UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 724

Protein Folding and DNA Origami

MARK MARVIN SEIBERT

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2010

ISSN 1651-6214
ISBN 978-91-554-7756-1
urn:nbn:se:uu:diva-121549

Dissertation presented at Uppsala University to be publicly examined in B21, Husargatan 3, 751 24 Uppsala, BMC, Tuesday, April 20, 2010 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract
Seibert, M M. 2010. Protein Folding and DNA Origami. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 724. 43 pp. Uppsala. ISBN 978-91-554-7756-1.

In this thesis, the folding process of the de novo designed polypeptide chignolin was elucidated through atomic-scale Molecular Dynamics (MD) computer simulations. In a series of long timescale and replica exchange MD simulations, chignolin's folding and unfolding was observed numerous times and the native state was identified from the computed Gibbs free-energy landscape. The rate of the self-assembly process was predicted from the replica exchange data through a novel algorithm and the structural fluctuations of an enzyme, lysozyme, were analyzed.

DNA's structural flexibility was investigated through experimental structure determination methods in the liquid and gas phase. DNA nanostructures could be maintained in a flat geometry when attached to an electrostatically charged, atomically flat surface and imaged in solution with an Atomic Force Microscope. Free in solution under otherwise identical conditions, the origami exhibited substantial compaction, as revealed by small angle X-ray scattering. This condensation was even more extensive in the gas phase.

Protein folding is highly reproducible. It can rapidly lead to a stable state, which undergoes moderate fluctuations, at least for small structures. DNA maintains extensive structural flexibility, even when folded into large DNA origami.

One may reflect upon the functional roles of proteins and DNA as a consequence of their atomic-level structural flexibility. DNA, biology's information carrier, is very flexible and malleable, adopting to ever new conformations. Proteins, nature's machines, faithfully adopt highly reproducible shapes to perform life's functions robotically.

Keywords: protein folding, Molecular Dynamics simulations, DNA origami

Mark Marvin Seibert, Molecular biophysics, Box 596, Uppsala University, SE-75124 Uppsala, Sweden

ISSN 1651-6214
ISBN 978-91-554-7756-1
urn:nbn:se:uu:diva-121549 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-121549)

Dedicated to those who made it possible

List of Papers

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

I **Reproducible polypeptide folding and structure prediction using Molecular Dynamics simulations.**
Seibert, M.M., Patriksson, A., Hess, B., van der Spoel
J. Mol. Biol., 354:173–183 (2005)

II **Protein folding kinetics and thermodynamics from atomistic simulations.**
van der Spoel, D., Seibert, M.M.
Phys. Rev. Letters, 96:238102 (2006)

III **On the precision of solvent-accessible surface areas.**
Novotny, M., Seibert, M., Kleywegt, G.J.
Acta Cryst. D, 63:270-274 (2007)

IV **Flexibility of scaffolded DNA origami structures.**
Seibert, M.M., Svenda, M., Bogan, M.J., Grossmann, G., Benner, W.H., Göthelid, E., Maia, F., Chapman, H.N., van der Spoel D., Hajdu, J.
Manuscript

List of additional publications:

V **Femtosecond diffractive imaging with a soft-X-ray free-electron laser.**
Chapman HN, Barty A, Bogan MJ, Boutet S, Frank M, Hau-Riege SP, Marchesini S, Woods BW, Bajt S, Benner WH, London RA, Plönjes E, Kuhlmann M, Treusch R, Düsterer S, Tschentscher T, Schneider JR, Spiller E, Möller T, Bostedt C, Hoener M, Shapiro DA, Hodgson KO, van der Spoel D, Burmeister F, Bergh M, Caleman C, Huldt G, Seibert MM, Maia FRNC, Lee RW, Szöke A, Timneanu N, Hajdu J.
Nature Physics (2) 839-843 (2006)

VI **Femtosecond Time-Delay X-ray Holography.**
Chapman HN, Hau-Riege SP, Bogan MJ, Bajt S, Barty S, Boutet S, Marchesini S, Frank M, Woods BW, Benner WH, London RA, Rohner U, Szöke A, Spiller E, Möller T, Bostedt C, Shapiro DA, Kuhlmann M, Treusch R, Plönjes E, Burmeister F, Bergh M, Caleman C, Huldt G, Seibert MM, Hajdu J.
Nature 448:676-679 (2007)

VII **Protein folding properties from molecular dynamics simulations.**
van der Spoel D, Patriksson A, Seibert MM.
Lecture Notes in Computer Science. 4699: 109-115 (2007)

VIII **Single particle X-ray diffractive imaging.**
Bogan MJ, Benner WH, Boutet S, Rohner U, Frank M, Barty A, Seibert MM, Maia F, Marchesini S, Bajt S, Woods B, Riot V, Hau-Riege SP, Svenda M, Marklund E, Spiller E, Hajdu J, Chapman HN.
Nano Letters 8(1): 310-316. (2008)

IX **Ultrafast single-shot diffraction imaging of nanoscale dynamics.**
Barty A, Boutet S, Bogan MJ, Hau-Riege S, Marchesini S, Sokolowski-Tinten K, Stojanovic N, Tobey R, Ehrke H, Cavalleri A, Duesterer S, Frank M, Bajt S, Woods BW, Seibert MM, Hajdu J, Treusch R, Chapman HN.
Nature Photonics 2(7): 415-419. (2008)

X **Massively parallel X-ray holography.**
Marchesini S, Boutet S, Sakdinawat AE, Bogan MJ, Bajt S, Barty A, Chapman HN, Frank M, Hau-Riege SP, Szoke A, Cui C, Shapiro DA, Howells MR, Spence JCH, Shaevitz JW, Lee JY, Hajdu J and Seibert MM.
Nature Photonics 2(9): 560-563. (2008)

XI **Ultrafast soft X-ray scattering and reference-enhanced diffractive imaging of weakly scattering nanoparticles.**
Boutet S, Bogan MJ, Barty A, Frank M, Benner WH, Marchesini S, Seibert MM, Hajdu J, Chapman HN.
J. Elec. Spectr. 166-167: 65-73 (2008)

XII **Aerosol Imaging with a Soft X-Ray Free Electron Laser.**
Bogan MJ, Boutet S, Chapman HN, Marchesini S, Barty A, Benner WH, Rohner U, Frank M, Hau-Riege SP, Bajt S, Woods B, Seibert MM, Iwan B, Timneanu N, Hajdu J, Schulz J. *Aerosol Sci. Tech.*, 44: 3, i — vi (2010)

XIII **Sacrificial Tamper Slows Down Sample Explosion in FLASH Diffraction Experiments.**
Hau-Riege SP, Boutet S, Barty A, Bajt S, Bogan MJ, Frank M, Andreasson J, Iwan B, Seibert MM, Hajdu J, Sakdinawat A, Schulz J, Treusch R, Chapman HN.
Phys. Rev. Let. in press (2010)

Contents

Abbreviations

AFM	Atomic Force Microscope
ASA	solvent-accessible surface area
CPC	Condensation Particle Counter
DMA	Differential Mobility Analyzer
EM	Electron Microscopy
GB	Generalized Born
GEMMA	Gas-phase electrophoretic molecular mobility analysis
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
OPLS	Optimized Potentials for Liquid Simulations
pdb	protein data bank
PME	Particle Mesh Ewald
REMD	Replica exchange Molecular Dynamics
rmsd	root-mean-square deviation
SA	surface area
SAXS	Small angle X-ray scattering
SAXS	Small-angle X-ray scattering
ssDNA	single stranded DNA
TEM	Transmission Electron Microscopy
<V>	average NOE violations
WC	Watson-Crick
3D	three-dimensional

Introduction

Nucleic acids and proteins are two of the most important molecules in biology. The simplest biological entities, viruses, can consist exclusively of these two. Both are linear polymers of heterogeneous building blocks and the combinatorial possibilities of their feasible assemblies quickly reach astronomical dimensions. The resulting molecular diversity enables both the range of biological creatures and functions on earth, and the individual genetic variation of each organism.

Revealing the three-dimensional structure of DNA and thousands of proteins has been one of the most fundamental contemporary scientific accomplishments. Our understanding of biological structure at the atomic level has become so extensive that we can begin to apply this knowledge to build, *de novo*, sequences of protein and DNA that form predictable structures.

Both DNA and proteins undergo a spontaneous self-assembly or folding process. Long double-stranded nucleotide chains coil themselves into double helices and polypeptide chains fold into compact, reproducible and functional structures. The energy of the secondary interactions that stabilize the three-dimensional structure of a macromolecule is commensurate with the thermal energy under physiological conditions, and as a consequence, macromolecular structures are dynamic and fluctuate between conformers (Linderstrøm-Lang et al. 1959). Accordingly, the folding process is frequently reversible both *in vivo* and *in vitro*.

Protein folds are uniquely defined by their underlying amino acid sequences, whereas any polynucleotide sequence can form the famous double helix, given a complementary sequence with which Watson-Crick base-pairs can be made. Nevertheless, the specific hydrogen bonding between complementary Watson-Crick base-pairs can also give rise to complex secondary structure based on the arrangement of double helical segments. This has become the basis for the molecular art of folding DNA to make nanostructures, known as DNA origami.

Designed polypeptide and polynucleotide sequences present the opportunity to explore the folding process that organizes these linear polymers into molecular architecture.

Protein folding is highly reproducible. It can rapidly lead to a stable state, which undergoes moderate fluctuations, at least for small structures. DNA maintains extensive structural flexibility, even when folded into large DNA origami.

One may reflect upon the functional roles of proteins and DNA as a consequence of their atomic-level structural flexibility. DNA, biology's information carrier, is very flexible and malleable, adopting to ever new conformations. Proteins, nature's machines, faithfully adopt highly reproducible shapes to perform life's functions robotically.

Subject of this thesis

In this thesis, the folding process of the *de novo* designed polypeptide chignolin was elucidated through atomic-scale Molecular Dynamics (MD) computer simulations. In a series of long timescale and replica exchange MD simulations, chignolin's folding and unfolding was observed numerous times and the native state was identified from the computed Gibbs free-energy landscape. The rate of the self-assembly process was predicted from the replica exchange data through a novel algorithm and the structural fluctuations of an enzyme, lysozyme, were analyzed.

DNA's structural flexibility was investigated through experimental structure determination methods in the liquid and gas phase. DNA nanostructures could be maintained in a flat geometry when attached to an electrostatically charged, atomically flat surface and imaged in solution with an Atomic Force Microscope. Free in solution under otherwise identical conditions, the origami exhibited substantial compaction, as revealed by small angle X-ray scattering. This condensation was even more extensive in the gas phase.

Protein Folding

Definition

The self-assembly process that turns a polypeptide chain of a foldable sequence into a stable three-dimensional conformation is called protein folding.

Anfinsen's Dogma

An organism's genome is frequently referred to as the 'blueprint' for that organism. Like an engineering drawing, it contains the 3D structural information for a complex entity. Unlike engineered designs, genomes are not accompanied by a set of detailed assembly instructions. Biology instead relies on a self-assembly process to create structure out of parts. Protein folding is perhaps the most prominent example of biological self-assembly. After being synthesized as linear polypeptide chains by the ribosome and released into solution, proteins obtain their predetermined shape and function, frequently without the involvement of other molecules. Anfinsen expressed the hypothesis that

> "… the particular conformation that a protein assumes, under any set of specific conditions, is the one that is thermodynamically most stable." (Epstein et al. 1963)

Through a series of elegant experiments that demonstrated how a small, soluble protein can be reversibly denatured *in vitro*, followed by spontaneous refolding, this thermodynamic hypothesis of protein folding became established as Anfinsen's Dogma. Summarized in his Nobel lecture (Anfinsen 1973) it states that the self-arrangement process that a linear polypeptide chain undergoes in order to assume the compact, functional, three-dimensional shape that defines a native protein is a spontaneous, energetically favorable transition that is defined exclusively by the intrinsic properties of the polypeptide chain, namely its sequence. Sequence defines structure.

Levinthal's Paradox

The configurations that a protein can adopt can theoretically be thought of as the set of all physically attainable dihedral angles along the peptide backbone and within each amino acid's sidechain. Such combinatorial configurations naturally lead to an exorbitant number of conceivable states. Even if one discretizes each dihedral angle coarsely to only a few fixed rotamer positions, the exponential growth of the possible combinations gives rises to an accessible configurational landscape of near-infinite expanse. Considering just four distinct values for each phi and psi backbone dihedral angle and sidechain rotation for a short polypeptide of 50 residues leads to approximately 10^{90} different hypothetical conformations. There is not enough time in the lifespan of any organism for a protein to explore its entire configurational landscape before settling into its native conformation. The phenomenon that proteins nevertheless select their native state rapidly, discarding a practically infinite number of alternative conformations without having had an opportunity to sample them is known as Levinthal's paradox (Levinthal 1968). Protein folding cannot proceed as a random search through conformational landscapes.

The genetic code, once transcribed and translated into an amino acid sequence must not only define the final protein structure, but also provide navigable assembly routes along which a nascent chain will autonomously fold into a functional protein, e.g. as it is slowly being synthesized on the ribosome with one residue added at a time. How is this route traversed, e.g. in denaturation/renaturation experiments and the assembly achieved? Cyrus Levinthal asked: "Are there pathways to protein folding?"

Molecular Dynamics

> "Certainly no subject or field is making more progress on so many fronts at the present moment, than biology, and if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that *all things are made of atoms* and everything that living things do can be understood in terms of the jigglings and wigglings of atoms." – Richard P. Feynman (Feynman 1963)

Richard Feynman's famous words may serve as the inspirational foundation for applying Molecular Dynamics (MD) techniques to the study of biomolecules. The physical foundation upon which MD rests is provided by Isaac Newton's laws of motion. Although Newton may have been more concerned with the paths taken by celestial bodies (or falling apples), the laws he established to describe the motion of astronomical objects are equally applicable at the other extreme of the observable size range. On the nanoscale, they can

be used to calculate and predict the motion of atoms and molecules. By computing the forces acting between all atoms and displacing each particle accordingly in a stepwise fashion, a MD simulation generates a trajectory for each particle in the simulation. The four sources of forces that are generally considered in an atomic-scale MD simulation are:

1. The force transmitted along a bond, where bond lengths are constrained to a fixed value (bond lengthening and shortening vibrations happen on time- and size scales shorter than most iterative MD time-steps and are consequently ignored).
2. The periodic forces associated with rotation around a single bond.
3. Forces resulting due to contact of two (or more) atoms' electron clouds as defined by their van der Waals radii.
4. Electrostatic attraction or repulsion forces that arise from charged atoms.

These forces are described mathematically as potential functions, those employed here are from the optimized potentials for liquid simulations (Kaminski et al. 2001; Jorgensen et al. 2005).

In principle, every single particle in a system can exert a force on every other particle. Therefore, in order to calculate the net force acting upon any one atom, it has to be paired with every other atom and the force resulting from all pairwise interactions need to be determined. This leads to a computational load that scales with the square of the number of particles. Fortunately, the forces transmitted along bonds and van der Waals interactions are of relevant magnitude only within a short range from the interacting particle. Their computation can therefore be truncated at some distance from each particle. Electrostatic forces do not drop off to negligible values within a short interaction radius. In order to reduce the computational cost and prevent the problem from growing as the system size squared, these forces can be calculated with a grid based Particle Mesh Ewald (PME) algorithm (Darden et al. 1993; Essmann et al. 1995). Briefly, particle charges are distributed over nearby grid positions, the grid is Fourier transformed, the long range Coulomb interaction between particles is computed in Fourier space and reverse Fourier transformed to obtain real-space forces. This algorithm scales as O(n log n) where n is the number of particles in the simulation.

The parameters for the potential functions are derived from a combination of first principles, quantum chemical simulations and empirically obtained values and bundled in a force field. Force fields have improved substantially in recent years, reaching a level of accuracy that is sufficient for predictive simulations in many cases. This achievement, in combination with the highly efficient and sophisticated MD software GROMACS (Lindahl et al. 2001; Van der Spoel et al. 2005; Hess et al. 2008) and the ever-increasing perfor-

mance of computational hardware has made the simulations presented here possible.

Replica Exchange Molecular Dynamics

The biggest gap between MD models and biophysical reality exists in the accessible timescales and ensemble sizes. Biophysical experiments are increasingly performed on single molecules, but even with the fastest hard- and software available, the simulation of atomic-scale motion is far slower than movement in reality. Current MD simulations generate trajectories at a rate equivalent to a slow-motion movie playing with a speed approximately 10-15 orders of magnitude slower than reality. Even if this speed doubles every year for the next 30 years, simulations will not reach parity with real time. Therefore, no efforts are spared in making the best possible use of limited simulation time. In addition to speeding up the simulations, one may attempt to accelerate the process being simulated. Replica Exchange MD (REMD) is a technique that exploits the physics of Brownian motion to achieve this: Warmer atoms jiggle and wiggle faster. In any given (simulation) time warmer atoms thus explore a larger range of their available conformation space, which itself increases as higher energy state become accessible at higher temperatures. The downside to increased thermal motion is reduced stability. Proteins unfold at elevated temperatures because their thermal energy exceeds the strengths of the interatomic interactions that constrain structure. A warm polypeptide chain is thus more likely to encounter states similar to its cold folded conformation, but is likely to traverse this state quickly to continue exploring the vast conformational landscape accessible at high temperature.

REMD (Hukushima et al. 1996) seeks to combine the benefits of high temperature rapid exploration with low-temperature stability. Multiple trajectories of identical systems (replicas) are simulated in parallel at different temperatures. At certain intervals, the potential energies of replicas with neighboring temperatures are compared. If the structure at the higher temperature has found a lower potential energy conformation, it is exchanged with the other structure and continues its trajectory at the lower temperature. If the high temperature structure has the higher potential energy (which one expects to be the case most of the time) the structures may still be exchanged, with a probability given by the potential energy difference between the two replicas. Mathematically this is expressed as a Metropolis criterion: (Okabe et al. 2001)

$$P(\leftrightarrow) = \min\left(1, e^{-(\beta_2-\beta_1)(U_1-U_2)}\right)$$

Where P is the probability of an exchange between two neighboring replicas, $\beta_1 = 1/k_bT_1$ and $\beta_2 = 1/k_bT_2$ where k_b is Boltzmann's constant, T_1 and T_2

are the temperatures and U_1 and U_2 the potential energies of replicas 1 and 2, respectively.

The REMD algorithm cycles the replicas through the entire range of temperatures used in the simulation set and effectively sorts low energy structures to low temperatures where they may be maintained and refined, and sends unfavorable, high energy conformations to a warm environment, where they may rapidly move trough conformational space (see Figure 1 below).

Schematic of the REMD algortithm

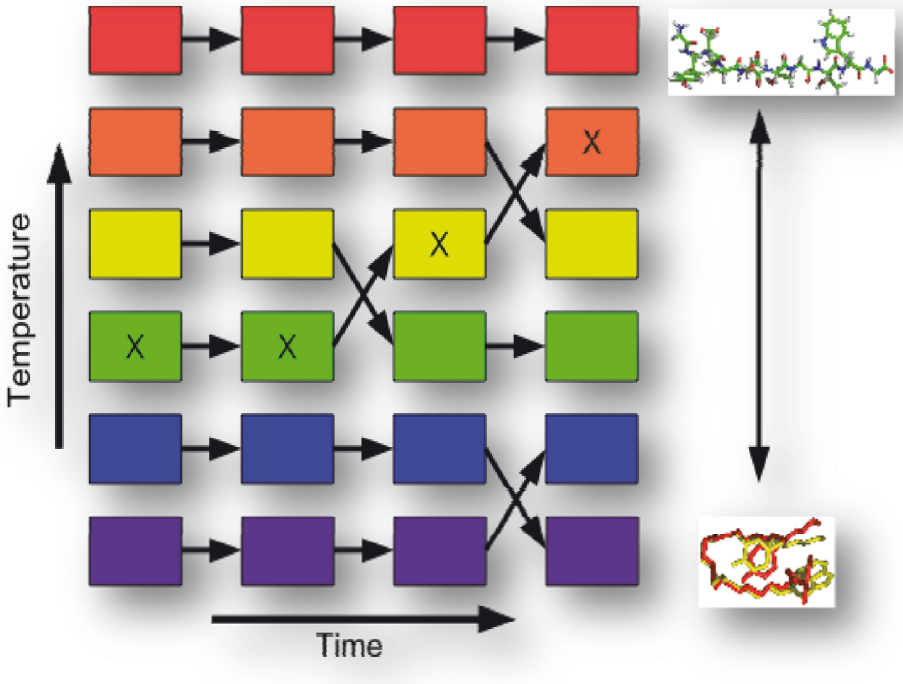

Figure 1. Replicas are simulated in parallel at different temperatures. At specified intervals, exchange probabilities are calculated based on a Metropolis and neighboring replicas are swapped accordingly. Stably folded structures gravitate towards lower temperatures were they are maintained, while misfolded conformations are further denatured by moving to higher temperatures.

Energy Landscapes

The conformational landscape a protein explores is a high-dimensional space. The number of degrees of freedom is proportional to the number of amino acids in the chain. For ease of visualization this space can be projected onto a few continuous or discrete parameter axis. This visualization

can be used to reveal the topology of the conformational landscape a protein experiences. The energy landscape concept can be used to reconcile various aspects of protein folding:

- Folding is a spontaneous process (Anfinsen)
  - that cannot occur as a random search (Levinthal)
    - and does not necessarily proceed along a predefined path
      - but reaches a well-defined, stable end-state.

The exploration of a Gibbs free-energy surface provides a framework that is compatible with all these aspects. A global energy minimum defines the native state and local minima are intermediate meta-stable states that can, but need not be encountered on the folding path. Re-arrangement of the protein along any path in the direction of decreasing energy is (part of) a folding path. Depending on the number of local minima and the slope of the energy gradient it may be characterized as either a smooth folding 'funnel', or a rough free-energy landscape.

A free-energy landscape can be calculated from simulation data based on the probability (i.e. abundance of snapshots) of finding a conformation with a given set of parameter combinations.

$$\Delta G(x,y,z) = -k_B T \ln\left[\frac{P(x,y,z)}{P(\min)}\right]$$

Where $\Delta G(x,y,z)$ is Gibbs free-energy, x,y,z are coordinates along parameter axes, k_b is Boltzmann's constant, T is temperature, P(x,y,z) is the probability (or relative abundance) of states at any given set of coordinates, and P(min) is the probability of finding a conformation in the global energy minimum, which hence has $\Delta G = 0$.

Protein Design

The goal of *de novo* protein design is the creation of a new amino acid sequences that fold into a stable, protein-like conformations with desired functions. The set of conceivable polypeptide combinations is gigantic. There are not enough atoms in the universe to synthesize only a single molecule of every possible sequence of amino acids the length of a normal protein – even a short chain of 50 residues can be any one of $20^{50} \cong 10^{65}$ different sequences. X-ray protein crystallographers, Nuclear Magnetic Resonance (NMR) spectroscopists and Electron Microscopist have solved the 3D structures of tens of thousands of proteins. Molecular biologists, biochemists, and biophysicists have been able to translate the structures into understanding of

sequence-structure and structure-function relationships. Recently, the knowledge has become so extensive, that attempts at *creating* peptide sequences that form desired, designed structures have become successful.

Observations of structural features exploitable for design can be as restrictive as the pattern of hydrophilic and hydrophobic residues in secondary structure elements (Kamtekar et al. 1993). More sophisticated analysis of sequences of the numerous proteins with known three-dimensional structures has enabled the generation of statistical propensities of each amino acid for certain structures. These analyses can be based either upon structural homology (e.g. Top7 by (Kuhlman et al. 2003)) or multiple sequence alignments (Russ et al. 2005; Socolich et al. 2005). Using these quantitative structure-sequence relation probabilities, it becomes possible to suggest and test stretches of sequences that are likely to fold into a predefined conformation. This approach to *de novo* protein design is often referred to as knowledge-based, since it takes advantage of the ensemble of naturally occurring proteins with known structures to generate novel sequences (Russ et al. 2002; Poole et al. 2006).

Knowledge-based protein design may be assisted by potential functions that employ the same parameters as force fields for MD, or a subset thereof. These physics-based models provide an assessment of the likely stability of a particular amino acid sequence in the conformation defined in the design. A score based upon the potential function may be used to evaluate the fitness of different sequences for a particular conformation, and candidates for synthesis can be selected. This approach to protein design makes use of structure-sequence relationships established from solved structures of naturally occurring proteins. It thus uses the information that is available before and after the self-assembly, without considering the protein folding process.

Chignolin

Chignolin is a *de novo* designed decapeptide with the sequence GYDPETGTWG that folds into a β-hairpin conformation (Honda et al. 2004). It is the smallest member of the novel class of proteins created by sequence design. Chignolin's sequence is not shared with any naturally occurring protein, but its geometric structure is substantially identical to the β-hairpin turn of the hexadecapeptide G-peptide (Honda et al. 2000), on which chignolin's design was based. Unlike stretches of α-helices and β-sheets, which are examples of continuous structures of repeating units, a turn such as a β-hairpin is a unique structural element of a defined size that introduces a reversal in the polypeptide's backbone structure, which is not arbitrarily repeatable. Such turns are frequently located in loop regions, which may be either flexible or stabilized by interactions with geometrically nearby amino acids that may come from segments separated by a large distance in the se-

quence. G-peptide, which contains residues 41-56 of the B1 domain of protein G, was the first isolated stretch of a protein to fold spontaneously into a β-hairpin in aqueous solution (Blanco et al. 1994). In addition to folding in isolation, the GB1 hexadecapeptide also binds specifically to a complementary fragment of the B1 domain of protein G (compromising residues 1-40). This complex exhibits a structure similar to the native GB1 (Kobayashi et al. 1995).

For the design of chignolin, structures from 100 non-homologous proteins were analyzed for structural motifs similar to G-peptide (Honda et al. 2004). Only the central 8 amino acids were considered in this comparison as it has been shown that they contribute more to the structural stability than residues close to the termini (Dinner et al. 1999; Pande et al. 1999; Kobayashi et al. 2000). Scores for each amino acid type were assigned based on the probability of each individual amino acid to exhibit backbone dihedral angles similar to those in the G-peptide hairpin. The most frequently identified type of amino acid was then chosen for each of the central eight positions, leading to the sequence YDPETGTW, which was flanked by two glycine residues, one at each terminus. The structure of chignolin was solved by Nuclear Magnetic Resonance (NMR), employing distance restraints derived from Nuclear Overhauser Effects to constrain molecular models.

The Folding of Chignolin

In **Paper I** chignolin's folding from an extended state to the native conformation was simulated. Substituting the computer for a microscope with spatial resolution on the atomic scale and time resolution of femtoseconds (the length of one discrete MD timestep) reveals the jiggling and wiggling of chignolin's atoms. As Feynman, Anfinsen and Levinthal envisioned, the protein folding path becomes observable through the application of Newton's laws of motion.

In the simulations, 16 out of 18 trajectories reached the native state at least once. At the time of publication of **Paper I** the continuous simulations were some of the longest MD trajectories recorded with simulation times of 1.6 and 1.8 μseconds. The trajectories showed the folding, unfolding and refolding of chignolin for the first time. Subsequent folding simulations of chignolin were reported in 2006 (Satoh et al. 2006), 2007 (Suenaga et al. 2007), 2008 (Dou et al. 2008; Terada et al. 2008; Xu et al. 2008), 2009 (Kannan et al. 2009; Roy et al. 2009) and 2010 (Rakshit et al. 2010).

The Free-energy Landscape of Chignolin

Chignolin explores a substantial area of conformational space in the MD trajectories. It can be observed in numerous structural arrangements, some of which greatly differ from the native state and many of which satisfy most NOE restraints and are substantially similar to the native state (see Figures 2 & 3 in **Paper I**). Order parameters can be used to group conformations into sets distinguished by geometric properties and create a free-energy surface derived from the relative abundance of structures within each set. **Paper I** employs three order parameters for a 3D free-energy landscape: 1. The distance between the amino to carboxy termini. 2. The distance between the hydrophobic sidechains of Tyr2 and Trp9. 3. The number of backbonde hydrogen bonds. Each of these coordinates is an indicator of protein folding: The termini distance is reflective of the compactness of the fold. The aromatic sidechains of Tyr2 and Trp9 form the 'hydrophobic core' of chignolin, and the distance between them gives an indication to which extend this core has formed. Backbonde hydrogen bonds define secondary structure.

The resulting 3D energy landscape is smooth, flat and has a steep and deep global minimum without any pronounced local minima (see Fig. 2).

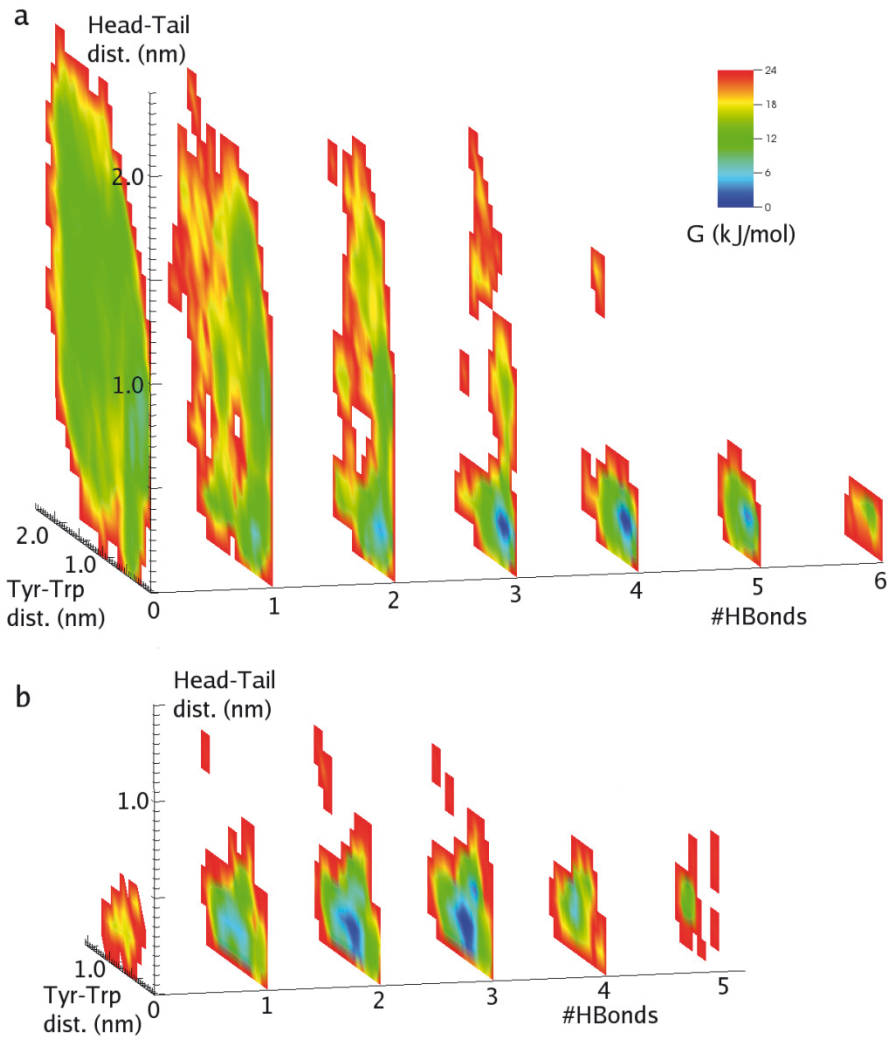

Figure 2. Panel (a) shows the 3D free-energy landscape in water and panel (b) *in vacuo*. In solution there is a single, small global minimum well, which corresponds to the experimentally determined native state. The sampled conformation space is much smaller *in vacuo*. Extended, unfolded conformations are practically never explored and the global minimum structure is distinct from the solvated structure.

The ensemble of structures found at the global minimum exhibit average NOE violations of <V> = 0.9 Å and Cα rmsd with respect to the first structure in the NMR ensemble of 1.0 Å. They unambiguously belong to the fully folded native state.

Two-dimensional energy landscapes for chignolin were determined by (Satoh et al. 2006; Xu et al. 2008; Roy et al. 2009). Xu et al. (2008) and Roy et al. (2009) use similar coordinate axis and finds an energy landscape with very similar defining characteristics and a single deep global minimum corresponding to the native structure. The hydrophobic interaction distance is an order parameter axis that is shared between these two works and **Paper I**, allowing direct, absolute comparisons. Along this axis, the global minimum is located in a practically identical position at c. 4Å, i.e. in excellent quantitative agreement in Xu et al and **Paper I**. In Roy et al. (2009) the global minimum stretches from 4-8Å along the Tyr2-Trp9 axis, which is very compatible with the value in Xu et al and Paper I. The energy landscape presented in Satoh et al is substantially different from the others. It has two almost identically deep minima separated by a substantial energy barrier. The global minimum, into which contains 25.1% of the sampled structures corresponds to the native state. In the deep local minimum, 24.9% of all structures are trapped in a misfolded configuration (Satoh et al. 2006). This is attributed to implicit treatment of solvent molecules in the simulation and discussed in the following section.

The role of water in protein folding

The solvent surrounding a protein has an essential influence on protein folding and dynamics. Water molecules act as hydrogen-bond donors and acceptors to polar amino acids and cluster around charged sidechains. By avoiding aromatic and other hydrophobic amino acids water molecules generate the hydrophobic effect, a force that causes the clustering of hydrophobic amino acids in the interior core of the protein, where they are shielded from the water.

Simulations performed *in vacuo* and shown in Figure 5b and 8b of **Paper I**, demonstrate that chignolin does not fold in the absence of a solvent. In MD simulations one consequently faces the choice of how to model water. Two solutions have emerged: Explicit and Implicit solvation.

Explicit solvation treats water molecules like all other atoms in the simulation systems and calculates forces and movements for each particle individually. This approach ensures that the same calculation principles and force field parameters are applied to solute (i.e. protein) and solvent. It avoids the introduction of additional modeling uncertainties, since all molecules are treated equally at the cost of additional computational load. For a typical small protein in solution, the number of water atoms greatly outnumbers the number of protein atoms (in the case of chignolin in **Paper I** by a factor of almost 20:1). Hence, in MD simulations with explicit solvent most CPU cycles are dedicated to simulating the Brownian motion of water, not the

folding of protein. GROMACS includes special optimizations to calculate water-water interactions particularly efficiently to reduce this effect.

Implicit solvent models, such as the frequently used Generalized Born (GB) continuum approximation of water eliminate solvent molecules from MD simulations and thereby greatly reduce the size of the system to be simulated. Additionally, as Terada et al. note, in the GB surface area (SA) model, "the protein does not experience the solvent's viscosity, and the conformational transitions of the protein occur much more frequently than in the explicit solvent (Ishizuka et al. 2004; Snow et al. 2005). This characteristic is advantageous for conformational sampling. However, the use of the GB/SA model inevitably causes an error in the calculation of the solvation free energy. (Zhou et al. 2002; Zhou 2003; Ishizuka et al. 2004)."

Solvation Effects

Satoh et al. employed the GB/SA model of (Still et al. 1990) in their MD folding simulations of chignolin. In these simulations, a native-like beta-hairpin conformation that satisfies the NOE restraints well is observed in the largest cluster of structures at the global minimum in the free-energy landscape with a probability of existence of 25.1%. However, a local minimum whose clustered structures have an abundance of 24.9% appears close to the global minimum. These structures show a non-native fold, where Tyr2 and Trp9 are located on opposite sides of the beta-sheet. These two aromatic residues therefore cannot make contact, and this misfolded chignolin conformation lacks a hydrophobic core. The authors note "The simulation and the experiment have two significant discrepancies. One is the lack of tight contact between the aromatic rings of Tyr2 and Trp9 in the simulation, and the other is the large fraction of the population made up by the misfolded species." Further "…we conclude that the aromatic rings actually made tight contact [from experimental data] and that the attractive interaction between the aromatic rings was underestimated in the simulation.". This attractive interaction is caused by the hydrophobic effect. Satoh et al. continue "To eliminate the first discrepancy, the accuracies of the force-field parameters and the models of solvation free energy must be improved". Explicit solvation achieves this by applying consistent force field parameters to both solute and solvent and eliminating additional solvation free energy models altogether.

"Therefore, if the interaction between the aromatic rings could be calculated accurately, the misfolded species might become less stable than the conformations with the native hydrogen bonds. As a result, the native structure with the correct aromatic ring arrangement might occupy the majority of the whole ensemble, and the second discrepancy, as well as the first one, might be eliminated. This furthermore results in only one free-energy well in the free-energy landscape with the Asp3O-Gly7N hydrogen bond. The fold-

ing mechanism proposed above is then consistent with the two-state, cooperative thermal transition observed in the experiment (Honda et al. 2004)." (Satoh et al. 2006).

Just how much water is required around chignolin was quantified by Suenega et al. Using the special purpose MDGRAPE hardware, they simulated chignolin in spherical drops of water of various sizes. "In our simulations, the peptides were immersed in the spherical water droplet, and water molecules at the surface of the droplet were constrained with a harmonic potential to prevent diffusion of the water molecules into vacuum." "The effect of the size of the water sphere was investigated in solvent depths of 9 to 27Å. We found that the water-sphere system with a solvent depth of 9Å heavily influenced the structural and dynamic properties of proteins, and these properties were not supported by experimental data. On the other hand, water-sphere systems with a solvent depth beyond 15Å showed similar behaviour in protein structure and dynamics, and this time the properties are consistent with experimental data. Accordingly, from the viewpoint of the structure and dynamics of proteins in water droplets, it was proved that the water-shere system with a solvent depth of at least 15Å was required for accurate representation of the protein dynamics." (Suenaga et al. 2007).

Simulations of hydrated and encapsulated proteins in vaccum have been performed (Iavarone et al. 2007; Patriksson et al. 2007; Friemann et al. 2009; Marklund et al. 2009; Wang et al. 2009). A small amount of water prevents unfolding over a broad temperature range, and hydrated macromolecules retain their conformational integrity in the gas phase. A single water layer mimics bulk solvent, and protects the structure. Evaporation is faster from hydrophobic surface areas than from hydrophilic patches, leading to reproducible surface patterns with reproducible hydrogen bonding.

Terada et al. (2008) compared explicit and implicit solvation models for simulations of chignolin by starting simulations in explicit solvent on folded structures obtained with implicit solvent simulations. An analysis of the hydrogen bonding pattern showed: "The characteristic hydrogen bonds (Asp3O-Gly7N and Asp3N-Thr8O) were stable in two runs. The average distance of the hydrogen bonds were 2.89Å ± 0.14Å and 2.96Å ± 0.18Å. In the third run, they were stably maintained for 28.5ns with average distances of 3.06Å ± 0.37 Å and 3.02Å ± 0.32Å, although the Asp3N-Thr8O hydrogen bond was broken after that."

Protein structure prediction from MD simulations

In order to employ MD simulations to predict the structure of unknown proteins, the native state has to be identified without any experimental data. The energy landscape analysis allows this, as an energy landscape can be drawn with order parameters that are applicable to any arbitrary peptide chain. The

correspondence between experimental NMR structure factors and the Free-energy calculated from MD is shown below.

There is no general relationship between Free-energy and degree of NOE violation of any given structure, but excellent concurrence in the minima of each, as Figure 3 shows. Consequently, it is possible to correlate the Free-energy minimum with the native structure, just as Anfinsen has postulated.

Roy et al. (2009) find very similar agreement. In order to refine the global energy minimum ensemble of structures, a cluster analysis can be performed to excludes structural outliers. In **Paper I** 97.8% of the structures at the global minimum (11,643 snapshots) formed a single cluster with pairwise rmsds of less than 2Å. Cluster analysis performed on an entire folding trajectory also leads to the identification of a correct native structure (Xu et al. 2008).

Correlation between NOE violations and Gibbs free-energy

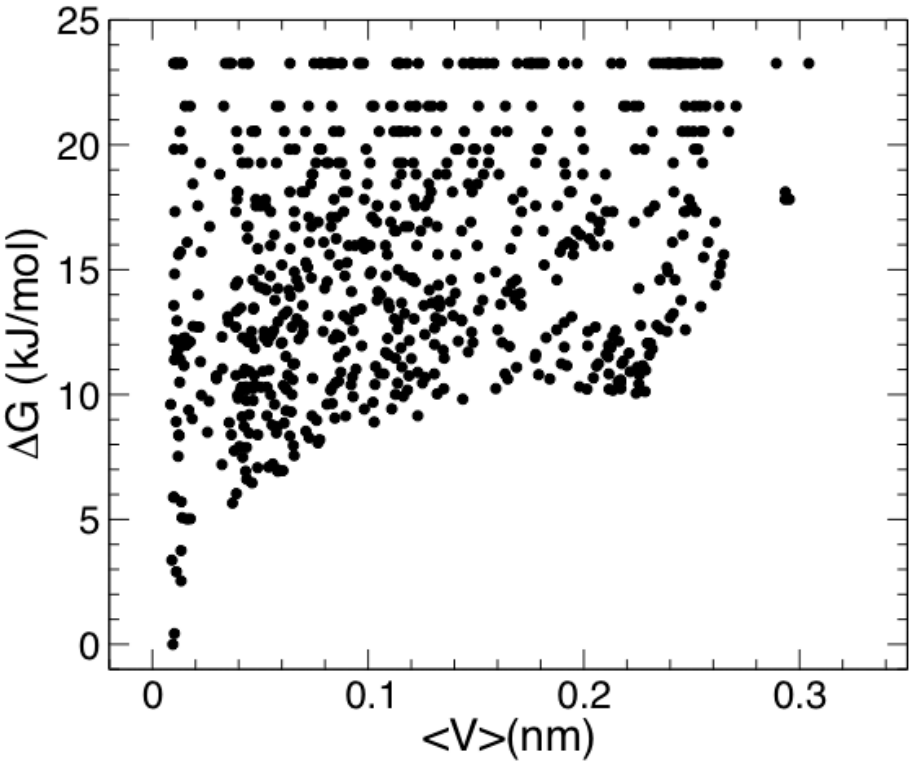

Figure 3. The relationship between free-energy calculated from the energy landscape (Figure 2) and experimentally determined structure restraints, or simply simulation vs. experiment. The calculated energy landscape identifies the experimental native conformation, as structures at the calculated global minimum have the lowest restraint violation values. Low <V> do not guarantee a low-energy structure.

The Folding Kinetics of Chignolin

Determining the rate at which a protein folds from a MD trajectory is in principle a straightforward procedure. A criterion needs to be established to determine whether any given snapshot within a trajectory represents a folded (i.e. native-like conformation) or unfolded structure. In cases where the protein structure is known from experiments, this can simply be a cut-off based on the root-mean-square deviation (rmsd) from a crystallographic structure or violations of NOEs from NMR spectra. For predictive simulations, when no experimental data are available, the structure identified at the global minimum on the Gibbs Free-energy landscape may serve as the reference for rmsd calculations. Every snapshot along the trajectory is then classified as either folded or unfolded and an estimate of the folding rate can be made from the simulation time that elapses before a folded conformation is reached.

Using this approach, chignolin's folding time was estimated to 1-2μs in **Paper I** at 300K. The large uncertainty in this estimate is due to the scarcity of folding events observed in the constant temperature simulations. In the REMD simulations far more folding events are observed and some trajectories reach folded structures within 50ns. However folding rates cannot be estimated from these simulations directly due to the frequent temperature changes, which are intended to speed-up the folding process. **Paper II** presents an algorithm that resolves this problem by explicitly taking into consideration the temperature jumps. Applying this algorithm to the chignolin REMD trajectories predicts a folding time of 1.0±0.3μs, classifying it as a relatively fast folder compared to other polypeptide sequences of comparable length (Eaton et al. 1997; Munoz et al. 1997; Xu et al. 2003).

Chignolin's folding rate has not yet been measured experimentally, but an additional folding rate estimate of ≈0.5μs was provided by (Suenaga et al. 2007). These authors state: "The folding time constant of about 0.5μs is larger [*sic*] than the (1.0±0.3) μs predicted by van der Spoel and Seibert [**Paper II**], because our analysis of folding events observed from our simulations were statistically poor." These estimates were obtained from constant temperature simulation trajectories, like those reported in **Paper I**. Those likewise lacked the number of folding events required to achieve small error margins and were the motivation to exploit the more frequent REMD folding events to estimate folding rates.

An additional algorithm for tackling the challenge of extracting kinetic information from REMD trajectories has recently been proposed (Buchete et al. 2008). Employing a different technique and testing it on a different model system (a penta-alanine α-helical peptide), these authors likewise reach the conclusion that faithful protein folding kinetics can be deduced from thermally accelerated REMD simulations.

When an energy landscape has been established, Daidone et al. provide an alternative approach (Daidone et al. 2005a; Daidone et al. 2005b) to obtain this information.

Proteins are dynamic structures

Under physiological conditions, a protein is constantly in motion and from a MD trajectory this is immediately obvious. To a structural biologist working with 3D protein data obtained by X-ray crystallography, the structure may appear more static. A crystal structure is the average representation of a unit cell within the crystal, an average of 10^9 or more copies of a protein. Often, crystals are frozen to reduce susceptibility to radiation damage and minimize Brownian molecular motion. Consequently, it may be of little surprise that structural biologists sometimes report some properties of the structures they obtained as numerical quantities with fixed values and no variability.

Even when derived from a high-resolution crystal structure, numerical values do come with some uncertainty. When extrapolated to the physiological state, these errors combine with the dynamic motion of proteins to lead to even greater variability. **Paper III** therefore suggests that physical parameters derived from protein structures be reported with an estimate of their variation. Taking the solvent-accessible surface area (ASA) as an example of a structure-derived quantitative property, its variation with experimental parameters, its fluctuations under simulated physiological conditions and the precision of methods to calculate it were investigated.

ASA is commonly defined by the rolling-probe method as the area covered by the center of a spherical probe with a given diameter that is rolled along the entire protein in contact with the van der Waals surface of the protein structure (Lee et al. 1971). The ASA can be divided into hydrophobic and hydrophilic parts. During folding the hydrophobic effect leads to the burial of hydrophobic surface area in the core of a protein, where it is shielded from the solvent. Residues exposed on the surface and those lining pores or cavities are frequently polar or charged and form hydrogen bonds with surrounding water or solute molecules. The ASA and changes thereof are thus useful for following the folding or unfolding process. They have also been employed in assessing the role of active-site residues (Mazumder-Shivakumar et al. 2005), to define side-chain conformational entropy at protein interaction sites (Cole et al. 2002), to characterize protein-nucleic acid recognition locations (Nadassy et al. 1999) or to improve rankings of docking solutions (Duan et al. 2005).

The ASA was found to vary by more than 10% during a 20ns simulation of the small lysozyme protein, from a minimum of 6577Å to a maximum of 7391Å. Additionally, there was a systematic increase in ASA observable

during the first 5ns of the simulation, when the protein relaxed from the crystallographically determined native state to the solution equilibrium conformation.

DNA Origami

Principles

For proteins, sequence defines structure and only a small subset of polypeptide sequences can form stable proteins. Most other polypeptides are unstructured and flexible. For DNA (and RNA) the opposite is the case. Any polynucleotide sequence can form the double helix, a highly regular, well-defined and stable structure, given a complementary sequence with which Watson-Crick base-pairs can be made. Nevertheless, the specific hydrogen bonding between complementary Watson-Crick base-pairs can also give rise to complex secondary structure based on the arrangement of double helical segments. This occurs naturally in RNA structures such as tRNAs and ribozymes and forms the secondary structure of single stranded plasmids or viral RNA and DNA genomes. It is also the basis for the molecular art of folding DNA into nanostructures, known as DNA origami.

The interactions between different nucleotides along a polynucleotide chain can be explained entirely through hydrogen bonding and the resulting backbone geometry can be deduced from the base-pairing pattern. This allows the creation of DNA structures through the application of a few conceptually simple design rules, in stark contrast to the algorithmic complexity of protein design, where intra- or interchain interactions are not predictable from simple sequence-based rules. The discoverers of these design principles have employed them to create numerous two- and three dimensional structures, including cubes (Chen et al. 1991), stick-figures (Seeman 1991), Serpinski triangles (Rothemund et al. 2004), smileys (Rothemund 2006), dolphins with wiggling tails (Andersen et al. 2008) and even rationally-designed macroscopic crystals arranged with sufficient precision to diffract an X-ray beam, producing Bragg peaks (Zheng et al. 2009).

Scaffolded DNA origami

A particularly versatile technique to direct the self-assembly of DNA origami is the scaffolded DNA origami method (Rothemund 2006). An example is shown in Figure 4. Here, a long single strand of DNA (the genome of the M13 bacteriophage is frequently used) is folded into a desired shape by constraining it with the help of a set of short oligonucleotide staples. These

staples have sequences that are complementary to two sections along the long ssDNA strand separated by a long intervening sequence. Thereby the staples force two distant parts of the long ssDNA into close geometric proximity. Supplying sufficient staples to cover the entire ssDNA scaffold constrains the resulting structure to one unique shape, which is entirely defined by the chimeric complementary sequences of the short staple oligonucleotides.

Scaffolded DNA origami

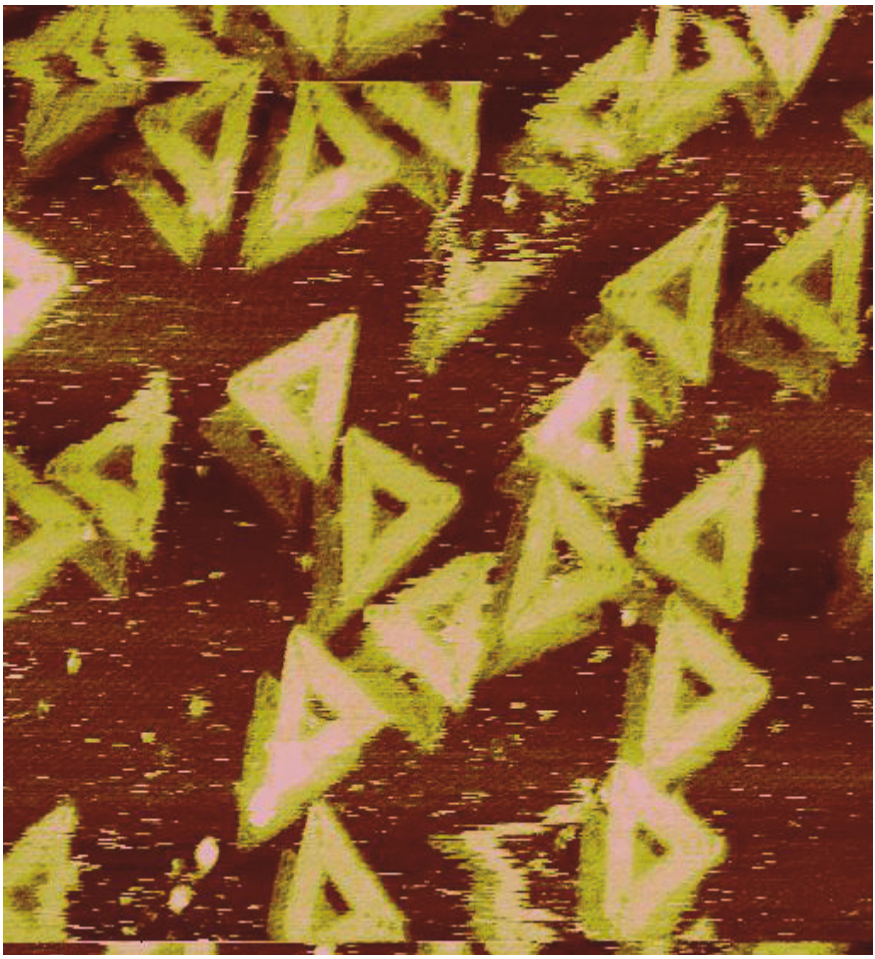

Figure 4. AFM image of triangular scaffolded DNA origami on a freshly cleaved mica surface. Image size 1×1 μm, height 2nm.

DNA origami are flexible structures

Two dimensional DNA origami, such as those made by the scaffolded DNA origami technique (Rothemund 2006) can be imaged by Atomic Force Microscopy (AFM). AFM is a mechanical surface sensing technique. A sharp tip mounted at the end of an elastic cantilever is scanned across the surface of a sample and the deflection of the cantilever is translated into a topographic image of the sample. AFM requires an atomically flat surface on which the sample is held. Mica offers such a surface, with the additional advantage that a freshly cleaved mica surface layer is negatively charged. The P atoms in the DNA backbone are likewise negatively charged. Divalent cations (commonly Mg^{++}) in the buffer solution facilitate the attachment of DNA to the freshly cleaved mica surface. This attraction anchors the DNA origami with sufficient strength such that they are (usually) not displaced by the scanning motion of the AFM tip. DNA origami imaged by this technique appear to be uniformly bound to the surface and correspondingly flat. TEM images which employ a positively charged poly-L-ornithine surface to anchor the DNA origami similarly show flat structures [**Paper IV**].

This leads to the question of whether scaffolded DNA origami are inherently planar or whether they assume this shape when affixed to suitably charged surface. Imaging individual free macromolecules is difficult, because few techniques can probe single molecules at high resolution. However, the unique geometry of the scaffolded DNA origami, specifically their unusual aspect ratio with a width-to-height relation of c. 50:1 makes inferences from low-resolution imaging techniques particularly potent. Maintenance of or deviation from the planar structure can be measured both in solution and in the gas phase.

Small-angle X-ray scattering (SAXS) was employed to probe the structure of triangular scaffolded DNA origami (Figure 4) in solution [**Paper IV**]. The observed radius of gyration Rg = 30.8 ± 0.4 nm and maximum overall dimension Dmax = 90 ± 2 nm is in stark contrast to that expected for a flat structure of Rg = 45 nm and outer edge length Dmax = 125 nm. The scattering profile is more consistent with a globular, partially collapsed structure. When the same sample used for the SAXS analysis was imaged by AFM afterwards, the structures appeared as normal, flat triangles.

In the gas phase, the electrophoretic mobility diameter can be measured with a GEMMA instrument. First the DNA origami are aerosolized with a charge-reduced electrospray unit, and this aerosol is then sent through a dif-

ferential mobility analyzer (DMA) which can be tuned to selectively transmit particles of specific aerodynamic sizes. Particles transmitted through the DMA are detected and counted by a condensation particle counter. The highest transmission is observed when the DMA is tuned to an electrophoretic mobility diameter of 28.1nm, comparable to the radius of gyration observed in solution, but again substantially different from the value expected for a flat structure.

These results indicate that despite their highly regular and reproducible structure, DNA origami retain flexibility that allows a substantial conformational transition between surface-supported, free-in-solution and aerosolized states. It remains to be determined whether the solution and gas-phase conformations are as unique as the planar conformation – the unprecedented sharpness of the DMA peak suggests this for the aerosol structure – or whether an ensemble of structures exists.

Conclusions and future perspectives

When the MD simulations of chignolin were first reported [**Paper I**] they were some of the longest continuous MD trajectories ever. Since then, special purpose hardware has been build and used to compute MD trajectories of 1 millisecond simulation time, an improvement of more than 500 times. The impressive rate of gene and gene variation discovery leads to an ever-increasing pool of protein sequence knowledge. This information will become even more valuable when it is translated into 3D structural information. The extraction of kinetic information from large-scale MD simulations [e.g. **Paper II**] and the appreciation of the dynamics of protein structures [**Paper III**] may provide additional insight not only into the folding process but also into protein function in general. The ability to make quantitative predictions as in **Paper II** may enhance our ability to establish universal engineering design principles for biomolecules that could lead to far more complex and functional protein assemblies than chignolin or other *de novo* designed proteins built to date.

The investigation of the physical properties of designed DNA assemblies as reported in **Paper IV** may enable the exploitation of the observed flexibility exhibited by these nanoscale objects. More than half a century after the discovery of the double helical structure of DNA, this molecule still surprises and fascinates researchers seeking to unravel the secrets it carries from generation to generation.

One may speculate that functional aspects of life's most important molecules are reflected in their structural properties. Proteins can be synthesized with practically infinite variety, reproducibly and spontaneously fold into functional conformations and carry out their tasks employing dynamic rearrangements of their structures as needed. They are nature's robots. DNA's structural predictability, stability and flexibility proves that not all its secrets are contained in a linear list of nucleotide letters. If scientists can create stunning structural assemblies with DNA one has to wonder what other architectural wonders life is performing with biology's defining molecules.

Sammanfattning på Svenska – Summary in Swedish

DNA och proteiner är livets viktigaste molekyler. De enklaste biologiska enheterna, virus, kan bestå av enbart dessa två makromolekyler. Både DNA och proteiner är sammansatta som långa kedjor av subenheter. Dessa subenheter benämns monomerer, och i DNAs fall så kallas monomererna nukleotider. Proteiner byggs upp av aminosyror, sammankopplade till en polypeptid.

Proteiner, som utför många av livets mest essentiella funktioner fungerar bara om de har en viss speciell tredimensionell struktur. Strukturen är individuell för varje protein och kallas proteinets energimässiga grundtillstånd. Proteinets struktur kan vara mycket komplicerad eller mycket enkel. Vissa proteiner har en otroligt fin arkitektur som t.ex. att vara en tunnel genom en cells membran. Andra är designade som en motor som kan vrida sig eller pumpa joner. Flertalet är små, globulära och lösliga i cellens interiör. Ett protein kan även fungera som en fysiologisk signal till andra proteiner.

De flesta proteiner består av hundratals eller tusentals aminosyrer. Även i de mest komplexa proteiner är aminosyrorna alltid ihopkopplade i till lång kedja. Det är enbart det sätt som denna kedja veckar sig som bestämmer hur proteinet ser ut. För många proteiner, speciellt små, lösliga, så sker veckningen spontant utan någon inverkan från andra molekyler.

Denna självveckningsprocess är inte magisk, utan är en konsekvens av de fysikaliska krafter som verkar på atomär nivå. Vi kan inte observera processen direkt eftersom det inte finns några mikroskop som kan visa individuella atomer och hur dessa rör sig. Därför simulerar vi proteinveckningen med hjälp av datorer. Vi programmerar datorer så att de kan räkna ut de krafter som verkar på atomerna i ett protein och de vattenmolekyler som omger proteinet. Newtons rörelselagar beskriver vilka krafter och på vilket sätt dessa krafter påverkar en partikel och dess rörelser. Sedan flyttar vi atomerna till de nya positioner som Newtons rörelselagar anvisar och bestämmer återigen vilka krafter som verkar på atomerna i sina nya positioner. Efter att vi har gjort detta ett par miljoner gånger får vi många bilder av processen som vi kan sätta ihop till en film, en animation av veckningen. Denna metod att matematiskt simulera partiklars rörelse kallas Molekylär Dynamik (MD).

Vi har använt MD simulationer för att ser hur en liten polypeptid veckar sig för att nå sitt energimässiga grundtillstånd. Chignolin består av enbart tio

aminosyrer. Det är inget naturligt protein (som brukar vara mycket större) utan är designat av forskare i Japan. Chignolin är en av de minsta polypeptider som har ett stabilt grundtillstånd. Det gör den till ett idealt modellsystem för simulationer. Simuleringarna börjar med alla tio aminosyrer i en linjär kedja, och under simulationen beter den sig som ett spagettistrå i kokande vatten. Men till skillnad från spagetti kan chignolin inte anta vilken struktur som helst. Efter en viss tid så hittar den en konformation, ett energimässigt grundtillstånd. Målet för simulationerna är inte bara att skapa en film av veckningsprocessen utan också att identifiera proteinets strukturella grundtillstånd utan att använda experimentellt data. Detta möjliggör nämligen att simulationer till viss del kan ersätta experiment i situationer där det är svårt eller omöjligt att genomföra experiment, t.ex. om man vill bestämma strukturer vid ovanligt höga temperaturer, under vacuum eller högt tryckt, etc.

Christian Anfinsen fick Nobelpriset i Kemi 1972 för sin upptäckt att ett proteins grundtillstånd helt enkelt är den struktur som är termodynamiskt mest stabil. Vi kan använda denna teori för bestämning av ett proteins grundtillstånd och detta har gjorts i den här avhandlingen. Chignolin veckar sig till en unik, robust men dynamisk struktur som kan identifieras med hjälp av MD simulationer och utan användande av experimentellt data.

Ett proteins struktur bestäms som tidigare nämnts av sin sekvens av aminosyror. DNA, å andra sidan kan forma den berömda dubbelhelixen med vilken sekvens av nukleotider som helst. Man kan bygga olika strukturer som består av sådana dubbelhelixsegment. Det är en molekylär konst som innebär att man kan skapa roliga eller nyttiga nanoobjekt av DNA, kallat DNA-origami. Forskare har till exempel byggt smileys, delfiner med rörliga svansar, kuber och makroskopiska kristaller som diffrakterar röntgenstrålning av DNA. Trots att dubbelhelixarna är sammansatta till sådana komplexa strukturer, så är dessa strukturer ganska mjuka och flexibla. I denna avhandling beskrivs experiment som visar hur en sådan DNA-origamistruktur kan anpassa sig till olika miljöer.

Först bestämdes strukturen av en DNA-origamikonstruktion när den var bunden till en mycket jämn yta. Ett nykluvet micaflak är atomiskt flat och DNA kan fås att binda till den. Ett Atom Kraft Mikroskop (AKM) är ett sorts mekaniskt mikroskop som kan avbilda ytan av ett objekt till nästan atomär upplösning. Mikroskopet fungerar genom att den skannar av en yta med en ytterst skarp (fåtal nanometer spetsdiameter) spets och mäter hur mycket spetsen deflekteras av provet. Med den tekniken, det vill säga DNA-origami bunden till en atomärt flat yta och observerat med ett AKM så visar det sig att DNA-origamikonstruktionen kan anta en helt plan konformation.

När man undersöker strukturen i vätska, vilket man kan göra med hjälp av small angle X-ray scattering, visar det sig att samma DNA-origamikonstruktion rullar ihop sig till en kompakt struktur. I luft blir den ännu mer kompakt. Detta har bestämts genom att aerosolisera DNA-origamikonstruktionen med en electrosprayenhet och mäta storleken (eller så

kallad elektroforetisk mobilitets diameter) med en differential mobility analyzer. I luft beter sig en DNA triangel med 125nm kantlängd och 2nm tjocklek som ett klot med 28nm diameter. Det innebär att DNA-origami har en stor förmåga att ändra sin struktur, någonting som förmodligen kan utnyttjas av framtidens DNA arkitektur. Eftersom naturen vanligen är lite före mänsklig kunskap att skapa unika strukturer kan man fundera över hur DNAs flexibilitet redan utnyttjas inom biologin.

De krafter som påverkar proteiners och DNAs struktur är de samma, men effekten de har på dessa två makromolekylers struktur är annorlunda. DNA, naturens informationsbärare är mjukt och flexibelt och anpassar sig olika miljöer. Proteiner däremot utför livets funktioner reproducerbart, som robotar.

Acknowledgements

Special thanks are first and foremost due to David van der Spoel for enabling the science presented here. Throughout the years of working with you I have not only had the privilege of learning Molecular Dynamics from one of its fathers, but also gotten an opportunity to appreciate the art of *investigative science*. Janos Hajdu has taught me the meaning of *explorative research*, whose value is hard to overestimate (but you are trying). I am particularly grateful to have learned the difference between management and leadership from you.

The members of the biophysics lab, xray, ICM and the flash imaging collaboration have been a fantastic gang, who I like to think of more as friends than colleagues. Together we have had many memorable adventures, for which I will forever be grateful. To name just a few random examples: Martin, I do appreciate that you prevented me from buying a Bandvagn, an amphibious car, a 6x6 military vehicle and god knows what else. (Nevertheless, I'm also glad that you weren't around when I found the truck.) It was another great idea of yours to check my stochiometry when I first attempted to make DNA triangles (and double check most of everything I have done since – it probably wouldn't have worked otherwise). Freddie, thank you very much for mothering our Antarctic creatures, they can't live without you! Emmanuelle, thanks for making endless AFMing fun and entertaining. Sebastien, Anton, Mike, Matthias, Henry, Stefano, Florian, Urs, Stefan, Bruce and everyone else on the beamtime team for truly original experiences, every time.

Kerstin, thank you for romance at the nuclear weapons lab (and everywhere else). Mama and Papa and Arnd, I do appreciate your constant, unconditional and unlimited support, even if I don't always show it.

This is only the beginning, not the end.

Bibliography

1. Andersen, E. S., M. Dong, M. M. Nielsen, K. Jahn, A. Lind-Thomsen, W. Mamdouh, K. V. Gothelf, F. Besenbacher and J. Kjems (2008). "DNA origami design of dolphin-shaped structures with flexible tails." ACS Nano **2**(6): 1213-1218.
2. Anfinsen, C. B. (1973). "PRINCIPLES THAT GOVERN FOLDING OF PROTEIN CHAINS." Science **181**(4096): 223-230.
3. Blanco, F. J., G. Rivas and L. Serrano (1994). "A SHORT LINEAR PEPTIDE THAT FOLDS INTO A NATIVE STABLE BETA-HAIRPIN IN AQUEOUS-SOLUTION." Nature Structural Biology **1**(9): 584-590.
4. Buchete, N. V. and G. Hummer (2008). "Peptide folding kinetics from replica exchange molecular dynamics." Physical Review E **77**(3): -.
5. Chen, J. H. and N. C. Seeman (1991). "Synthesis from DNA of a molecule with the connectivity of a cube." Nature **350**(6319): 631-633.
6. Cole, C. and J. Warwicker (2002). "Side-chain conformational entropy at protein-protein interfaces." Protein Science **11**(12): 2860-2870.
7. Daidone, I., A. Amadei and A. Di Nola (2005). "Thermodynamic and kinetic characterization of a beta-hairpin peptide in solution: An extended phase space sampling by molecular dynamics simulations in explicit water." Proteins-Structure Function and Bioinformatics **59**(3): 510-518.
8. Daidone, I., M. D'Abramo, A. Di Nola and A. Amadei (2005). "Theoretical characterization of alpha-helix and beta-hairpin folding kinetics." Journal of the American Chemical Society **127**(42): 14825-14832.
9. Darden, T., D. York and L. Pedersen (1993). "Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems." Journal of Chemical Physics **98**(12): 10089-10092.
10. Dinner, A. R., T. Lazaridis and M. Karplus (1999). "Understanding beta-hairpin formation." Proceedings of the National Academy of Sciences of the United States of America **96**(16): 9068-9073.
11. Dou, X. H. and J. H. Wang (2008). "Folding Free Energy Landscape of the Decapeptide Chignolin." Modern Physics Letters B **22**(31): 3087-3098.
12. Duan, Y., B. V. Reddy and Y. N. Kaznessis (2005). "Physicochemical and residue conservation calculations to improve the ranking of protein-protein docking solutions." Protein Science **14**(2): 316-328.
13. Eaton, W. A., V. Munoz, P. A. Thompson, C. K. Chan and J. Hofrichter (1997). "Submillisecond kinetics of protein folding." Current Opinion in Structural Biology **7**(1): 10-14.
14. Epstein, C. J., R. F. Goldberger and C. B. Anfinsen (1963). "GENETIC CONTROL OF TERTIARY PROTEIN STUCTURE - STUDIES WITH MODEL SYSTEMS." Cold Spring Harbor Symposia on Quantitative Biology **28**: 439-&.

15. Essmann, U., L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen (1995). "A Smooth Particle Mesh Ewald Method." Journal of Chemical Physics **103**(19): 8577-8593.
16. Feynman, R. P. (1963). Six Easy Pieces, Penguin Books.
17. Friemann, R., D. S. Larsson, Y. Wang and D. van der Spoel (2009). "Molecular dynamics simulations of a membrane protein-micelle complex in vacuo." J Am Chem Soc **131**(46): 16606-16607.
18. Hess, B., C. Kutzner, D. van der Spoel and E. Lindahl (2008). "GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation." Journal of Chemical Theory and Computation **4**(3): 435-447.
19. Honda, S., N. Kobayashi and E. Munekata (2000). "Thermodynamics of a beta-hairpin structure: Evidence for cooperative formation of folding nucleus." Journal of Molecular Biology **295**(2): 269-278.
20. Honda, S., K. Yamasaki, Y. Sawada and H. Morii (2004). "10 residue folded peptide designed by segment statistics." Structure **12**(8): 1507-1518.
21. Hukushima, K. and K. Nemoto (1996). "Exchange Monte Carlo method and application to spin glass simulations." Journal of the Physical Society of Japan **65**(6): 1604-1608.
22. Iavarone, A. T., A. Patriksson, D. van der Spoel and J. H. Parks (2007). "Fluorescence probe of Trp-cage protein conformation in solution and in gas phase." J Am Chem Soc **129**(21): 6726-6735.
23. Ishizuka, T., T. Terada, S. Nakamura and K. Shimizu (2004). "Improvement of accuracy of free-energy landscapes of peptides calculated with generalized Born model by using numerical solutions of Poisson's equation." Chemical Physics Letters **393**(4-6): 546-551.
24. Jorgensen, W. L. and J. Tirado-Rives (2005). "Potential energy functions for atomic-level simulations of water and organic and biomolecular systems." Proceedings of the National Academy of Sciences of the United States of America **102**(19): 6665-6670.
25. Kaminski, G. A., R. A. Friesner, J. Tirado-Rives and W. L. Jorgensen (2001). "Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides." Journal of Physical Chemistry B **105**(28): 6474-6487.
26. Kamtekar, S., J. M. Schiffer, H. Y. Xiong, J. M. Babik and M. H. Hecht (1993). "Protein Design by Binary Patterning of Polar and Nonpolar Amino-Acids." Science **262**(5140): 1680-1685.
27. Kannan, S. and M. Zacharias (2009). "Simulated annealing coupled replica exchange molecular dynamics-An efficient conformational sampling method." Journal of Structural Biology **166**(3): 288-294.
28. Kobayashi, N., S. Honda, H. Yoshii and E. Munekata (2000). "Role of side-chains in the cooperative beta-hairpin folding of the short C-terminal fragment derived from streptococcal protein G." Biochemistry **39**(21): 6564-6571.
29. Kobayashi, N., S. Honda, H. Yoshii, H. Uedaira and E. Munekata (1995). "COMPLEMENT ASSEMBLY OF 2 FRAGMENTS OF THE STREPTOCOCCAL PROTEIN-G B1 DOMAIN IN AQUEOUS-SOLUTION." Febs Letters **366**(2-3): 99-103.
30. Kuhlman, B., G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard and D. Baker (2003). "Design of a novel globular protein fold with atomic-level accuracy." Science **302**(5649): 1364-1368.
31. Lee, B. and F. M. Richards (1971). "Interpretation of Protein Structures - Estimation of Static Accessibility." Journal of Molecular Biology **55**(3): 379-&.

32. Levinthal, C. (1968). "Are There Pathways for Protein Folding." Journal De Chimie Physique Et De Physico-Chimie Biologique **65**(1): 44-&.
33. Lindahl, E., B. Hess and D. van der Spoel (2001). "GROMACS 3.0: a package for molecular simulation and trajectory analysis." Journal of Molecular Modeling **7**(8): 306-317.
34. Linderstrøm-Lang, K. U. and J. A. Shellman (1959). Protein structure and enzyme activity. The Enzymes. P. D. Boyer, Academic Press. **1:** 443-510.
35. Marklund, E. G., D. S. Larsson, D. van der Spoel, A. Patriksson and C. Caleman (2009). "Structural stability of electrosprayed proteins: temperature and hydration effects." Physical Chemistry Chemical Physics **11**(36): 8069-8078.
36. Mazumder-Shivakumar, D. and T. C. Bruice (2005). "Computational study of IAG-nucleoside hydrolase: Determination of the preferred ground state conformation and the role of active site residues." Biochemistry **44**(21): 7805-7817.
37. Munoz, V., P. A. Thompson, J. Hofrichter and W. A. Eaton (1997). "Folding dynamics and mechanism of beta-hairpin formation." Nature **390**(6656): 196-199.
38. Nadassy, K., S. J. Wodak and J. Janin (1999). "Structural features of protein-nucleic acid recognition sites." Biochemistry **38**(7): 1999-2017.
39. Okabe, T., M. Kawata, Y. Okamoto and M. Mikami (2001). "Replica-exchange Monte Carlo method for the isobaric-isothermal ensemble." Chemical Physics Letters **335**(5-6): 435-439.
40. Pande, V. S. and D. S. Rokhsar (1999). "Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G." Proceedings of the National Academy of Sciences of the United States of America **96**(16): 9062-9067.
41. Patriksson, A., E. Marklund and D. van der Spoel (2007). "Protein structures under electrospray conditions." Biochemistry **46**(4): 933-945.
42. Poole, A. M. and R. Ranganathan (2006). "Knowledge-based potentials in protein design." Current Opinion in Structural Biology **16**(4): 508-513.
43. Rakshit, S. and G. K. Ananthasuresh (2010). "A novel approach for large-scale polypeptide folding based on elastic networks using continuous optimization." Journal of Theoretical Biology **262**(3): 488-497.
44. Rothemund, P. W., N. Papadakis and E. Winfree (2004). "Algorithmic self-assembly of DNA Sierpinski triangles." PLoS Biol **2**(12): e424.
45. Rothemund, P. W. K. (2006). "Folding DNA to create nanoscale shapes and patterns." Nature **440**(7082): 297-302.
46. Roy, S., S. Goedecker, M. J. Field and E. Penev (2009). "A Minima Hopping Study of All-Atom Protein Folding and Structure Prediction." Journal of Physical Chemistry B **113**(20): 7315-7321.
47. Russ, W. P., D. M. Lowery, P. Mishra, M. B. Yaffe and R. Ranganathan (2005). "Natural-like function in artificial WW domains." Nature **437**(7058): 579-583.
48. Russ, W. P. and R. Ranganathan (2002). "Knowledge-based potential functions in protein design." Current Opinion in Structural Biology **12**(4): 447-452.
49. Satoh, D., K. Shimizu, S. Nakamura and T. Terada (2006). "Folding free-energy landscape of a 10-residue mini-protein, chignolin." Febs Letters **580**(14): 3422-3426.
50. Seeman, N. C. (1991). "Construction of three-dimensional stick figures from branched DNA." DNA Cell Biol **10**(7): 475-486.

51. Snow, C. D., E. J. Sorin, Y. M. Rhee and V. S. Pande (2005). "How well can simulation predict protein folding kinetics and thermodynamics?" Annual Review of Biophysics and Biomolecular Structure **34**: 43-69.
52. Socolich, M., S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner and R. Ranganathan (2005). "Evolutionary information for specifying a protein fold." Nature **437**(7058): 512-518.
53. Still, W. C., A. Tempczyk, R. C. Hawley and T. Hendrickson (1990). "Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics." Journal of the American Chemical Society **112**(16): 6127-6129.
54. Suenaga, A., T. Narumi, N. Futatsugi, R. Yanai, Y. Ohno, N. Okimoto and M. Taiji (2007). "Folding dynamics of 10-residue beta-hairpin peptide chignolin." Chemistry-an Asian Journal **2**(5): 591-598.
55. Terada, T., D. Satoh, T. Mikawa, Y. Ito and K. Shimizu (2008). "Understanding the roles of amino acid residues in tertiary structure formation of chignolin by using molecular dynamics simulation." Proteins-Structure Function and Bioinformatics **73**(3): 621-631.
56. Van der Spoel, D., E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen (2005). "GROMACS: Fast, flexible, and free." Journal of Computational Chemistry **26**(16): 1701-1718.
57. Wang, Y., D. S. Larsson and D. van der Spoel (2009). "Encapsulation of myoglobin in a cetyl trimethylammonium bromide micelle in vacuo: a simulation study." Biochemistry **48**(5): 1006-1015.
58. Xu, W. X., T. F. Lai, Y. Yang and Y. G. Mu (2008). "Reversible folding simulation by hybrid Hamiltonian replica exchange." Journal of Chemical Physics **128**(17): -.
59. Xu, Y., R. Oyola and F. Gai (2003). "Infrared study of the stability and folding kinetics of a 15-residue beta-hairpin." Journal of the American Chemical Society **125**(50): 15388-15394.
60. Zheng, J., J. J. Birktoft, Y. Chen, T. Wang, R. Sha, P. E. Constantinou, S. L. Ginell, C. Mao and N. C. Seeman (2009). "From molecular to macroscopic via the rational design of a self-assembled 3D DNA crystal." Nature **461**(7260): 74-77.
61. Zhou, R. H. (2003). "Free energy landscape of protein folding in water: Explicit vs. implicit solvent." Proteins-Structure Function and Genetics **53**(2): 148-161.
62. Zhou, R. H. and B. J. Berne (2002). "Can a continuum solvent model reproduce the free energy landscape of a beta-hairpin folding in water?" Proceedings of the National Academy of Sciences of the United States of America **99**(20): 12777-12782.
63.
64.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 724

Editor: The Dean of the Faculty of Science and Technology

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

Distribution: publications.uu.se
urn:nbn:se:uu:diva-121549

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2010