure elements, and highly non-local hydrogen bonds between the N- and C-terminal strands. CFr is known to form extremely stable homodimers, which retain their secondary structure till very high temperatures like 371 K and high concentrations of denaturants [59].

In [13, 14], an earlier version of our model was used to study the folding of CFr. The simulations pointed to an unexpected folding mechanism. The N-terminal strand initially folds as a non-native continuation of the adjoining α-helix. After the other secondary structure elements form and diffuse to an approximately correct tertiary organization, the non-native extension of the helix unfolds and frees the N-terminal residues. These residues then attach to an existing β-hairpin to complete the three-stranded β-sheet of the native structure. Premature fastening of the chain ends in β-sheet contacts puts the

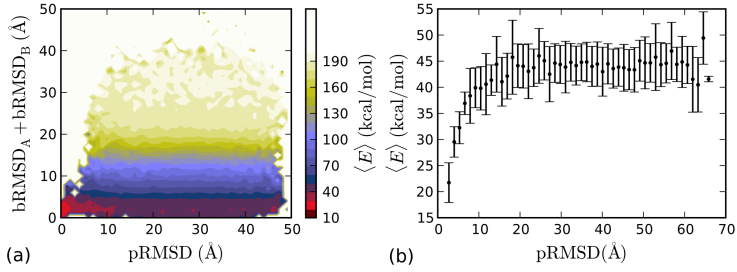

Figure V.8: The heterodimeric AB zipper. (a) Mean energy as a function of pRMSD over both chains and the sum of individual bRMSDs. The direction of the energy gradients implies that a system with two folded monomers is energetically favorable compared to unfolded monomers. The proper dimeric form is the area closest to the origin, and has a lower energy. (b) Mean energy of all states in which both chains have bRMSD < 5 Å, shown as a function of the dimer RMSD measure pRMSD.

molecule in a deep local energy minimum, in which the folding and proper arrangement of the other secondary structure elements is hampered by large steric barriers. The above "caching" mechanism, spontaneously emerging in the simulations, accelerates folding by helping the molecule avoid such local minima.

The folding properties of CFr, including the above mentioned caching mechanism, are preserved under the current modifications of the interaction potential. The center of the native free-energy minimum shifts from bRMSD (all residues) of 1.7 Å as reported in [13] to about 2.2 Å. This state remains the minimum energy state, although the new energy function changes the energy ordering of the other low energy states. The runs made for this study (see Table V.6) found 22 independent folding events. The free-energy landscape observed in the simulations is rather complex with a plethora of deep local minima sharing one or more secondary structure elements with the native structure. They differ in the registry and ordering of strands and the length of the helix. Longer runs are required for the MC simulations to correctly weight these different minima. Temperature dependence of the properties of CFr can therefore not be reliably obtained from these runs.

We note that the simulations ran on twice as many processors but were only about one sixth the length of those used for [13], in which 15 independent folding events were found. The improved efficiency is partly due to the changes in the energy function presented here, and partly due to the optimization of the parallel tempering described in [26].

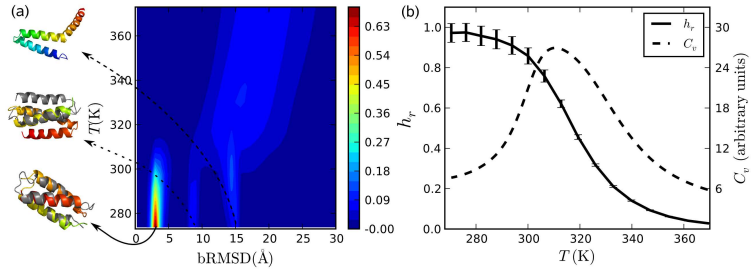

Figure V.9: The three-helix-bundle protein GS-α_3W. (a) Variation of histogram of bRMSD with temperature. At high temperatures, there is a broad distribution of bRMSD with values > 10 Å. At lower temperatures there are three clearly separated clusters. Representative structures from these clusters are also shown (color) aligned with the native structure (gray). (b) Temperature dependence of specific heat, C_v, and the ratio h_r of the observed helix content and the helix content of the native structure.

GS-α_3W. GS-α_3W is a designed three-helix-bundle protein with 67 residues [60], whose structure was characterized by NMR [61]. The stability was estimated to be 4.6 kcal/mol in aqueous solution at 298 K, based on CD data [60].

It turns out that this protein is very easy to fold with our model. Our results are based on extensive sampling of the conformation space with $64 \times 3.5 \times 10^9$ Monte Carlo updates, resulting in about 800 independent folding events to the native state. For this estimate, structures with bRMSD (all residues) under 5 Å were taken to be in the native minimum (see Figure V.9 for justification). Two visits to the native state were considered statistically independent (i) if they occurred in independent Markov chains, or (ii) if the two visits to the native state were separated by at least one visit to the highest temperature in the simulation. For the entire run, we spent about 10 days of computing time on 64 AMD Opteron processors running at 2.0 GHz.

In Figure V.9a, we show how the probabilities for structures with different bRMSD vary with temperature in the simulations. Clearly, the protein makes a transition from a rather continuous distribution of bRMSD at high temperatures to a distribution dominated by three well separated clusters. Analysis of the structures at the lower temperatures shows that all three free-energy minima consist almost exclusively of structures with all three helices of GS-α_3W formed. The plot of the ratio of the observed helix content and the helix content of the native state, shown in Figure V.9b, further supports this idea. The average value of this ratio approaches 1 as the temperature decreases

below 300 K. The specific heat curve, also shown in Figure V.9b, indicates that the formation of these structures correlates with the steepest change in energy.

The cluster with a center at bRMSD ~ 3 Å dominates at the lowest temperatures. The structures contributing to the cluster with ~ 8–9 Å bRMSD superficially look like well folded three-helix bundles. But as illustrated in the figure, the arrangement of the helices is topologically distinct from the native arrangement. The cluster seen at larger bRMSD values is broader and consists of a host of structures in which two of the helices make a helical hairpin, but the third helix is not bound to it. The unbound helix could be at either side of the chain.

According to our model therefore, the population at the lowest temperatures consists of ~ 80% genuinely native structures, ~ 10% three-helix bundles with wrong topology, and ~ 10% other structures with as much helix content as the native state. In order to experimentally determine the true folded population of the protein, the experimental probe must be able to distinguish the native fold from the other helix rich structures described here.

Discussion

The model presented here is intrinsically fast compared to many other all-atom models, because all interactions are short range. By exploiting this property and using efficient MC techniques, it is possible to achieve a high sampling efficiency. We could, for example, generate more than 800 independent folding events for the 67-residue GS-α_3W. The speed of the simulations thus permits statistically accurate studies of the global free-energy landscape of peptides and small proteins.

In developing this potential, a set of 17 peptides with 10–37 residues was studied. The peptides were added to this set one at a time. To fold a new sequence sometimes required fine-tuning of the potential, sometimes not. A change was accepted only after testing the new potential on all previous sequences in the set. In its final form, the model folds all 17 sequences to structures similar to their experimental structures, for one and the same choice of potential parameters.

Also important is the stability of the peptides. A small polypeptide chain is unlikely to be a clear two-state folder, and therefore its apparent folded population will generally depend on the observable studied. For β-sheet peptides, we used the hydrophobicity energy E_{hp} and the hydrogen bond-based nativeness measure q_{hb} to monitor the melting behavior. The extracted T_{m} values indeed showed a clear probe dependence; the E_{hp}-based value was

always larger than that based on q_{hb}. For the β-hairpins studied, we found a good overall agreement between our E_{hp}-based results and experimental data. For the three-stranded β-sheets, instead, the q_{hb} results agreed best with experimental data. The reason for this difference is unclear. One contributing factor could be that interactions between aromatic residues play a more important role for the β-hairpins studied here than for the three-stranded β-sheets. These interactions may influence spectroscopic signals and are part of E_{hp}. Probe-dependent T_{m} values have also been obtained experimentally, for example, for trpzip2 [53].

The probe dependence makes the comparison with experimental data less straightforward. Nevertheless, the results presented clearly show that the model captures many experimentally observed stability differences. In particular, among related peptides, the calculated order of increasing thermal stability generally agrees with the experimental order, independent of which of our observables we use.

It is encouraging that the model is able to fold these 17 sequences. However, there is no existing model that will fold all peptides, and our model is no exception. Two sequences that we unsuccessfully tried to fold are the β-hairpins trpzip4 and U_{16}, both with 16 residues. Trpzip4 is a triple mutant of GB1p with four tryptophans [52]. For trpzip4, our minimum energy state actually corresponded to the NMR-derived native state [52], but the population of this state remained low at the lowest temperature studied ($\sim 14\,\%$ at 279 K, as opposed to an estimated T_{m} of 343 K in experiments [52]). U_{16} is derived from the N-terminal β-hairpin of ubiquitin [62]. It has a shortened turn and has been found to form a β-hairpin with non-native registry [62]. In our simulations, this state was only weakly populated ($\sim 8\%$ at 279 K, as opposed to an estimated $\sim 80\%$ at 288 K [62]). Instead, the main free-energy minima corresponded to the two β-hairpin states with the registry of native ubiquitin, one with native hydrogen bonds and the other with the complementary set of hydrogen bonds.

Our calibration of the potential relies on experimental data with non-negligible uncertainties, on a limited number of peptides. It is not evident that this potential will be useful for larger polypeptide chains. Therefore, as a proof-of-principle test, we also studied three larger systems, with very good results. Our simulations showed that, without having to adjust any parameter, the model folds these sequences to structures consistent with experimental data. Having verified this, it would be interesting to use the model to investigate the mechanisms by which these systems self-assemble, but such an analysis

is beyond the scope of this article. The main purpose of our present study of these systems was to demonstrate the viability of our calibration approach.

The potential can be further constrained by confronting it with more accurate experimental data and data on new sequences. The challenge in this process is to ensure backward compatibility — new constraints should be met without sacrificing properties already achieved.

Conclusion

We have described and tested an implicit solvent all-atom model for protein simulations. The model is computationally fast and yet able to capture structural and thermodynamic properties of a diverse set of sequences. Its computational efficiency greatly facilitates the study of folding and aggregation problems that require exploration of the full free-energy landscape. A program package, called PROFASI [28], for single- and multi-chain simulations with this model is freely available to academic users.

Acknowledgements. We thank Stefan Wallin for suggestions on the manuscript. This work was in part supported by the Swedish Research Council. The simulations of the larger systems were performed at the John von Neumann Institute for Computing (NIC), Research Center Jülich, Germany.

References

1. Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. Protein Sci 11:739–756.
2. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6:197–208.
3. Yoda T, Sugita Y, Okamoto Y (2004) Secondary-structure preferences of force fields for proteins evaluated by generalized-ensemble simulations. Chem Phys 307:269–283.
4. Shell MS, Ritterson R, Dill KA (2008) A test on peptide stability of AMBER force field with implicit solvation. J Phys Chem B 112:6878–6886.
5. Irbäck A, Samuelsson B, Sjunnesson F, Wallin S (2003) Thermodynamics of α- and β-structure formation in proteins. Biophys J 85:1466–1473.
6. Irbäck A, Mohanty S (2005) Folding thermodynamics of peptides. Biophys J 88:1560–1569.
7. Cheon M, Chang I, Mohanty S, Luheshi LM, Dobson CM, et al. (2007) Structural reorganisation and potential toxicity of oligomeric species formed during the assembly of amyloid fibrils. PLoS Comp Biol 3:e173.
8. Irbäck A, Mitternacht S (2008) Spontaneous β-barrel formation: an all-atom Monte Carlo study of Aβ_{16-22} oligomerization. Proteins 71:207–214.

9. Li D, Mohanty S, Irbäck A, Huo S (2008) Formation and growth of oligomers: a monte carlo study of an amyloid tau fragment. PLoS Comp Biol 4:e1000238.
10. Irbäck A, Mitternacht S, Mohanty S (2005) Dissecting the mechanical unfolding of ubiquitin. Proc Natl Acad Sci USA 102:13427–13432.
11. Mitternacht S, Luccioli S, Torcini A, Imparato A, Irbäck A (2009) Changing the mechanical unfolding pathway of FnIII-10 by tuning the pulling strength. Biophys J 96:429–441.
12. Mohanty S, Hansmann UHE (2006) Folding of proteins with diverse folds. Biophys J 91:3573–3578.
13. Mohanty S, Meinke JH, Zimmermann O, Hansmann UHE (2008) Simulation of top7-cfr: a transient helix extension guides folding. Proc Natl Acad Sci USA 105:8004–8007.
14. Mohanty S, Hansmann UHE (2008) Caching of a chameleon segment facilitates folding of a protein with end-to-end β-sheet. J Phys Chem B 112:15134–15139.
15. Ponder JW, Case DA (2003) Force fields for protein simulation. Adv Protein Chem 66:27–85.
16. Hubner IA, Deeds EJ, Shakhnovich EI (2005) High-resolution protein folding with a transferable potential. Proc Natl Acad Sci USA 102:18914–18919.
17. Herges T, Wenzel W (2005) In silico folding of a three helix protein and characterization of its free-energy landscape in an all-atom force field. Phys Rev Lett 94:018101.
18. Ding F, Tsao D, Nie H, Dokholyan NV (2008) Ab initio folding of proteins with all-atom discrete molecular dynamics. Structure 16:1010–1018.
19. Hovmöller S, Zhou T, Ohlsson T (2002) Conformations of amino acids in proteins. Acta Cryst D 58:768–776.
20. Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 256:623–644.
21. Li H, Tang C, Wingreen NS (1997) Nature of driving force for protein folding: a result from analyzing the statistical potential. Phys Rev Lett 79:765–768.
22. Lyubartsev AP, Martsinovski AA, Shevkunov SV, Vorontsov-Velyaminov PN (1992) New approch to Monte Carlo calculation of the free energy: Method of expanded ensembles. J Chem Phys 96:1776–1783.
23. Marinari E, Parisi G (1992) Simulated tempering: a new Monte Carlo scheme. Europhys Lett 19:451–458.
24. Swendsen RH, Wang JS (1986) Replica Monte Carlo simulation of spin glasses. Phys Rev Lett 57:2607–2609.
25. Hukushima K, Nemoto K (1996) Exchange Monte Carlo method and application to spin glass simulations. J Phys Soc (Jap) 65:1604–1608.
26. Meinke J, Mohanty S, Nadler W (2009) Manuscript in preparation .
27. Favrin G, Irbäck A, Sjunnesson F (2001) Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. J Chem Phys 114:8154–8158.
28. Irbäck A, Mohanty S (2006) PROFASI: a Monte Carlo simulation package for protein folding and aggregation. J Comput Chem 27:1548–1555.
29. García AE, Sanbonmatsu KY (2002) α-helical stabilization by side chain shielding of backbone hydrogen bonds. Proc Natl Acad Sci USA 99:2782–2787.
30. Miller RG (1974) The jackknife – a review. Biometrika 61:1–15.

31. DeLano WL (2002). The PyMOL molecular graphics system. San Carlos, CA: DeLano Scientific.
32. Gnanakaran S, Nymeyer H, Portman J, Sanbonmatsu KY, García AE (2003) Peptide folding simulations. Curr Opin Struct Biol 13:168–174.
33. Meinke J, Hansmann UHE (2009) Submitted manuscript .
34. Neidigh JW, Fesinmeyer RM, Andersen NH (2002) Desiging a 20-residue protein. Nat Struct Biol 9:425–430.
35. Liu Y, Liu Z, Androphy E, Chen J, Baleja JD (2004) Design and characterization of helical peptides that inhibit the e6 protein of papillomavirus. Biochemistry 43:7421–7431.
36. Qiu L, Pabit SA, Roitberg AE, Hagen SJ (2002) Smaller and faster: the 20-residue trp-cage protein folds in 4 μs. J Am Chem Soc 124:12952–12953.
37. Streicher WW, Makhatadze GI (2007) Unfolding thermodynamics of Trp-cage, a 20 residue miniprotein, studied by differential scanning calorimetry and circular dichroism spectroscopy. Biochemistry 46:2876–2880.
38. Bierzynski A, Kim PS, Baldwin RL (1982) A salt bridge stabilizes the helix formed by isolated c-peptide of rnase a. Proc Natl Acad Sci USA 79:2470–2474.
39. Scholtz JM, Barrick D, York EJ, Stewart JM, Baldwin RL (1995) Urea unfolding of peptide helices as a model for interpreting protein unfolding. Proc Natl Acad Sci USA 92:185–189.
40. Shoemaker KR, Kim PS, York EJ, Stewart JM, Baldwin RL (1987) Tests of the helix dipole model for stabilization of α-helices. Nature 326:563–567.
41. Lockhart DJ, Kim PS (1992) Internal Stark effect measurement of the electric field at the amino acid terminus of an α helix. Science 257:947–951.
42. Lockhart DJ, Kim PS (1993) Electrostatic screening of charge and dipole interactions with the helix backbone. Science 260:198–202.
43. Thompson PA, Eaton WA, Hofrichter J (1997) Laser temperature jump study of the helix⇋coil kinetics of an alanine peptide interpreted with a 'kinetic zipper' model. Biochemistry 36:9200–9210.
44. Williams S, Causgrove TP, Gilmanshin R, Fang KS, Callender RH, et al. (1996) Fast events in protein folding: Helix melting and formation in a small peptide. Biochemistry 35:691–697.
45. Chakrabartty A, Ananthanarayanan VS, Hew CL (1989) Structure-function relationships in a winter flounder antifreeze polypeptide. J Biol Chem 264:11307–11312.
46. Steinmetz MO, Jelesarov I, Matousek WM, Honnappa S, Jahnke WA, et al. (2007) Molecular basis of coiled-coil formation. Proc Natl Acad Sci USA 104:7062–7067.
47. Honda S, Yamasaki K, Sawada Y, Morii H (2004) 10 residue folded peptide designed by segment statistics. Structure 12:1507–1518.
48. Pastor MT, López de la Paz M, Lacroix E, Serrano L, Pérez-Payá E (2002) Combinatorial approaches: a new tool to search for highly structured β-hairpin peptides. Proc Natl Acad Sci USA 99:614–619.
49. Blanco F, Rivas G, Serrano L (1994) A short linear peptide that folds into a native stable β-hairpin in aqueous solution. Nat Struct Biol 1:584–590.
50. Fesinmeyer RM, Hudson FM, Andersen NH (2004) Enhanced hairpin stability through loop design: the case of the protein G B1 domain hairpin. J Am Chem Soc 126:7238–7243.
51. Muñoz V, Thompson PA, Hofrichter J, Eaton WA (1997) Folding dynamics and mechanism of β-hairpin formation. Nature 390:196–199.

52. Cochran AG, Skelton NJ, Starovasnik MA (2001) Tryptophan zippers: stable monomeric β-hairpins. Proc Natl Acad Sci USA 98:5578–5583.
53. Yang WY, Pitera JW, Swope WC, Gruebele M (2004) Heterogeneous folding of the trpzip hairpin: full atom simulation and experiment. J Mol Biol 336:241–251.
54. Kortemme T, Ramírez-Alvarado M, Serrano L (1998) Design of a 20-amino acid, three-stranded β-sheet protein. Science 281:253–256.
55. López de la Paz M, Lacroix E, Ramírez-Alvarado M, Serrano L (2001) Computer-aided design of β-sheet peptides. J Mol Biol 312:229–246.
56. de Alba E, Santorio J, Rico M, Jimenez MA (1999) De novo design of a monomeric three-stranded antiparallel β-sheet. Protein Sci 8:854–865.
57. Marti DN, Bosshard HR (2004) Inverse electrostatic effect: electrostatic repulsion in the unfolded state stabilizes a leucine zipper. Biochemistry 43:12436–12447.
58. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302:1364.
59. Dantas G, Watters AL, Lunde BM, Eletr ZM, Isern NG, et al. (2006) Mis-translation of a computationally designed protein yields an exceptionally stable homodimer: Implications for protein engineering and evolution. J Mol Biol 362:1004–1024.
60. Johansson JS, Gibney BR, Skalicky JJ, Wand AJ, Dutton PL (1998) A native-like three-α-helix bundle protein from structure-based redesign: a novel maquette scaffold. J Am Chem Soc 120:3881–3886.
61. Dai QH, Thomas C, Fuentes EJ, Blomberg MRA, Dutton PL, et al. (2002) Structure of a de novo desined protein model of radical enzymes. J Am Chem Soc 124:10952–10953.
62. Jourdan M, Griffiths-Jones SR, Searle MS (2000) Folding of a β-hairpin peptide derived from the N-terminus of ubiquitin. conformational preferences of β-turn residues dictate non-native β-strand interactions. Eur Biophys J 267:3539–3548.

PAPER VI

Effects of mutations on the folding of the Alzheimer's Aβ42 peptide

Simon Mitternacht[1], Iskra Staneva[1], Torleif Härd[2] and Anders Irbäck[1]

[1]Computational Biology and Biological Physics, Department of Theoretical Physics, Lund University, Sweden. [2]Department of Molecular Biology, Swedish University of Agricultural Sciences (SLU), Biomedical Center, Uppsala, Sweden.

LU TP 09-04

We investigate the folding properties of the 42-residue amyloid-β peptide, Aβ42, by implicit solvent all-atom Monte Carlo simulations. In addition to the wild-type sequence, we study the three mutants F20E, E22G and E22G/I31E, which are known to have different aggregation properties. We find that all three mutations significantly alter the free-energy landscape in global properties like hydrophobic surface area and β-sheet content. The largest changes are found in the 17–32-residue segment, part of which forms a bend structure centered around residues 23–26. For the double mutant E22G/I31E, the population of this structure is markedly reduced. The E22G and F20E mutations have distinct but less drastic effects on this bend. The aggregation-accelerating E22G mutation makes the bend conformationally more diverse, which supports the previously proposed hypothesis that reduced stability of this bend correlates with increased aggregation propensity. We further find that the aggregation-decelerating F20E mutation has the opposite effect, thus suggesting an extension of this hypothesis.

Introduction

Aβ peptides are a main component of amyloid plaques in the brains of patients with Alzheimer's disease (AD). Understanding the formation and character of different Aβ aggregates, from oligomers to amyloid fibrils, and their roles in AD is currently the focus of intense research efforts.

Aβ is present in two main forms, Aβ40 and Aβ42, with 40 and 42 residues, respectively. Of these, Aβ42 is most strongly linked to AD. Aβ42 is more neurotoxic [1–3] and aggregates more rapidly into fibrils, protofibrils and oligomers [4–7].

Some familial early-onset forms of AD are associated with single amino acid mutations of Aβ. It is well-known from in vitro studies that such mutations can significantly alter the propensity of Aβ to aggregate. It is not obvious, however, how much one can learn from in vitro aggregation studies about the behavior of Aβ under complex in vivo conditions. This issue was recently addressed by studies of Drosophila flies expressing different Aβ42 variants [8, 9]. A clear correlation was found between a variant's in vitro aggregation rate and its influence on fly longevity and locomotion.

A first step toward a molecular understanding of Aβ aggregation is to characterize the Aβ monomer. The solution behavior of Aβ has been studied by both NMR [10–12] and atomic-level computer simulations [13–16]. The NMR results suggest that Aβ40 and Aβ42 both are largely unstructured in aqueous solution [10–12]. The main difference seems to be that Aβ42 is more rigid than Aβ40 at the C-terminus [12]. This conclusion is supported by explicit water molecular dynamics simulations at the microsecond time scale [15].

Another region of Aβ that might be crucial for its aggregation properties is the segment 21–30, which has been identified as a protease-resistant part of the molecule [17]. NMR [17] and computational [18–20] studies found the excised Aβ(21–30) fragment to adopt a bend structure in solution. It was further found that several familial AD mutations reduce the stability of this bend [21, 22].

A bend in this region has also been identified in fibrils of full-length Aβ [23, 24]. In the fibrils each monomer unit has two β strands, in Aβ40 connected by a bend at residues 25–29 [23], and in $^{35\mathrm{Mox}}$Aβ42 at residues 27–30 [24]. Furthermore, a recent NMR study determined the structure of Aβ40 bound to an affibody protein dimer and found a hairpin conformation with a turn spanning residues 24–29 [25]. In common for this structure and the two fibril models are two β-strands that include residues 17–21 and 31–36 and a bend in between.

In addition to Aβ(21–30), there are several other Aβ fragments that have been extensively studied, both experimentally and by computer simulations. Perhaps best studied is the strongly hydrophobic and fibril-forming [26] 7-residue fragment Aβ(16-22), which might play a driving role in the aggregation of full-length Aβ [27]. Residues 17–21 are usually called the central hydrophobic cluster (CHC).

Here we use implicit solvent all-atom Monte Carlo (MC) simulations to investigate the folding properties of four variants of full-length Aβ42, all of which were included in the above-mentioned Drosophila study [8]. The four variants studied are wild-type (WT), the two single mutants F20E and E22G, and the double mutant E22G/I31E. The E22G and F20E mutations are chosen because they have distinct and opposite effects on aggregation. The E22G mutation is associated with the familial Arctic form of AD [28] and enhances aggregation, whereas F20E Aβ42 shows a markedly reduced aggregation propensity compared to WT Aβ42 [8]. For E22G/I31E Aβ42, it is important to distinguish between global aggregation propensity and the propensity for forming protofibrils. Its global aggregation propensity was found to be almost as high as that of the E22G variant, whereas its propensity for forming protofibrils was found to be almost as low as that of the F20E variant [8].

Methods

Model. The model we use is an implicit solvent all-atom model with torsional degrees of freedom. It has the advantage of being computationally very fast compared to many other all-atom models. A detailed description of the model can be found elsewhere [29, 30]. Briefly, the interaction potential is composed of four major terms:

$$E = E_{\rm loc} + E_{\rm ev} + E_{\rm hb} + E_{\rm sc}\,.$$

The first term, $E_{\rm loc}$, contains local interactions between atoms separated by only a few covalent bonds. The other three terms are non-local in character: $E_{\rm ev}$ represents excluded-volume effects, $E_{\rm hb}$ is a hydrogen-bond potential, and $E_{\rm sc}$ contains residue-specific interactions between pairs of sidechains. The $E_{\rm sc}$ potential contains two components: effective hydrophobic attraction and interactions among charged sidechains.

As with any other existing model, given an arbitrary sequence with a native fold, there is no guarantee that this model will fold it to its proper structure. Characterizing an unstructured polypeptide chain like Aβ42 is a very different problem. Here the task is to identify relevant conformational (sub)ensembles

and find their relative populations, rather than to single out one particular conformation. Because of this difference, it is possible that some model details are less crucial when studying a natively unfolded protein. Other model properties might, instead, be more critical for such proteins. One fundamentally important and nontrivial requirement when studying a natively unfolded protein is that the model must not be biased toward either α-helix or β-sheet structure. The model above has previously been shown to capture both structural and thermal stability properties of a diverse set of peptides and small proteins [30], for one and the same set of model parameters, which makes it well suited for our study. To avoid introducing unphysical biases, we study Aβ42 using exactly the same model parameters as in the previous studies of other sequences [30].

The strong propensity of Aβ42 to aggregate makes its monomer properties difficult to determine experimentally. To validate a model by direct comparison with experimental data is therefore nontrivial. Nevertheless, such comparisons have been carried out with some success by Sgourakis et al. [15], Yang and Teplow [16], and Lam et al. [31]. It turns out that several important properties observed in our simulations agree well with the results of these studies (see below). It is worth noting that these groups used models completely different from ours; Sgourakis et al. used the OPLS force field [32] combined with the TIP3P explicit water model [33], Yang and Teplow used the Amber version PARM99SB [34] and the generalized Born implicit solvent model [35], and Lam et al. used a coarse-grained four-bead model.

Simulation techniques. We study the folding properties of the four Aβ42 variants by MC simulations of the above model. The simulations are performed using simulated tempering [36, 37]. The simulations cover six temperatures in the range 297–367 K.

Three different conformational updates are used in the simulations: single variable updates of sidechain and backbone angles, respectively, and Biased Gaussian Steps (BGS) [38]. The BGS move is semi-local and updates up to eight consecutive backbone degrees of freedom in a manner that keeps the ends of the segment approximately fixed. The ratio of sidechain to backbone updates is the same at all temperatures, whereas the relative frequency of the two backbone updates depends on the temperature. At high temperatures the single variable update is the only backbone update used, and at low temperatures only BGS is used. At intermediate temperatures both updates are used.

For each of our four Aβ42 variants, we performed a set of ten independent runs, using the open source C++-package PROFASI [39]. Each run contains $16 \cdot 10^9$ elementary MC steps and required $\sim$20 days of computing time on an Intel Xeon 5160 (3.0 GHz) processor. All runs are started from random initial conformations.

Analysis. In the simulations, we monitor several different properties. Secondary-structure content is measured using the program STRIDE [40]. The accessible hydrophobic surface area, HSA, is calculated using the algorithm by Shrake and Rupley [41] with 2000 test points. Atomic radii and probe radius are set according to Ooi et al. [42]. All carbon atoms are defined as nonpolar and all other heavy atoms as polar [43]. Hydrogen atoms are excluded from the HSA calculation. A contact between two residues, which are not nearest or next-nearest neighbors along the chain, is defined as follows. A pair of heavy atoms within 4.5 Å of each other, one from each residue, is said to provide a link between the residues. The residues are defined as being in contact if there exist at least two links between them.

Results

The low-energy conformations seen in our simulations are typically compact with a significant β-sheet content and little or no α-helix structure. Figure VI.1 shows the MC evolution of the energy in a representative run for WT Aβ42. During the course of the run, the system makes several independent visits to low-energy states. The figure also indicates how the temperature, a dynamical variable in simulated tempering, fluctuates during the run. For each sequence, ten trajectories of this length were generated.

Figure VI.2 shows secondary-structure profiles for all the four sequences at 310 K. The α-helix probability varies smoothly along the sequence and is small (< 0.15) everywhere, for all sequences. The β-sheet probability is significantly higher than the α-helix probability for most residues, but varies widely along the chain. One region of high β-sheet probability is residues 17–19, which is part of the CHC. Comparing the different sequences, we see that their β-sheet profiles share a similar overall shape. For example, all profiles show three sharp minima near residues 14, 25 and 37, respectively. These minima correspond to turn regions, as will be seen below. While the overall shape of the profiles is similar, some interesting differences can also be seen. The main differences between the four peptides are found in the 17–32-residue segment. Perhaps somewhat surprising are the effects of the mutation F20E.

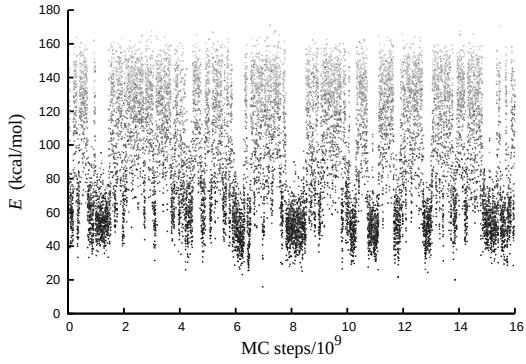

Figure VI.1: MC evolution of the energy E in a typical run for Aβ42 WT. In the simulated-tempering method, the temperature is a dynamical variable. To indicate how the temperature fluctuates during the course of the simulation, data points corresponding to lower temperatures have been made darker

Naively, one might expect the loss of the strongly hydrophobic F20 to have a destabilizing effect. In addition, glutamic acid has a high helix-forming propensity. Note, therefore, that we find that this mutation increases the β-sheet probability at several positions along the chain. At the positions 20, 21 and 26–29, the β-sheet probability is ~ 0.2 higher for F20E than for WT.

To elucidate the tertiary structure of the peptides, we construct contact maps, which indicate the probability of contact formation for all possible residue pairs (except nearest and next-nearest neighbors). Figure VI.3 shows our calculated contact maps for all the four sequences at 310 K. For WT, most of the frequent contacts are found within the four chain segments 4–12, 9–20, 17–32 and 28–42. Three major bend structures can be identified, centered around residues 13–16, 23–26 and 35–38, respectively. These structures are represented by bands extending perpendicularly from the main diagonal. The turn regions coincide with the three regions of low β-sheet probability mentioned above (see Figure VI.2).

In the contact map, as in the secondary-structure analysis, the main effects of the mutations are found in the 17–32-residue segment (see Figure VI.3). A major part of the frequent contacts within this segment are associated with the bend centered around residues 23–26. In addition, there is a smaller group of frequent contacts representing interactions between the CHC and residues 22–24. Several contacts in this second group are weakened by all the three mutations studied; examples of this are the (18,23) and (17,24) contacts.

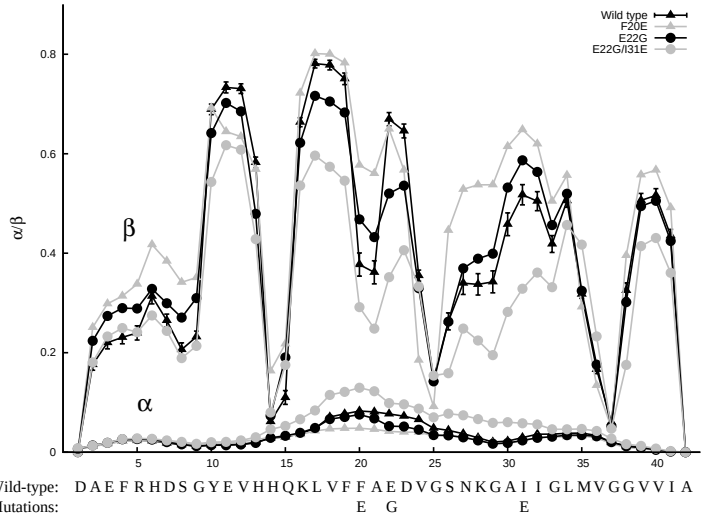

Figure VI.2: Secondary-structure profiles for WT Aβ42 and the three mutants F20E, E22G and E22G/I31E, at 310 K. The probabilities of each residue to participate in α and β secondary structure were measured using STRIDE [40]. The WT amino acid sequence and the positions of the different mutations studied are indicated along the x-axis. For clarity, statistical errors are shown only for the WT β-sheet profile. The error bars are similar for the other three sequences. The statistical errors on the α-helix probabilities are smaller than the plot symbols.

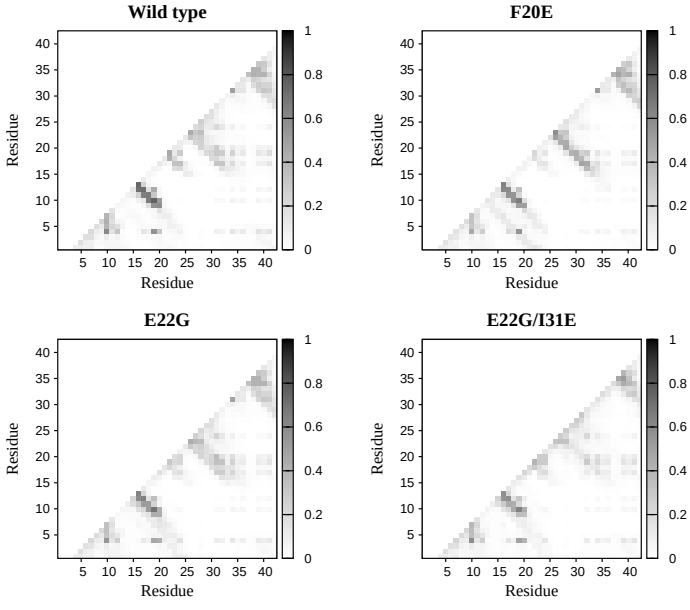

Figure VI.3: Probability map of residue contact formation for WT Aβ42 and the three mutants F20E, E22G and E22G/I31E, at 310 K.

The bend region, on the other hand, is affected differently by the different mutations. The F20E mutation makes this band of frequent contacts more narrow, and therefore reduces the conformational diversity of the bend. This finding is consistent with the above observation of an increased β-sheet probability in this region for this mutant. The E22G mutation instead increases the conformational diversity of the bend. The E22G/I31E mutation makes many contacts associated with the bend much weaker, and therefore has a strongly destabilizing effect on this structure.

The mutations influence not only the secondary-structure profiles and the contact map, but also the free energy $F(\mathrm{HSA},\beta)$, calculated as a function of hydrophobic surface area, HSA, and β-sheet content, β. Figure VI.4 shows this free energy for all the four sequences, at the same temperature as before (310 K). For WT, there are two shallow but discernable free-energy minima of similar depth. One of the minima has higher HSA and higher β and corresponds to structures with single, extended β-sheets. The other minimum is of lower HSA and β and the structures are more compact, often with two layers shielding a hydrophobic core. In the figure, examples of WT structures from the two minima are shown. For the mutant peptides, the structures

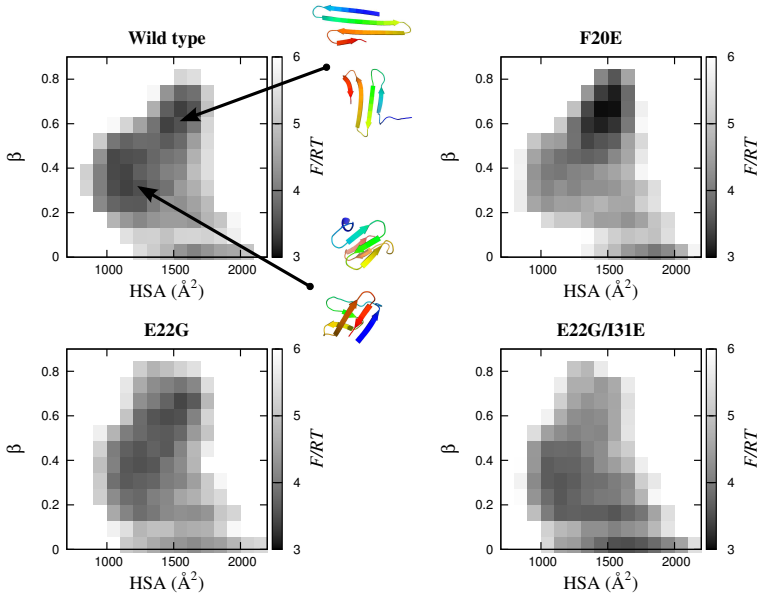

Figure VI.4: The free energy $F(\text{HSA}, \beta)$ calculated as a function of hydrophobic surface area, HSA, and β-sheet content, β, for WT Aβ42 and the three mutants F20E, E22G and E22G/I31E, at 310 K. Two representative conformations (drawn with PyMOL [44]) are shown for each of the two free-energy minima for WT Aβ42.

found in these two regions of the (HSA,β)-plane are similar to those of WT – the difference lies in the balance between the minima. The F20E mutation increases the population of the high-HSA state, at the expense of the low-HSA state. The E22G/I31E mutation instead favors the state with lower HSA and lower β. The E22G mutation only marginally changes the relative population of the two states, but lowers the barrier between them. For all four sequences, there is also a horizontal band of significantly populated low-β states.

Discussion

Our simulations suggest that these four Aβ variants form a variety of compact and β-sheet rich structures. Two major types of conformations can be identified. One corresponds to single β-sheets and the other to more compact structures, typically with two layers. Within each of these two families, large variations occur in the detailed structure. Nevertheless, three major bend

regions can be identified, centered around residues 13–16, 23–26 and 35–38, respectively. The first of these bends gives rise to the most frequent contacts (see Figure VI.3), indicating a relatively well defined structure. The C-terminal bend is, by contrast, conformationally diverse. The character of the central bend, around residues 23–26, varies among the four mutants.

Comparing our results with previous simulations for WT Aβ42, one finds both differences and similarities. Our overall β-sheet content is higher than what Sgourakis et al. [15] and Yang and Teplow [16] found, but comparable to the results of Lam et al. [31]. The α-helix content was small in these simulations and is small in ours, too. Interestingly, although based on completely different models, all these studies found turns in the 23–26-residue region, where we find one of our three major turns. Sgourakis et al. actually observed turns in all our three major turn regions.

Our study of the three mutants suggest that the strongest effects of these mutations are found in the 17–32-residue segment. The most prominent structural feature in this region is the bend centered around residues 23–26. We also observe a group of frequent contacts representing interactions between the CHC and residues 22–24. Several of these interactions are more common in the WT peptide than in the three variants. This picture is consistent with recent results obtained by Baumketner et al. [45]. These authors studied the effects of the E22Q mutation, associated with familial Dutch AD, on Aβ(15–28), by explicit solvent all-atom simulations. The mutation was found to weaken the interactions between the CHC and a bend spanning residues 22–28.

The response of the bend centered around residues 23–26 to the mutations is intriguing, because we find that the F20E and E22G mutations have opposite effects on this structure. The E22G mutation increases the conformational diversity of the bend, which probably can be attributed to the high intrinsic flexibility of glycine and the loss of a favorable electrostatic interaction between E22 and K28. That the F20E mutation would decrease the conformational diversity in this region is less easy to anticipate, as discussed above. However, it is conceivable, especially since F20 belongs to the CHC, that this mutation makes the system less frustrated by reducing the number of energetically favorable conformations and thereby the conformational diversity. This behavior is what we observe.

In the above-mentioned Drosophila study [8], the mutations F20E and E22G were found to increase and decrease, respectively, fly longevity and locomotion. It is further known that the mutations F20E and E22G decelerate and accelerate, respectively, the aggregation of Aβ42 [8]. These different aggregation properties can probably be explained, at least partially, in terms of

primary structure alone; the mutation F20E introduces a charged residue into the CHC region, whereas the mutation E22G removes a charged residue near the CHC. Do the mutations have additional effects on the aggregation propensity, stemming from altered folding properties? A study of Grant et al. [21] suggests that this might be the case. These authors experimentally investigated the effects of several disease-associated mutations on Aβ(21–30), and found that they all destabilized the bend structure that this fragment adopts in solution [17]. It was further found that the destabilizing effect of a mutation correlated with its influence on full-length Aβ oligomerization propensity [21]. Our finding that E22G Aβ42 shows an increased conformational diversity in this region supports this hypothesis. Interestingly, the same picture seems to hold for the F20E mutation as well. Experiments found that this mutation decelerates aggregation, and we find that it reduces the conformational diversity in this region.

The strong destabilization of the bend that we observe for E22G/I31E Aβ42 does not conform with the above picture, because E22G/I31E Aβ42 does not aggregate faster than E22G Aβ42 [8]. It is possible that this double mutation affects the folding properties too much for this simple picture to remain valid. The fact that E22G/I31E Aβ42 has been found to have a high global aggregation propensity but only a low propensity for forming protofibrils [8] may indicate that the aggregation mechanisms indeed are different for this variant.

In Aβ42 fibrils, as mentioned above, each Aβ42 molecule is believed to participate in two tightly packed β-sheets [24]. The turn separating the two strand regions is overlapping with or near residues 23–26 [24, 46, 47], where a turn can be seen in both our and previous simulations of the Aβ42 monomer. However, the precise shape of the turn need not be the same in fibrils as in the monomer, and the turn might thus have to be reorganized upon fibril formation. Increased conformational diversity in this region should facilitate such a reorganization, which provides a possible explanation of how destabilization might lead to a higher aggregation propensity [21]. It is also worth noting that a bend centered at residues 23–26 may partially protect the CHC from the solvent. Increased conformational diversity of the bend could speed up aggregation by increasing the solvent exposure of the CHC.

Finally, it is interesting to compare our results with those of Cheon et al. [48], who simulated oligomerization for two Aβ fragments, the hydrophobic Aβ(16–22) and the less hydrophobic Aβ(25–35). Aβ(16–22) rapidly formed disordered oligomers, which were subsequently converted into ordered oligomers through a reorganization process. By contrast, Aβ(25–35) formed ordered

oligomers directly, in a one-step process. The difference found between these peptides is somewhat reminiscent of the difference we find in folding properties between the E22G and F20E Aβ42 variants. The less hydrophobic systems, Aβ(25–35) and F20E Aβ42, seem to share a less frustrated behavior.

Conclusion

We have investigated the folding properties of four Aβ42 variants with different aggregation properties: WT, F20E, E22G and E22G/I31E. We find that the three mutations in particular have a significant impact on a bend centered around residues 23–26. Bends in this region have been observed in previous WT Aβ42 simulations [15, 16, 31]. Our study suggests that the E22G/I31E mutation, which leads to heterogeneous aggregation properties [8], strongly destabilizes this bend. We further find that the aggregation-accelerating E22G mutation increases the conformational diversity of the bend. This finding supports the previously proposed hypothesis that reduced stability in this region accelerates aggregation [21]. Our study of the slowly aggregating F20E variant provides additional support for this picture, by suggesting that this mutation instead decreases the conformational diversity of the bend. That a mutation that slows down rather than speeds up aggregation may have this reverse effect on the Aβ42 monomer remains, as far as we know, to be verified experimentally.

Acknowledgment. This work was in part supported by the Swedish Research Council.

References

1. Selkoe DJ (1999) Translating cell biology into therapeutic advances in Alzheimer's disease. Nature 399:A23–A31.
2. Zhang Y, McLaughlin R, Goodyer C, LeBlanc A (2002) Selective cytotoxicity of intracellular amyloid β peptide$_{1-42}$ through p53 and Bax in cultured primary human neurons. J Cell Biol 156:519–529.
3. Dahlgren KN, Manelli AM, Stine Jr WB, Baker LK, Krafft GA, et al. (2002) Oligomeric and fibrillar species of amyloid-β peptides differentially affect neuronal viability. J Biol Chem 277:32046–32053.
4. Jarrett JT, Berger EP, Lansbury Jr PT (1993) The carboxy terminus of the β amyloid protein is critical for the seeding of amyloid formation: implications for the pathogenesis of Alzheimer's disease. Biochemistry 32:4693–4697.
5. Harper JD, Wong SS, Lieber CM, Lansbury PT (1997) Observation of metastable Aβ amyloid protofibrils by atomic force microscopy. Chem Biol 4:119–125.

6. Walsh DM, Lomakin A, Benedek GB, Condron MM, Teplow DB (1997) Amyloid β-protein fibrillogenesis. Detection of a protofibrillar intermediate. J Biol Chem 272:22364–22372.
7. Bitan G, Kirkitadze MD, Lomakin A, Vollers SS, Benedek GB, et al. (2003) Amyloid-β protein (Aβ) assembly: Aβ40 and Aβ42 oligomerize through different pathways. Proc Natl Acad Sci USA 100:330–335.
8. Luheshi LM, Tartaglia GG, Brorsson A, Pawar AP, Watson IE, et al. (2007) Systematic in vivo analysis of the intrinsic determinants of amyloid β pathogenicity. PLoS Biol 5:e290.
9. Iijima K, Chiang HC, Hearn SA, Hakker I, Gatt A, et al. (2008) Aβ42 mutants with different aggregation profiles induce distinct pathologies in Drosophila. PLoS One 3:e1703.
10. Riek R, Güntert P, Döbeli H, Wipf B, Wüthrich K (2001) NMR studies in aqueous solution fail to identify significant conformational differences between the monomeric forms of two Alzheimer peptides with widely different plaque-competence, Aβ40^{ox} and Aβ42^{ox}. Eur J Biochem 268:5930–5936.
11. Hou L, Shao H, Zhang Y, Li H, Menon NK, et al. (2004) Solution NMR studies of the Aβ40 and Aβ42 peptides establish that the Met35 oxidation state affects the mechanism of amyloid formation. J Am Chem Soc 126:1992–2005.
12. Yan Y, Wang C (2006) Aβ42 is more rigid than Aβ40 at the C terminus: implications for Aβ aggregation and toxicity. J Mol Biol 364:853–862.
13. Baumketner A, Bernstein SL, Wyttenbach T, Bitan G, Teplow DB, et al. (2006) Amyloid β-protein monomer structure: a computational and experimental study. Protein Sci 15:420–428.
14. Flöck D, Colacino S, Colombo G, di Nola A (2006) Misfolding of the amyloid β-protein: a molecular dynamics study. Proteins 62:183–192.
15. Sgourakis NG, Yan Y, McCallum SA, Wang C, García AE (2007) The Alzheimer's peptides Aβ40 and 42 adopt distinct conformations in water: a combined MD/NMR study. J Mol Biol 368:1448–1457.
16. Yang M, Teplow DB (2008) Amyloid β-protein monomer folding: free-energy surfaces reveal alloform-specific differences. J Mol Biol 384:450–464.
17. Lazo ND, Grant MA, Condron MC, Rigby AC, Teplow DB (2005) On the nucleation of amyloid β-protein monomer folding. Protein Sci 14:1581–1596.
18. Cruz L, Urbanc B, Borreguero JM, Lazo ND, Teplow DB, et al. (2005) Solvent and mutation effects on the nucleation of amyloid β-protein folding. Proc Natl Acad Sci USA 102:18258–18263.
19. Baumketner A, Bernstein SL, Wyttenbach T, Lazo ND, Teplow DB, et al. (2006) Structure of the 21–30 fragment of amyloid β-protein. Protein Sci 15:1239–1247.
20. Chen W, Mousseau N, Derreumaux P (2006) The conformations of the amyloid-β (21–30) fragment can be described by three families in solution. J Chem Phys 125:084911.
21. Grant MA, Lazo ND, Lomakin A, Condron MM, Arai H, et al. (2007) Familial Alzheimer's disease mutations alter the stability of the amyloid β-protein monomer folding nucleus. Proc Natl Acad Sci USA 104:16522–16527.
22. Krone MG, Baumketner A, Bernstein SL, Wyttenbach T, Lazo ND, et al. (2008) Effects of familial Alzheimer's disease mutations on the folding nucleation of the amyloid β-protein. J Mol Biol 381:221–228.

23. Petkova AT, Ishii Y, Balbach JJ, Antzutkin ON, Leapman RD, et al. (2002) A structural model for Alzheimer's β-amyloid fibrils based on experimental constraints from solid state NMR. Proc Natl Acad Sci USA 99:16742–16747.
24. Lührs T, Ritter C, Adrian M, Riek-Loher D, Bohrmann B, et al. (2005) 3D structure of Alzheimer's amyloid-β(1-42) fibrils. Proc Natl Acad Sci USA 102:17342–17347.
25. Hoyer W, Grönwall C, Jonsson A, Ståhl S, Härd T (2008) Stabilization of a β-hairpin in monomeric Alzheimer's amyloid-β peptide inhibits amyloid formation. Proc Natl Acad Sci USA 105:5099–5104.
26. Balbach JJ, Ishii Y, Antzutkin ON, Leapman RD, Rizzo NW, et al. (2000) Amyloid fibril formation by Aβ(16-22), a seven-residue fragment of the Alzheimer's β-amyloid peptide, and structural characterization by solid state NMR. Biochemistry 39:13748–13759.
27. Tjernberg LO, Näslund J, Lindqvist F, Johansson J, Karlström AR, et al. (1996) Arrest of β-amyloid fibril formation by pentapeptide ligand. J Biol Chem 271:8545–8548.
28. Nilsberth C, Westlind-Danielsson A, Eckman CB, Condron MM, Axelman K, et al. (2001) The 'arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced Aβ protofibril formation. Nat Neurosci 4:887–893.
29. Irbäck A, Mohanty S (2005) Folding thermodynamics of peptides. Biophys J 88:1560–1569.
30. Irbäck A, Mitternacht S, Mohanty S (2009) An effective all-atom potential for proteins. Submitted manuscript .
31. Lam AR, Teplow DB, Stanley HE, Urbanc B (2008) Effects of the arctic ($E^{22} \rightarrow G$) mutation on amyloid β-protein folding: discrete molecular dynamics study. J Am Chem Soc 130:17413–17422.
32. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105:6474–6487.
33. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935.
34. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, et al. (2006) Comparison of multiple Amber force fields and development of improved backbone parameters. Proteins 65:712–725.
35. Mongan J, Case DA, McCammon JA (2004) Constant pH molecular dynamics in generalized Born implicit solvent. J Comput Chem 25:2038–2048.
36. Lyubartsev AP, Martsinovski AA, Shevkunov SV, Vorontsov-Velyaminov PN (1992) New approch to Monte Carlo calculation of the free energy: Method of expanded ensembles. J Chem Phys 96:1776–1783.
37. Marinari E, Parisi G (1992) Simulated tempering: a new Monte Carlo scheme. Europhys Lett 19:451–458.
38. Favrin G, Irbäck A, Sjunnesson F (2001) Monte Carlo update for chain molecules: biased Gaussian steps in torsional space. J Chem Phys 114:8154–8158.
39. Irbäck A, Mohanty S (2006) PROFASI: a Monte Carlo simulation package for protein folding and aggregation. J Comput Chem 27:1548–1555.
40. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23:566–579.

41. Shrake A, Rupley JA (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol 79:351–371.
42. Ooi T, Oobatake M, Némethy G, Scheraga HA (1987) Accessiable surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc Natl Acad Sci USA 84:3086–3090.
43. Miller S, Janin J, Lesk AM, Chothia C (1987) Interior surface of monomeric proteins. J Mol Biol 196:641–656.
44. DeLano WL (2002). The PyMOL molecular graphics system. San Carlos, CA: DeLano Scientific.
45. Baumketner A, Krone MG, Shea JE (2008) Role of familial dutch mutation E22Q in the folding and aggregation of the 15–28 fragment of the alzheimer amyloid-β protein. Proc Natl Acad Sci USA 105:6027–6032.
46. Olofsson A, Sauer-Eriksson AE, Öhman A (2006) The solvent protection of Alzheimer amyloid-β-(1–42) fibrils as determined by solution NMR spectroscopy. J Biol Chem 281:477–483.
47. Masuda Y, Uemura S, Ohashi R, Nakanishi A, Takegoshi K, et al. (2009) Identification of physiological and toxic conformations in Aβ aggregates. Comput Biol Chem 10:287–295.
48. Cheon M, Chang I, Mohanty S, Luheshi LM, Dobson CM, et al. (2007) Structural reorganisation and potential toxicity of oligomeric species formed during the assembly of amyloid fibrils. PLoS Comp Biol 3:e173.

2	5			3		9		1
	1				4			
4		7				2		8
		5	2					
				9	8	1		
	4				3			
			3	6			7	2
	7							3
9		3				6		

3					5			8
4	8				1	6		
		2						
		4		9			2	
	3	7				8	9	
	5			6		4		
						2		
		9	5				8	7
5			8					3