PROTEIN FOLDING AND MACROMOLECULAR DYNAMICS: FUNDAMENTAL LIMITS OF LENGTH AND TIME SCALES

Thesis by

Milo M. Lin

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2012

(Defended May 16, 2012)

ACKNOWLEDGEMENTS

I would not be the scientist I am today were it not for my advisor, Prof. Ahmed H. Zewail. It was he who challenged me, at the critical crossroads in my professional career, to reflect deeply about my true passion as a physicist; with his help, I found with great clarity that this passion resides in the areas of complexity and biology. In my time at Caltech, we have explored many topics revolving around these two themes. Despite the variety of systems and methods we studied, his questions to me were always the same: what is the first-order physics of the problem and can the essence of its complexity be understood in a predictive way? Often, the solution was not truly clear until it was reduced to its fullest possible simplicity, but no simpler (to paraphrase Einstein). His way of looking at problems in "Fourier space," whereby the most essential factors are first isolated before proceeding onto more detailed analysis, has informed how I think about and present concepts. Under his guidance I have learned to appreciate and hone many other important skills, including the ability to extract and act on the important developments across many fields, to evaluate new links between fields, and to collaborate with experimentalists on a meaningful level rather than to just use the fruits of their labor. Perhaps most important of all, his infectious excitement to pursue, with great success despite the early skeptics, the really big ideas, has infused me with the optimism and determination to tackle the difficult problems of science. For his profound influence on my scientific capabilities and outlook, I will always be grateful.

In addition, I would like to acknowledge my long-time colleague, Dr. Dmitry Shorokhov, from whom I have learned more than he realizes. His lessons, by example, on the effectiveness of rigor, consistency, and organization have stuck with me,

and I will always remember fondly our numerous discussions and joint expeditions into uncharted scientific waters.

I would like to thank my examining committee: Professors Michael C. Cross, Thomas F. Miller III, Douglas C. Rees, and Thomas A. Tombrello for their encouragement and helpful feedback. Prof. Tombrello, especially, has been a constant source of support and inspiration since I was in his famous Physics 11 freshman course a full decade ago. This course has been the research gateway for generations of aspiring physicists, including me. Over the years, Prof. Tombrello has helped each to find his own, sometimes surprising, passion.

I have gained much from my immersion in a collaborative and exciting research environment, and I am grateful to all of my colleagues, mentors, and mentees who have stimulated me to explore new ideas and fields beyond my own local minimum. They are too numerous to list, but Reuben Britto, Prof. Jeffrey Evanseck, Andreas Gahlmann, Adam Jermyn, Lars Meinhold, Omar Mohammed, Ebrahim Najafi, and Prof. Dongping Zhong are among them. These gifted scientists span a huge swath of scientific interests, and the truly interdisciplinary nature of Caltech has allowed my pursuits to evolve by intellectual excitement rather than by the expectations of any field. Indeed, the entire community at Caltech has made my life here very pleasant and productive. In particular, Ms. De Ann Lewis, by compensating for my bouts of absentmindedness and showing a genuine concern for my well-being, has acted above and beyond the level of professionalism and efficiency that distinguishes a great administrator.

Lastly, but just as importantly, this Thesis is dedicated to my friends and family, who believed in me, sacrificed for me, and reminded me that my scientific pursuits do not happen in a vacuum. To my grandfather, who lit the spark of curiosity before I could even walk, taking me to the outdoor markets and showing me that everything has a name and a function; to Mr. Farley, my middle school teacher, who nurtured this flame and showed me that intellectual discovery can be a lifelong adventure**;** to Elaine, who is my haven and my beacon of joy; to Juhwan, who stayed on with me at Caltech to fight the good fight; to Tony who came back to join in the fight; to Austin who found me more or less in the same state as when he last saw me; to my sister Kathy who is paving her own path and making me proud; to my parents, who gave me form and function, pouring more love into me than I can ever repay. These people in my life continue to remind me that curiosity can only take me so far, that in the long run I must have a purpose and that this purpose is inevitably shaped by the possibilities and limitations of the wider world. The way that we choose this purpose is a very human process despite the objective nature of science, and it is only this sense of purpose that can compel us to push over and over against seemingly hopeless obstacles, to nudge the world closer to its realization.

I would like to express my sincere gratitude to the National Science Foundation for funding of this research at Caltech, and to the Krell Institute and the US Department of Energy for my graduate fellowship.

ABSTRACT

In the present Thesis, physics-based models of protein folding at the secondary and tertiary levels are developed to resolve long-standing issues of protein folding kinetics. As discussed in the Introduction, the main objective is to explore fundamental limits imposed on the length and time scales involved in protein folding. Protein folding is also placed within the broader context of macromolecular dynamics, which is extensively studied in the unfolded, folded, and unfolding regimes for the key molecular motifs of cellular biochemistry, including lipids, nucleic acids, and proteins. The effect of the water hydration and temperature are systematically probed to elucidate the crucial role of the environment in macromolecular stability and dynamics. For a wide range of bio-molecular phenomena, the observed collective behavior is shown to arise directly from first principles. Throughout, the emphasis is on analytic results free of tunable parameters, supported by ensemble-convergent computational simulations, and corroborated by experimental evidence.

TABLE OF CONTENTS

NOMENCLATURE

Space metric. Angstrom (**Å**) = 10^{-10} meters; nanometer (**nm**) = 10^{-9} meters; micron (**μm**) = 10^{-6} meters.

Time metric. Femtosecond (**fs**) = 10^{-15} seconds; picosecond (**ps**) = 10^{-12} seconds; nanosecond (**ns**) = 10^{-9} seconds; microsecond (**μs**) = 10^{-6} seconds.

Ultrafast process. A process occurring in less than one nanosecond.

Energy metric. Kilocalories per mol (**kcal/mol**) = number of kilocalories per 6.022×10^{23} (Avogadro's number) copies of the species. 1 kcal/mol = 4184 joules/mol ≈ 0.04336 electron volts.

Nucleotide. Molecules composed of a nucleobase (nitrogenous base), a five-carbon sugar (either ribose or 2-deoxyribose), and one phosphate group.

Deoxyribonucleic acid (DNA). Polymers composed of four types of 2-deoxyribose nucleotides (A, T, C, G) linked by ester bonds. The sequence of bases in the polymers encodes the genetic information. The DNA segments carrying this genetic information are called genes. Due to complementary hydrogen bonding between A and T, and C and G, two anti-parallel polymers of DNA form a complementary DNA double helix.

Ribonucleic acid (RNA). Polymers composed of four types of ribose nucleotides (A, U, C, G) linked by ester bonds. RNA is used primarily to transcribe and translate the genetic sequence from DNA to proteins.

Amino acid. Molecules containing an amine group, a carboxylic acid group, and a side chain that is specific to each amino acid.

Polypeptide. A single linear polymer chain composed of twenty types of naturally occurring amino acids bonded together by peptide bonds between the carboxyl and amino groups of adjacent amino acids.

Protein. A polypeptide present in living organisms. The sequence of amino acids in a protein is defined by the sequence of DNA called the gene.

Cell. The basic structural and functional unit of all known life. It is the smallest unit of life that is classified as a living thing. In plants and animals, cell sizes range from 1 to 100 μm.

Cell nucleus. A membrane-enclosed organelle containing most of the cell's genetic material.

Chapter 1

INTRODUCTION

Starting at the ecosystem level, and magnifying to the population, organism, organ, cellular, and macromolecular scales, the lengthscale of biological complexity spans 12 orders of magnitude from nanometer to kilometer. In conjunction, the time scale of biological dynamics ranges over 17 orders of magnitude, from (sub)nanoseconds to years. Amazingly, life maintains both complexity and robustness, which is defined as structural and functional reproducibility, over this vast length and time domain (Figure 1.1, top). The root of this complex robustness can be traced, at the smallest scale, to the nano-machines called proteins, which are the cleaners, builders, motors, messengers and transporters of the cell (1). Proteins are linear chains composed of 20 types of amino acids; the number of possible sequences increases exponentially with the length of the protein. The particular sequence of every protein is encoded by a corresponding segment of DNA, which, in the eukaryotic cell, is stored in the nucleus. When a particular protein is needed, its DNA blueprint is copied and the code translated into the matching amino acid sequence (Figure 1.1, bottom left). The linear chain is then molded into a 3D conformation, or fold, that will carry out the protein's particular function (Figure 1.1, bottom center and right). Called the *native fold*, this structure is stabilized by chemical forces both within the protein and between the protein and its surrounding environment (mostly water) (2).

Life requires proteins to be functionally reliable, so each protein sequence has evolved to possess a unique fold (or set of folds) that is more stable than all other possible

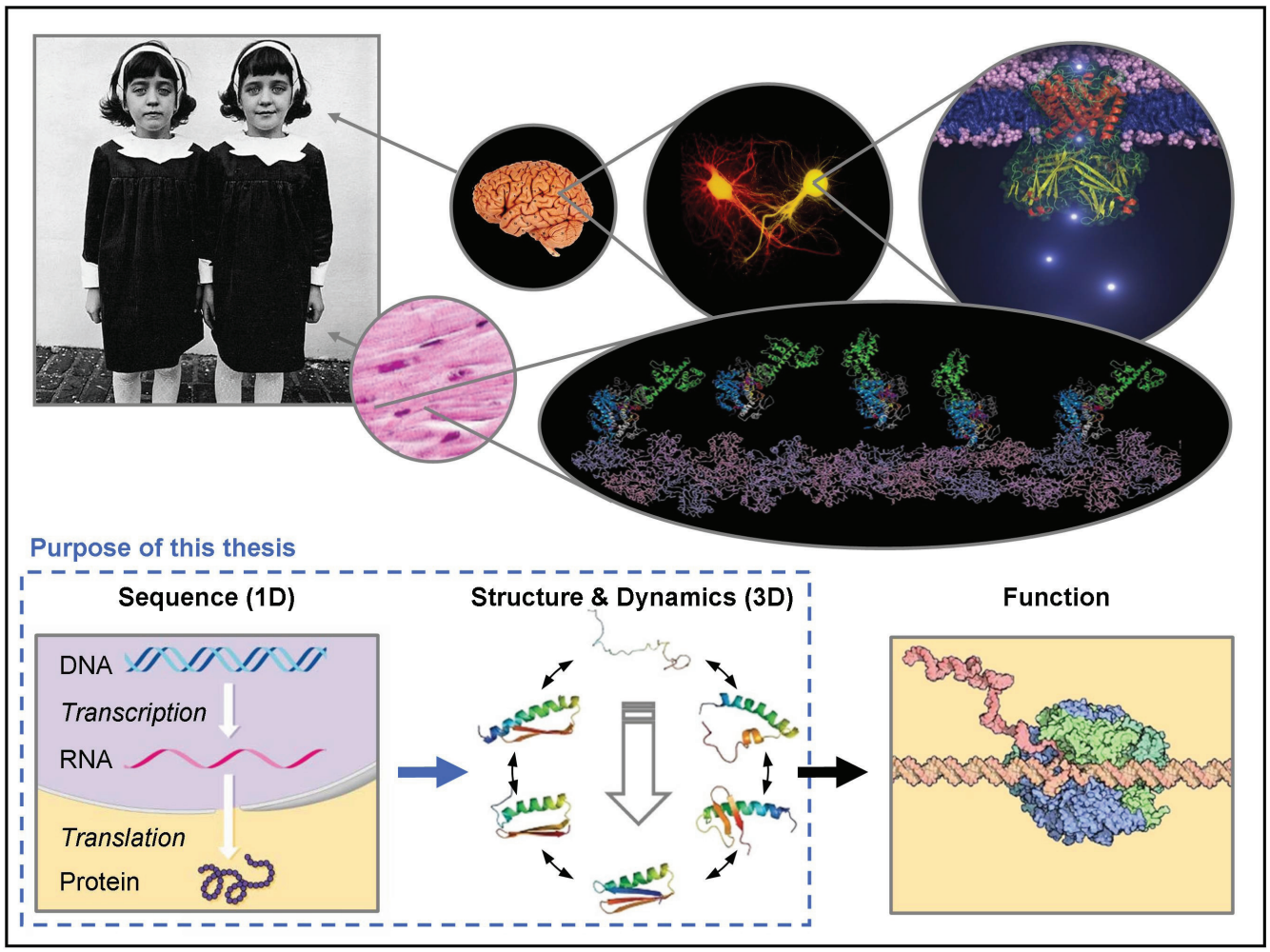

Fig. 1.1. Proteins: the molecules underpinning biological complexity. Being largely responsible for the structural and functional processes within and between cells, directly coded for via the genetic sequence, and often able of self-assembly (folding) into their highly inhomogeneous functional forms with high fidelity, proteins are distinguished as the building blocks of life. Building upon proteins' functional diversity and genetic programmability, progressively larger scales of biological structure are able to inherit the unique combination of precision, robustness, and complexity which originates with proteins and extends ultimately to the scale of the organism (top). Biomolecules such as proteins share a common theme of being intrinsically one-dimensional chains composed of a sequence of subunits (bottom left), which via conformational rearrangement (folding) attains a functional form (bottom center) optimized for its function (bottom right). In the case of proteins, whose structure is most complex and function most ubiquitous, the folding process often occurs spontaneously in aqueous environment. The purpose of this thesis within the context of molecular biology is to understand the dynamical process by which one-dimensional sequence information is transformed into three-dimensional structural information (blue inset), as well as to elucidate the roles of intrinsic (dis)order, entropy, temperature and solvent.

folds. Because native folds are highly complex and irregular, their stability is one of the most awesome examples of how natural selection can lead to creations that are both intricate and robust. It therefore came as a shock when Anfinsen discovered that many proteins can fold by themselves without the aid of any cellular machinery (3). There are simple things that self-organize (freezing water) and complicated things that external machinery reliably makes (airplanes). There are also many examples of complicated phenomena that non-reliably self-organize (the weather). Proteins are unique because they are complicated structures that *reliably* self-organize. This seemingly impossible engineering feat at the molecular level is the scaffolding that allows life to be simultaneously complex and efficient. In this sense, proteins are responsible for robust biological complexity at all larger length scales. We limit our focus in this Thesis to the self-folding of macromolecules such as proteins, and will not discuss the equally interesting subject of their function.

The means by which a protein attains its secondary and tertiary structure, called the protein folding problem (4), is still unclear. Because the fundamental physics of the folding mechanism involves the self-interactions of a polymer with twenty kinds of different constituent monomers (and the solvent), it seems tantalizingly plausible that protein structure can be predicted from first principles. Predicting protein structure would facilitate understanding of protein functions in the cell. On the practical side, a solid physical theory for folding would allow the design of protein sequences that can fold into desirable structures useful in medicine, nano-machinery, or materials science. Equally importantly, a complete theory of protein folding would provide valuable insight into the general problem

of finding stable equilibrium solutions with which to describe a heterogeneous system characterized by a very large and complex conformational space.

The problem of protein folding can be decomposed into two complementary tasks. The first one implies finding the equilibrium ground state(s) corresponding to the protein sequence under study. With exponentially large number of degrees of freedom, and a lack of symmetry, it seems that the ground state should be degenerate and therefore the protein should be in a large number of different structural states at equilibrium rather than a unique native structure crucial for effective protein function (5). The second problem involves identifying the mechanisms and time scales for folding. Levinthal noted that a random search for the native-fold structure within the conformational state space would result in longer sampling times than the age of the universe even for small proteins, rather than the microsecond-to-second time scale typical of protein folding (6). Thus, proteins constitute a subset of all possible polypeptide sequences which possess both native state specificity and fast folding. The molecular basis for these crucial properties is the subject of ongoing investigation.

The objective of this Thesis is to help address, from a theoretical perspective, the latter problem: how do proteins fold (Figure 1.1, bottom inset)? We will tackle this problem at all the scales involved, as well as place the process of protein folding in the general context of macromolecular dynamics. Specifically, the following questions will be asked: (a) what are the characteristic rates and rate limiting mechanisms of structure formation on different length scales within proteins? (b) what is the quantitative resolution of Levinthal's paradox? (c) how is protein folding different from other types of

conformational dynamics? (d) what is the role of water in protein folding and dynamics? In the process of answering these four questions, we find that they are intimately connected.

We review the essential experimental and theoretical background in Chapter 2, and summarize the recurrent methodologies employed in this research in Chapter 3. In Chapter 4, we elucidate the nature of structural dynamics on different length scales within macromolecules across a broad range of conditions and species. The conditions include the unfolded and folded states as well as the unfolding transition at various temperatures and in the presence and absence of solvent; the species we studied represent the following classes of biological macromolecues: fatty acids, nucleic acids, and proteins. The temperature-induced processes and the accompanying temporal behavior are associated with qualitatively different transformations taking place at different length scales across macromolecules as well as in different solvation states. Stimulated by elucidation of these intrinsic properties of macromolecular dynamics in Chapter 4, in Chapter 5 we tackle the unresolved issues of protein folding at both the secondary and tertiary structural levels, identifying both the speed limit and length limit of protein folding. In particular, it is found that the hydrophobic force quantitatively resolves the Levinthal paradox. Together, these two Chapters clarify the current understanding of macromolecular dynamics and protein folding, as well as the essential and active role of water. The resulting picture is based on theoretical results, complemented by state-of-the-art computational simulations, and supported by experimental evidence.

Chapter 2

BACKGROUND

2.1 EMPIRICAL

2.1.1 Synthesis and Structure of the Key Biomolecules

Nucleic acids. Deoxyribonucleic acid (DNA) encodes for proteins and is therefore the genetic blueprint for all life, responsible for both diversities and similarities throughout the biosphere. The molecular structure of DNA is a chain composed of four types of 2-deoxyribose nucleotide bases, adenine (A), thymine (T), cytosine (C), and guanine (G), linked together by ester bonds, with the sequence of nucleotides in the chain referred to as the genetic sequence. Two strands of DNA chain combine to form a double helix stabilized by complimentary interstrand hydrogen bonds between adenine and thymine (A-T base pair) and cytosine and guanine (C-G base pair). Because the DNA sequence is organism-specific, and because it does not vary noticeably from one cell to another under normal circumstances, DNA macromolecules represent an invaluable source of biological, medical and forensic information which has been subject to intense investigation since the discovery of their spatial structure (7) from X-ray fiber diffraction patterns (8, 9) in 1953. With single-crystal X-ray diffraction (10, 11), the macromolecular architecture was resolved at the atomic level, revealing, e.g., the two hydrogen bonds characteristic of the A-T base pair, as well as the three hydrogen bonds characteristic of the C-G base pair. In addition to *base pairing* between the two strands of the double helix, *base stacking* of π-

orbitals along each strand helps stabilize the macromolecular structure. Manipulations of DNA sequence and structure via direct genetic engineering are now widely used to improve crops and livestock quality (12), as well as to produce biological tissues and substances with desired characteristics (13). In addition, the A-T and C-G base-pair specificity renders DNA an ideal material for computing (14-16) and structural self-assembly at the nano-scale (17-19). The double stranded helix of the DNA is unique in many aspects, including chemical and thermodynamic stability, packing efficiency, and site-specific strand separation which prevent harmful mutations, facilitate folding, and allow for transcription, respectively. For example, there are 23 pairs of DNA macromolecules packed into the micron-sized nucleus of a human cell (20), despite the fact that, if pulled, they would stretch to 3 meters (21). Despite such tight packing, DNA is easily accessed and avoids tangling due to its fractal folding topology (22). Both macromolecular structure and (un)folding dynamics of DNA are, therefore, central to our understanding of a variety of processes taking place in vivo.

The ribonucleic acids (RNA), which are mainly involved in the transcription and translation process, but also have other roles, e.g., in signaling or catalysis, are structurally identical with DNA, except for the presence of a hydroxyl group at the 2’ position of the ribose sugar site of each nucleotide. This chemical distinction causes the RNA double helix to adopt the A-form geometry characterized by a very deep and narrow major groove and a shallow and wide minor groove (23), rather than the more even B-form most commonly observed in DNA (24). RNA macromolecules, in addition to forming the double helix motif, also generally possess heterogeneous base-pairing patterns and structures (25).

Proteins. Largely responsible for the mechanical and chemical work done in the cell, proteins are linear chains of amino acids connected by peptide (covalent) bonds. Although there are only twenty different types of amino acids available to life, a protein of length L may be characterized by 20^L amino acid sequences; this exponential variability in sequence is ultimately responsible for the variability of function among proteins. The particular sequence of a protein is encoded by a corresponding sequence of nucleic acids of the DNA in the nucleus of the cell. To synthesize the protein, mRNA transcribes the nucleic acid sequence from the DNA. The information from the mRNA is translated into a protein sequence within large proteins called ribosomes, with tRNA translating between nucleic-acid and amino-acid sequences by mapping every three-nucleic-acid segment on the mRNA to a unique amino acid on the protein chain. The sequence of amino acids that defines the protein architecture is called the primary structure. The 3D spatial structure of the protein is characterized by periodic and mostly local interactions, called secondary structure, global topological and geometrical morphology called tertiary structure, as well as multi-protein aggregation called quaternary structure (Figure 2.1). Large proteins are typically composed of structural subunits, called domains, which tend to fold independently (26). Similarly to DNA, folded proteins are tightly packed, with cavities accounting for less than 2% of their total volume (27), such that the interiors of globular proteins are as dense as those of crystals of their constituent amino acids (28).

In 1951, based on the structures (29) of amino acids and the planar nature of the peptide bond, Pauling and Corey correctly proposed the existence of the α-helix and β-strand motifs, both stabilized by hydrogen bonds (Figure 2.1, left), as the fundamental building blocks in protein secondary structure (30, 31). Almost all proteins are

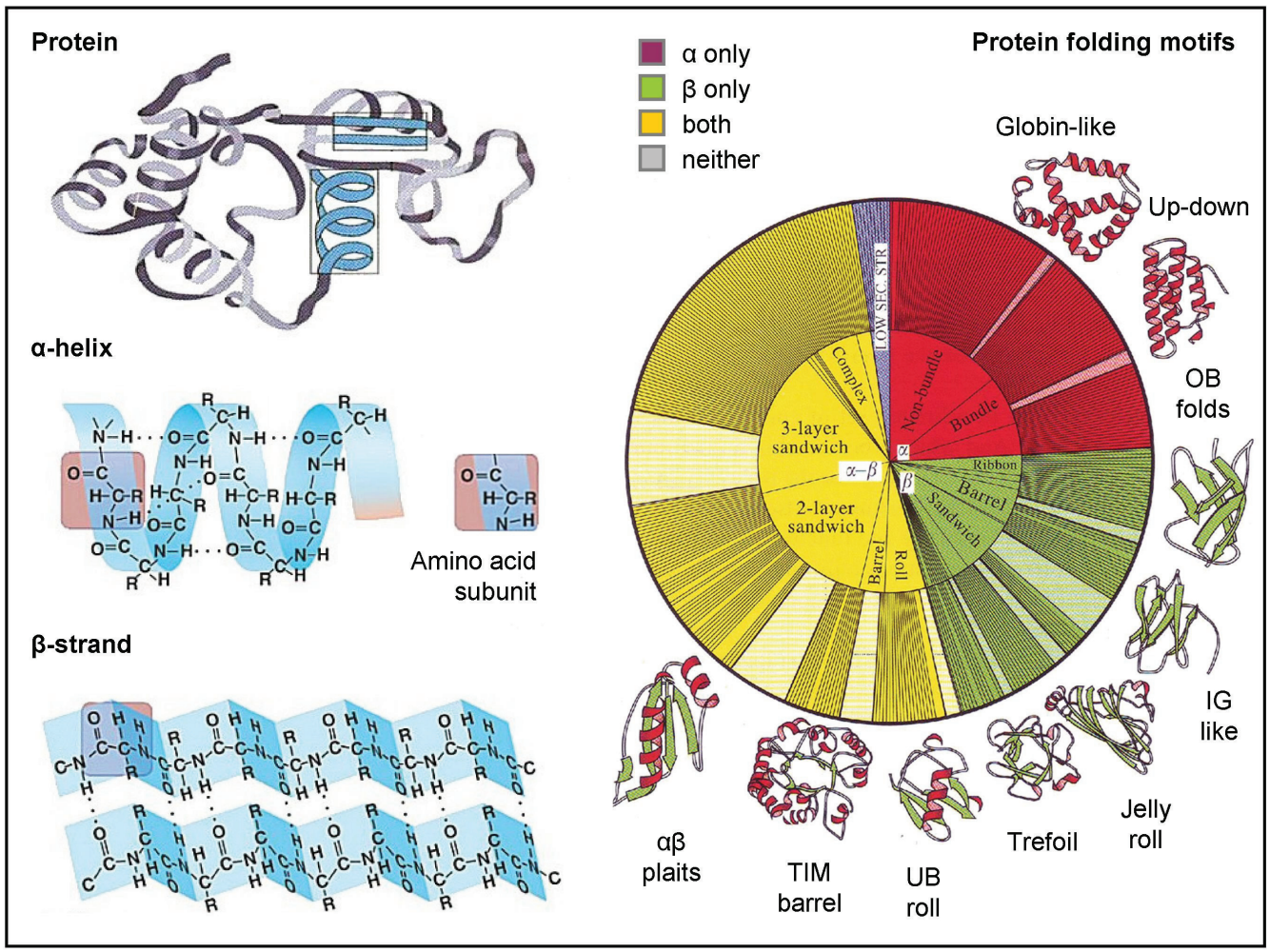

Fig. 2.1. Protein structure. Proteins possess secondary structure consisting of repeating periodic motifs stabilized by hydrogen bonds along the chain backbone (the two main types being α-helices and β-strands), as well as tertiary structure defined by the overall topology and geometry (left). The α-helix is the only prevalent periodic local. Virtually all proteins contain at least one type of secondary structure (red, green, yellow), and over half contain both (yellow; right). In contrast, there are numerous possible tertiary structures, with only a few topological classes shown (right).

characterized by at least one type of secondary structure (5), whereas the number of possible tertiary structures is vast and grows with protein size (Figure 2.1, right). The native protein fold is stabilized by a range of interactions including the covalent disulfide bridge, hydrogen bonds, purely electrostatic interactions, and van der Waals forces (5). In addition, the forces governing internal rotation along the chain backbone, as well as solvent-solute hydrogen bond formation, are known to be powerful forces influencing the folding process (5, 32). Despite the exponentially large conformational space and complicated interactions involved, proteins can fold into their complex 3D native folds with high fidelity, and many can do so without the aid of molecular chaperones guiding this process. The task of constructing a predictive model of protein folding, called the *protein folding problem*, is one of the central and elusive goals of modern science (4). The unique native fold(s) of each protein are structurally and electronically suited to the fine-tuned functional purpose of the protein, which may include signaling, metabolism, degradation, catalysis, infrastructure, mechanical motion, or regulation (5). Protein misfolding not only leads to inefficiencies in these tasks, but may also result in potentially toxic side-effects such as aggregation, e.g., in amyloid formation (33).

There are currently over 75,000 protein structures in the online Protein Data Bank (PDB) depository (34), which are categorized into about 1,300 topologically distinct folding conformations. The known proteins range in size from tens of amino acids to over 5,000 amino acids, with mean and standard deviation of 640 and 960 amino acids, respectively; in contrast, most domains that comprise proteins are less than 200 amino acids in size (34). In addition to studying the naturally occurring proteins, engineering novel proteins with desired functionality, especially for drug design (35, 36), is also a

major avenue of research (37-39). However, the complex interplay of many different forces giving rise to both secondary and tertiary structure, in contrast to the dominant Watson-Crick interactions in the case of nucleic acids, makes protein structural prediction and design much more difficult than that of DNA (40); the bottleneck of protein engineering is therefore the protein folding problem.

2.1.2 Experimental Measurements

To observe the structure and dynamics relevant to macromolecular (un)folding, experimental techniques with sufficient spatial and temporal resolutions are needed. Generally, different techniques involve trade-offs between these two desired properties, and one or another is chosen depending on the specific system under study and question of interest. For example, X-ray crystallography can capture the (static) ground-state structures of proteins to atomic resolution, whereas fluorescence spectroscopy can probe the dynamics of coarse-grained structural features, such as intra-chain contacts, with picosecond temporal resolution. Although the work reported here is purely theoretical, understanding the possibilities offered by, as well as limitations imposed on, the experimental methods is instrumental in evaluating the predictions made using analytical or computational methods. As we have found, experimental methods can also aid in identifying attractive theoretical coarse-graining approaches (see Sections 3.3 and 4.2 for the method and its application, respectively). Below, we briefly outline the experimental techniques relevant to the studies reported here, which broadly fall under the two major categories: spectroscopy and diffraction.

Spectroscopy. Spectroscopic techniques rely on the dependence of the electromagnetic absorption or emission properties on the state of the macromolecule under study. During the course of **fluorescence spectroscopy** measurements, the distance between two regions of a macromolecule is revealed by enhancement (or quenching) of fluorescence of a probe attached to one of the regions, upon close proximity to the other. A wealth of information about temporal evolution of the conformational state of the macromolecule can thus be obtained (41, 42). **Infrared (IR) spectroscopy** relies on the fact that the vibrational spectra of macromolecules depend on their conformations. To probe the secondary structure, the band in the IR spectra corresponding to the stretching frequency of the C=O bond of the peptide backbone is monitored. Because formation of, e.g., α-helices changes the resonant frequency of this vibrational mode, the resultant spectral shift can act as a gauge of the secondary structure content. In addition, by engineering peptides with isotope substitutuents occupying desired sites within a macromolecule, the (un)folding behavior of a specific site can be pinpointed. For example, by substituting for ^{13}C at a C-site of interest, the heavier mass of the isotope results in a shift of the vibrational frequency in the amide I frequency band, and any change of this shifted component of the spectrum can be attributed to the conformational dynamics at the substitution site (43-45). Rather than directly inducing vibrational motions in the IR domain, **ultraviolet resonance Raman spectroscopy (UVRS)** excites the studied macromolecule at a frequency that overlaps with a particular electronic absorption band. If the probing laser frequency is resonant with the excited electronic state, these excitations will selectively transmit their energy to certain amide vibrational modes, depending on the molecular conformation. For example, the intensity and frequency

change of these selectively enhanced vibrational modes are characteristic of secondary structure content, as well as solvent exposure of aromatic residues (46, 47). **Circular dichroism (CD) spectroscopy** takes advantage of the fact that protein backbone conformation determines the relative absorbance of right- and left-circularly polarized light in the far-UV spectral domain. Right-handed and left-handed α-helices, as well as β-strands and random-coil configurations, are all characterized by distinctive absorption vs. frequency profiles (or CD "fingerprints"). As a result, CD spectroscopy is the most commonly used method to determine the (equilibrium) secondary structure content. During the course of time-resolved CD measurements, rapid-mixing stopped-flow techniques are typically used to induce protein (un)folding by adjusting the solvent composition (48). Due to the (diffusion-limited) solvent mixing rate, the temporal resolution of CD is confined to the microsecond time scale. **Nuclear magnetic resonance (NMR)**, which utilizes knowledge of magnetic spin-spin coupling characteristic of different nuclei to translate the measured spin relaxation time following excitation by a radio-frequency magnetic field into information about the inter-nuclear distances, is the most prevalent alternative to crystallography in determining protein structure. Although the method is restricted to small proteins (tens of kilodaltons), it is the only method available to date with which to determine the atomic-resolution structures of macromolecules characterized by poor crystallinity and/or large disorder in the solid state. Temporal resolution of NMR measurements is typically limited to time scales longer than 100 microseconds (49). Finally, the ability to move and manipulate nanoparticles by attaching them to optically trapped beads has opened up the way to studying individual macromolecules. The main advantage of such studies is the ability to

tease out step-wise or reversible motions associated with, e.g., DNA transcription (50) or rotation of protein motors (51), that would otherwise be smoothened out by ensemble averaging. Notably, combining this technique with fluorescence spectroscopy has enabled conformational probing of individual molecules (52).

Recent studies of time-resolved macromolecular (un)folding using the above spectroscopic methods have been made possible by advances in the laser-based experimental techniques leading to the modern temperature-jump (*T*-jump) instrumentation. During the course of a laser-induced *T*-jump experiment (53), a heavy-water (D_2O) solution containing the biomolecules under study is heated by a near-IR nanosecond laser pulse, typically by tuning the exciting-pulse frequency to match that of the first vibrational overtone of water (the H–O stretching frequency). The water molecules are excited within picoseconds, and they subsequently relax back to the ground state through a non-radiative process. Because proteins do not directly absorb the near-IR radiation energy, their heating occurs via ultrafast (thermal) absorption of energy from the water, which precludes undesirable side effects such as photo-ionization or localized superheating. The temperature change is calibrated by monitoring changes in the infrared absorption of water, of which the temperature-dependence is well known. Alternatively, changes in fluorescence emission of a dye with known temperature-dependent quantum yield can also be used.

Diffraction. The diffraction-based experimental techniques take advantage of the fact that the superposition of waves scattered from particles, such as atoms, is the Fourier transform of the inter-nuclear distances ("scattering terms") defining the geometric arrangement of

the atoms. Consequently, the geometric structure can, in principle, be obtained by inverse-Fourier transforming the resulting wavefunction. Because charge-coupled device (CCD) detectors are known to register only the amplitude of the function (the complex wavefunction is projected onto the space of a real observable during the course of photon detection), the loss of the phase information implies that the inverse Fourier transform can yield only a pair-wise distance density distribution (the so-called *radial distribution function*), rather than a set of exact atomic coordinates; to extract information about scattering phases, the analysis needs to be augmented by using heavy-atom substituents and/or system-specific constraints (54). The relatively weak scattering of electromagnetic radiation limits the practical usage of diffraction measurements to crystalline samples, for which periodically repeating inter-nuclear distances give rise to sharp (Bragg) diffraction peaks in reciprocal space. Ever since the pioneering work of von Laue (55) and Bragg (56) in harnessing the above properties of wave interference, X-ray diffraction, which utilizes electromagnetic radiation of wavelength comparable to the inter-atomic spacing, has been used to reveal the spatial structure of matter. In regard to biomolecules, beginning with the pioneering study of myoglobin (57), X-ray crystallography has produced the vast majority of protein and nucleic acid structures known to date (58). However, the need to perform X-ray diffraction measurements on crystalline samples restricts the study of biological macromolecules to static, highly artificial environments (which are often difficult to prepare), and therefore limits practical applications of the method in the study of macromolecular dynamics and function.

Fortunately, because of the particle-wave duality ($\lambda = h/p$, where λ is the wavelength of a particle of momentum p and h is Planck's constant), diffraction

experiments are not limited to electromagnetic radiation. Notably, in the case of electrons being scattered, electromagnetic lenses can be constructed which inverse-Fourier transform the signal before detection to generate a real-space image, thereby capturing the structural information (59). Due to the increased momentum associated with λ approaching the magnitude of inter-atomic spacing, the scattering cross-section of electrons is much larger than that typical of X-rays. This may be a serious disadvantage from the perspective of radiation damage and multiple scattering processes (60), but can be advantageously used to study non-crystalline samples of, e.g., isolated or membrane-bound macromolecules for which dynamical information can be obtained (61). In contrast, the X-ray scattering cross-section is too small to yield a satisfactory signal-to-noise ratio for such systems. In addition, the accompanying radiation damage can be reduced by performing experiments on samples embedded in vitrieous (non-crystalline) ice and/or performing studies of large 3D macromolecular structures with the aid of other sources of information, such as atomic-resolution X-ray or NMR structures of some of the constituent structural domains (62). Recent experiments aiming at delivering intact biological macromolecules into the gas-phase, based on the technique of laser desorption developed in this laboratory (63-65) and in part motivated by the simulations reported below (Section 3.3 and 4.2), pave the way for prospective electron diffraction studies of biomolecular dynamics in vacuo.

2.2 THEORETICAL

2.2.1 Statistical Mechanics of Macromolecules

Much of the theoretical understanding of protein folding comes from applying statistical mechanical analysis to some coarse-grained representations of proteins. Usually, the goal is to construct a representation of the system such that a physically meaningful partition function can be obtained. For biological systems, the conditions of interest are constant temperature and pressure (298 K and 1 atm, respectively). The number of solvated macromolecules may be fixed at one if the effects of crowding and aggregation are disregarded. The solvent molecules are either treated as part of the temperature bath or constitute a unit cell around the solute which is propagated in space using periodic boundary conditions. Thus, structural dynamics of biological macromolecules are generally modeled using canonical ensembles of constant number, temperature, and pressure (*NPT*). After dividing up the macromolecular conformational space into N discrete sub-states (where N is chosen based on the level of coarse graining desired for the system under study), the canonical partition function, Z, is represented by the sum of the statistical weights of all sub-states:

$$Z = \sum_{i=1}^{N} \exp\left(-G_i / kT\right), \qquad [2.1]$$

where k and T are Boltzmann's constant and absolute temperature, and the Gibbs free energy is given by $G_i = H_i - TS_i$, with H_i and S_i, being the enthalpy and entropy of the i-th sub-state. Each sub-state is typically defined as consisting of N_i distinct microstates, each of which is characterized by (identical) enthalpies H_i; then, the entropy of the i-th sub-state is

$S_i = k \ln N_i$. For example, each sub-state of a chain that can make contact bonds with itself can be defined by a unique set of contact bonds called a contact map; in this case, the number of microstates in a given sub-state of the chain equals the number of distinct conformational states consistent with the corresponding contact map. For a protein with a single folded (native) sub-state F, the partition functions of the folded and unfolded ensembles are given by $Z_F = \exp(-G_F/kT)$ and $Z_U = Z - Z_F$, respectively; the equilibrium fraction of properly folded proteins is therefore Z_F/Z. The free energies of states, such as the folded and unfolded states, are $G_F = -kT\ln Z_F$ and $G_U = -kT\ln Z_U$, respectively. A macromolecule is considered folded when $Z_F/Z > 1/2$ which corresponds to $G_F < G_U$. Often, the main challenge in calculating the free energy differences between individual states of the system under study is related to defining the sub-states and calculating their entropies, whether analytically (Sections 4.2.1, 5.1.2, and 5.2) or by explicitly sampling the states computationally to obtain the population fractions and, by taking the logarithm, the free energies (Sections 4.2.1, 5.2). The analytical method generally involves some form of coarse-graining approximation in mapping to the sub-states, such as, e.g., representing the continuous conformation space on a discrete lattice (66). Alternatively, using the Jarzynski equality theorem (67):

$$\left\langle \exp\left(-W/kT\right)\right\rangle = \exp(-\Delta G/kT), \qquad [2.2]$$

where W and ΔG are the work applied to transition between two states and the free energy difference between the states, respectively, and the angular brackets denote ensemble averaging. In the adiabatic limit in which the transition is taken to be slow enough to be reversible, the work equals the free energy difference, whereas for more

realistic irreversible work, the sum over the exponent of the work ensures that the average work exceeds the free energy difference, with the excess energy dissipated as heat. Qualitatively, this is identical to the conclusion obtained by invoking the second law of thermodynamics. Yet, quantitatively, in the form of an equality, Equation 2.2 can be used to computationally obtain the ΔG between two states of a potentially complicated system by repeatedly inducing transitions from one state to the other and averaging the exponent of the work performed. The above theorem, which has been extended beyond thermodynamic ensembles to general microscopically (but not necessarily macroscopically) reversible stochastic systems (68), may be used to accurately determine free energy differences for ensembles that are too large or too complex for analytical modeling or unbiased computational sampling**.**

For polymer chains such as proteins, the different conformational sub-states, which are defined in accordance with the coarse-graining representation chosen (e.g., denatured vs. extended, folded vs. unfolded, or lower-energy vs. higher-energy native conformations), are favored or disfavored depending on the intrinsic properties of the chain, and other commonly varied parameters. The phase diagram representing the boundaries between the sub-states as a function of parameters such as solvent environment, chain density, temperature, and pressure, is generally nontrivial to obtain. For example, when distinguishing the sub-states by geometric size (extended vs. compact), the radius of a macromolecular chain of length L will scale as $R \propto L^{\nu}$. In the absence of solvent, the chain will form a globule consistent with a random walk: $\nu > 1/2$, with the inequality arising from excluded-volume (steric-repulsion) effects that tend to

swell the globule and decrease ν. When placed in a solvent in which solvent-chain interactions are disfavored relative to intra-chain interactions, the chain will tend to contract to avoid contact with the solvent. When the solvent-induced contraction exactly balances the excluded volume effect, $\nu = 1/2$ and the system is said to be at the θ-point (69). If the solvent-chain interaction becomes even more unfavorable the chain is said to be in a poor solvent, and the globule contracts as the solvent-chain interaction energy increases until the globule squeezes out all the solvent and its volume is roughly equal to that of the chain, $\nu = 1/3$. On the other hand, starting from the θ-point, if the relative solvent-chain interaction free-energy is decreased, the chain is said to be in a good solvent and will swell. In the limit of the energy of solvent-chain interactions approaching that of the intra-chain interactions, i.e., when the solvent is composed of chain monomers, $\nu = 3/5$ (69).

The nature and intensity of solvent-chain interactions can be manipulated not only by changing the solvent type, but also by adjusting the temperature of the system. For example, many biological macromolecules experience poor solvent conditions when placed in aqueous solution at room temperature (70), leading to compact hydrophobic collapse; this phenomenon is instrumental in regard to both the thermodynamic and kinetic aspects of protein folding (Sections 2.2.2 and 5.2, respectively). The poor solvent condition arises from an effective unfavorable solvent-chain interaction free energy penalty due to maximization of *solvent entropy*, which favors compact chains because they minimize trapping of solvent molecules. If the temperature is decreased, the statistical weight of this solvent entropy contribution decreases, thereby favoring the extended chain. At a critically low temperature, the chain enters the good solvent regime

and unfolds to favor the extended ensemble: $G_{compact} > G_{extended}$. As the temperature is increased the free energy component due to conformational *chain entropy* increases, and, above the unfolding temperature, the free energy of the extended sub-state again becomes lower than that of the compact sub-state: $G_{compact} > G_{extended}$. Therefore, proteins experience both cold and heat denaturation below and above room temperature, respectively (71).

Importantly, the many degrees of freedom characteristic of large biopolymers often give rise to dynamical conformational fluctuations that determine the mechanism and time scale of the thermodynamic transitions between individual sub-states of interest. These fluctuations occur on many length scales ranging from (atomic-scale) local rotations and vibrations to (molecular-scale) global conformational motions. Concerted local motions can drive global dynamics; conversely, the global motions can transfer energy to ever-smaller scales until it is dissipated as heat on the atomic scale, similarly to the hierarchical energy cascade of turbulent flow (32, 72). Because larger-scale motions encounter more of a steric frustration, at a sufficiently low temperature, called the *glass transition temperature*, such motions are "frozen out," so that the conformational barrier prevents the kinetic accessibility of the lowest free energy states (73).

2.2.2 Protein Folding

The considerations of Section 2.2.1 are applicable to macromolecules in general. Yet, protein folding requires thermodynamic stabilization and kinetic accessibility of a single (or, in some cases, a few) unique native state(s), which usually includes only local

perturbations around *one* microstate. This presents significant challenges beyond those associated with the more typically studied transitions between high-entropy, coarse-grained sub-states (e.g. “compact” or “extended”). A comprehensive solution to the protein folding problem outlined in Section 2.1.1 would be a method which takes the amino acid sequence (primary structure) of a protein as input, and provides the secondary and tertiary structure, the folding time, and the dominant folding mechanisms and pathways as output. In contrast, because of the difficulty of the protein folding problem, this area of research is currently dominated by “means-based” and “ends-based” theories and methodologies. Whereas the former seek to understand the process from physical principles, the goal of the latter is to enable native state prediction by any reliable means. The latter approach includes both physicochemical considerations as well as bioinformatics, which employs sequence and structure databases to extract useful patterns. For example, the *sequence alignment* technique exploits the idea that two regions of the sequence that are evolutionary co-conserved are likely to be close to each other in the native state structure (74). By systematically analyzing these evolutionary correlations, sequence alignment has been shown to produce enough pair-wise contacts to constrain the folding topology, which, when combined with geometrical and electrostatic constraints, can lead to accurate prediction of macromolecular structure (75). The most successful structure prediction algorithms make use of effective heuristics to intelligently search the conformational space as well as to score the candidate structures; for example, a successful technique is based on finding local sub-sequences that match sequences in a database of known structures whereas the degrees of freedom of the sub-sequence are fixed during the search (76, 77). Here, we are predominantly concerned with the means-based (physics) investigations rather

than with the ends-based statistical algorithms, although it is hoped that attaining ever-deeper understanding of the underlying physics will one day lead to accurate macromolecular structure prediction, thereby unifying the field of protein folding.

It is remarkable that structural and energetic changes caused by conformational interconversions in biological macromolecules are controlled by a subtle balance of weak forces such as hydrogen bonding, electrostatics, dispersion, and hydrophobic interactions (78). The net result is the emergence of a unique function out of complexity. One example is the rotational motion in biopolymers, which is known to largely determine the stability of the secondary and higher-order molecular structures as well as to control the dynamics of their (un)folding. The potential energy barriers to rotation about the backbone torsional (Ramachandran) angles (79), which define the conformation of a protein, are surmounted and structures are stabilized by formation of bonding intramolecular interactions (80). These interactions, in their turn, are weak enough to be disrupted at the expense of a few kcal/mole, or ~0.1 eV, underscoring the fluctional nature of the native fold. The conformational dynamics, which is at the heart of complex macromolecular function, is largely determined by the topology of the free energy landscape guiding the (un)folding process (81). Depending on their size and structural complexity, proteins are known to fold on time scales ranging from microseconds to minutes (5). In the remainder of this Section, the preexisting consensus in regard to the dominant thermodynamic and kinetic properties of protein secondary and tertiary structure is summarized.

Secondary structure: α-helices. The discovery of the spatial structure and chemical bonding patterns of α-helices (Figure 2.1, left), which has become a cornerstone of

structural biology (82), laid the ground work for further investigations of their thermodynamic properties that are typically defined in the context of *order–disorder* (or helix–coil) transitions. Among the elements of such thermodynamically stable configurations (native and/or misfolded structures of biological significance) (83), right-handed α-helices are the most abundant structural motifs. In an α-helix, the C=O group of an amino acid located at the position *i* in the polypeptide (backbone) chain of a protein forms a hydrogen bond with the N–H group of another amino acid that occupies the position $i + 4$, which results in a compact and mechanically robust molecular packing characterized by a well-defined pitch (rise per single turn of the helix) of ~5.5 Å (1). Unlike β-strands, helices are the result of *local* contact formation, which qualitatively distinguishes their (un)folding dynamics from those of tertiary structure.

To propagate (or grow) an existing α-helix segment, a non-helical amino acid on the boundary of the helical segment needs to attain the α-helical configuration; the torsional angles that define the backbone conformation of the amino acid are thereby fixed, thus reducing the overall conformational entropy. Subsequently, a hydrogen bond is formed between the backbone of the amino acid and a residue of the preexisting helical segment. At the melting temperature, the enthalpy change due to the hydrogen bond formation is exactly balanced by the entropy change associated with the backbone restriction, i.e., helices can grow and decay freely. In contrast, to nucleate a helix (i.e., to form a new helical island), the backbone of *three* consecutive amino acids must attain the helical configuration, whereas only one hydrogen bond is gained. Thus, helix nucleation carries an extra free energy penalty, and the polypeptide will tend to minimize the number of domain boundaries between helices and coils. Notably, this is equivalent to the enthalpy versus

entropy balance in the 1D Ising model of interacting spins (Figure 2.2). Similarly to the Ising model, the helix–coil transition can be described using the transfer matrix method to obtain percent helicity as a function of temperature, as follows (84). Let ΔG_{prop} and ΔG_{nuc} be the free energy changes of propagating and nucleating one turn of the helix, respectively. The statistical weights of these two processes are $s \equiv \ln(-\Delta G_{prop}/kT)$ and $\sigma \equiv \ln(-\Delta G_{nuc}/kT)$. The partition function, Z, is then the sum of all possible products of s and σ, corresponding to all possible helix/coil combinations. This summation is the result of expanding the following matrix product:

$$Z = \begin{bmatrix} 0 & 1 \end{bmatrix} \cdot \prod_{n=1}^{L} W_n \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad [2.3]$$

where $W_n = \begin{bmatrix} s_n & 1 \\ \sigma_n s_n & 1 \end{bmatrix}$ is the weight matrix of contact *n,* whose free energy change upon helix formation is dependent upon the identity of the residues in the vicinity of the contact. When expanded, the matrix multiplication gives all of the possible helix/coil permutations. Note that if algebraically expanded, the number of terms in Z grows exponentially with L, and corresponds to the number of possible permutations. From the partition function, the ensemble-wide percent helix content can be calculated as:

$$H = \frac{1}{LZ} \sum_{i=1}^{C} s_i \frac{\partial}{\partial s_i} Z \,, \qquad [2.4]$$

where C is the (sequence-dependent) number of contact types. In the case of a homogeneous sequence (C = 1), $W_n = W$ for all n; by diagonalizing $W = UD\,U^{-1}$, $Z = \begin{bmatrix} 0 & 1 \end{bmatrix} \cdot UD^L U^{-1} \cdot \begin{bmatrix} 0 & 1 \end{bmatrix}^T$, where D^L denotes the diagonal elements of D taken to the power of L because D is a diagonal matrix. In this case, H can be calculated by hand for

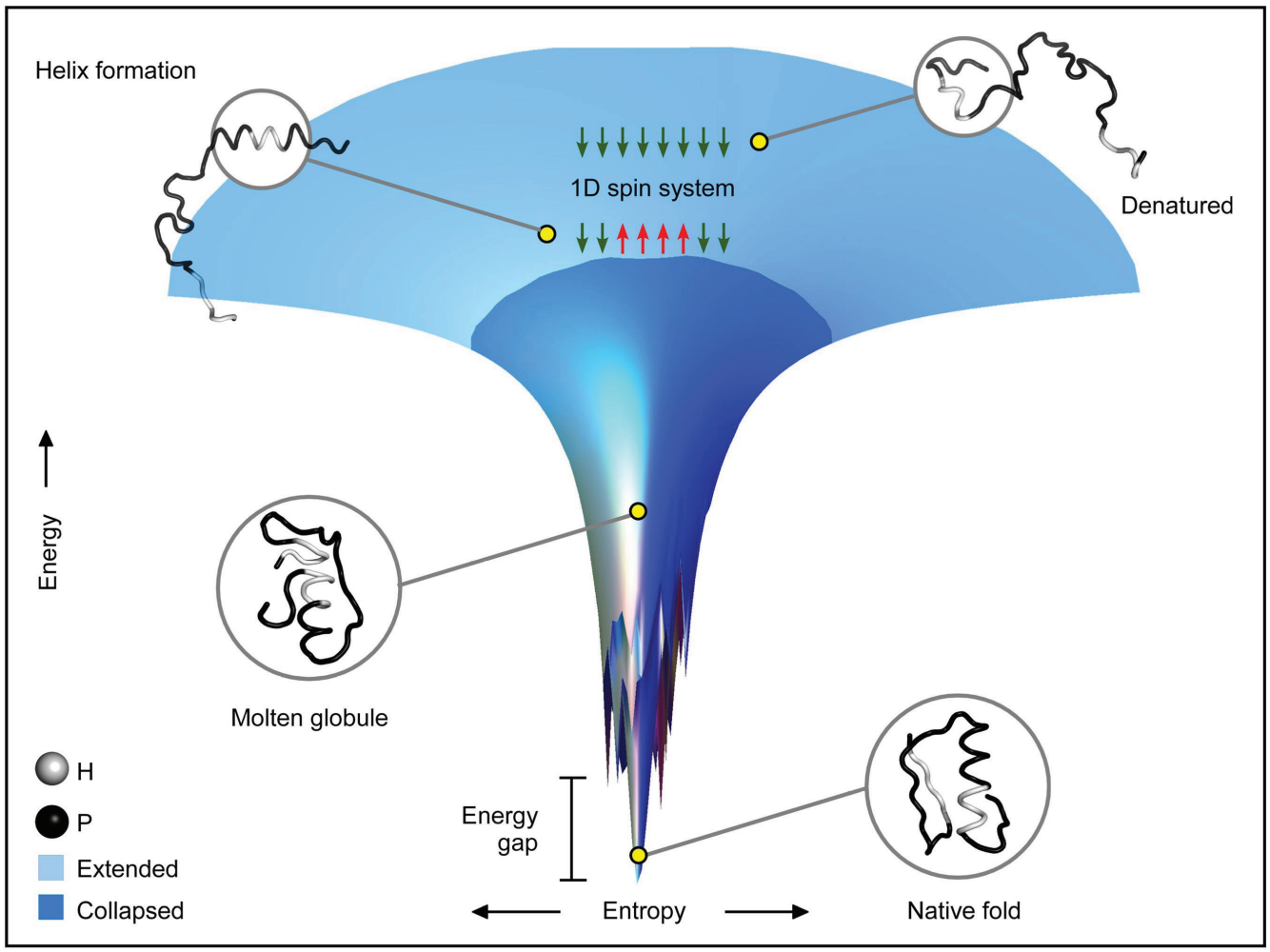

Fig. 2.2. Energy landscape picture of protein folding thermodynamics. Starting with the denatured state (light blue), helix formation is a spontaneously favored event, analogous to the proliferation of domain boundaries in a chain of magnetic spins, which occurs before or concomitantly with hydrophobic collapse into the molten globule state (dark blue). Within this state, the structure can anneal to the final (native) conformation, which is intrinsically, as well as evolutionarily, separated from all other folds by a significant energy gap. Light gray and dark regions of the schematic structures correspond to hydrophobic (H) and hydrophillic (P) regions. Parts of the protein may also form in a hierarchical manner starting from local contacts (see section 2.2 in Text).

arbitrarily long helices using Equation 2.4. The transfer matrix method is a general technique with which to enumerate the permutation space of linear systems, and it will be used in Section 4.1 to calculate the persistence length of macromolecules.

In general, for the *D*-dimensional Ising model, the free energy penalty due to each instance of helix nucleation is incurred for every boundary between helix and coil. This boundary is of dimension $D - 1$ because it is a plane in a *D*-dimensional space. On the other hand, the entropy of creating a boundary is proportional to the logarithm of the number of different positions one can place the boundary at. Since the number of different ways to create a boundary is exponentially related to the dimension *D*, we have the general result that the free energy of creating a boundary is

$$\Delta G = \alpha L^{D-1} - TD \ln L , \qquad [2.5]$$

where α is the enthalpy difference associated with forming a boundary (e.g., helix nucleation). For $D = 1$, the magnitude of the entropy term grows with the chain length *L*, whereas the enthalpy term remains constant. Therefore, sufficiently large systems will never be homogeneous (i.e., free of domain boundaries) at any non-zero temperature and, consequently, first-order phase transitions between the two states are precluded for such systems. This result, formally proven as the Mermin-Wagner theorem (85), implies that it is thermodynamically unfavorable for a protein to be all helix or all coil. During the course of the folding process, proteins can nucleate a fluctuating number of helix islands that dynamically vanish or merge with other helical islands. Indeed, non-native helix content has been experimentally observed during early and/or intermediate stages of protein folding

(86-88). Only with the collective effect of long-range interactions accompanying the formation of native structure will this dynamical nature of helix (un)folding be overridden.

In contrast to the well-understood thermodynamical properties of the coil-to-helix transition, the kinetic picture of the process has only recently been elucidated. The actual rates of helix formation have only been obtained in the last few decades, both by theory and experiment (41), because the (sub)nanosecond temporal resolution required for such measurements, typically via IR or fluorescence spectroscopy following a *T*-jump, was not achievable until the 1990s (41, 42, 89). It was not until the picosecond-resolved (un)folding studies carried out recently in this laboratory that the fundamental ultrafast steps of helix nucleation and propagation could be separately measured and theoretically modeled (90, 91). In contrast to the then-prevailing paradigm, these studies, reported in Section 5.1, demonstrated for the first time that the kinetic picture of α-helix formation does not obey the simple nucleation–propagation model; rather, non-native misfolding intermediates dominate the dynamics.

Tertiary structure. The ability of proteins to attain unique native folds is remarkable because the size of the conformational space increases exponentially with protein size, which presents challenges to (i) the stability of the native fold (the entropy penalty) and (ii) the search for the native fold in a reasonable time (the Levinthal paradox). A significant part of the problem is due to the conformational variability of the tertiary structure, which is stabilized by a wide variety of weak forces. It is through the cooperative culmination of these weak forces that the native fold is distinguished from the other possible conformations below the melting temperature, typically through a first-

order phase transition, as expected for a 3D system with numerous degrees of freedom. However, unlike typical disorder-to-order transitions such as ice formation, the native structures of proteins are flexible as they can undergo large conformational fluctuations (92).

Mutually correlated intra-protein interactions result in excluded-volume effects dependent on the nature of individual conformational microstates. Because these microstate-specific dependencies cannot be incorporated into a closed-form (i.e., not simulated) partition function calculation, analytical methods generally ignore them. For example, the random energy model (REM) reduces all pair-wise interactions within a protein heteropolymer to those taking place between two randomly chosen monomer types (93), which is equivalent to the mean-field-theory representation of a randomly chosen sequence. The monomer types involved may either correspond to the twenty known types of amino acids or be coarse-grained to be either hydrophobic (H) or polar (P) as in the HP model (94). It has been found that a random sequence composed of a two-letter monomer alphabet does not produce a stable native state, whereas a random sequence composed of a twenty-letter alphabet produces a unique native state with an energy gap separating it from all other states (95, 96). Although the accuracy of the REM has been questioned due to its neglect of correlated behavior (97), its usefulness in reducing protein folding to a tractable statistical mechanical framework justifies its use as a coarse-grained theory describing the intrinsic properties of protein folding thermodynamics.

Regardless of the coarse graining level assumed, a number of fundamental questions pertinent to the kinetics of tertiary structure formation remain unanswered. A major issue is whether timely protein folding (i.e., the resolution of the Levinthal paradox) is attained by evolutionarily selected optimal sequences or by the intrinsic physical forces. The answer to this question has relevance for the issue of likelihood of spontaneous life, among other problems of fundamental significance such as, e.g., designing protein structure prediction algorithms, or creating artificial proteins with novel functions. The two main candidates for this intrinsic guiding force are (early) helix formation and hydrophobic collapse, both of which significantly reduce the conformational entropy.

It is apparent that helix formation alone does not resolve the Levinthal paradox. For example, it has been demonstrated that 80% of a 100-residue chain need to be helical at all times in order for the search time to be brought down to 1 minute (98). It is, therefore, clear that even with an overwhelming fraction of the chain fixed to be helices, folding of a 200-residue protein domain would be impossible within a reasonable time period. Moreover, the above assessment does not take into account the transient nature of helix formation, cf. the Mermin-Wagner theorem (see above), which would further increase the search time in real proteins. In addition, because the native structure of the protein is determined by many weak forces acting in competition and cooperation, helices are not the only dominant structural motif. For example, β-strands, van der Waals interactions, and side-chain–side-chain interactions between distant parts of the chain can all be dynamically sampled in their numerous combinations. In particular, non-native conformations that are accessible kinetically will be sampled. This is the essential hurdle that makes protein structure prediction difficult. Indeed, there is evidence obtained from lattice simulations that tight

packing constraints imposed on the system spontaneously lead to helix formation, implying that secondary structure may be the result of, rather than the cause of, tertiary collapse (99).

Although the mechanism of hydrophobic collapse had long been recognized as a universal force constraining the protein folding process (100), it was found that the number of possible conformations of a collapsed chain of length L scales as approximately 2^L (101). Therefore, the reduction of the conformational search space due to the protein chain collapse was insufficient to accelerate protein folding to a reasonable rate. Because the intrinsic physical constraints by themselves did not help resolve the Levinthal paradox, to date, the consensus has been a folding mechanism combining these forces with sequence-specific evolutionary selection. In the 1990s, Frauenfelder, Wolynes and coworkers approached the issue of attaining realistic protein folding times from the energy landscape perspective. Figure 2.2 reflects the free energy bias that disfavors high-entropy, extended protein conformations (lighter blue) against compact globular folds resulting from the hydrophobic collapse (darker blue) (102). The free energy landscape of Figure 2.2, which exhibits some roughness due to conformational frustration, appears to capture a number of key concepts pertinent to protein folding such as, e.g., the energy gap between native fold and nearby states as well as the dynamical nature of α-helix formation. Within the collapsed state, which, as stated above, is characterized by ~ 2^L individual conformations, the proteins is "funneled" toward its native fold via both thermodynamic and kinetic mechanisms optimized by evolutionary selection (Section 5.2). In contrast to the above picture predominant in the literature, a major result of this work is the finding that the hydrophobic force (which includes both chain collapse *and* segregation of hydrophobic residues in the protein interior), by itself, can reduce the conformational space of a protein

to an extent sufficient to achieve feasible folding times, and therefore that the process of protein folding is *inherently fast* (Section 5.2).

In regard to specific pathways leading to the native state, there is still an ongoing debate concerning the fundamental steps proteins take to fold. There exist two prominent models suggesting either "bottom-up" or "top-down" folding. Within the framework of the *diffusion–collision model* of folding (103), smaller sub-structures achieve stability and collide to form larger scaffolds of stability. The sub-structures do not need to have the exact secondary structure of the native state, but the overall topological arrangement of sub-structures should be correct such that annealing can take place locally without too much frustration (104). This way, once the individual domains have come together, a molten globule structure with the correct topology is formed, thereby leading to fast annealing to the native state. The above viewpoint is attractive because folding is modeled to occur hierarchically, thereby quickly achieving the native conformational topology.

The top-down *nucleation–condensation model* stipulates fast hydrophobic collapse to an ensemble of near-native global conformations, followed by slower annealing of both the global topology and the local geometry. This mechanism is attractive because the hydrophobic force has a physical basis in the reduction of the free energy due to compaction (maximized solvent entropy) and aggregation (clustering) of hydrophobic residues, thereby reducing the conformational space, and accelerating the search for the native fold. Up to now, the above reduction of the sampled sub-space had not been quantitatively assessed, and it was not clear to what extent the hydrophobic force would help resolve the Levinthal paradox. There is evidence to suggest that, depending on the

protein sequence, one or both mechanisms are applicable to protein folding. A number of experimental surveys report a preponderance of either the diffusion–collision (104), or the nucleation–condensation (105) mechanism, the latter being typically associated with the chain collapse taking place on a microsecond time scale.

Regardless of whether most proteins fold hierarchically or via an annealing mechanism, there is strong evidence suggesting that the formation of long-range native contacts must be the rate limiting step. Over the years, topological parameters have proven to be useful single-variable structural predictors of folding rates. For example, Plaxco *et al*. (106) devised a phenomenological parameter called the contact order (*CO*), which is defined as:

$$CO = \frac{1}{NL} \sum_{(i,j)\in N} d_{ij} \ , \qquad [2.6]$$

where L is the protein length, N is the number of tertiary contacts, (i, j) are all pairs of amino acids that make tertiary contacts with each other, and d_{ij} is the sequence separation between i and j. Thus, the contact order parameter is larger with proteins which are characterized by longer sequence separations between contacting amino acids. The strong empirical dependence of the logarithm of the folding rate on *CO* indicates that an important component of the rate-determining mechanism is the formation of key tertiary contacts, regardless of whether these contacts form in a bottom-up or top-down fashion. Interestingly, Dagget *et al.* performed point mutations on a globular protein that affected the stability of the mutant but not the *CO* of the mutant, which resulted in four orders of magnitude difference in folding rates due only to contact stability. Therefore, although the

CO is important for determining the entropy barrier required to form long-range contacts, the strength of long-range intramolecular interactions has a substantial effect on the folding rate as well (107). A more thorough understanding of the relationship between *CO* and the folding rate will require a first-principles-based theory as well as explicit simulations of protein dynamics and folding (108, 109).

Although many general aspects of protein thermodynamics are well understood, the kinetics of secondary and tertiary structures remains elusive. For simple chemical reactions, a free energy landscape representation can easily provide an overall picture of the kinetic pathways, and the time scales involved can be estimated from the Arrhenius barrier crossing probabilities. For macromolecules with numerous degrees of conformational freedom this is impossible for two reasons. First, the mapping of an exponentially large number of degrees of macromolecular freedom onto at most two effective variables (or reaction coordinates) necessarily implies that any given microstate is composed of individual conformations that are not kinetically accessible to each other. As a consequence, (fast) kinetic pathways which would be visible on the hyperspace landscape comprising all possible degrees of freedom are hidden in the coarse grained 2D representation, leading to under-estimation of reaction rates. Second, the numerous misfolded states that dramatically slow down the overall folding kinetics are often impossible to identify using the chosen reaction coordinates, which results in over-estimation of reaction rates. Taken together, these two sources of error may significantly distort the true folding mechanism and time scales. Therefore, advanced methods that are tailor-made to investigate kinetics of specific systems under study are required to address these challenges. One recent approach has been to abandon the landscape description in

favor of a network description, mapping individual conformations onto discrete Markov states which are not limited in the number of order parameters (110-112). In addition, these algorithms have the advantage of learning the dominant pathways and intermediates directly from the simulation data, rather than mapping the data onto pre-conceived models such as diffusion–collision or nucleation–condensation. However, such methods are suited for the analysis of simulation data rather than for the creation of a predictive model from first principles. In contrast, as demonstrated below for certain types of macromolecular systems, high-dimensional free energy landscapes can be mapped onto a limited number of properly chosen effective dimensions while preserving the influence of the collapsed dimensions on the macromolecular kinetics (Sections 4.2.1 and 5.1). The main advantages of the approach suggested here are the ability to predict both (un)folding mechanisms and rates using experimentally measured thermodynamic parameters (i.e., without carrying out explicit kinetics simulations), and the preservation of a conformational metric space on which the dynamics can be clearly interpreted.

Chapter 3

METHODOLOGY

In the present Chapter, we justify the selection of macromolecules studied and introduce methods and techniques which we use throughout our work. Due to the numerous degrees of conformation freedom involved in biological transformations, the macromolecular ensembles of interest exhibit a wide range of spatial structures and dynamical behavior. Because individual dynamical trajectories may deviate greatly from the ensemble scale behavior, the techniques we used, although diverse, all employ ensemble averaging. As illustrated in Figure 3.1, a large macromolecular ensemble representative of the system under study is obtained, often with atomic-scale spatial and temporal resolutions, and individual microstates constituting the ensemble are subsequently mapped onto a coarse-grained representation. Importantly, the coarse-graining method must collapse the (non-additive) structural information of each microstate into simplified (additive) form which maintains the structural and temporal resolutions required to observe phenomena of interest. The data obtained using the coarse-grained representation can then be averaged over the ensemble. We note that methods which do not map each microstate to an additive representation before ensemble averaging, such as, e.g., mean-field theories, may give a distorted picture of the ensemble behavior. Much of the work reported here involves developing novel coarse-graining approaches which exploit the properties of the class of systems they are designed for. These system-specific approaches are outlined in detail when they are applied in Chapters

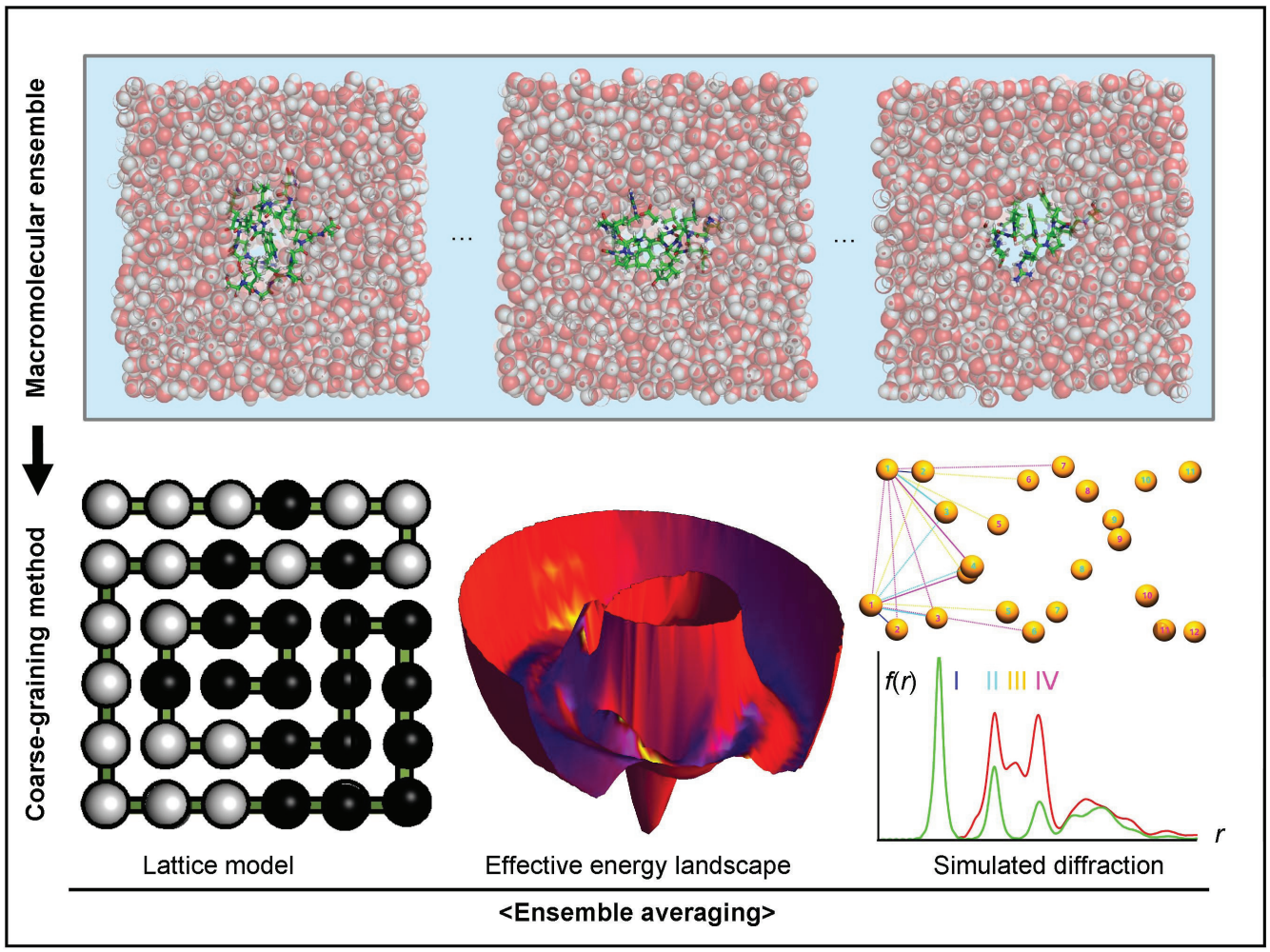

Fig. 3.1. Analysis of dynamics at the ensemble level. Despite the wide range of theoretical and computational methods used, the general methodology involves the generation of the time-dependent ensemble (top), followed by mapping each microstate of the ensemble, which consists of non-additive information, onto a coarse-grained additive representation (bottom), and finally averaging the coarse-grained representation over the entire ensemble. The choice and design of the coarse-graining method is essential to enable proper ensemble averaging without losing the most significant information. Examples of time-dependent ensemble generation include calculating the partition function or ensemble-converging molecular dynamics (MD) simulations. Examples of coarse-graining methods leading to additive results include mapping onto a lattice model (bottom left), a free energy landscape parameterized by system-specific order parameters (bottom center), and the radial distribution function of electron diffraction (bottom right). This methodology allows for analysis of dynamics with statistical significance.

4 and 5. In what follows, we focus on more general methodological aspects of this work and, therefore, limit the discussion to the techniques that have been repeatedly used in our investigations.

3.1 SELECTION OF MACROMOLECULES

Despite the overwhelming structural complexity characteristic of biological macromolecules, there exist persistent structural motifs which serve as fundamental building blocks of biomolecular complexity; the way these ubiquitous components form and aggregate largely determines the overall structure of proteins and nucleic acids. Here, we mainly focus on the two iconic structural motifs of molecular biology: the protein *α-helix* and the *DNA double helix*. For proteins, we also address the issue of formation of the overall (global) structure. Although the aim of this research was to develop predictive theories that apply to protein folding or macromolecular dynamics in general, we found it useful to explore concrete manifestations of the studied phenomena using a number of macromolecules which contain the fundamental structural motifs ubiquitous to biochemistry.

The biological macromolecules studied in this Thesis are depicted in Figure 3.2. To represent conformational dynamics characteristic of the denatured state, a long-chain linear alkane (arachidic acid) was chosen for its structural homogeneity and ease of analytic modeling. A DNA duplex and a number of DNA hairpins of various lengths, loop sizes, and sequences were studied as well, both in vacuo and in solution. To elucidate elementary

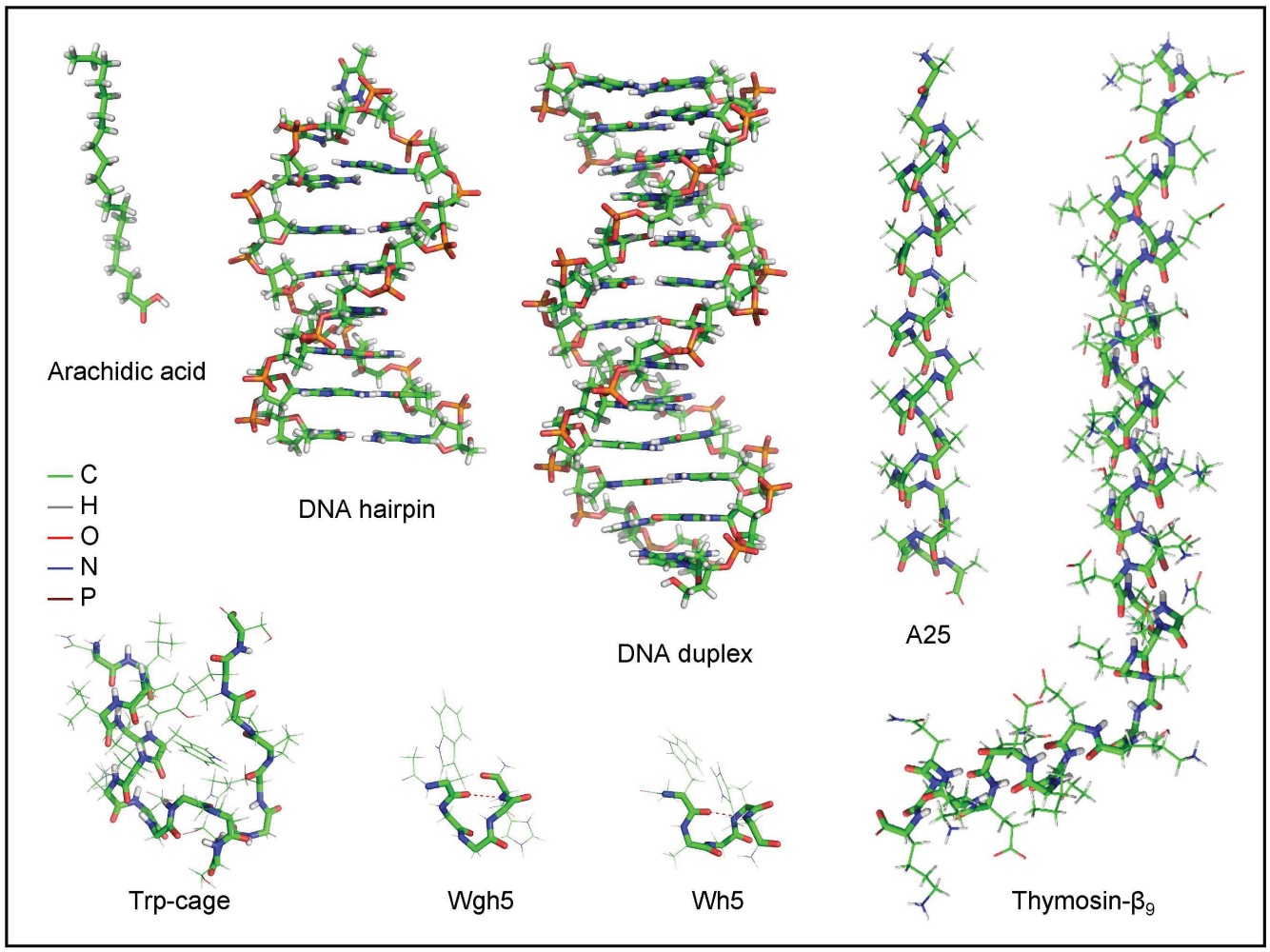

Fig. 3.2. Macromolecular motifs studied. Although the focus is to gain insight into general mechanisms of protein folding and macromolecular dynamics, specific structures are also studied to provide concrete examples for analysis and simulation. These structures are chosen to represent the key motifs of biomolecular structure, including the double helix of nucleic acids (formed either by two separate strands or a single strand forming hairpin loop), the α-helix nucleus, full α-helices (bottom), and simple examples of tertiary structure.

steps of α-helix formation in proteins, the five-residue peptides Wh5 and Wgh5 were studied to isolate the process of helix nucleation, whereas the homo-polypeptide A_{25}, which consists of 25 Ala residues characterized by the highest helix-forming propensity among the 20 known types of amino acids, was used to obtain insights into the details of helix growth. In addition, α-helix formation in hetero-polypeptides was studied using the 41 residue-long thymosin-β_9 (113). Thymosin is one of several polypeptide hormones secreted by the thymus, and functions to modulate the growth of actin filaments via binding of actin monomers (114). In probing protein tertiary structure, we studied the Trp-cage mini-protein, which, in addition to helical content, also possesses a hydrophobic core. Despite its small size (20 residues), the Trp-cage is known to fold in 4 μs (115).

3.2 MOLECULAR DYNAMICS SIMULATIONS

In the absence of breaking or formation of chemical bonds, molecular motions can be adequately described using classical (Newtonian) mechanics. Therefore, it is possible to simulate the dynamics of a molecule given its initial coordinates and an accurate description of the intramolecular forces. This molecular dynamics (MD) methodology began to gain practical traction at the dawn of the computer era (116), and shortly thereafter it was employed in simulating protein motions (108, 117). Although MD methods per se are not the focus of the work presented here, they are used extensively throughout this Thesis to (i) confirm theoretical findings, (ii) justify the use of methodologies or approximations constituting the basis of a variety of theoretical approaches, and (iii) generate and monitor large-scale ensembles of biological

macromolecules with atomic-scale spatial and temporal resolutions. A survey of concepts, uses and limitations pertinent to MD methods is, therefore, given below in brevity. The potential energy of a system of N atoms is usually defined in the form (118):

$$U(R) = U_{\text{local}}(R) + \sum_{ij} \left\{ \left[\left(\frac{A_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{B_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \right\}, \qquad [3.1]$$

where R is the set of generalized coordinates of the system and $U_{\text{local}}(R)$ represents the potential energy terms associated with (local) interactions, such as internal rotations and vibrations, which are conveniently parameterized using bond distances, bond angles, and torsional angles. These interactions, except for torsional potentials, are typically approximated using harmonic potentials with spring constants corresponding to the experimentally determined energies of stretching and bending vibrations. The non-local term involves the van der Waals and electrostatic interactions between individual atoms, which are summed over all non-bonded pairs of nuclei i and j ($i, j \leq N$) separated by distance r_{ij}; here, A_{ij}/B_{ij} and q_i/q_j are the van der Waals parameters and electrostatic charges associated with indicated atoms, respectively. During the course of MD simulations, the atomic positions are integrated forward in time t using $U(R,t)$ and a chosen sampling interval Δt (119). Because accounting for the long-range electrostatic interactions constitutes the rate-limiting step in obtaining updated atomic coordinates, the Ewald summation method (120), which scales approximately as $N\ln N$ using the particle mesh Ewald (PME) algorithm (121, 122), is typically employed. Due to the electrostatic calculation bottleneck, MD simulations become particularly expensive when they employ explicit water molecules rather than a continuum solvent model. Ever since their initial

use (123), explicit water simulations have justified their higher computational cost by showing the important role that water geometry plays in macromolecular stability and dynamics. For example, explicit water simulations revealed that the hydration shell surrounding biopolymers is characterized by unique structural features (124) which play a crucial role in their dynamics and folding. At present, the computational cost for simulating medium-to-large proteins that are known to fold on the millisecond time scale (109, 125, 126) is still beyond all but the most Herculean computational efforts. Consequently, a variety of techniques are invoked to accelerate the calculations. Thus, an important factor limiting the efficiency of conformational sampling is the roughness of the free energy landscape, which results in formation of local minima (or saddle-points). The aim is therefore to enable macromolecules to release themselves from local minima while ensuring that the sampling is not biased in the ergodic limit (i.e., in the assumption of the infinite sampling time). Umbrella sampling (127) is perhaps the most popular method of exploring large conformational changes in computational biochemistry, whereby unfavorable states, such as barriers separating low-energy conformational basins on the free energy landscape, are artificially exaggerated but the thermodynamic weights of such states are decreased to compensate for the fact that they are over-sampled. Another technique, called replica exchange (128), relies on carrying out massively distributed simulations which are performed at a variety of temperatures, and periodically swapping the resulting conformations between simulations of adjacent temperatures. The latter approach allows lower-temperature simulations to exploit wider conformational sampling characteristic of higher temperatures. Importantly, the ergodicity is preserved by ensuring that the swapping probability is weighted according to the relative energies

of the conformational states. Although such methods are of limited value in finding the true kinetic trajectories, they nevertheless provide an accurate thermodynamic picture of the conformational space. To date, MD simulations have been successfully used to model protein folding (109, 125, 126, 129), protein-ligand interactions (130), protein motions (127, 131, 132), as well as the energetics of protein function (133).

The generic MD protocol we employed to simulate (un)folding of macromolecular ensembles is as follows. A cutoff of 14 Å was used for the van der Waals calculations, and electrostatic interactions in systems with periodic boundary conditions were computed using the PME method (121, 122) with the direct-sum cutoff and Fourier grid spacing typically being 9 Å and 1.2 Å, respectively. To simulate an aqueous environment, macromolecules were surrounded with explicit TIP3P (three-point) representations of water molecules, typically in cubic unit cells with periodic boundary conditions imposed. The number of water molecules surrounding the solute was usually chosen such that at 1 atm pressure and room temperature, the cell was at least 1.5 times longer than the largest macromolecular dimension. Prior to carrying out the actual simulation, the system was first energy minimized to a root-mean-square (RMS) force gradient of 0.12 $kJ \cdot mol^{-1} \cdot Å^{-1}$. For the studies of macromolecular unfolding, the system was subsequently heated to 500 – 1200 K for up to 1 ns. For all the MD studies reported here, the system was heated or cooled for 100 ps to reach the (un)folding temperature, and thereafter evolved with the number of particles, pressure, and temperature kept constant (*NPT* ensemble). Temperature and pressure coupling were enforced using the extended-ensemble Nose-Hoover/Parrinello-Rahman algorithms with a coupling time constant of 1 ps (134-137).

During all simulations, bonds involving hydrogen atoms were constrained using the LINCS algorithm (138) and rigidity of the TIP3P water molecules was enforced by the SETTLE algorithm (139). The above constraints are justified because protein dynamics should be insensitive to the (very fast) hydrogen vibrations, allowing for an integration time step of 2 fs. Coordinates were typically saved with a sampling interval of Δt = 1 ps.

3.3 DIFFRACTION SIMULATIONS AND ENSEMBLE CONVERGENCE

There are two distinct challenges pertinent to understanding macromolecular dynamics. The first challenge—*structural complexity*—is inherent to the study of any inhomogeneous structure with numerous degrees of mechanical freedom. The second challenge—*cooperative dynamics*—is a hurdle characteristic of biomolecular systems. Addressing the above challenges requires developing novel coarse-graining approaches with which to elucidate the essential ensemble-wide aspects of structural change, while preserving the mechanistic nature of the dynamics (e.g., cooperative motion or resonance phenomena). The need for appropriate coarse-graining methods becomes apparent as we ascend the complexity ladder. The overwhelming complexity associated with the ensemble-wide macromolecular (un)folding behavior makes computational simulation an essential tool for understanding proteins at the molecular level. A number of coarse-graining techniques have been developed to investigate the complex energy landscapes of protein (un)folding (81). For example, in the case of helix-coil transitions, disconnectivity graphs have been successfully applied to energetically sort the discrete states of native, part-native, and non-native populations that dominate the folding dynamics (140, 141).

Here, we introduce the *ensemble-averaged radial distribution function* which is frequently used to present the results of electron diffraction measurements as a novel coarse-graining technique for analyzing the data obtained from ensemble-convergent MD simulations. We note that, whereas other coarse graining approaches have been developed during the course of the studies reported in Chapters 4 and 5, the latter approaches were tailored to the specific systems under consideration. In contrast, the ensemble-convergent radial distribution function was applied as a standard all-purpose coarse-graining method throughout the entire line of research presented in this Thesis. We, therefore, found it useful to outline corresponding methodological details below rather than in the subsequent Chapters devoted to specific biomolecular sytems.

The (time-dependent) ensemble-averaged radial distribution function, $\langle f(r, t)\rangle_n$, can be interpreted as a weighted internuclear distance histogram distribution. As stated above, $\langle f(r, t)\rangle_n$, has been directly adapted from its use in electron diffraction experiments, and, due to its fast ensemble-convergence to statistical signal-to-noise ratio comparable to that typical of experimental measurements, it is well-suited to analyzing the *simulated* structural dynamics of complex systems in real time. Importantly, $\langle f(r, t)\rangle_n$ provides structural information on both the local and global length scales. For example, the locally periodic motifs such as α-helices are manifest as *structural resonance* peaks centered at the characteristic repeating distances in the $\langle f(r, t)\rangle_n$ curve, and the global radius of gyration can be obtained from the (weighted) sum of the area under the curve; the local resonance features we explored include the α-helix pitch and the base-pairing/base-stacking distances of the DNA duplex (see below and Section 4.2 for pertinent definitions and practical applications, respectively). The ensemble-averaged radial distribution function

complements and, in many cases, supersedes the information content of standard coarse-graining metrics such as percent native contact and root-mean-square deviation (RMSD) with respect to the native structure. The results, in addition to being directly comparable to those of electron diffraction measurements, can be used to suggest new experiments with promising signatures in $<f(r, t)>_n$. Thus, experimental efforts are currently underway in this laboratory to observe the structural resonances seen in the simulated diffraction patterns. It is, indeed, remarkable that an *experimental methodology* has stimulated theoretical exploration, which in turn has directly guided the next generation of experimental studies. In what follows, the method of obtaining $<f(r, t)>_n$ is, therefore, described in the original context of ultrafast electron diffraction experiments.

Because elementary events of macromolecular dynamics often occur, or are triggered, on the ultrafast time scale, their inherent structural dynamics is elusive to the conventional methods of experimental probing that are typically limited to the nanosecond temporal resolution. Such elementary events (142) span a time window—from femtoseconds to nanoseconds—which turns out to be orders of magnitude shorter than the typical time scale of the globular motions in biopolymers (microseconds or longer). Ultrafast electron diffraction (UED), crystallography (UEC), and microscopy (UEM) have the potential for direct visualization of biostructural change as they provide atomic-scale spatial and temporal resolutions (143, 144). Of these, UED is unique for its possible elucidation of macromolecular dynamics in the absence of solvent.

To date, studies of biological species in vacuo have been made using various experimental methods, including, e.g., mass spectrometry which was shown to preserve

macromolecule structures (145-147). However, in contrast to spectrometric and spectroscopic investigations into gas-phase behavior of isolated biopolymers, UED measurements could, in principle, allow for direct probing of both macromolecular geometry and its temporal change. Because recent experimental efforts undertaken in this laboratory have already enabled nondestructive delivery of biomolecules into the gas phase, the way is now open for UED measurements to explore their conformational dynamics with ultrafast temporal resolution. Therefore, $<f(r, t)>_n$ is more than just another computational coarse grained order parameter—it is also an experimental observable in potential UED studies of biomolecular behavior in the absence of water.

During the course of a typical UED experiment, the molecular sample is usually excited by an (initiating) ultrafast laser pulse, followed, for probing, by a series of ultrashort electron pulses which map the spatiotemporal changes induced in the macroscopic, gas-phase molecular ensemble under study (148). Two-dimensional electron-diffraction patterns are thus obtained at each particular point in time for a series of chosen time points. As stated above, the UED data represent an average over the molecular ensemble $\{l\}$, $l \in [1, N]$ and, at the processing stage, they are radially averaged to yield 1D experimental scattering intensities (149). Because of the much larger scattering cross-section of electrons on atoms, as compared to that of X-ray radiation, it is possible to achieve the ultrafast temporal resolution when studying molecular systems in the gas phase (a nontrivial task because of the lack of crystallinity and low molecular density of the sample). The above experimental methodology (150, 151) has been applied in the numerous studies of chemical reactions, excited-state structure dynamics, and

nonequilibrium conformational changes on their native (ultrafast) temporal scales (144, 152). Examples of UED studies from this laboratory (153-156) include determination of transient structures in radiationless (dark) relaxation processes, relaxation pathway bifurcations, and intramolecular structural rearrangements. At present, the challenge is in the study of systems with many conformers. For such complex systems, the traditional analyses of gas electron diffraction, which have been successful in the determination of thousands of structures (157), need to be revisited.

Detailed theoretical accounts of (conventional) gas electron diffraction methodology may be found in a number of sources (158, 159). Briefly, a 2D electron-scattering pattern is radially averaged to yield the one-dimensional ensemble-averaged scattering intensity $< I(s) > = <^{\mathrm{M}} I(s) > + ^{\mathrm{B}} I(s)$. The structural information is contained in:

$$<^{\mathrm{M}} I(s) > \sim \sum_{l} \sum_{i \neq j} {}^{\mathrm{M}} I_{ij}^{l}(s), \qquad [3.2]$$

which is a sum over all the internuclear distances accross a molecule $\{r_{ij}\}$, $i \neq j$, and over the ensemble N. The term $^{\mathrm{B}}I(s)$ is a monotonic background scattering function, which is the sum of atomic and inelastic scattering and other experimental factors contributing to the background, and s is the magnitude of the momentum transfer vector between an incident electron and an elastically-scattered electron; $s = (4\pi/\lambda)\cdot\sin(\theta/2)$, where λ is the de Broglie wavelength (0.069 Å at 30 keV), and θ is the scattering angle.

By Fourier-transforming the modified molecular intensity $<sM(t; s)>$ (158), we obtain the ensemble-averaged radial distribution function, $<f(t; r)>$, which provides a

snapshot of the density distribution of internuclear separations across the ensemble at a particular point in time. Finally, by subtracting $\langle f(t; r)\rangle$ obtained at "time zero" from that of later times, we obtain the ensemble-averaged diffraction differences, $\langle \Delta f(t; r)\rangle = \langle f(t; r)\rangle - \langle f(0; r)\rangle$, which map the spatiotemporal evolution of the sample (152, 160).

For small and medium-sized molecules, quantum chemical calculations followed by normal-coordinate analyses of the resulting Cartesian force fields are normally used to compute $\{r_{ij}\}$ as well as first- and higher-order vibrational corrections (161, 162). Alternatively, RMS vibrational amplitudes may be obtained from spectroscopic data, or estimated using empirical equations (163-165). For a (large) conformationally flexible biomolecule, however, the ensemble averaging involves the entire landscape of quasi-random conformations, which may, in principle, be generated using a variety of different methods. Throughout the line of research reported here, our method of choice involved applying the UED program developed in this laboratory (158) onto (simulated) macromolecular ensembles to obtain the ensemble-averaged electron diffraction patterns. Although the atomic-scale *spatial* resolution is lost by this method (notably, the picosecond *temporal* resolution characteristic of the original simulations is preserved), the coarse-graining approach based on $\langle f(t; r)\rangle$ produces distinct structural fingerprints which turn out to be useful for analyzing both secondary and tertiary structure dynamics (Section 4.2). By examining the actual (simulated) conformational ensemble jointly with its diffraction pattern, a wealth of information on (un)folding transitions in biological macromolecules can be obtained.

For example, we note that from the accurate internuclear distance density distribution $P(r)$, $P(r)dr$ being the probability of finding $r \in [r, r + dr]$, we can, in principle, obtain the *radius of gyration* $R_g = (I/M)^{1/2}$, where M is the molecular mass and I is the moment of inertia. Specifically (166),

$$<R_g^2> = \frac{\int_0^\infty <P(r)> r^2 dr}{2\int_0^\infty <P(r)> dr} = \frac{\int_0^\infty < f(r) > r^3 dr}{2\int_0^\infty < f(r) > r dr}, \qquad [3.3]$$

where $<f(r)> \sim <P(r)>/r$ is the ensemble-averaged radial distribution function calculated on the interval $s \in [0, \infty]$ with no artificial damping applied ($k = 0$). In Section 4.2.3, we utilize the ensemble-averaged radius of gyration to monitor order-disorder transitions in polypeptides.

Because $<f(r; t)>$ which provides both local and global structural information is a projection of 3D molecular structures onto a 1D coarse-graining representation, it achieves ensemble-convergence with a surprisingly small ensemble size. Shown in Figure 3.3 is the convergence behavior of both the decay of a base pair contact in a DNA duplex (left) and the radial distribution function of the duplex (right) as obtained from MD simulations following a temperature jump of $\Delta T = 1200$ K in vacuo. Both the base pair contact survival probability $<P_i(t)>_n$ and the ensemble-averaged radial distribution function of the duplex $<f(r, t)>_n$ reach convergence when averaged over $n = 100$ independent MD trajectories. Notably, whereas the information from individual trajectories (top left) may lead one to speculate that, for the native Watson-Crick contact $i = 3$, the base pairing may be quasi-

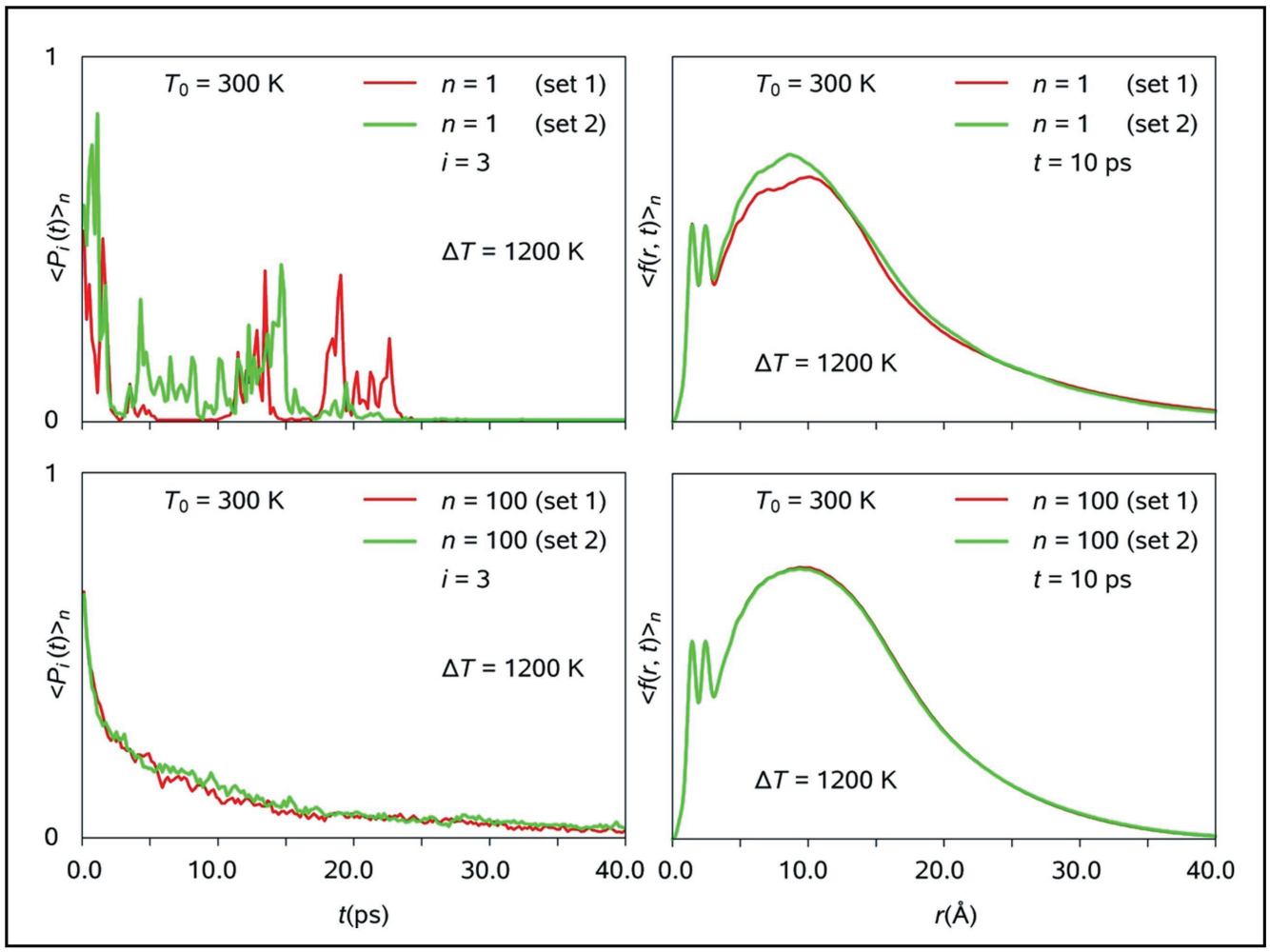

Fig. 3.3. Ensemble convergence illustrated. Generating the time-dependent ensemble of DNA double helix unfolding via molecular dynamics (MD) simulations, the convergence to a smooth distribution can be seen by contrasting ensemble sizes of n = 1 (top) versus n = 100 (bottom) microstates, for the probability that the third base-pair is intact as a function of time (left) and the radial distribution function of the ensemble 10 ps after a temperature jump (right). Red and green denote two separate ensembles. Two different n = 1 ensembles are not only different from each other (top) but also fail to provide a picture of the decay characteristics of the base-pair (top left). In contrast, for n = 100 (bottom), the two ensembles are almost indistinguishable and a time constant can be obtained from the base-pair decay (bottom left), indicating that ensemble-convergence is achieved at n = 100.

stable for more than 20 ps after the temperature jump with the possibility of dynamically reforming the base pair at longer times, the information from averaging over one hundred trajectories (bottom left) indicates that the characteristic time scale for the base pairing contact rupture is about 5 ps. Likewise, the $\langle f(r, t)\rangle_n$ profile averaged over n = 100 trajectories has the signal-to-noise ratio comparable to that typically obtained during the course of UED experiments (167). In a similar fashion to UEC (168) experiments, which utilize the spatial and temporal coherence in the specimen to reveal the structural dynamics, the "*computational microscopy*" experiments reported here make use of both spatial resonance *and* ensemble averaging to extract accurate and reliable structural data at each particular point in time.

3.4 PROGRAMS USED

Throughout the studies summarized in the subsequent Chapters, quantum chemical calculations were performed using the GAUSSIAN program package (169). The majority of ensemble-convergent MD simulation were carried out using the CHARMM (118) suite of programs, although the GROMACS (170) program suite was also employed (Section 4.2.1). The resulting MD trajectories were analyzed with the help of the VMD (171) software, whereas macromolecular visualization, manipulation, and rendering were performed using the program PYMOL (172). Data plotting was executed with the aid of MATHEMATICA (173) and GNUPLOT (174). Electron diffraction simulations were performed using the in-house UEDANA program.

Chapter 4

RESULTS: MACROMOLECULAR DYNAMICS

Structural dynamics plays a variety of important roles in and between all three states of a biological macromolecule: native, unfolded, and intermediate. In the present Thesis, the protein folding dynamics are separately addressed in Chapter 5 because of their pivotal significance to our understanding of biological function and self-organization in complex systems. Yet to understand the process of protein folding with clarity, a number of key issues pertinent to macromolecular dynamics in general must first be elucidated. These issues include the nature of the unfolded state (and the extent to which such states may be considered "random"), the types of conformational dynamics at different length and time scales, the effect of temperature on the dynamics characteristic of various length scales, and the role of solvent in facilitating specific macromolecular motions. Notably, structural dynamics plays a crucial role in many areas other than macromolecular folding, such as, e.g., function (175), regulation (176), and aggregation (177). In these biological processes, both unfolding and dynamical interconversion between a number of folded states are the relevant mechanisms. Protein function is often associated with a concerted motion of the native fold, which may range from (large-amplitude) conformational changes resembling those involved in the myosin power cycle of muscle contraction (178) to evolutionarily conserved pathways inside proteins that transmit dynamical signals, a mechanism facilitating protein allostery (179, 180). Perhaps more surprisingly, intrinsically disordered (unfolded) proteins also play a role in protein function and quality control processes, e.g.,

by trapping multi-domain antibodies in the endoplasmic reticulum until all the protein domains are assembled (181). In the present Chapter, we investigate biomolecular structural dynamics which occur under a variety of circumstances. In so doing, we aim to elucidate general properties of the dynamics taking place both between and within all the relevant states. The pertinent processes include (un)folding, misfolding, and conformational interconversions characteristic of the folded and (partially) unfolded macromolecular ensembles. The studies reported below have been carried out on a variety of biological macromolecules ranging from fatty acids, to nucleic acids, to polypeptides, both in vacuo and in solution, and over a wide range of temperatures. The phenomena that we describe, such as temperature-dependent temporal bifurcations taking place at a variety of length scales and the effect of the solvent on conformational changes in polypeptides, help to construct a general theoretical framework for understanding and manipulating macromolecular dynamics.

In anticipation of the pivotal significance of the polar solvent to protein folding feasibility revealed in Chapter 5, of special interest here is the role of water in the shaping of macromolecular structure and dynamics. Studies of structural and conformational changes which are free of the additional effects of solvation, crystallization or external ordering imposed on the specimen have attracted increasing attention (145, 182). Although biological macromolecules are known to be extremely sensitive to the effects of the environment, certain structural motifs abundant in vivo turn out to be relatively robust, persisting in a wide variety of environments. This gives rise to a number of questions of fundamental significance; in particular: what structural features are preserved in vacuo and are therefore “inherent” to the physics of isolated macromolecules? Equally importantly,

how do the presence and nature of the solvent affect the conformational freedom of biomolecular systems? Answering these questions will further elucidate the extent to which biological functionality relies on the aqueous environment. We begin our exploration with perhaps the simplest example of macromolecular dynamics which underscores the *non-random* nature of the unfolded state.

4.1 DYNAMICS OF THE UNFOLDED STATE

Persistence length: a polymer descriptor. To fully understand macromolecular (un)folding, both the folded and unfolded states must be well defined and understood. In contrast to the folded state, the unfolded state is typically characterized by a vast ensemble of microstates. Therefore, properties of the unfolded state can only be accounted for in a probabilistic manner. Understanding the dynamics and biases of the unfolded state is not merely a philosophical issue. Such understanding is required, e.g., when distinguishing whether a given folding population arises through thermally-driven interconversions within the unfolded ensemble or is pre-determined by the unique macromolecular sequence and geometry. In this sense, the key question pertinent to studies of flexible macromolecules is the extent to which the structural features abundant at low temperatures (such as quasi-periodic motifs) will be preserved at elevated (including physiological) temperatures. To address this question, we assessed the ensemble-averaged *persistence length* of a long-chain linear alkane (arachidic, or eicosanoic, acid) at a variety of temperatures using classical statistical thermodynamics as well as ensemble-convergent numerical simulations. These extended hydrocarbon chains, which are the chief components of cellular

membranes, possess a quasi-1D spatial periodicity, self-assemble on surfaces (substrates) and can also be made as "2D crystals" (183, 184). Chosen to be the starting point of our investigations into macromolecular dynamics, arachidic acid is used to give the simplest possible benchmark analysis of the interplay between long-chain connectivity, (low) torsional rotation barriers, and steric repulsion, which are the inherent features of complex biomolecular systems.

The persistence length, L_p, is a basic statistical property indicating the characteristic distance within a polymer for which directional coherence is lost. For an ensemble of infinite chains of covalently bound structural sub-units, it is defined as the projection of the (averaged) end-to-end vector onto the axis of the first covalent bond in the chain (185). For fragments of the chain that are shorter than the persistence length, the molecule behaves rather like a flexible elastic rod, whereas the ends of fragments that are longer than the persistence length have essentially no correlated motion. We shall compute the end-to-end persistence length (denoted L_p^*) for $C_{19}H_{39}COOH$. Thus, L_p^* will serve as a coherence length over which a quasi-periodic structural motif characteristic of arachidic acid will be preserved over the large-scale ensemble of structures. The dependence of the above coherent behavior on chain length and temperature will also be addressed.

In the following, bold letters and angular brackets denote vectors and ensemble averaging, respectively, and all lengths are given in units of C–C bond distances ($r_{C\text{-}C} \approx$ 1.533 Å). Suppose that $\boldsymbol{r}_1$, the vector aligned with the first C–C bond in the chain, is defined to be in the direction of the z-axis, $\boldsymbol{e}_z$. Then, L_p^* is given by the z-component of the ensemble-averaged end-to-end vector $\boldsymbol{R}$:

$$L_{\mathrm{p}}^{*} \equiv \left\langle \mathbf{R} \cdot \mathbf{e}_z \right\rangle = \sum_{i=1}^{N} \left\langle \mathbf{r}_i \cdot \mathbf{e}_z \right\rangle, \qquad [4.1]$$

where $\langle \boldsymbol{R} \rangle$ is the sum of the individual (ensemble-averaged) bond vectors $\langle \boldsymbol{r}_i \rangle$, and polymer length is represented by the number of C–C bonds, N. As an illustration, L_{p}^{*} is shown for the molecular conformation depicted in Figure 4.1, top left. In the limit of N approaching infinity, L_{p}^{*} approaches the formally defined persistence length L_{p} (185). To calculate $\langle \boldsymbol{r}_i \rangle$, we define a series of local Cartesian coordinates for every C–C bond (Figure 4.1, top left inset). The z-axis is defined to be in the direction of the bond itself, whereas the x-axis perpendicular to the z-axis lies in the plane formed by the bond and the *previous* bond in the chain. The y-axis is then uniquely defined according to the right-hand convention for Cartesian coordinates.

To distinguish between a specific bond and the coordinate system associated with it, we designate $\langle \boldsymbol{r}_i \rangle_j$ to be the i-th C–C bond described in the j-th coordinate system. For example, in its own local coordinates, each bond by definition points in the z direction: $\langle \boldsymbol{r}_i \rangle_i = [0, 0, 1]$. In order to define the (x,y)-coordinates of the first bond, which has no prior bonds to establish a coordinate system, we fix the coordinate frame of the first bond to the global coordinate frame: $\langle \boldsymbol{r}_i \rangle_i \equiv \langle \boldsymbol{r}_i \rangle$. Thus, $\langle \boldsymbol{r}_i \rangle$ can be computed by writing $\langle \boldsymbol{r}_i \rangle_i$ in the coordinates of $\langle \boldsymbol{r}_i \rangle_1$. This is accomplished recursively by the transformation matrix:

$$M = \begin{bmatrix} -\cos\theta\langle\cos\psi\rangle & \cos\theta\langle\sin\psi\rangle & \sin\theta\langle\cos\psi\rangle \\ -\cos\theta\langle\sin\psi\rangle & -\cos\theta\langle\cos\psi\rangle & \sin\theta\langle\sin\psi\rangle \\ -\sin\theta & -\sin\theta & -\cos\theta \end{bmatrix} \qquad [4.2]$$

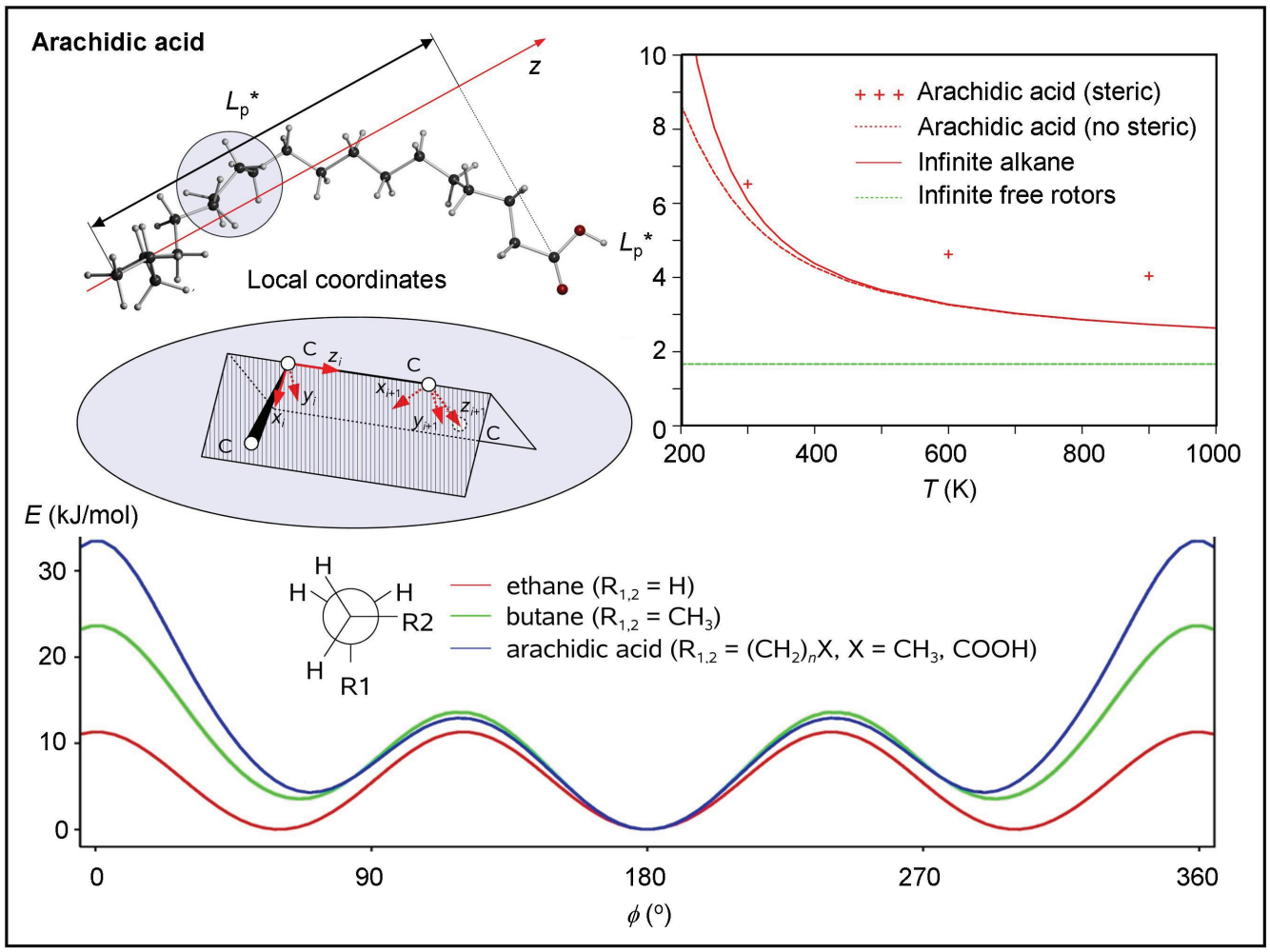

Fig. 4.1. Persistence length of arachidic acid. The persistence length is the projection of the end-to-end vector onto the direction of the first C−C bond (top left). The analytical transfer matrix method, which ignores steric repulsion, gives the ensemble-averaged end-to-end vector as a function of temperature. In this method, each C−C bond is the *z*-axis of its own local Cartesian coordinate frame (top left inset), and the end-to-end vector is the sum of the individual bond vectors, each of which can be calculated in the thermodynamic ensemble by multiplying the appropriate power of the transfer matrix. The persistence length is plotted as a function of temperature (top right), for an infinite chain with free rotation (green line), an infinite alkane chain (solid red line) as well as for arachidic acid (dashed red line). Alternatively, the persistence length with steric repulsion can be explicitly computed by averaging the simulations of 2^{10} self-avoiding arachidic acid chains generated at T = 300, 600, and 900 K (red cross). The backbone torsional potential of an alkane chain used in the persistence length calculations is calculated by rotating the torsion angle incrementally and evaluating the energy of the quantum mechanical electronic ground state at that angle (bottom).

which represents any vector in the *i*-th coordinate system in terms of its averaged coordinates in the (i − 1)-th coordinate system. Here, ψ is the torsional angle of rotation of the *i*-th C–C bond relative to the *i*-th positive *x*-axis, and θ = 113.6º is the C–C–C valence angle characteristic of a linear-alkane chain. Averaging over the torsional angle yields:

$$\langle \cos\psi \rangle \approx \frac{1}{Z} \sum_{\psi=-\pi}^{\pi} \exp\left(\frac{-E(\psi)}{kT} \right) \cos(\psi), \qquad [4.3a]$$

$$\langle \sin\psi \rangle \approx \frac{1}{Z} \sum_{\psi=-\pi}^{\pi} \exp\left(\frac{-E(\psi)}{kT} \right) \sin(\psi), \qquad [4.3b]$$

where *Z* is the normalizing partition function and the torsional potential $E(\psi)$ is determined from our quantum chemical calculations (see below). The accuracy of Equation 4.3 increases as the ψ binning interval is reduced (i.e., the number of ψ points to be summed over is increased). Although <sin ψ> = 0 due to symmetry about the *x* axis, <cos ψ> is nonzero and depends on $E(\psi)$ and the ambient temperature.

With the aid of the transformation matrix *M* of Equatiion 4.2, we obtain $\langle \mathbf{r}_i \cdot \mathbf{e}_z \rangle = [0\ 0\ 1]\, M^{i-1}\, [0\ 0\ 1]_i^T$, where the superscript and subscript on a vector denote transposition and coordinate system index, respectively. From right to left, we take the *i*-th bond vector in its own coordinate system, transform it (i − 1) times until it is in the coordinate system of the first bond, and extract the *z*-component of the final vector to obtain the ensemble-averaged projection of the *i*-th C–C bond onto the very first bond. Note that when Equations 4.3 are substituted into *M*, ensemble averaging arises from the

distributive multiplication of *M*. Finally, the end-to-end persistence length is obtained by summing over all bonds:

$$L_p^{*} = \sum_{i=1}^{N} [0\ 0\ 1]\, M^{i-1} [0\ 0\ 1]_i^{T} \ ; \qquad [4.4]$$

L_p^{*} approaches L_p as N approaches infinity.

To calculate $E(\psi)$ (and therefore M), the molecular structure of $C_{19}H_{39}COOH$ was optimized at our standard B3LYP/6–311G(d,p) computational level using the quantum chemical suite of programs GAUSSIAN 98 (169), and imposing the C_s point-group symmetry constraints which implied planarity of the backbone chain of the molecule throughout the structure optimization (Figure 4.1, bottom). The backbone-averaged values of bond distances and valence angles, $\langle r_{C\text{-}C} \rangle$ and $\langle \alpha_{C\text{-}C\text{-}C} \rangle$, were equal to 1.533 Å and 113.6°, respectively. A single rotational coordinate, $\phi = \tau_{C9\text{-}C10\text{-}C11\text{-}C12}$ ranging from 0 to 180° was chosen to define the relative orientation of the two equally large molecular moieties, and a ground state energy calculations were performed using the $\Delta\phi$ binning of 10° (all molecular structure parameters except ϕ were kept fixed to their optimized values throughout the potential energy surface scan). Although the resulting potential to internal rotation closely resembles that of *n*-butane on going from *trans*- to *gauche*-configuration ($60° < \phi < 180°$), the rotational barrier in the vicinity of $\phi = 0°$ is somewhat higher in the case of arachidic acid (E_0 = 34 kJ/mol, cf. 24 and 11 kJ/mol as obtained for *n*-butane and ethane, respectively; Figure 4.1, bottom).

Results. The torsional potential characteristic of arachidic acid as obtained from the above density functional theory (DFT) calculations was used to compute $\langle \cos \psi \rangle$ as a function of temperature by Boltzmann weighting for an infinitely large molecular ensemble. The results, both for an infinitely long linear alkane and $C_{19}H_{39}COOH$ (N = 19), are plotted in Figure 4.1, top right. These results asymptote to the value of $(1 + \cos \theta)^{-1}$, or 1.67, in the limit of infinite temperature, which is the persistence length of a freely rotating chain, $E(\psi) \equiv 0$ (185). In the limit of zero temperature, the *trans*-configurations dominate. Thus, $L_p^* = \boldsymbol{R} \cdot \boldsymbol{e}_z$, with $\|\boldsymbol{R}\|$ approaching the length of the extended chain. In the latter case, as N approaches infinity, L_p diverges. We note that Equation 4.4 does not take into account steric (i.e., excluded volume) interactions which render certain domains of the conformational space inaccessible to the ensemble under study. The discrepancy between the temperature dependence of L_p^* estimated using Equation 4.4 and the three discrete values of $L_p^*(T)$, T = 300, 600, and 900 K, as obtained from averaging over 2^{10} molecular structures generated via explicit Monte Carlo simulations taking the excluded volume effects into account, increases with temperature. The above discrepancy (Figure 4.1, top right) arises because compact conformers characterized by lower persistence lengths are more likely to be sterically excluded.

Summary. Although simplified, the approach summarized above elucidates the non-random nature of unfolded states of biological macromolecules. At physiological temperatures, the persistence length of unbranched long-chain macromolecules appears to be significantly higher than that of a chain of free rotors; this is mainly due to the cumulative effects of the (weak) torsional potentials restricting internal rotations within the

carbon backbone of the chain, with steric repulsions playing a secondary role in the absence of large "side chain" substituents. We conclude that the unfolded (or "random coil") states characteristic of the actual biological macromolecules may possess significant intrinsic structural correlations and, as such, are far from being truly random.

4.2 EFFECT OF TEMPERATURE AND SOLVENT ON MACROMOLECULAR DYNAMICS

4.2.1 Unfolding of DNA Hairpins

KIS model of hairpin unfolding. Hairpins are common structural motifs of nucleic acids and are crucial for their tertiary structure and function (186). RNA and DNA hairpins play important regulatory roles in transcription and replication as well as mutagenesis facilitation (187-189). Understanding their stability and (un)folding kinetics is, therefore, likely to shed light on the relationship between hairpin structure and functional dynamics. Furthermore, due to their small size and simplicity relative to proteins and multi-loop nucleic acids, the DNA hairpin structures explored here are ideal benchmark systems for the development of robust theories of macromolecular dynamics.

From the experimental perspective, melting curves at equilibrium globally exhibit a two-state behavior. Recent work, however, suggests that DNA/RNA hairpin (un)folding may involve intermediate state(s). For example, master equation methods, and MD simulations predict multiple pathways as well as misfolded traps for RNA hairpin kinetics (190-192). Fluorescence correlation spectroscopy, or FCS (193-195), has inferred the

presence of intermediates and, given the flow and diffusion rates of the experiments, established a sub-millisecond time scale for the intermediate state (194). Studies involving time-resolved spectroscopy following a laser-induced *T*-jump, typically with nanosecond or longer time resolution, also find evidence of intermediate states (196, 197). For example, UV absorbance following a *T*-jump on short RNA hairpins suggested non-two-state microsecond unfolding kinetics for a range of temperatures and loop sequences (196). Recently, with ultrafast (sub-nanosecond) temporal resolution, utilizing both absorption of the bases and fluorescence probes to elucidate the roles of stacking and loop closure, respectively, the state-of-the-art *T*-jump study of Ma *et al.* provided direct evidence of collapsed intermediate state(s) for a DNA hairpin at temperatures higher than the melting temperature (196, 198). Such states, "collapsed but not folded," are also important for protein folding and may involve hydrophobic and/or secondary structure collapse (199, 200).

In what follows, we introduce an analytical model with which to elucidate the (un)folding kinetic pathways and intermediate states characteristic of quasi-1D biological motifs such as DNA/RNA hairpins (201), as well as protein secondary structures (Section 5.1.2). As such, it is termed the kinetic intermediate structure (KIS) model. We note that although the model describes macromolecular *kinetics*, only experimentally-determined *thermodynamic* parameters and diffusion time scales are required as input.

The (temperature-dependent) free energy of any macromolecular state relative to a reference state (usually either the initial or the final state representative of the interconversion process under study) is, generally, given by: $\Delta G(\boldsymbol{p}) = \Delta H(\boldsymbol{p}) - T\Delta S(\boldsymbol{p})$,

where ΔG, ΔH, and ΔS are the differences in free energy, enthalpy and entropy, respectively, between the chosen state and the reference state, T is the absolute temperature, and $\boldsymbol{p}$ is a set of order parameters. Ideally, $\boldsymbol{p}$ is chosen such as to partition the entire state space of the studied macromolecule into a complete (i.e., comprehensive) set of structurally distinct states. The resulting (coarse grained) free energy landscape describing the process of interest is then a projection of the (complete) free energy hyperspace onto $\boldsymbol{p}$. Within the framework of the free energy landscape representation, the (equilibrium) statistical weight of the macromolecular ensemble populating state $\boldsymbol{p}$ is proportional to the logarithm of $\Delta G(\boldsymbol{p})/kT$. In contrast, to study the *non-equilibrium* structural dynamics following, e.g., an external perturbation such as an ultrafast T-jump, which cannot be accounted for using the conventional Boltzmann formalism, the following steps can be undertaken. First, the topology of the free-energy landscape is to be identified for both the initial and final states. (We note that although a generalized free-energy landscape changes continuously in time following perturbation, in the present work we are mostly concerned with structural relaxation processes for which the time scale of the initiating perturbation is significantly shorter than that of the processes ensued.) Second, the possible transitions and their associated barriers are to be determined. Third, a macromolecular ensemble distributed according to the initial (unperturbed) conditions can be placed onto the free-energy landscape representing the final conditions and allowed to evolve with time. In this way, kinetic intermediates, dominant pathways and the associated time scales can all be obtained as the ensemble equilibrates to the final state.

Notably, the kinetic processes taking place on a generic free energy landscape are highly non-linear. In addition, although a major strength of computational methods is in the

elucidation of the actual spatial structures of studied macromolecules, the challenge is to consolidate the vast amount of resulting information in a comprehensive yet clear manner. To represent the ensemble-level evolution of macromolecular structures possessing hundreds of thousands of degrees of freedom, coarse graining of the atomic-scale details to two or three variables is often required. For example, MD trajectories are typically projected onto the effective order parameters such as percent native, or non-native, base contacts (NC/NNC) or the RMSD from the native structure (191, 192, 202). However, we note that dissimilar structures may be characterized by very similar values of NC/NNC or RMSD. Therefore, to properly account for the actual structural changes induced by external perturbations, achieving the correct balance between comprehensiveness and structural specificity is of supreme importance.

In regard to the secondary structure (un)folding, the single sequence approximation (SSA), which implies that there exists no mote than a single continuous island of order per macromolecule, is often a realistic description of the structural dynamics involved due to the nucleation barrier associated with the formation of each separate ordered island. In nucleic acids, this barrier is due to the multiple destacking events necessary to initiate an internal loop. The SSA, which serves as a foundation for the equilibrium and kinetic studies of helix-to-coil transitions, is known to successfully describe (un)folding in polypeptides (41) as well as nucleic acids (203); it only breaks down for relatively long chains (for which there are many possible interior nucleation sites). Thus, Ares *et al.* demonstrated, via Monte Carlo simulations carried out on double stranded DNA, that internal bulges are only significant for continuous internal A/T stretches of length $l = 20$ or

more (204). In what follows, we limit the analysis to DNA hairpins characterized by the stem length $l = 6$, for which the SSA appears to hold.

Unlike the case of *folding*, for which the interstate barriers appear to dominate the associated kinetics (Section 5.1.2), the non-equilibrium populations of states forming favored *unfolding* trajectories are determined solely by the equilibrium free energies of those states. This assumption, denoted the *reversible sampling approximation* (RSA), is legitimate to the extent that, at a given time during melting, the zipping and unzipping processes are frequent enough to locally reach detailed balance away from the unfolded state. Therefore, the RSA allows the kinetics of interest to be determined directly from the (relative) free energies of intermediate states, which can be calculated using tabulated thermodynamic parameters. Importantly, our ensemble-converging MD simulations (described below) indicate that both the SSA and RSA appear to hold for all temperatures reported here.

Making use of the validity of the SSA and RSA, in the following we introduce the KIS model of DNA/RNA unfolding kinetics, which utilizes experimentally-measured thermodynamic parameters. In Section 5.1.2, the model is further extended to describe folding of α-helical proteins. For DNA hairpins, we consider the (native) Watson-Crick base pairs, and choose the reaction coordinates i and j to be the number of broken (unzipped) base pairs on the loop and free ends of the stem, respectively (Figure 4.2; notably, the choice of coordinates i and j implicitly constrains the model to the SSA). All intermediate states are then represented by unique points on a 2D coordinate grid (i, j), with the native state of the hairpin located at (0, 0). The only state that does not have a unique

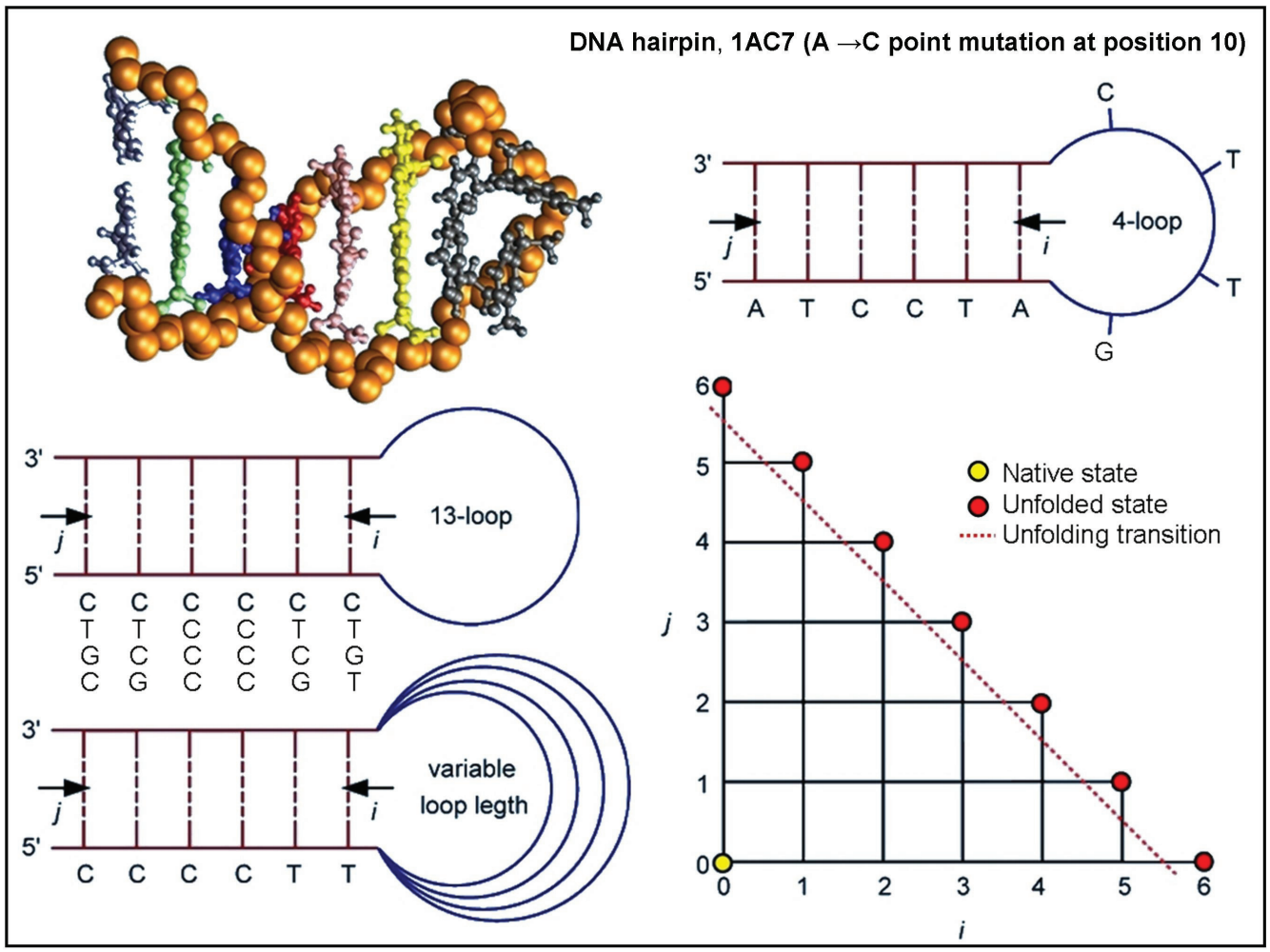

Fig. 4.2. KIS model of DNA hairpin unfolding. The structures studied are the 1AC7 hairpin: 5′-ATCCTA-GTTC-TAGGAT-3′, with is shown structurally (top left) and schematically (top right). In addition, a wide range of stem sequence permutations (center left) and loop lengths (bottom left) were studied. The KIS model parameterizes the unfolding free energy landscapes of these hairpins on the (*i*,*j*)-coordinate space (bottom right). Native states of the hairpins reside at (0,0) and (partially) unfolded states (*i*,*j*) correspond to *i* broken base pairs on the loop end and *j* broken base pairs on the free end of the stem. Note that all of the points (*i*,6 − *i*), *i* ≤ 6, situated on the diagonal of the grid are degenerate within the framework of our model as they represent the ensemble of totally unfolded states. The unfolding predictions of the KIS model for 1AC7 are compared with ensemble-scale explicit-atom molecular dynamics simulations.

representation on the above landscape is the unfolded-state ("denatured") structure ensemble, which is represented by the points on the diagonal boundary of the coordinate space. We note that every given state (i, j) corresponds to a population of structures that share the same base pairing but may differ in their detailed atomic coordinates. The coarse grained representation of the (complete) free-energy hypersurface $\Delta G(i, j)$ is obtained by calculating the free energy for every (i, j)-state with respect to that of the native state (0, 0), using the experimentally-measured thermodynamic parameters employed by Kuznetsov *et al.* (205).

Following the assumption made by Poland and Scheraga (206), each base pair is allowed to be either broken or intact, with the overall energetics determined by base pairing contacts, nearest-neighbor stacking, and the length of the loop (207, 208). The relative free-energy of the state (i, j) is calculated by:

$$\Delta G(i, j) = \Delta H_{\mathrm{p,s}}(i, j) - T\Delta S_{\mathrm{p,s}}(i, j) + \Delta G_{\mathrm{init}}(i, j) + \Delta G_{\mathrm{loop}}(i, j)\,. \qquad [4.5]$$

Individual terms in Equation 4.5 are defined as follows. $\Delta H_{\mathrm{p,s}}(i, j)$ and $\Delta S_{\mathrm{p,s}}(i, j)$ are the differences in pairing-stacking enthalpies and entropies, respectively, between the state (i, j) and the native state (0, 0); each term represents the summation over all base pairs of state (i, j). The stacking parameters were obtained from the studies of Benight and coworkers (207, 208), and the pairing parameters from Klump and Ackermann (209) and Frank-Kamenetskii (210). Although there exist empirical corrections for calculating the thermodynamic parameters at any given salt concentration (211), in the present work the simulations were performed for 100 mM NaCl solutions, for which the thermodynamic

parameters were obtained. Because the above parameters are temperature-independent to a good approximation (212), free energies were obtained for a wide range of temperatures with the aid of Equation 4.5.

The free energy changes associated with the native-contact rupture and formation were computed as follows. The free energy difference due to initiating a new contact can be expressed as:

$$\Delta G_{\text{init}}(i,j) = \begin{cases} \dfrac{kT}{2}\ln < \sigma >, & i + j = 6, \\ 0, & i + j < 6, \end{cases} \qquad [4.6]$$

where k is the Boltzmann constant and the initiation parameter $<\sigma> = 4.5 \cdot 10^{-5}$ is averaged over the 10 unique types of base-stacking interactions; given the 4 DNA bases (G, C, A, and T), there exist 16 base-stacking permutations, with 6 of those permutations being redundant. Finally, $\Delta G_{\text{loop}}(i,j) = -kT\ln[w(n+2i)] + kT\ln[w(n)]$ is the free energy difference that arises from changing the loop size upon unzipping in the i-direction, where n is the number of bases forming the loop in the native state of the hairpin. Note that for each base pair unzipped from the loop end (i.e., in the i-direction), the loop length increases by two residues. The end-loop weighting function $w(n)$, obtained by Kuznetsov *et al.*(205), is given by: $w(n) = V_r g(n)\sigma_{\text{loop}}(n)\left[(2/3)\pi b^2\right]^{-3/2}$, where b is the Kuhn length for a single-stranded DNA polymer, $V_r = 4\pi r^3/3$ is a reaction volume with a characteristic radius r in units of nm, within which the bases at the two ends of the loop can form hydrogen bonds (205), and $g(n)$ is the Yamakawa-Stockmayer probability of loop-closure for a worm-like chain with n bases (213), which can be expressed as:

$$g(n) = \begin{cases} \dfrac{1}{N_b^{3/2}}\left(1 - \dfrac{5}{8N_b} - \dfrac{79}{640N_b^2}\right), & N_b > 1, \\ \dfrac{1.51 \cdot 10^3}{N_b}(1 - 0.81242N_b)\exp\left(\dfrac{-7.0266}{N_b}\right), & N_b \le 1. \end{cases} \quad [4.7]$$

The numerical coefficients for $N_b \le 1$ in Equation 4.7 are chosen to give a smooth function for all n, and $N_b = h(n+1)/b$ is the number of statistical segments (Kuhn lengths) in a hairpin loop with n bases, where h is the distance between adjacent nucleotides. For single-stranded DNA, $b \approx 2.6$ nm, $r = 1$, and $h = 0.52$ nm (205, 214). With these values, $N_b > 1$ for $n > 4$. The free energy parameters employed in the KIS model are loop-sequence independent for hairpins with loop sizes greater than 4 (215). To account for the higher stability of smaller loops, which arise from intra-loop interactions (216-218), a correction given by $\sigma_{loop}(n) = <\sigma>^{1/2} + C_{loop} N_b^{-\gamma}$ is introduced (205). The empirical parameters for a hairpin with six stem bases are $C_{loop} = 9.0$, and $\gamma = 6$ (205). Although there is some experimental uncertainty associated with these parameters, the resulting errors mostly affect the free energy difference between the partially folded states and the unfolded state. Relative free energies of intermediate states with respect to the native state of the hairpin, and therefore the (un)folding trajectories, are not sensitive to the errors in these parameters.

Results: KIS unfolding trajectories. The KIS model introduced above was used to construct the free energy landscape that guides temperature-induced unfolding of the DNA hairpin with sequence 5'-ATCCTA-GTTC-TAGGAT-3' (Figure 4.2, top right). The native-state structure of the hairpin was obtained from NMR measurements (PDB entry 1AC7) (34, 219), except for a point mutation in which the adenine residue at position 10 was

replaced with a cytosine residue. The hairpin sequence was chosen to enable comparison of the KIS model predictions with MD simulations starting from the experimentally obtained structure of the hairpin (see below). Because the chosen hairpin is characterized by loop length $n = 4$, the point mutation was performed to obtain a tetraloop sequence that would preclude significant loop sequence-dependent stabilizing interactions.

For the studied DNA hairpin, the free-energy landscapes as obtained for the temperature range of $300 \leq T \leq 400$ K using the KIS model are presented in Figure 4.3. From these results, the melting temperature T_m, as defined by the temperature at which the population of the native state and the totally unfolded state are equal, was estimated to be 320 K. We note that all intermediate states (i, j) have a higher free energy than that of (0, 0) for temperatures in the vicinity of T_m. Thus, for $T \approx T_m$, (un)zipping has no kinetic intermediates on the free-energy landscape due to the barrier formed by partially unfolded states (Figure 4.3, top left). The free-energy barrier decreases with increasing temperature (Figure 4.3, bottom right). However, instead of leading directly to monotonic unfolding at some threshold temperature, the energy landscape develops a kinetic intermediate state at (0, 2) which is lower in free energy than (0, 0), but must surmount a barrier to completely unfold to the global-minimum free-energy state (Figure 4.3, top right). The locally-stable (0, 2) intermediate state persists for $340 \leq T \leq 365$ K. Within this temperature range, a fast unzipping of A-T base pairs from the free end (j) of the hairpin leading to the intermediate state (0, 2) is followed either by a slower unzipping of the G-C base pairs from the free end (j) or unzipping of A-T base pairs at the loop end (i).

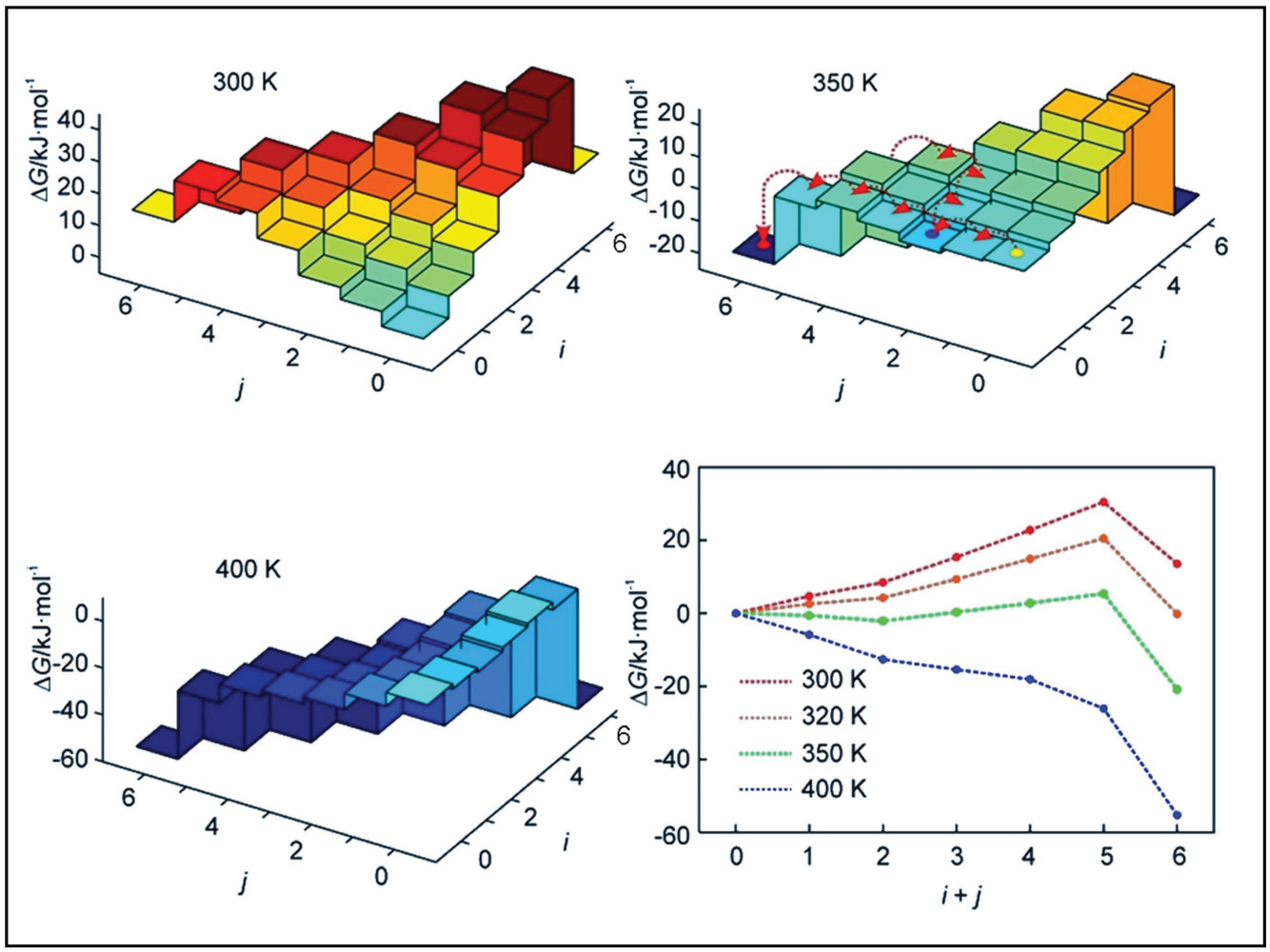

Fig. 4.3. Free-energy landscapes of the 1AC7 DNA hairpin as obtained from the KIS model. The landscapes are computed (see section 4.2.1 in Text) for the folded (top left), kinetic intermediate-mediated unfolding (top right), and downhill unfolding (bottom left) temperatures. Note the dramatic temperature dependence of $\Delta G(i,j)$. At T = 350 K, likely dynamic trajectories visiting the intermediate state at (0,2) are superimposed on the landscape (top right). Most likely (un)folding pathways characteristic of the above landscapes are represented by 1D profiles with the adjacent (i,j)-states connected by dotted lines, and magnified to illustrate the onset of the kinetic intermediate state; note that the barrier for (un)zipping between the states, which may contribute to the overall barrier, is unknown (bottom right).

For $T > 365$ K, the above barriers vanish and the hairpin exhibits monotonic unfolding at $T = 400$ K (Figure 4.3, bottom left). For the temperature range of $300 \leq T \leq 400$ K, the most likely (un)folding pathway can be determined (Figure 4.3, bottom right). This pathway is traced from the native state (0, 0) to the unfolded state of the hairpin by choosing, at each point, the (un)zipping direction associated with the greatest loss (or least gain) of free energy. With increasing temperature, the pathway evolves from a barrier crossing ($T = 320$ K) to an unfolding valley ($T = 350$ K) to monotonic unfolding ($T = 400$ K). Furthermore, for $T = 350$ K the intermediate state (0, 2) has lower free energy than the native state (0, 0), with a barrier of 8 kJ·mol^{-1} (2.7 kT) separating (0, 2) from the unfolded-structure ensemble, which indicates that (0, 2) is a kinetic intermediate state. In the following, we assess the validity of the assumptions as well as the kinetic predictions of the KIS model using ensemble-convergent MD simulations.

Results: MD unfolding trajectories. To test the findings of the KIS model, we performed MD simulations on the hairpin. The number of trajectories was sufficiently large to achieve ensemble convergence, i.e., such that the unfolding behavior of the ensemble would not significantly change when varying the number ($100 \leq n \leq 500$) of trajectories included in the data analysis. For the MD simulations, the starting-point structure of the hairpin was obtained from the PDB, as described above. The hairpin was centered in a rhombic-dodecahedron primary simulation cell with an initial box length of 60 Å. In addition to the hairpin, 4,856 TIP3P water molecules (220), 24 sodium ions and 9 chloride ions were added as a 100 mM salinity solvent yielding an electrically neutral system comprising 15,109 atoms. MD simulations were performed with the GROMACS suite of programs

using the all-atom AMBER99 force field and periodic boundary conditions imposed on the system under study (170, 221-223).

The SSA implicit in the KIS model was first verified by the MD simulations. Bonding contacts between any pair of nucleotides were determined for all trajectories. Two nucleotides were denoted to be in contact if at least one of the two (A-T pairs) or three (G-C pairs) Watson-Crick hydrogen bonds were formed. For the purpose of the analysis, a hydrogen bonding contact was defined by a donor-acceptor distance of 3.5 Å and an acceptor-donor-hydrogen angle of 30° (or less); the G_HBOND routine of the GROMACS suite of programs was used to identify the contacts. At lower temperatures ($T \leq 350$ K) almost all the MD trajectories satisfy the SSA. Variations occur due to the increased mobility of nucleotides at both ends of the stem leading, e.g., to out-of-plane bending of a nucleotide which then induces the displacement of the neighboring (stacked) nucleotide from its Watson-Crick position. The non-SSA fluctuations occur on time scales ranging from a few to hundreds of picoseconds. At higher temperatures ($T \geq 400$ K), mobility of individual nucleotides is further increased, leading to increased structural variability and a consequently reduced fraction of SSA-like structures. However, at all the temperatures reported here the SSA correctly describes the topology of at least 94% of the MD configurations obtained.

To enable comparison with the KIS model, the entire set of MD trajectories is projected onto the (i, j) coordinate space by calculating the probability, $p_{MD}(i, j)$, of the (i, j)-state being occupied. Since the SSA is valid for at least 94% of all trajectories, this

projection accounts for 94% (or larger fraction) of the MD ensemble. Knowing $p_{MD}(i, j)$, the effective coarse grained free-energy landscape, $\Delta G_{MD}(i, j)$, can be calculated using:

$$\Delta G_{MD}(i, j) = -kT \ln[p_{MD}(i, j)]. \quad [4.8]$$

We note that ΔG_{MD} is often denoted as the potential of mean force and, in the present work, it is associated with the non-equilibrium process of hairpin relaxation ($T \leq 350$ K) or unzipping ($T \geq 400$ K) following a 100 ps T-jump from the initial ($T_0 = 300$ K) to the final temperature T, with the initial state being identical, or in proximity, to the fully folded (0, 0)-state. In examining $\Delta G_{MD}(i, j)$, the results of Figure. 4.4 indicate that the unique intermediate state that arises at $400 \leq T \leq 500$ K is indeed (0, 2). This is in contrast to the findings made using the irreversible unzipping hypothesis, which would predict (0, 2) and (1, 1) being equally likely to be populated. Consequently, the reversible (un)zipping observed in our MD simulations supports the validity of the RSA.

Having found that the SSA and the RSA are both valid approximations for the DNA hairpin under study, we now evaluate the kinetic predictions of the KIS model using MD results. Shown in Figure 4.4 are $\Delta G_{MD}(i, j)$ landscapes as obtained from ensemble-convergent MD simulations for a range of temperatures. At $T_0 = 300$ K, the hairpin appears to populate the states (0, 0) and (0, 1) with approximately equal probabilities. With increasing temperature, the number of (i, j)-states available for sampling increases and, above $T = 400$ K, effectively all the (i, j)-states are sampled in the MD simulations. The question arises as to how to compare the landscapes $\Delta G(i, j)$ and $\Delta G_{MD}(i, j)$ for a given temperature T. Due to non-equilibrium sampling characteristic of the MD simulations, $\Delta G_{MD}(i, j)$ will not accurately reflect the equilibrium free energies $\Delta G(i, j)$, especially for

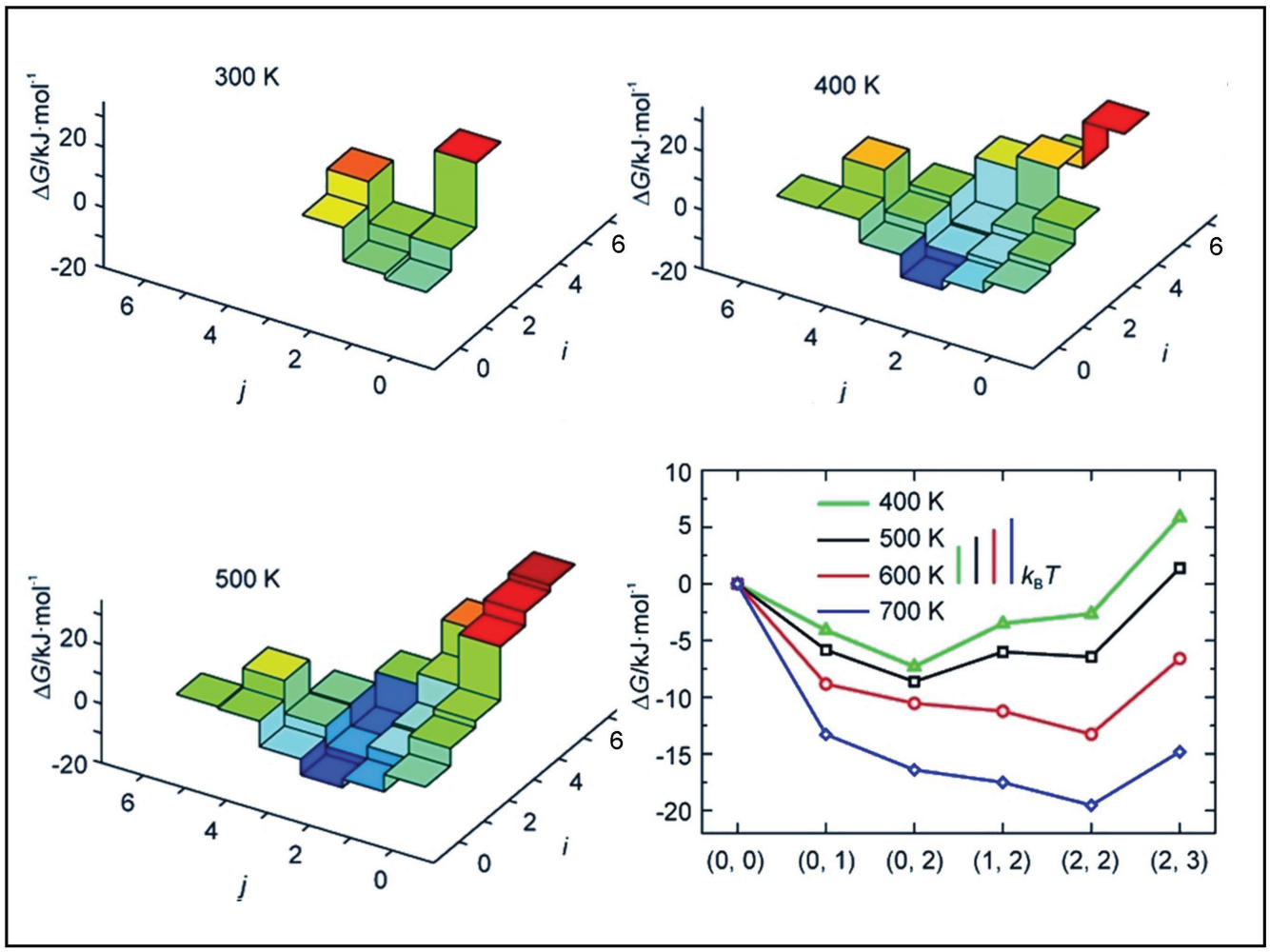

Fig. 4.4. Free-energy landscapes of the 1AC7 DNA hairpin as obtained from ensemble MD simulations. The free energy $\Delta G_{MD}(i,j)$ of the hairpin (see Figure 4.2, top right) is calculated by taking the logarithm of the population from the MD simulations that correspond to each point of the KIS landscape, for the folded (top left), kinetic intermediate-mediated unfolding (top right), and downhill unfolding (bottom left) temperatures. 1D profiles of $\Delta G_{MD}(i,j)$ along the most likely unfolding pathways are shown as well (bottom right). At lower temperatures, the free energy of some points on the KIS landscape near the unfolded state are artificially high due to insufficient sampling.

higher *T*-jumps. However, a minimum along the unfolding valley in $\Delta G(i, j)$ will correspond to the highest values of $p_{MD}(i, j)$, leading to a corresponding minimum in $\Delta G_{MD}(i, j)$.

An examination of $\Delta G(i, j)$ and $\Delta G_{MD}(i, j)$, Figures 4.3 and 4.4, respectively, demonstrates that the topological features predicted by the KIS model and MD are very similar, although the temperatures at which certain features arise are somewhat different. The above discrepancy may be rationalized in terms of the limited sampling of the coordinate space, which is a general drawback of the MD simulations. For low *T*-jumps, both MD results and the KIS model predict a two-state behavior associated with the barrier formed by the partially unfolded states. For sufficiently high *T*-jumps ($T \approx 350$ K for the KIS model and $T = 400$ K for MD), both $\Delta G(i, j)$ and $\Delta G_{MD}(i, j)$ show the existence of the intermediate state (0, 2) that is lower in free energy than (0, 0). Furthermore, the barrier between the kinetic intermediate state and the unfolded-structure ensemble is estimated by the KIS model to be 2.7 *kT* at $T = 350$ K, being the same order of magnitude as 4 *kT,* the MD estimate for the barrier at $T = 400$ K.

The effect of the intermediate on the unfolding time scales can be derived from the MD trajectories by tabulating the average number of intact native contacts in the stem as a function of time following the *T*-jump. As stated above, at $T = 400$ K the (rate limiting) barrier for unfolding separates the kinetic intermediate state from the unfolded-structure ensemble, the barrier height between the two states being approximately 4 k*T* (Figure 4.4, bottom right). Upon averaging over the MD trajectories, the unfolding data as obtained for $T = 400$ Kwere fitted by the sum of two exponentials with time constants τ_1 and τ_2; $\tau_1 = 45$

ns is characteristic of the fast unzipping from the native state (0, 0) to the kinetic intermediate state (0, 2), whereas $\tau_2 = 9$ μs is the time scale on which the intermediate is populated after the *T*-jump. At $T = 400$ K, where only a fraction of the hairpin trajectories were found to unfold within the MD simulation window, the probability of observing the kinetic intermediate state (0, 2) peaks at 30% after time ~ τ_1. The probability then decreases to a plateau at approximately 15% for the remainder of the simulation window, which indicates that the state (0, 2) persists on a time scale much longer than the duration of the window (i.e., τ_2), suggesting that (0, 2) is, indeed, a long-lived kinetic intermediate state.

At $T > 365$ K and $T > 400$ K for the KIS model MD simulations, respectively, both methods predict monotonic unfolding. According to the results of Figure 4.3, bottom right, for $T \geq 400$ K no local minimum exists on the $\Delta G(i, j)$ landscape within the framework of the KIS model. Although the local minimum in $\Delta G_{MD}(i, j)$ persists and, at higher temperatures, is shifted to the (2, 2)-state (data not shown), it no longer corresponds to a kinetic intermediate because the barrier between the local minimum and the global minimum (i.e., the completely unfolded state) decreases to the order of magnitude of the thermal quantum (kT). Interestingly, for $T \geq 500$ K, we detected a rapid increase in the probability of observing the local minimum state subsequent to the *T*-jump across the entire MD ensemble, which was followed by a somewhat slower decay to zero on a very similar time scale. Because the population of the state increases and decays on similar time scales, no accumulation of kinetic intermediate structures occurs. Correspondingly, there is a single exponential decay of the number of native contacts for $T \geq 500$ K, which leads to a monotonic two-state unfolding.

Summary. We conclude that, for the studied DNA hairpin, the analytical KIS model and ensemble convergent MD simulations both predict the same temperature-dependent kinetic behavior: barrier-crossing kinetics on the free-energy landscape for lower *T*-jumps ($T \leq 340$ K for KIS, $T < 400$ K for MD), three-state kinetics due to the long-lived intermediate state (0, 2) for intermediate *T*-jumps ($340 \leq T \leq 365$ K for KIS, $T \approx 400$ K for MD), and monotonic unfolding for higher *T*-jumps ($T \geq 400$ K for KIS, $T \geq 500$K for MD).

The KIS model, despite its simplifications, can accurately predict the relevant, structure-specific, kinetic behavior for the macromolecule. For a range of final temperatures above the melting temperature, intermediate states representing collapsed (but not folded) structures emerge as local valleys on the $\Delta G(i, j)$ landscape with lower free energies than that of the native fold. The intermediate state(s) are separated from the global free energy minimum populated by the unfolded-structure ensemble by a significant barrier, leading to non-two-state dynamics. To conclude, (i) the unfolding kinetics can be non-two-state for a wide range of *T*-jumps, (ii) the stem sequence of the hairpin determines the identity of the kinetic intermediate(s) and the most likely unfolding pathways, and (iii) the hairpin loop length affects the depth of the local minima on the free-energy landscape. Unlike the case of the polyalanine homopolymer (Section 5.1.2), for which misfolding intermediates appear to dominate the folding dynamics, the kinetic-intermediate states characteristic of DNA hairpin unfolding are the result of sequence inhomogeneity and loop entropy. Following the benchmark comparison with ensemble-converging MD simulations, the KIS model can be used to generate free energy landscapes for all the hairpin sequence permutations as well as for varying loop and stem lengths, and to determine the temperature range for which the two-state unfolding hypothesis breaks down; the base

pairing configurations of the intermediate states on such landscapes are readily obtained as well.

4.2.2 Unfolding Dynamics of DNA Double Helix in Vacuo and in Solution

Introduction. Under physiological conditions, the stability of a DNA duplex stems from a delicate balance of a number of competing forces and mechanisms (224). In particular, the Coulomb repulsion between negatively charged phosphate groups is compensated by stacking and hydrogen bonding interactions between DNA bases and by the screening effect of water and surrounding ions. (De)hydration is expected to affect the stability of the duplex because the change from aqueous to less polar solvents reportedly leads to pronounced conformational transitions and/or disruption of the helical pattern (225, 226). Similarly, it has been demonstrated by X-ray fiber diffraction that, upon variation of the relative humidity of fiber environment, the molecular structures assumed by DNA fibers vary from A-DNA to Z-DNA, and that the hydration-driven transitions between the DNA conformers are fully reversible (227-230). A recent determination of the structure of DNA in single crystals of nucleosome core particles revealed that the DNA is predominantly in the B-form with local distortions and irregularities, which facilitate its superhelical path in the nucleosome (231). Generally, there seems to be little reservation about prevalence of B-DNA occurring in vivo, although in various situations (e.g., around histones) the molecule adapts a bent configuration (232).

In light of the structural integrity of B-form DNA in solution, the physical source of this stability becomes a highly relevant topic. In particular, what features of the duplex are preserved in vacuo and are therefore "inherent" to the physics of DNA alone?

Because the replication and transcription of DNA are dynamical processes involving strand separation, an equally important goal is the understanding of the effect of solvent on both conformational and unfolding dynamics. Although biological macromolecules often tend to undergo local distortions, the structures they commonly assume are relatively robust, persisting in a wide variety of environments. Thus, according to electrospray experimental observations, DNA retains its major structural features even in the absence of the hydrating water layer (233-236). However, because direct experimental determination of detailed molecular structures of large, flexible biopolymers in vacuo and in solution is not feasible at present, theoretical methods remain the guiding force in exploring the configurational space of DNA/RNA in various environments (201).

Remarkably, the double-helix architecture of nucleic acids gives rise to a number of persistent structural features which can be efficiently exploited in experimental and theoretical studies. For example, by analyzing the diffraction from spatially aligned DNA fibers, both the helical symmetry and macromolecular structure of the fibers can be deduced (232, 237). This is typically accomplished by calculating the diffraction pattern using a "theoretical" (anticipated) structure of a *single* fiber which is presumed to be known with a sufficiently high resolution. By comparing the theoretical pattern with the "observed" (experimental) one, which originates from a large number of well-oriented, *coherently scattering* fibers, an improved structural model of the fiber may be constructed. However, of special interest in molecular biology is the study of structural and conformational changes which are free of the effects of solvation, crystallization or external ordering imposed on the specimen (182). Because of their large size and

unprecedented flexibility, DNA macromolecules possess a myriad of quasi-random structural configurations during the course of an order–disorder transition, and this complexity may, naively, suggest the masking of any significant change in diffraction. However, as we demonstrate in detail below, an accurate theoretical mapping of macromolecular ensembles which consist of hundreds of DNA duplex microstates indicates that the pronounced features of the quasi-periodic structure (*spatial resonance*) in DNA may be used as a natural measure of the disruption of the double-helix ordering in both space and time.

For over a decade, MD simulations have been capable of reproducing the structure and dynamics of large DNA macromolecules in aqueous solutions (238, 239). Although MD simulations of gas-phase nucleic acids have been performed even earlier for sub-nanosecond dynamics (240-242), in a recent computational study, Rueda *et al*. demonstrated for the first time that a somewhat distorted DNA duplex might be stable in the gas phase on the (sub)microsecond time scale (224). Importantly, the conformational transition due to vaporization of DNA should occur very rapidly given the size of the macromolecule under scrutiny (the complete equilibration is reportedly achieved on a few-nanosecond time scale). However, the extended duplex structures retain many features characteristic of the canonical (hydrated) DNA configuration irrespective of both the temperature ($T \leq 448$ K) and the neutralization protocol employed in the numerical simulations. Despite the similarities which reportedly exist between the equilibrated duplexes and C-type DNA, the former are better described as mechanically-stretched (elongated) double helices (243). Interestingly, the vaporization does not have a dramatic effect on the conformational preferences of DNA nucleotides. Thus, 60 to 90% of

canonical hydrogen bonds are preserved in the gas phase at 298 K. However, the fraction of non-canonical hydrogen bonds, which is negligible in solution, increases to ≤ 40% in the gas phase. The DNA bases remained well-stacked throughout the simulations, but the stacking direction was no longer parallel to the helix axis. No strand separations were observed in any MD trajectories.

In what follows, we discuss ultrafast structural dynamics of DNA unfolding in the gas phase as obtained from ensemble-convergent MD simulations carried out for a number of charged states (Q = –6 or –12) and T-jumps (T_0 = 300 K; 300 ≤ ΔT ≤ 1200 K) and make a comparison with unfolding dynamics of the same macromolecule in aqueous solution. In contrast to earlier theoretical work, the focus here is on the dynamics of DNA duplex unfolding, including the influence of hydration on the mechanisms and time scales characteristic of the process. In addition to the use of native-contact metrics to measure base pairing disruption, we also investigate the conformational dynamics of the duplex via the radial distribution functions of UED simulations. Though certain apparent implications for future UED experiments are briefly outlined below, the above-mentioned radial distribution functions are only invoked here as an intuitively appealing coarse-graining approach. As all analyses were performed on a large number of independent unfolding trajectories, we elucidate the dramatic effect of solvation on the structural and dynamical properties of DNA in an ensemble-convergent manner.

Computational: MD and electron diffraction simulations. In the present work, large-scale macromolecular ensembles of 5'-d(CGCGGTGTCCGCG)-3' DNA duplexes (PDB entry 1LAI) (244) were generated both in vacuo and in aqueous solution, and their

spatiotemporal evolution was monitored following a variety of *T*-jumps during the course of distributed MD simulations. Radial distribution functions of the full duplex as obtained for the (native-state) NMR structure in aqueous solution are depicted on the right in Figure 4.5. Also shown in the Figure are corresponding patterns of individual (uncoupled) DNA strands as they appear in the duplex. We note that the differences between the two kinds of patterns are striking, which suggests a very pronounced change in diffraction upon the complete separation of the strands (the experimentally-observed difference would be even larger because the uncoupled DNA strands would also lose their structural ordering associated with both stacking and helical-structure periodicity characteristic of the B-DNA structure). Importantly, calculated diffraction patterns of the macromolecule and the uncoupled strands are almost indistinguishable at $s > 3$ Å^{-1} which indicates that the features arising from the interstrand r_{ij} distances are largely concentrated in the innermost area of the scattering pattern.

A comparison of the radial distribution functions characteristic of full DNA duplex and separate strands (Figure 4.5) reveals that the strand-to-strand internuclear distances result in pronounced distance-density accumulations (spatial resonance) at $r \approx$ 13, 16, and 18 Å. Because a typical hydrogen-bonded DNA base pair is about 10 Å in size, the resonant distance density accumulations at $r < 10$ Å (Figure 4.5, top right) are largely due to base pairing and stacking ($r \approx 5$ Å) in the duplex. For the single strand, the ($r \approx 5$ Å) base-stacking peak of $f(r)$ persists whereas the inter-strand ($r \approx 7$ Å) base pairing peak disappears, as expected. Thus, order–disorder transitions in DNA can be explored by monitoring the structural resonance arising from these base pairing and stacking distances (see below).

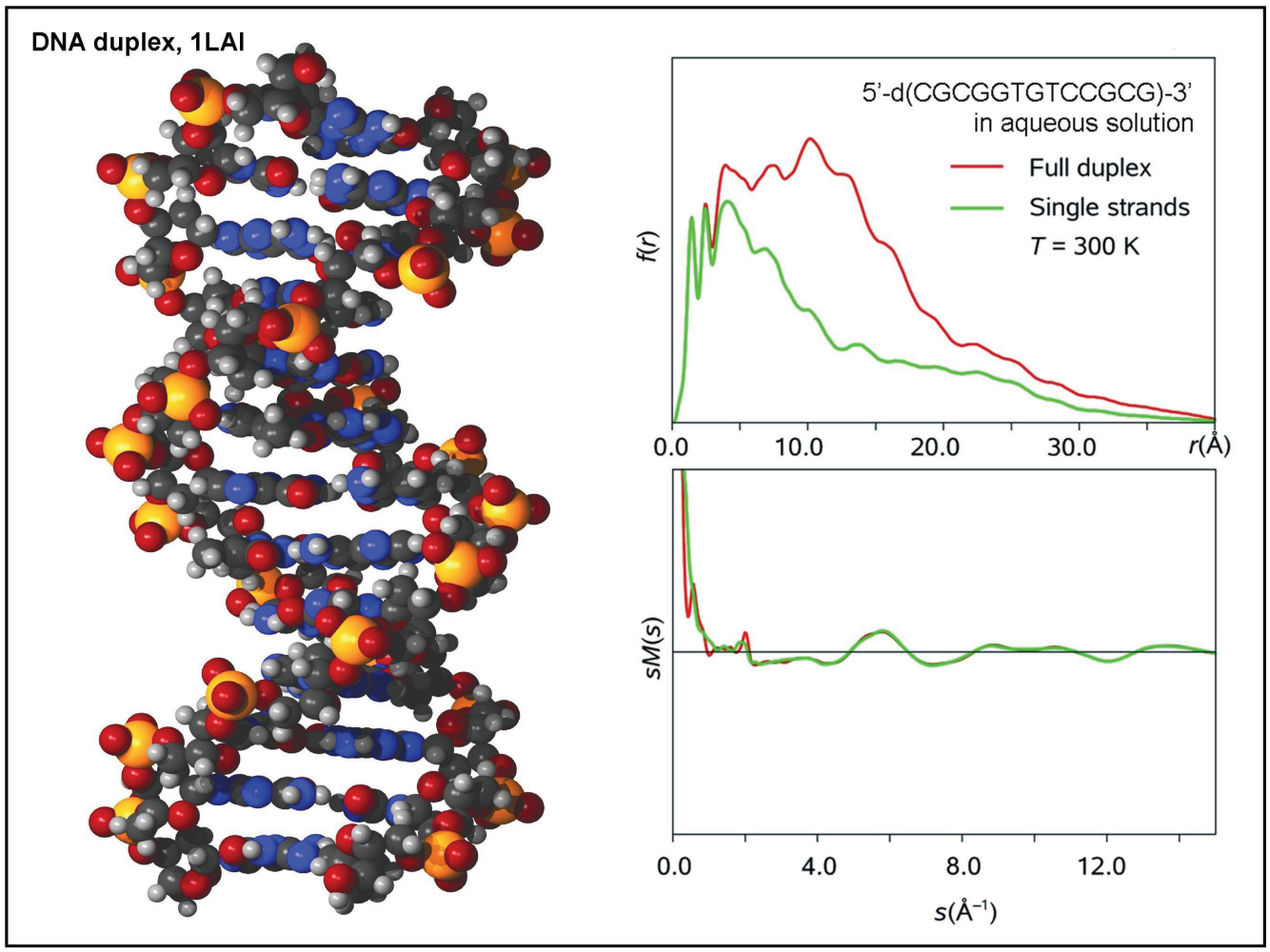

Fig. 4.5. Electron diffraction simulations of the DNA double helix. The experimentally determined B-type structure (left) can be projected onto a one-dimensional radial distribution function (top right), which is the Fourier transform of the scattered electron intensity in a typical gas electron diffraction experiment (bottom right). Intuitively, the radial distribution function is the histogram distribution of pair-wise atomic distances in the structure, weighted to accentuate local structure (see section 3.3). Note the peak at r = 5 Å corresponding to the base stacking distance along the DNA strand, which is present in both the single strand and the duplex, as well as the r = 7 Å peak corresponding to inter-strand base-pairing, which is only present in the duplex.

As stated above, the ensemble-convergent MD simulations were carried out for the DNA duplex 1LAI in the gas phase and in aqueous solution. The simulations were performed using the CHARMM (118) suite of programs with the all-atom CHARMM27 (245, 246) force field parameters describing nucleic acids. The starting point geometry of the duplex was that of the NMR experimental structure. Because DNA macromolecules are known to have an intrinsic negative charge which is responsible for their acidic character, and which is concentrated on the backbone phosphate groups, the DNA ensemble in vacuo is characterized by a distribution of charged states due to varying degrees of macromolecular protonation. According to the electrospray experiments (233), the charged state of a typical DNA duplex in vacuo should be about 1/4 of its intrinsic negative charge. Because the duplexes studied here possess an inherent charge of –24, the charged states of Q = –6 and –12 were assumed in vacuo in order to assess the impact of different values of Q on the macromolecular dynamics.

Rueda *et al* (224) manipulated the charged state of their DNA duplexes by either neutralizing selected phosphate groups or scaling down the negative charge of all phosphate groups by the appropriate fractional factor. It was found that the structural stability was essentially independent of neutralization scheme. In the present study, we applied the latter methodology and scaled the phosphate-group charges of the PDB structure by factors of 1/4 and 1/2 in order to obtain the desired charged states (see above). For each of the two charged states, the initial PDB structure was energy-minimized for 12000 steps in vacuo and then heated to T = 300 K and pre-equilibrated for 100 ps. In the gas phase, the structure was further equilibrated for 400 ns at T = 300 K. From the latter equilibration step, n = 200 random DNA configurations were obtained to represent the

(isolated) macromolecular ensemble at T_0 = 300 K. To assess the ensemble-averaged *T*-jump dynamics of the DNA duplexes, the above-mentioned equilibrated macromolecular ensemble was used as a starting point for three sets of n = 200 independent heating trajectories representing the 300 K, 600 K, and 1200 K *T*-jumps (Figure 4.6). For each of these trajectories, the starting-point macromolecular configuration was heated to the final temperature, $T = T_0 + \Delta T$, within 1 ps and was subsequently allowed to evolve for up to 100 ns to obtain the duplex-unfolding statistics.

For the solution-phase MD simulations, the original PDB structure was centered in the cubic primary-simulation cell with the initial box length of 60.5 Å, which contained 7007 TIP3P water molecules and 24 sodium atoms for a neutral system. For all solution-phase simulations, MD trajectories were obtained using periodic-boundary conditions with long-range electrostatics computed via the PME method (121). Following 20,000 steps of energy minimization, the box was heated to T_0 = 300 K within 10 ps. Subsequently, the system was allowed to evolve for 1 ns at constant temperature and 1 atm pressure. From this trajectory, n = 200 random water-DNA configurations were chosen to represent the solvated structure. Using these configurations as the starting-point structures, the *T*-jump MD trajectories were obtained as follows. The system was heated to T = 600 K within 10 ps and was subsequently maintained at constant temperature and 1 atm pressure for 6 ns to obtain the duplex-unfolding statistics (see below). The MD simulations and data analyses were then repeated for T = 500 K.

To assess the fraction of intact native (Watson-Crick) base pairing contacts as well as both local and global structural changes throughout the ensemble, two complementary

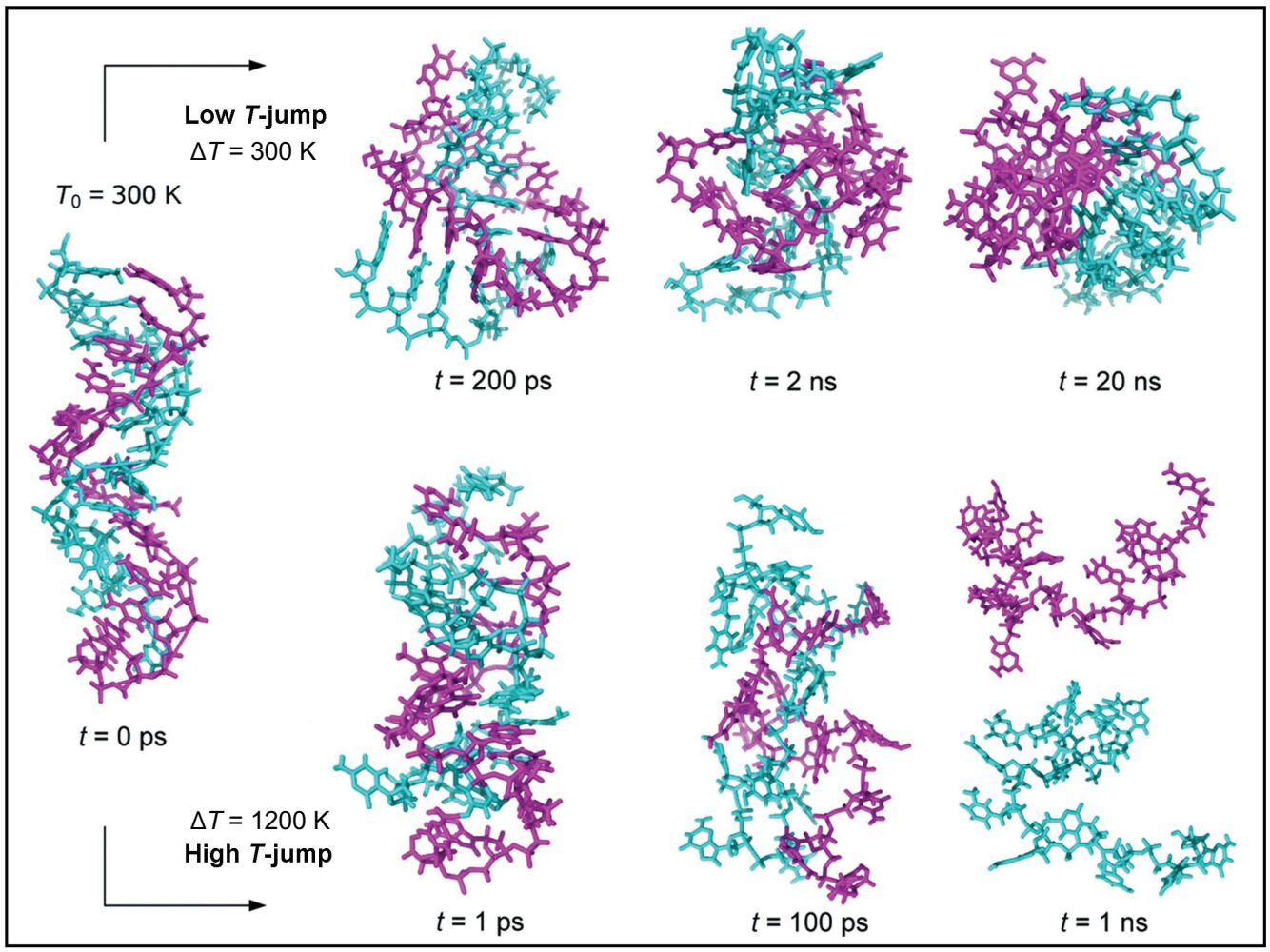

Fig. 4.6. Temperature-induced unfolding of the DNA double helix. Starting from the equilibrated gas phase ensemble, temperatures jumps of 300, 600, and 1200 K were applied on the entire ensemble. Representative unfolding snapshots are shown illustrating the different temperature dependences of local versus global unfolding. Although both the -6 and -12 charge states were studied, only the -6 charge state is reported due to the similarity of the results. The 600 K jump results, being qualitatively similar to those of the 1200 K jump, are not shown. In addition, double helix unfolding was performed in the solvated phase at much lower temperature jumps for comparison.

types of data were collected as a function of time. First, for all sets of independent trajectories, the fraction of each native base pairing contact remaining intact at time t was calculated as follows. The fraction of intact hydrogen bonds was obtained for every Watson-Crick base pair and further averaged over the n = 200 independent trajectories to obtain the average decay of each native contact as a function of time. A hydrogen bond was defined to be 100 % intact if the distance between the donated proton and the nitrogen or oxygen atom (the hydrogen acceptor) was less than 1.8 Å and the straight line joining the proton and the hydrogen acceptor was no more than 90 degrees out of the plane defined by the aromatic rings of the base pair. In addition, the smoothness of the transition between a fully intact and a fully broken hydrogen bond was enforced using an exponential attenuation of the bond strength such that the hydrogen bond would be 1/e-fold intact at a distance of 2.5 Å. These criteria are consistent with established conventions for geometry-based hydrogen bond determination (247), and it should be noted that the (fast) process of base pair disruption renders the results thus obtained insensitive to variance in the threshold values used.

Second, the ensemble-averaged radial distribution functions, $\langle f(r, t)\rangle_n$, were calculated for a variety of time points following the T-jumps using the in-house UEDANA diffraction simulation code with an artificial damping factor of k = 0.02 Å^2 to compensate for the unwanted oscillations induced by a finite data range ($s_{max} < \infty$). The RMS amplitudes of thermal vibrations of the DNA duplex were estimated using empirical equations (see Section 3.3 for methodological details).

Results: base pairing separation. For the lower (ΔT = 300 K; top) and higher (ΔT = 1200 K) *T*-jumps illustrated in Figure 4.6, the ensemble-wide DNA unfolding behavior is presented in Figure 4.7. The fraction of intact base pairs is plotted as a function of time (right), and the position dependence of base pair breaking is displayed using the color-coded representation (left). There are three noteworthy observations pertinent to DNA stability and (un)folding in the gas phase. First, the mechanism of DNA unzipping (or native-contact rupture) turns out to be robust with respect to both the charged state assumed throughout the ensemble and the *T*-jump experienced by macromolecules. As evidenced by our MD simulations, an increased Coulomb repulsion between the strands results in somewhat decreased backbone ordering and smaller gliding shifts of stacked DNA bases which are more reminiscent of the canonical structure of B-DNA. However, for both charged states used in our study, Q = –6 and –12 (the latter one is omitted from the discussion for conciseness), the DNA duplex tends to undergo very similar unfolding processes all the final temperatures considered, with larger *T*-jumps leading to shorter times required for complete base pair unzipping (5, 150, and 5000 ps for ΔT = 1200, 600, and 300 K, respectively). The order of base pair disruption is preserved for all values of ΔT, whereas the unfolding time scales and behavior are unaffected by the charged state. Importantly, the gas-phase unfolding induces the formation of a "bubble" (broken base pair sequence) in the interior of the duplex, such that the duplex is split into two distinct base paired regions in violation of the SSA.

This is in contrast to the unfolding behavior of the DNA duplexes in solution characteristic of $\Delta T \leq 200$ K (Figure 4.8, bottom), for which the SSA appears to hold. The

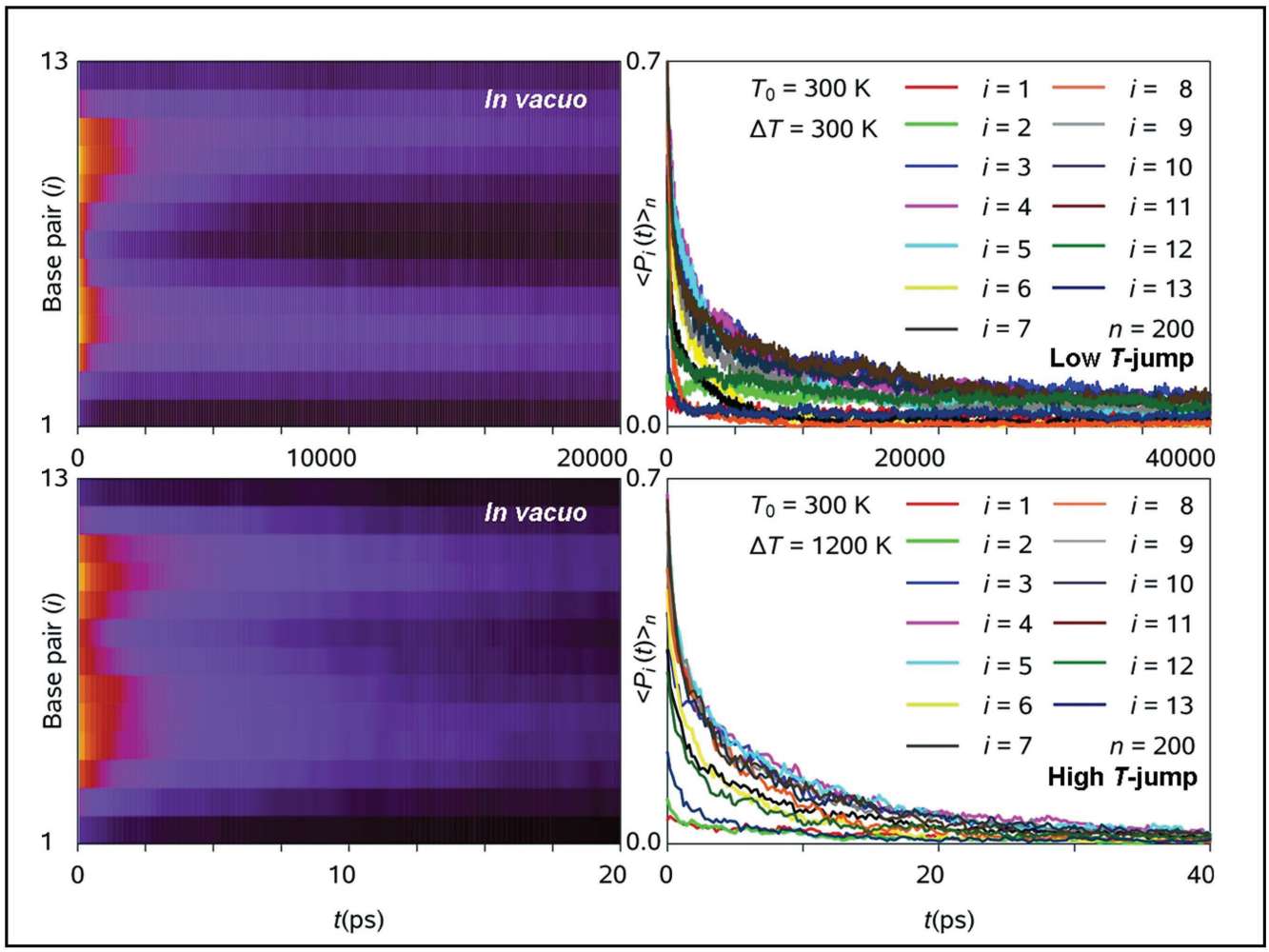

Fig. 4.7. DNA base-pair breakage as a function of temperature in vacuum. Upon heating with lower (top) and higher (bottom) temperature jumps , base-pair breaking as a function of time is shown along the DNA sequence (yellow: 100% intact, black: 0% intact; left), and the decay of each contact is plotted (right). In the gas phase, early breakage of the base-pairs in the center of the DNA sequence creates two separate base-pairing islands, and the drastic difference in unfolding timescales for lower and higher temperature jumps is consistent with crossing an enthalpic barrier when breaking the base-pairing.

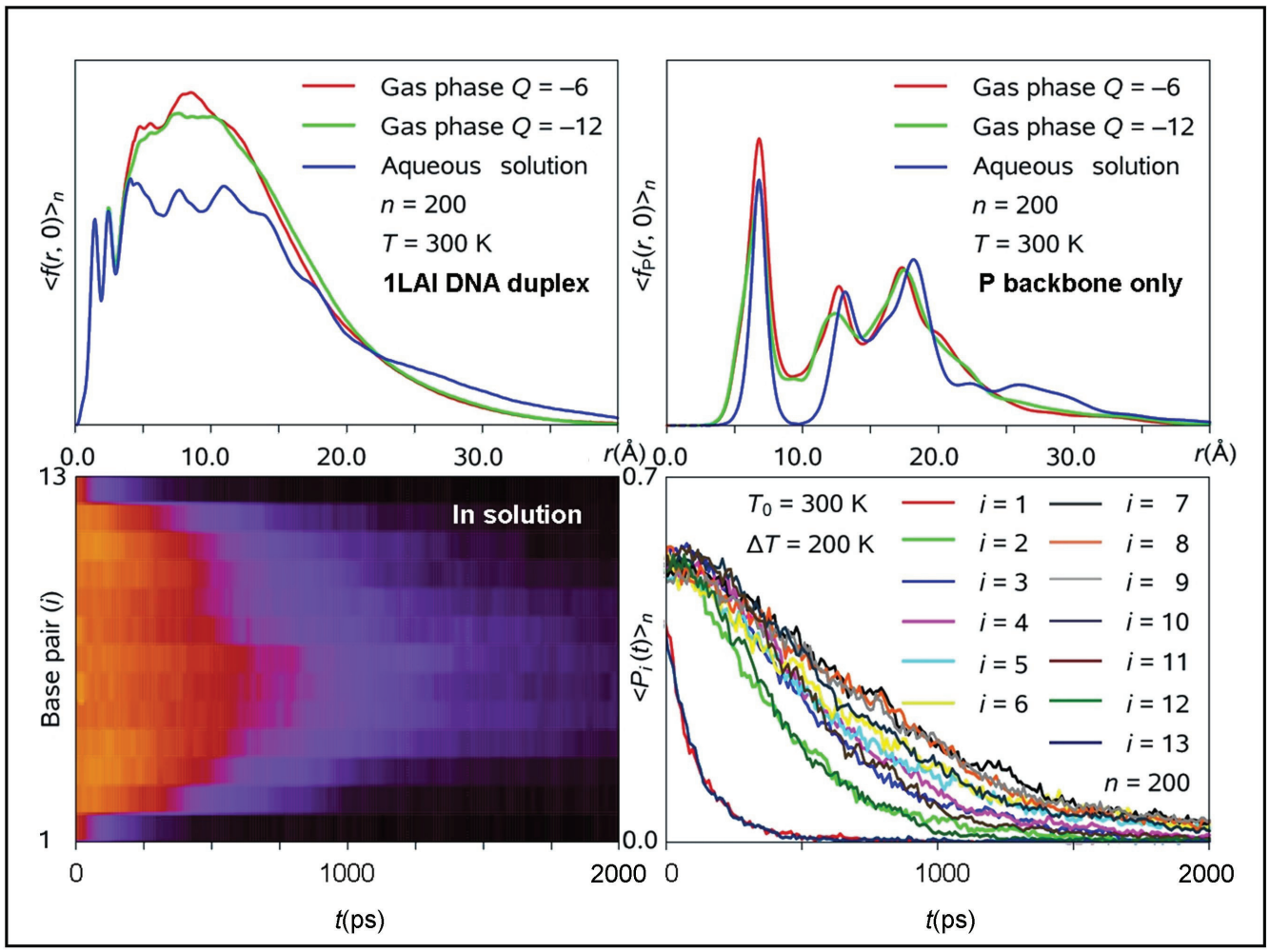

Fig. 4.8. Effect of water on DNA duplex structure and base-pair unfolding dynamics. Upon MD equilibration in solution, as well as in gas phase for two possible charge states for an ensemble of n = 200 structures, the ensemble-averaged radial distribution function gives a picture of the conformational distribution of all three states (top). The DNA double helix in aqueous solution maintains the classic B-form structure with periodically repeating base stacking and pairing distances, as seen in the resonance peaks at r = 5 and 7 Å in $\langle f(r)\rangle_n$, respectively; in contrast, the duplex in the absence of water loses its *global* structural periodicity due to deformations that tend to compactify the duplex (top left). Nevertheless, the local base-pairing is robustly maintained despite the global deformation, as can be seen in the preservation of the base pairing resonance peak $\langle f_P(r)\rangle_n$ of the phosphorus backbone (top right). Upon heating, base-pair breaking as a function of time is shown along the DNA sequence (yellow: 100% intact, black: 0% intact; bottom left), and the decay of each contact is plotted (bottom right). The base pair unfolding occurs from the ends of the double helix rather than forming a bubble in the center, in contrast to the vacuum case (Figure 4.7).

main reason for the validity of the SSA in solution is the necessity to disrupt an extra set of stacking interactions in order to allow for the formation of an interior bubble. In the case of DNA duplex in the gas phase, this is compensated for by the presence of (weaker) A-T base pairing contacts in the center of the duplex. In the solution phase, however, the hydrogen bonding with water decreases the stabilizing advantage of G-C base pairing as compared to A-T base pairing, thereby favoring unzipping from the ends of the duplex. The importance of hydrogen bonding within the duplex in the gas phase also contributes to the relatively long time required for complete strand separation, which was only observed for ΔT = 1200 K. At this (very high) temperature, although hydrogen bonding of the native contacts was broken in a few picoseconds, the formation of random non-native hydrogen bonds between the two "sticky" DNA strands delayed strand separation for a nanosecond or so despite a substantial Coulomb repulsion force.

Finally, in addition to the pronounced differences in the DNA unfolding mechanism, the presence of water, which competes for hydrogen bonding with DNA bases, results in a striking acceleration in duplex unzipping. Thus, at T = 600 K, the solvated duplexes lose their native base pairing contacts in about 100 ps as opposed to 5 nanoseconds in the gas phase, which may be rationalized in terms of (efficient) heat exchange between DNA and surrounding water molecules. However, despite shedding light on the dynamics of base pair contact disruption, the above insights do not provide a picture of local (or global) macromolecular contortions following the external stress. In particular, it is of interest to know if pronounced unzipping of Watson-Crick contacts is necessary for such contortions to occur or if the DNA structure can undergo significant deformations while preserving most of its native base pairs. Below, we address this issue in

detail using both the “duplex-as-a-whole” and “phosphate-backbone-only” representations of the studied macromolecular ensembles.

Results: conformational unfolding and radial distribution functions. Despite the rapid transition from canonical (B-DNA) to somewhat extended macromolecular structures which takes place upon DNA vaporization, the major structural motifs characteristic of hydrated DNA duplexes are largely preserved in the gas phase (Figure 4.8, top). For the DNA ensembles equilibrated at T_0 = 300 K ($t \equiv 0$), regardless of the presence of hydrating environment, multiple sets of spatially coherent intramolecular distances give rise to unique resonant features in the radial distribution function characteristic of the full DNA duplex, $\langle f(r, t)\rangle_n$, and the subset consisting of the backbone phosphorus atoms, $\langle f_P(r, t)\rangle_n$. A comparison of radial distribution features of hydrated and isolated ensembles of 1LAI (Figure 4.8, top) reveals that the spatial resonance associated with structural ordering in the duplexes is slightly weakened upon vaporization, which may be rationalized in terms of formation of locally ordered domains separated by somewhat distorted moieties. However, as evidenced by sharp resonant peaks in $\langle f_P(r, 0)\rangle_n$, these changes are not associated with a pronounced loss of helicity in the backbone. It is also noteworthy that the radial-distribution-based representation of isolated DNA ensembles is robust to variations in the charged state of the macromolecule. In the following we demonstrate that deterioration of the spatial resonance in $\langle f(r, t)\rangle_n$ and $\langle f_P(r, t)\rangle_n$ following a *T*-jump ($t > 0$) provides direct insight into the details of order–disorder transitions throughout the studied macromolecular ensemble.

For a T-jump of $\Delta T = 1200$ K, the complete unfolding of DNA macromolecules in the gas phase includes three distinct stages each of which is characterized by its specific time scale (Figure 4.9, bottom). First, the base pairing and base-stacking contacts are broken throughout the duplex in the order of increasing energy penalty, and corresponding base pairs swing out of their "slots" in the equilibrated structure. Notably, the resulting decrease of structural ordering throughout the DNA duplexes, which manifests itself through gradual deterioration and smoothing of resonant features in $<f(r, t)>_n$ (Figure 4.9, bottom), does not affect the structure of the sugar backbone (Figure 4.10, bottom). Indeed, for the initial steps of the order–disorder transition studied here, which occur on the time scale of several ps, $<f_P(r, t)>_n$ of the phosphorous chains remains virtually unchanged, whereas the (macromolecule-wide) $<f(r, t)>_n$ decreases monotonically and quasi-linearly with time, which is also true for the resonant base-stacking feature at $r \approx 5$ Å.

Second, from $t \approx 10$ ps and on, the sugar backbones of the duplexes lose their characteristic rigidity, which leads to a partial loss of the native (helical) structure. Thus, the second-largest peak in $<f_P(t, r)>_n$, $r \approx 18$ Å, starts to broaden and deteriorate, which is indicative of decreasing long-range ordering in the duplex (Figure 4.10, bottom). At $t \approx$ 100 ps, the process of non-specific "coiling" of the two strands is virtually complete, and the resonant peak at $r \approx 18$ Å evolves into a residual shoulder. Simultaneously, the resonant feature at $r \approx 13$ Å transforms into a higher and broader *nonresonant* peak, and the nearest phosphorous-phosphorus distances $r_{P...P}$ ($r \approx 7$ Å) become more spread due to the increasing randomization of the backbone structure. Finally, the two DNA strands separate at $t \approx 1$ ns, which is indicative of the completion of the order–disorder transition,

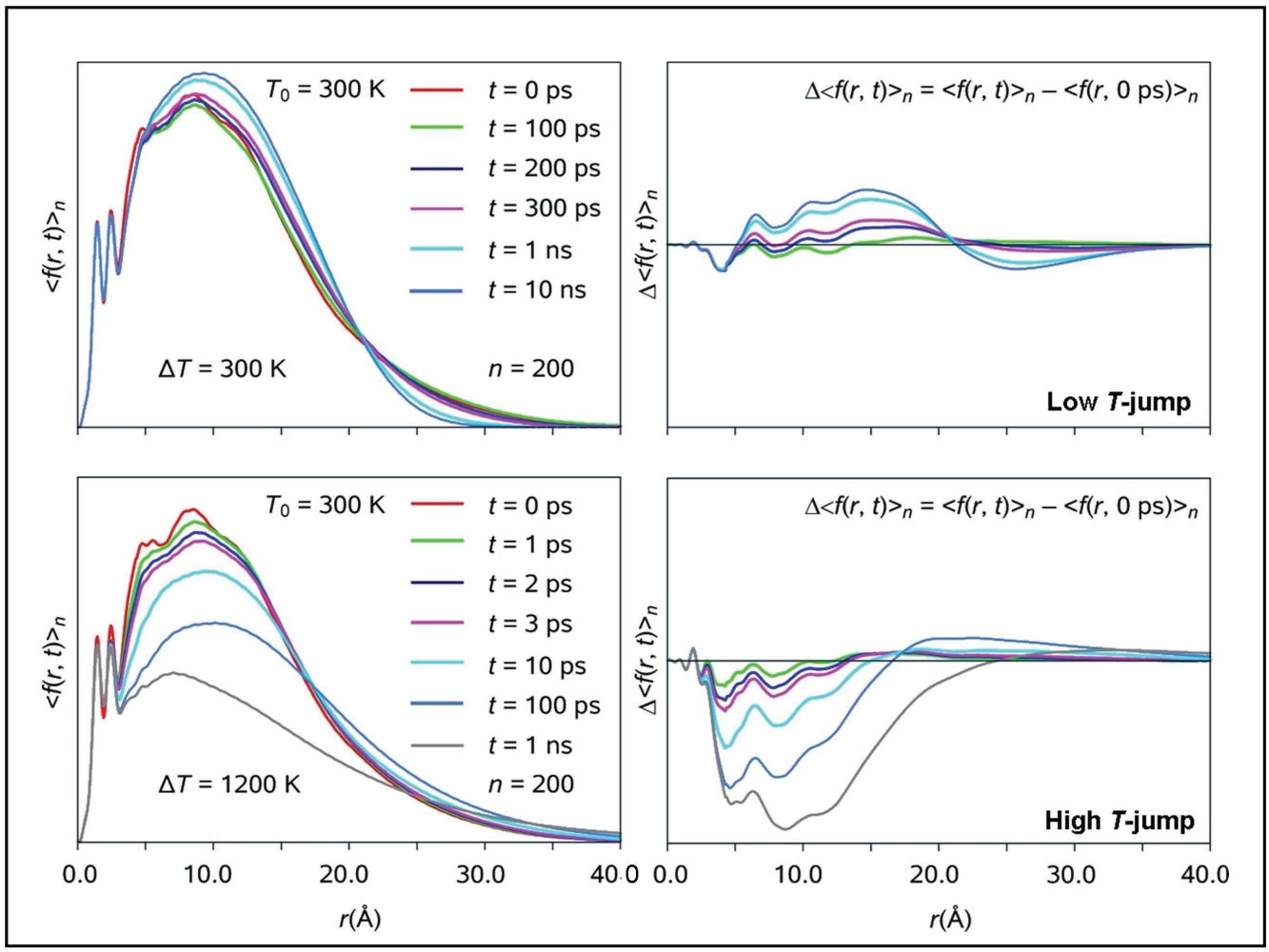

Fig. 4.9. DNA gas-phase unfolding monitored on the radial distribution function. The evolution of the ensemble-averaged pair-wise (radial) distribution function, $\langle f(r)\rangle_n$ as a function of t, during unfolding at different temperature jumps (left) gives structural information about both the global geometry (the growth of the tail), as well as local geometry (the resonance distances of ~ 5 and 7 Å corresponding to base stacking and pairing distances). The latter can be seen via the change of $\langle \Delta f(r)\rangle_n$ showing depletion around the resonance distances (right), although this effect is obscured by global structural diffusion happening at the same time-scale at lower temperature jumps (top right). This implies that ultrafast electron diffraction experiments should be done for ΔT > 300 K to observe the base pair unfolding dynamics.

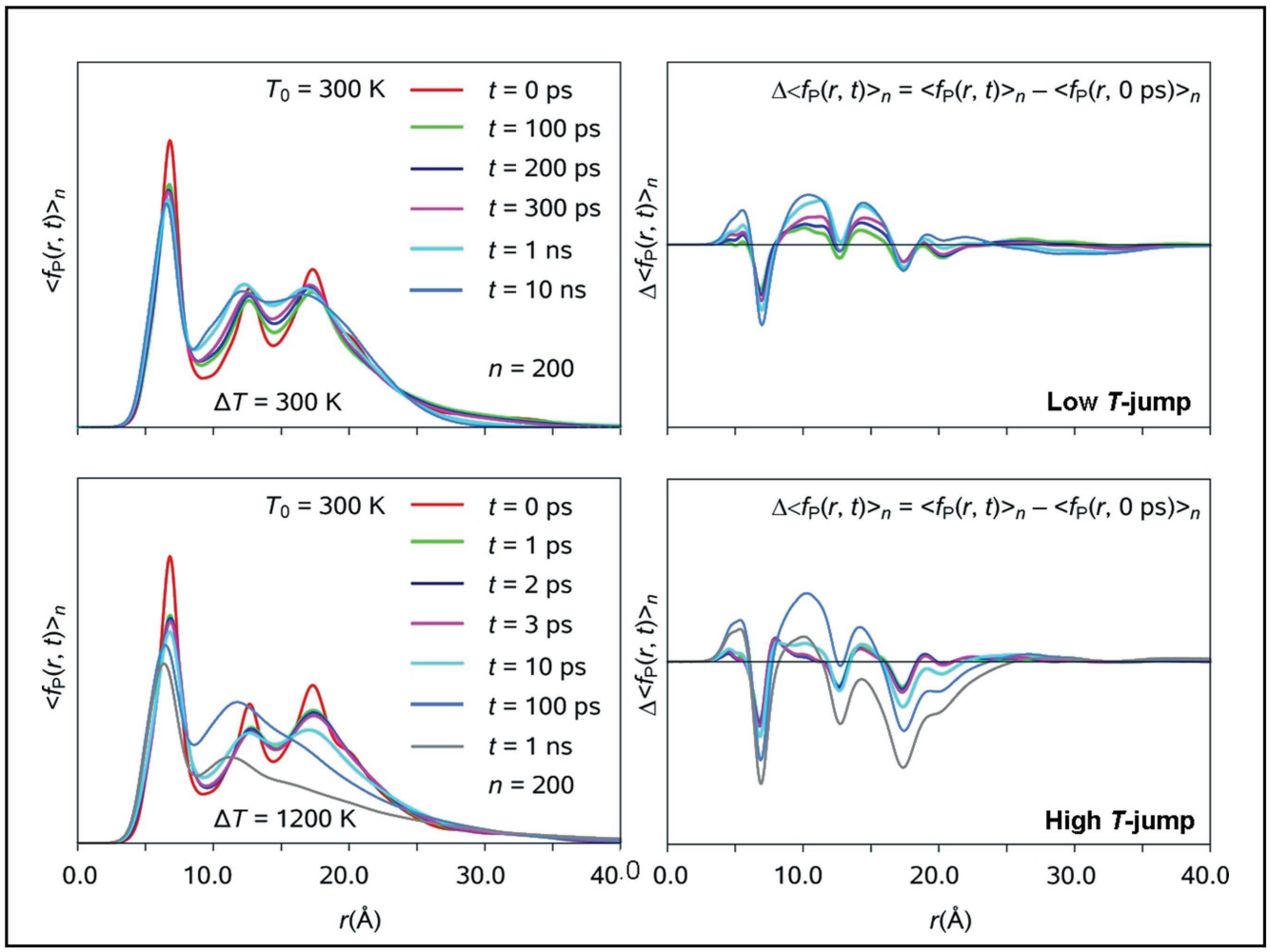

Fig. 4.10. DNA gas-phase unfolding monitored on the radial distribution function of the phosphorus backbone. In contrast to $\langle f(r)\rangle_n$ of the full molecule, the phosphorus backbone is free of the obscuring effects due to the nucleotides themselves, and shows a dominant peak at the inter-strand base-pairing distance of 7 Å (left). In this case, the unraveling of the double helical structure via breaking of the parallel base-pairing distance between the two DNA chains can be seen at all temperature jumps (right). Note that, as expected, the inter-nucleotide base stacking resonance peak evident in the all-atom $\langle f(r)\rangle_n$ is missing in these plots.

although large T-jumps are necessary in order for the strand separation to occur within a practically feasible simulation window. As for the case of base pair unzipping, the temporal responses of radial distribution functions to the T-jumps of ΔT = 600 K and ΔT = 1200 K are qualitatively similar. Despite some minor loss of helical ordering at shorter times, the overall structure of the backbone remains largely intact for tens of ps following the T-jump of ΔT = 600 K, and a (partial) coiling of the strands becomes pronounced on the time scale of hundreds of picoseconds.

In contrast, the order–disorder transition induced across the studied macromolecular ensemble by a T-jump of ΔT = 300 K is vastly different from that discussed above. For the 300 K T-jump in vacuo, both $\langle f(r, t)\rangle_n$ and $\langle f_P(r, t)\rangle_n$ were found to *increase* with time for shorter internuclear distances, indicative of ensemble-wide global bending and compaction (top panels of Figures 4.9 and 4.10). Because the majority of base pairing contacts remained intact on the shorter time scales, the overall molecular rigidity prevented significant backbone unfolding for the first 300 ps. Moreover, even at t = 1 ns the helical motif was well preserved. However, the DNA double helix undergoes significant global conformational contortions, changing its shape from a rod-like duplex to a spheroid despite maintaining the majority of its base pair contacts over time. Consequently, because the native contact rupture lags behind the global conformational changes such as backbone coiling, the (local) decay of base pairing and stacking interactions is no longer isolated to the shortest times and its representative signal is therefore convoluted with that of the (non-specific) global motion. Therefore, following the lowest T-jump studied here, the base pair (un)folding dynamics is not readily evident in the temporal changes of $\langle f(r, t)\rangle_n$.

The above behavior arises because base pairing/stacking disruption and global conformational change are enthalpy- and entropy-driven processes, respectively. At lower temperatures (ΔT ~ 300 K) conformational diffusion happens even if few disruptions occur. Because the (barrier-crossing) native contact disruption process speeds up exponentially with increasing temperature whereas the (diffusive) global conformational motion is weakly dependent on temperature, at a certain threshold temperature the time scales for these two processes cross. Consequently, at sufficiently high temperatures (ΔT ~ 600 K), all base pairs are broken prior to the characteristic diffusion time. Crucially, this observation also suggests that UED *T*-jumps be performed for $\Delta T >>$ 300 K to observe ultrafast local dynamics with clarity. Notably, neither 300 K nor 600 K *T*-jumps were sufficient to trigger a strand separation process in the gas phase within the temporal window characteristic of our MD simulations.

Summary. In the present study we have demonstrated for the first time that order–disorder transitions in gas phase DNA duplexes cannot be accounted for using a conventional "two-state" model. The folding-unfolding landscape of a free DNA macromolecule involves a number of intermediate structures which may be described as (partially) unfolded, or collapsed. The time scales characteristic of native-contact rupture and sugar backbone denaturation are strongly temperature dependent, so much so, in fact, that coiling of the strands may either precede (at low-enough temperatures) or follow the base pair separation. For the total strand separation to occur on the nanosecond time scale, the *T*-jump experienced by the ensemble under study has to be high enough to compensate for the "stickiness" (the intramolecular hydrogen-bonding bias) characteristic of DNA in the absence of water. For an isolated macromolecular ensemble of DNA

duplex 13-mers equilibrated at T_0 = 300 K and experiencing a T-jump of ΔT = 1200 K, the native contact rupture, backbone coiling, and strand separation were found to occur on 10 ps, 100 ps, and 1 ns time scales, respectively (in the case of Q = –12, the strands were found to move apart somewhat faster because of a stronger Coulomb repulsion). For ΔT = 600 K, the transitional behavior was qualitatively similar to that characteristic of ΔT = 1200 K (apart from the unpairing/destacking and backbone coiling time scales of ~60 ps and 400 ps, respectively), but for ΔT = 300 K the behavior was radically different: the DNA backbone would twist and compactify for hundreds of picoseconds prior to any substantial base pair disruption.

Although ensemble-averaged unfolding trajectories of free DNA macromolecules are largely insensitive to both charged-state and temperature changes, the denaturation dynamics in vacuo is vastly different from that characteristic of aqueous solution. The above-mentioned "stickiness" of free (dehydrated) DNA makes order–disorder transitions in the gas phase about two orders of magnitude slower than in solution, which is caused by repetitive bonding recombination of broken base pairs in the absence of water. The (extended) double-helical structures are robust in vacuo because they are stable on the sub-microsecond time scale. From the (time-dependent) base pairing picture analysis following the ensemble convergent MD simulations, DNA duplexes in vacuo were found to experience simultaneous end and interior (bubble) base pair disruption across the entire ensemble, regardless of both the charged state and temperature used. Thus, unlike MD simulations in solution which indicate that the A-T rich sequence in the center of the duplex unfolds at about the same time as the neighboring (inner) base pairs, the gas phase simulations reveal that the inner A-T contacts unwind almost simultaneously with the

ends. In the gas phase, the destacking energy required to form an internal bubble is compensated for by the extra stabilization of a C-G contact as compared to an A-T contact, which renders base pair unzipping in the middle of the sequence feasible. Contrastingly, in aqueous solution, the extra stabilization of the C-G base pairs is no longer able to compensate for destacking energies required to form an internal bulge because neighboring water molecules can compensate for any broken base pair hydrogen bonds.

The results reported above may suggest some biological implications. Structurally, although the double-helical motif appears to be stable on the sub-microsecond time scale independent of the charged state assumed, the high degree of conformational freedom and the loss of some base pairing and stacking in the gas phase, especially near the ends, indicate that water is necessary for maintaining robustness of the double-helical structure. Moreover, the striking contrast of the dynamical behavior observed in the presence and in the absence of water suggests that the free energy constituents contributed by hydrogen bonding, base stacking, and entropy that are characteristic of DNA in aqueous solution have all been fine-tuned to facilitate high-speed/low-error biological functionality. For example, the enhanced intra-strand hydrogen bonding propensity in vacuo results in higher levels of non-native contacts and dramatic slow-down of the strand separation. This also led to the formation of A-T rich internal bubbles, which were not detected in solution. The above effect may have consequences pertinent to transcription in vivo: because the initiation of strand separation (bubble formation) along DNA duplexes is less favored than its propagation, the chance that the strand separation can only occur in response to a site-specific interaction (e.g.,

with a transcription factor) rather than following random thermal fluctuations is increased. Once the strand separation process has been initiated, the relatively low energetic penalty for unzipping of the DNA duplex may facilitate its speedy transcription. The mechanisms and time scales reported here for a variety of large-scale macromolecular ensembles bring us one step further in the quest for a complete atomic-scale picture of the effect of solvent on DNA stability, dynamics, and function (248, 249).

4.2.3 Helix-Coil Transitions in Proteins

Introduction. Employing the same tools as in Section 4.2.2 for the study of nucleic acid (un)folding, it is also possible to directly observe helix-to-coil transformations in isolated proteins. Similarly to double-stranded DNA, there exists a set of characteristic periodically repeating internuclear distances which, in the case of a helical polypeptide, is associated with the helix pitch (Figure 4.11, left). The unique periodic spatial structure of the α-helix gives rise to a "resonance" peak in the ensemble-averaged radial distribution function which is more localized than that characteristic of the DNA double helix. In what follows, we investigate the conformational dynamics of a helical protein thymosin-β_9 ($T\beta_9$) (113) using the ensemble-averaged radial distribution functions and ensemble-convergent MD simulations. In doing so, we gain time-dependent insights into both local and globular structural changes associated with the helix–coil transformations in large-scale macromolecular ensembles of $T\beta_9$. Importantly, the decay of α-helical motifs and the (globular) conformational annealing in $T\beta_9$ occur consecutively or competitively, depending on the magnitude of the applied *T*-jump.

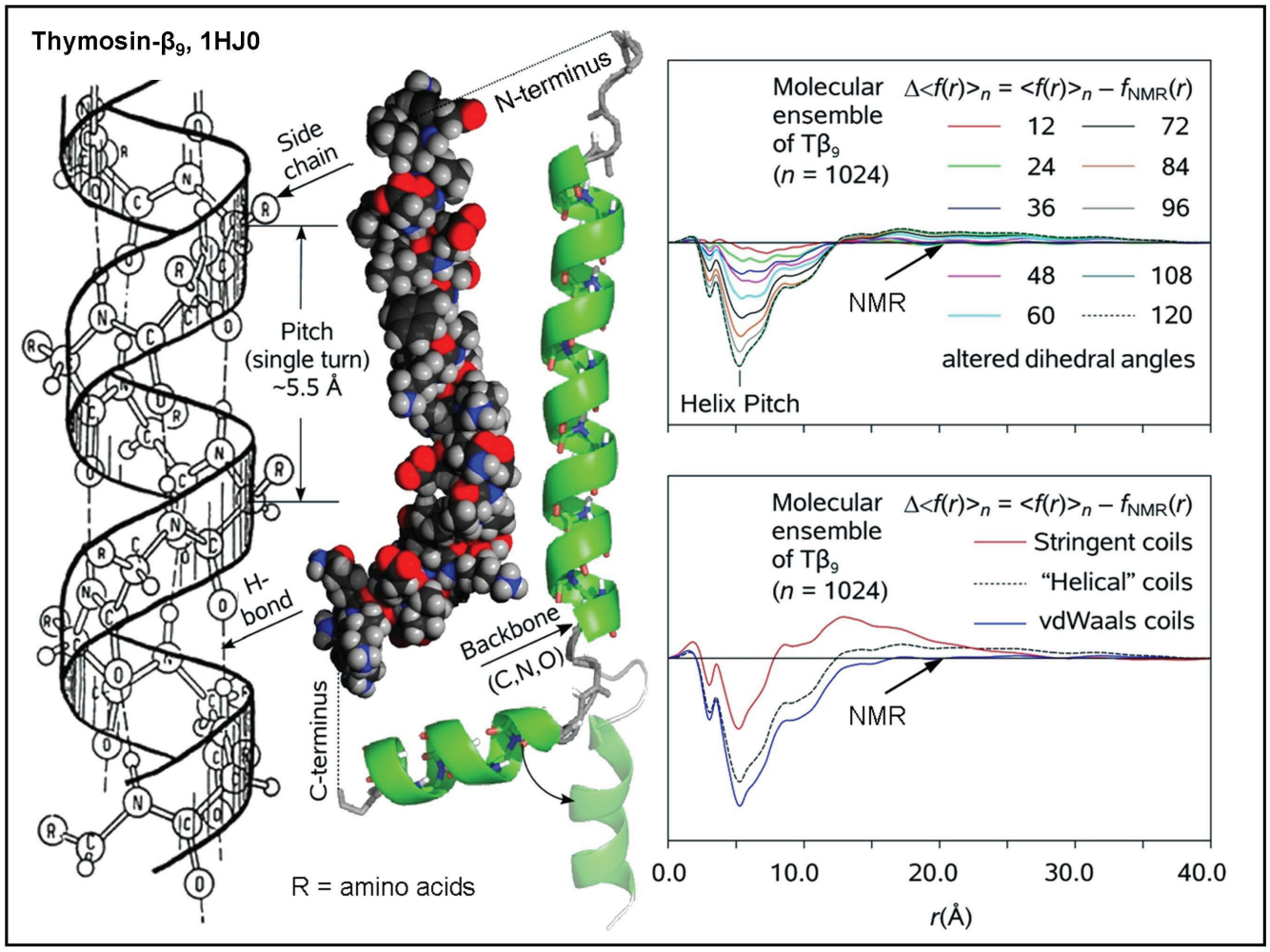

Fig. 4.11. Unfolding dynamics for a small α-helical protein. In canonical right-handed α-helices, the C=O group of an amino acid located at the position *i* in the backbone chain of a protein forms a hydrogen bond with the N–H group of another amino acid that occupies the position *i* + 4. Equivalent atomic positions recur every 5.4 Å along the chain which defines the *pitch* of the helix (left). Also shown in a variety of representations is the molecular structure of the protein thymosin β_9 ($T\beta_9$; PDB ID 1HJ0) as obtained from 2D NMR at T = 298 K; 667 atoms of $T\beta_9$ give rise to about 222,000 pair-wise interatomic distances; the distance-weighted frequency histogram of these distances defines the radial distribution function $f(r)$. By averaging $f(r)$ over an ensemble of size n, the radial distribution function difference (relative to the initial ensemble) was calculated for $T\beta_9$ for a series of ensembles whose helix content was systematically lowered ($\langle \Delta f(r) \rangle_n$, n = 1024; top right). The algorithm used to calculate the ensembles fixed the desired helix content, but treated other degrees of freedom as a random walk. The peak in $\langle \Delta f(r) \rangle_n$ at r = 5.5 Å corresponds to the structural resonance of the α-helix, and is robust to changes in the algorithm parameters such as the steric volume (bottom right). This peak grows quasi-linearly with the helix content, and therefore can be used as a signature for helix formation in analyzing molecular simulations, as well as potentially in diffraction experiments.

Preliminaries: the helicity fingerprint and helix unfolding algorithm. To quantitatively assess the extent of the resonance peak as a function of helix content, we found it useful to generate conformations of $T\beta_9$ with controlled fractions of helical and coil moieties. To (partially) unfold the α-helical structure of $T\beta_9$, the position of each atom in the protein backbone, except the first three, was specified in terms of the previous three backbone atoms using internal coordinates (r, θ, τ). Within the framework of our model, bond distances r and valence angles θ remained fixed throughout the backbone, whereas torsional angles τ were allowed to vary. Finally, the dihedral angles with respect to each pair of adjacent single bonds in the backbone (Ψ, ϕ) were chosen at random from a Ramachandran pair probability diagram representative of amino acids with large functional groups (2).

At the initial stage of generating a coil moiety, the structure thus obtained was typically characterized by a large number of close intramolecular contacts. To resolve the (numerous) steric collisions, every non-bonded interatomic distance which was shorter than a pre-defined cutoff distance was tagged. Then, the (Ψ, ϕ) pair of torsional angles associated with the first peptide unit in the longest polypeptide stretch that was at least 80% tagged was re-selected at random from the Ramachandran diagram. By adjusting the dihedral angles of the first peptide element in the tagged stretch, the majority of steric collisions were resolved; this proved effective because tagged polypeptide stretches would typically correspond to a randomly structured ("coiled") moiety overlapping with the remainder of the macromolecule. The above procedure was repeated until no peptide

elements were tagged. Finally, with the backbone shape fixed, side-chain–side-chain and side-chain–backbone collisions were resolved by rotating side chains in a similar fashion.

With the above algorithm employed to generate large ensembles of (partially) unfolded $T\beta_9$ macromolecules, the electron diffraction simulations were carried out as follows (250). First, n = 1024 (partially) randomized pseudoconformers of $T\beta_9$ were generated using the experimentally obtained (NMR) structure of the macromolecule as a starting point. The following cutoff distances were tested: (*i*) $r_{ij} > 2.6$ Å ("stringent" constraint); (*ii*) $r_{ij} > r_{ij}$(native) as obtained from the NMR structure of $T\beta_9$ ("helical" constraint), and (*iii*) $r_{ij} > 3.1$ Å (van der Waals-type constraint). Second, ensemble-averaged radial distribution functions, $\langle f(r, T)\rangle_n$, were calculated for the above ensembles at T = 300 K. Third, and finally, $\langle f_{NMR}(r, T)\rangle_n = f_{NMR}(r, T)$ of the native (NMR) structure of $T\beta_9$ was calculated in a similar fashion, and further subtracted from $\langle f(r, T)\rangle_n$ of the (partially) randomized macromolecular ensembles to obtain $\Delta\langle f(r, T)\rangle_n$ characteristic of the structural change. Using the above algorithm to construct ensembles with variable amounts of helical content, $\Delta\langle f(r, T)\rangle_n$ was plotted as a function of the ensemble-averaged helicity fraction (Figure 4.11, top). Notably, the amplitude of the (strongest) resonant peak associated with the helix pitch ($r_{ij} \sim 5.5$ Å) was found to be proportional to the residual helical content across the studied macromolecular ensemble. Depicted in Figure 4.11, bottom are the diffraction difference patterns as obtained by subtracting the radial distribution function of the native (α-helical) structure of $T\beta_9$ from those of the large (n = 1024) random-coil ensembles generated with the aid of the above algorithm using various steric cutoffs. As

seen in the Figure, the native (α-helical) structure ensembles can be reliably discriminated from the random-coil ensembles of $T\beta_9$ regardless of the steric-cutoff distance employed.

The diffraction peak centered at $r_{ij} \sim 5.5$ Å, which arises from the structural resonance associated with the periodically repeating helix pitch distance, may be used to monitor the residual helicity fraction across the ensemble at each particular point in time. However, the fact that the ensemble-averaged helicity fraction is reduced to about 20% in the gas phase, which results in a substantial reduction of the relevant diffraction difference signal, renders the diffraction data analysis significantly more difficult. A way to excise the majority of incoherently distributed scattering terms that will inevitably obscure helix-to-coil transitions in the gas phase is to perform UED simulations on the backbone atoms of $T\beta_9$. Indeed, as in the case of the phosphorus backbone of DNA in the previous Section, the resonant features become much sharper and more distinct when $\langle f_B(r, t)\rangle_n$ is employed (see below).

Following the proof-of-principle study summarized above (250), ensemble convergent MD simulations were used to picture the actual temporal evolutions of macromolecular structures of $T\beta_9$ induced by a number of *T*-jumps. During the course of MD simulations, ensemble convergence was monitored with the aid of ensemble-averaged radial distribution functions, helicity fractions, and unfolding trajectories. The focus, as in the case of the DNA duplex, was on the temperature dependence of different types of structural dynamics that occur on a spectrum of length- and time-scales. Selected implications of these results for experimental studies of order–disorder transitions in α-helical macromolecules are succinctly highlighted below.

Equilibration in vacuo at room temperature. The radial distribution functions of isolated macromolecular ensembles of $T\beta_9$ (the "time zero" curves in Figures 4.12 and 4.13 for $\langle f(r, t)\rangle_n$ and $\langle f_B(r, t)\rangle_n$, respectively) reveal that the spatial resonance associated with α-helical ordering in the protein is somewhat weakened upon vaporization. This may be rationalized in terms of formation of highly ordered (α-helical) local domains separated by distorted, loosely structured moieties such as loops. The ensemble-convergent MD simulations indicate that these structures are *transient* in nature. Despite non-native intramolecular hydrogen bonding contacts that arise at $300 \leq T \leq 600$ K in the absence of water molecules, which are known to efficiently compete with the intramolecular hydrogen bonding in aqueous solutions, the conformational dynamics associated with non-helical moieties remains significant not only due to their inherent flexibility, but also because of the continuous rupture and recombination of the non-native hydrogen bonds. Both the excessive intramolecular hydrogen bonding and (entropy-driven) contraction of highly extended macromolecular conformations lead to the dominance of more compact ("globular") structures across the ensemble, which are characterized by dramatically reduced radii of gyration. However, despite the transition from "canonical" (NMR-like) to partially randomized, somewhat more globular macromolecular structures which takes place upon vaporization, a substantial fraction of the original helix content is preserved by $T\beta_9$ in the gas phase. Thus, in the absence of water, the α-helix pitch resonance peak dominating both $\langle f(r, t)\rangle_n$ and $\langle f_B(r, t)\rangle_n$ indicates the presence of high helical content at room temperature ($T_0 = 300$ K).

Temperature-induced structural dynamics. As evidenced from the results of Figure 4.12 and, more clearly, Figure 4.13, the temporal decay of spatial resonance in $\langle f(r, t)\rangle_n$ and $\langle f_B(r, t)\rangle_n$ that is observed following a T-jump ($\Delta T > 0$; $t > 0$) provides direct insight into the details of order–disorder transitions throughout the studied macromolecular ensemble. In what follows we quantitatively describe the temporal variation of the sharp helix resonance feature of $\langle \Delta f_B(r, t)\rangle_n$ during the course of the helix-to-coil transition induced in macromolecular ensembles of $T\beta_9$ by a range of T-jumps.

For $\Delta T = 900$ K, the (local) helix "unzipping" and (global) conformational coiling of $T\beta_9$ in the gas phase include multiple stages, each of which is characterized by its own specific timing. Thus, the native (α-helical) hydrogen bonding contacts are first broken "dynamically" on an ultrafast time scale; i.e., the native-contact rupture starts to prevail over recombination throughout the entire ensemble of $T\beta_9$ within a few picoseconds. As a result, the inherent stiffness of α-helices that constitutes the basis for their singular mechanical properties (251) weakens, and the overall structure of the helices becomes more *diffuse* (hence the gradual broadening, and displacement towards higher values of r_{ij}, of the characteristic helicity fingerprints in the ensemble-averaged radial distribution function). Notably, the overall shape of $\langle f(r, t)\rangle_n$ and $\langle f_B(r, t)\rangle_n$ patterns remains intact for at least 40 ps (bottom panels of Figures 4.12 and 4.13). However, by $t = 400$ ps, the deterioration of α-helical motifs in $T\beta_9$ is complete, which is evidenced by the disappearance of any resonant features in both $\langle f(r, t)\rangle_n$ and $\langle f_B(r, t)\rangle_n$. On longer time scales, the molecule undergoes global conformational diffusion such that by $t = 4$ ns the formation of a globular coil is virtually accomplished, and further (non-resonant) transient structural changes average out across the studied macromolecular ensemble.

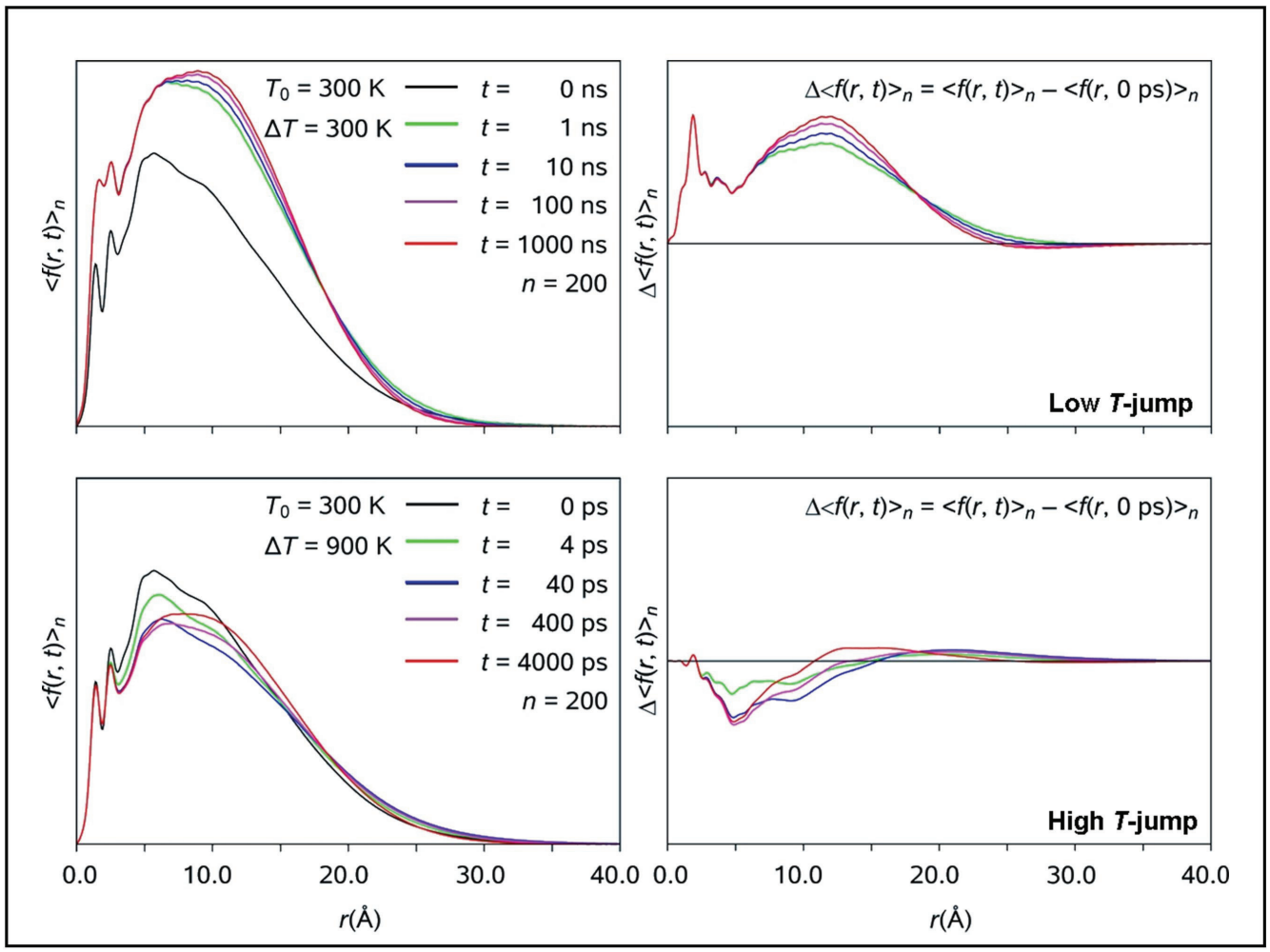

Fig. 4.12. Thymosine unfolding dynamics: radial distribution functions in vacuum. Shown are temporal evolutions of radial distribution functions (left) and corresponding diffraction differences (right) as obtained using a large (n = 200) macromolecular ensemble equilibrated in vacuo at room temperature and further subjected to 300 K (top), and 900 K (bottom) temperature jumps during the course of ensemble-convergent MD simulations. The results of the 600 K temperature jumps are not shown due to their qualitative similarity to the 900 K jumps. Note the dramatic difference between the ensemble-averaged unfolding behaviors at lower and higher temperature jumps, which arise due to the overlap of the global dynamics with the local dynamics of interest (namely α-helix unfolding monitored at the r = 5.5 Å resonance peak) at lower temperatures. As in the case of the DNA duplex, this implies that ultrafast electron diffraction experiments should be done for ΔT > 300 K to observe the change in secondary structure.

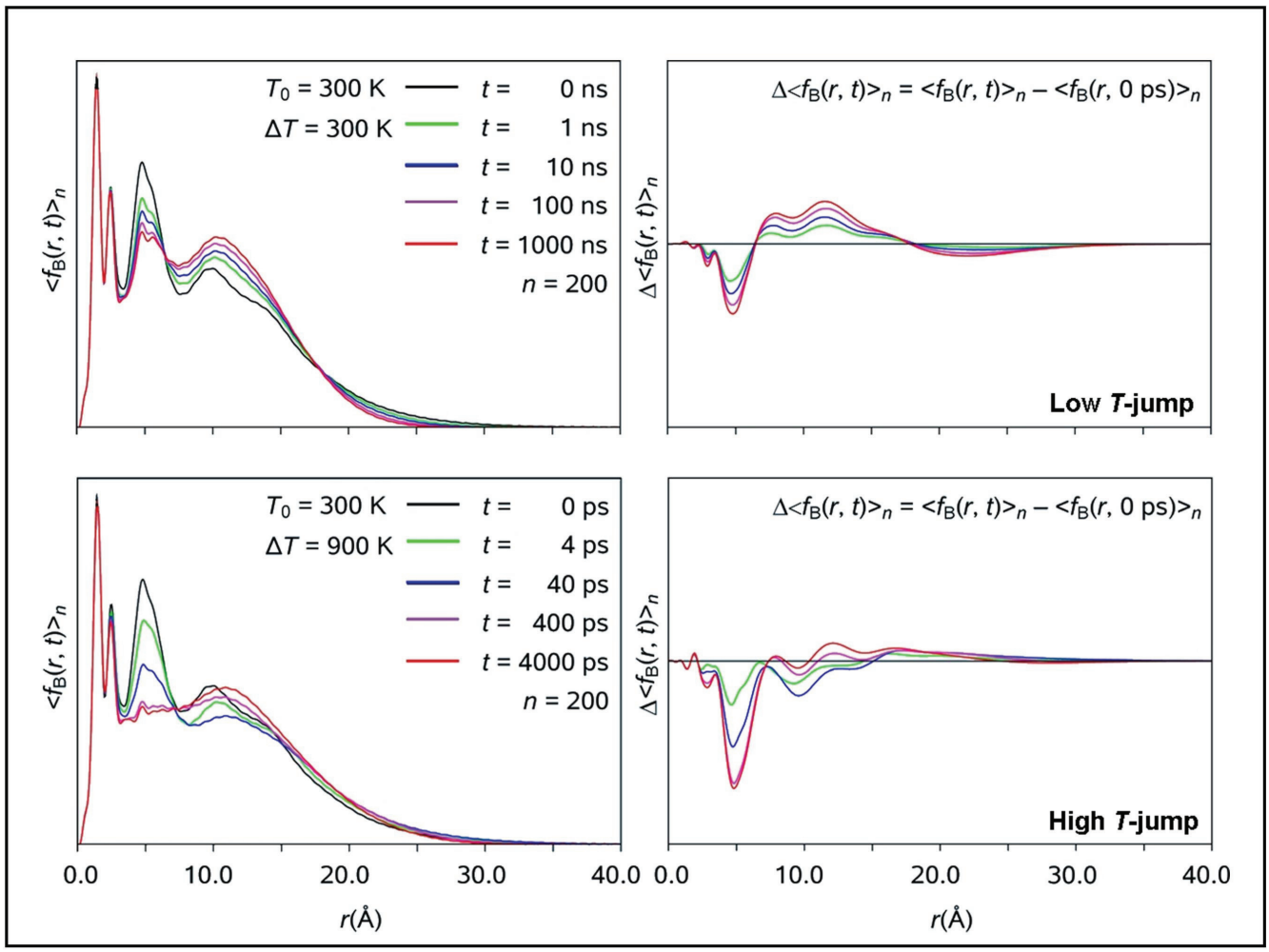

Fig. 4.13. Thymosine unfolding dynamics: backbone radial distribution functions in vacuum. Shown are temporal evolutions of backbone-specific radial distribution functions (left) and corresponding diffraction differences (right) as obtained using a large (n = 200) macromolecular ensemble equilibrated in vacuo at room temperature and further subjected to 300 K (top), and 900 K (bottom) temperature jumps during the course of ensemble-convergent MD simulations. The results of the 600 K temperature jumps are not shown due to their qualitative similarity to the 900 K jumps. Note that the helix resonance is remarkably sharp and the obscuring effects of the global dynamics at lower temperature (see Figure 4.12) are not present when restricting the analysis to the backbone atoms.

Thus, high *T*-jumps induce helix unzipping on the picosecond time scale, whereas the global structure of the macromolecule changes two orders of magnitude more slowly; therefore, the overall shape of $T\beta_9$ can be considered "frozen" during the course of local unfolding.

It is, perhaps, instructive to emphasize here that when the tight atomic packing characteristic of α-helices, which manifests itself through a resonant accumulation of relatively short (5–10 Å) interatomic distances across the macromolecular ensemble, becomes fully (or even partially) scrambled, the diffuse packing of the (collapsed) globular structure gives rise to a broad, smooth, hump-like incoherent feature centered at about 12 Å in $\langle f_B(r, t)\rangle_n$ (Figure 4.13, bottom); a similar effect is observed in $\langle f(r, t)\rangle_n$ as well, but it is less pronounced, especially at higher temperatures which virtually preclude the formation of stable intramolecular hydrogen bonds in the gas phase; cf. Figure 4.12, bottom. The overall pattern of structural changes as obtained for ΔT = 600 K (data not shown), despite being somewhat less dramatic, closely resembles that characteristic of ΔT = 900 K, with the only exception of higher residual-helicity fractions being preserved throughout the studied ensemble at longer times.

As opposed to (vibrationally hot) ensembles of $T\beta_9$ created by higher (ΔT = 600 K and ΔT = 900 K) *T*-jumps (see above), the ensemble-convergent behavior associated with the helix-to-coil transition triggered by ΔT = 300 K appears to be strikingly different insofar as the massive formation of (non-native) intramolecular hydrogen bonding contacts is no longer precluded by rapid vibrations and contortions of the macromolecule. The increased number of stable non-native hydrogen bonds manifests itself through a

dramatic increase in the heights of the two innermost peaks of $\langle f(r, t)\rangle_n$ (Figure 4.12, top), which are associated with nearest-neighbor (bonded) and second-nearest-neighbor intramolecular distances [we note that the corresponding peaks in $\langle f_B(r, t)\rangle_n$ remain virtually unchanged because, unlike the C, N, O atoms forming the side chains of $T\beta_9$, the atoms that constitute its backbone chain do not appear to participate in formation of the new hydrogen bonds (Figure 4.13, top)].

From the overall shape of $\langle f(r, t)\rangle_n$ (Figure 4.12, top) one may mistakenly conclude that the new hydrogen-bonded structures formed across the studied ensemble must be completely random. However, an examination of $\langle f_B(r, t)\rangle_n$ indicates that, despite somewhat increased contraction of the macromolecules triggered by formation of the new, non-native hydrogen bonds, the helical fingerprint in $\langle f_B(r, t)\rangle_n$ remains intact (Figure 4.13, top). The "stickiness" (or self-bonding bias) characteristic of $T\beta_9$ in the absence of water induces formation of a substantial number of non-native, side-chain-dominated, hydrogen-bonding contacts even in the presence of fairly long α-helices. The latter distort, but nevertheless retain their characteristic structural features throughout the duration of the entire simulation window. From comparing $\langle f(r, t)\rangle_n$ with $\langle f_B(r, t)\rangle_n$, it is clear that global reorganization across the ensemble is fully accomplished within a few nanoseconds (Figure 4.12, top), whereas the pronounced helicity peak in $\langle f_B(r, t)\rangle_n$ (Figure 4.13, top) indicates that the helices are still unzipping on this time scale. Therefore, for ΔT= 300 K, the local unfolding occurs at much longer times than the global unfolding and (non-native) hydrogen bond formation.

Similarly to the conformational and structural changes reported earlier for temperature-induced helix-to-coil transitions in DNA, the above-mentioned ensemble-convergent behavior arises because the native (α-helical) hydrogen-bonding contact disruption and the global conformational change are enthalpy- and entropy-driven processes, respectively. At lower temperatures ($\Delta T \leq 300$ K) conformational diffusion happens even if few disruptions occur. Unlike DNA macromolecules, which are highly negatively charged despite protonation (224), isolated macromolecules of $T\beta_9$ are neutral, which causes them to be even stickier in the absence of water. However, a temperature increase across the ensemble can unravel the intramolecular motions that facilitate formation of a globular coil stabilized by non-native hydrogen bonds. Because the (barrier-crossing-type) hydrogen-bond disruptions speed up exponentially with increasing temperature whereas the (diffusive) conformational interconversions are only weakly dependent on temperature, at a certain threshold temperature (specifically, on the interval $300 \leq \Delta T \leq 600$ K), the time scales for these two kinds of processes cross and the α-helices of $T\beta_9$ unzip *prior* to the characteristic diffusion time required for large-scale conformational changes to occur.

The above results suggest that it is important to perform UED *T*-jump experiments at $\Delta T >> 300$ K to prevent global dynamics from convoluting local helix unzipping in the patterns of $\langle f(r, t)\rangle_n$. It is also noteworthy that the admixture of residual helicity characteristic of $T\beta_9$ unfolding in the gas phase cannot be reliably detected using $\langle f(r, t)\rangle_n$ alone because the spatial resonance associated with the α-helical moieties is simply not strong enough to compete with incoherently distributed scattering terms

associated with side chains and randomly structured loops. Because backbone-specific diffraction differences have proven to be sensitive to the subtle local (and global) structural changes that constitute elementary steps of order–disorder transitions in biological macromolecules, $<f_B(r, t)>_n$ and its varieties constitute a novel coarse graining approach which is significant for conformational analysis of such transitions.

Rate constants of the helix-to-coil transition in Tβ$_9$. By measuring the height of the $<f_B(r, t)>_n$ helix resonance peak centered at ~ 5.5 Å, the results of biexponential fitting of as many as a dozen time-dependent profiles of $<\Delta f_B(r, t)>_n$ are presented in Figure 4.14, left, providing the time constants involved in the temperature-induced helix-to-coil transitions reported here. Importantly, α-helix unzipping appears to include two distinct phases, each of which is characterized by its own time scale. Cleavage of the native hydrogen-bonding contacts, which becomes an ultrafast process above a certain temperature threshold, is characterized by $\tau_{300} = 0.7 \pm 0.2$ ns, $\tau_{600} = 6.8 \pm 1.6$ ps and $\tau_{900} = 3.9 \pm 0.8$ ps, as assessed for the T-jumps of $\Delta T = 300$, 600, and 900 K, respectively. Surprisingly, the loss of residual helical structure following hydrogen bond rupture is a slower process characterized by $\tau_{300} = 90 \pm 40$ ns, $\tau_{600} = 440 \pm 80$ ps and $\tau_{900} = 50 \pm 16$ ps. Presented in Figure 4.15, bottom right, are the (log-scale) unfolding times of these two processes plotted as a function of the inverse temperature; the (Arrhenius) slopes of the lines are known to provide the associated energy barriers. Importantly, the slower (τ_2) process of global helicity loss displays Arrhenius behavior over the entire range of temperature studied with a characteristic barrier of 18 ± 1 kcal/mol. In contrast, the hydrogen-bond disruption process (τ_1), which is up to two orders of magnitude faster,

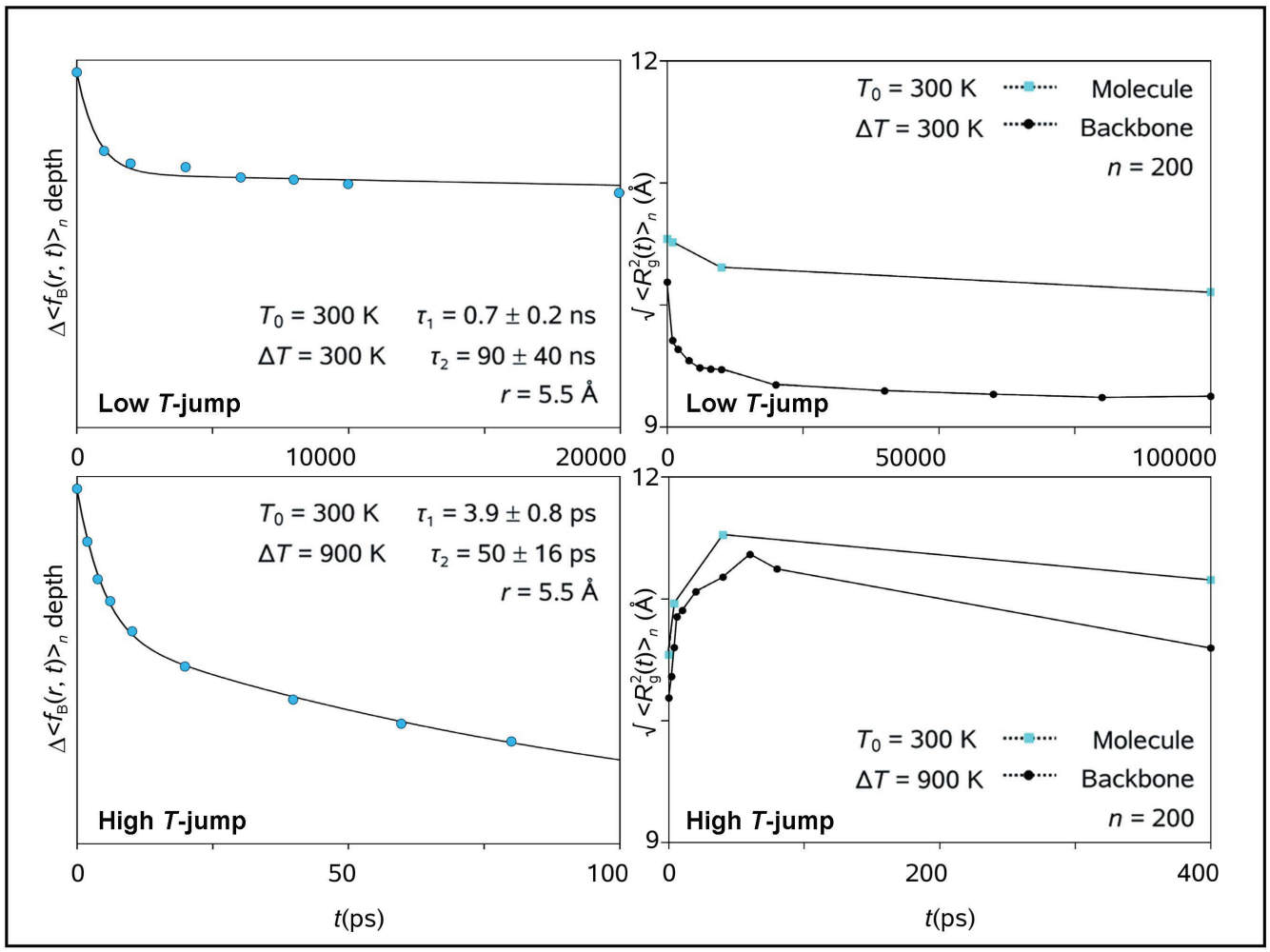

Fig. 4.14. Temporal dynamics of local and global structure. Temporal evolutions of the backbone radial distribution function at the α-helix resonance peak distance ($\Delta\langle f_B(r, t)\rangle_n$, $r \approx 5.5$ Å; left) and the gyration radius of the protein and its backbone (right) are obtained using large (n = 200) macromolecular ensemble equilibrated in vacuum at room temperature and further subjected to 300 K (top), 600 K (similar to the 900 K behavior; not shown), and 900 K (bottom) temperature jumps during the course of ensemble-convergent MD simulations. There is a pronounced difference between the ensemble-averaged unfolding behaviors characteristic of lower (ΔT = 300 K) and higher (ΔT = 600, 900 K) temperature jumps; the radius of gyration, which characterizes the overall size of the macromolecule, reflects heat-induced swelling prior to globular-formation-driven contraction for sufficiently large ΔT (right). This lag between contact rupture and conformational reorganization timescales is also present on the local scale, as manifest in the large differences between the bond breaking (τ_1) and the helicity loss (τ_2) time (left).

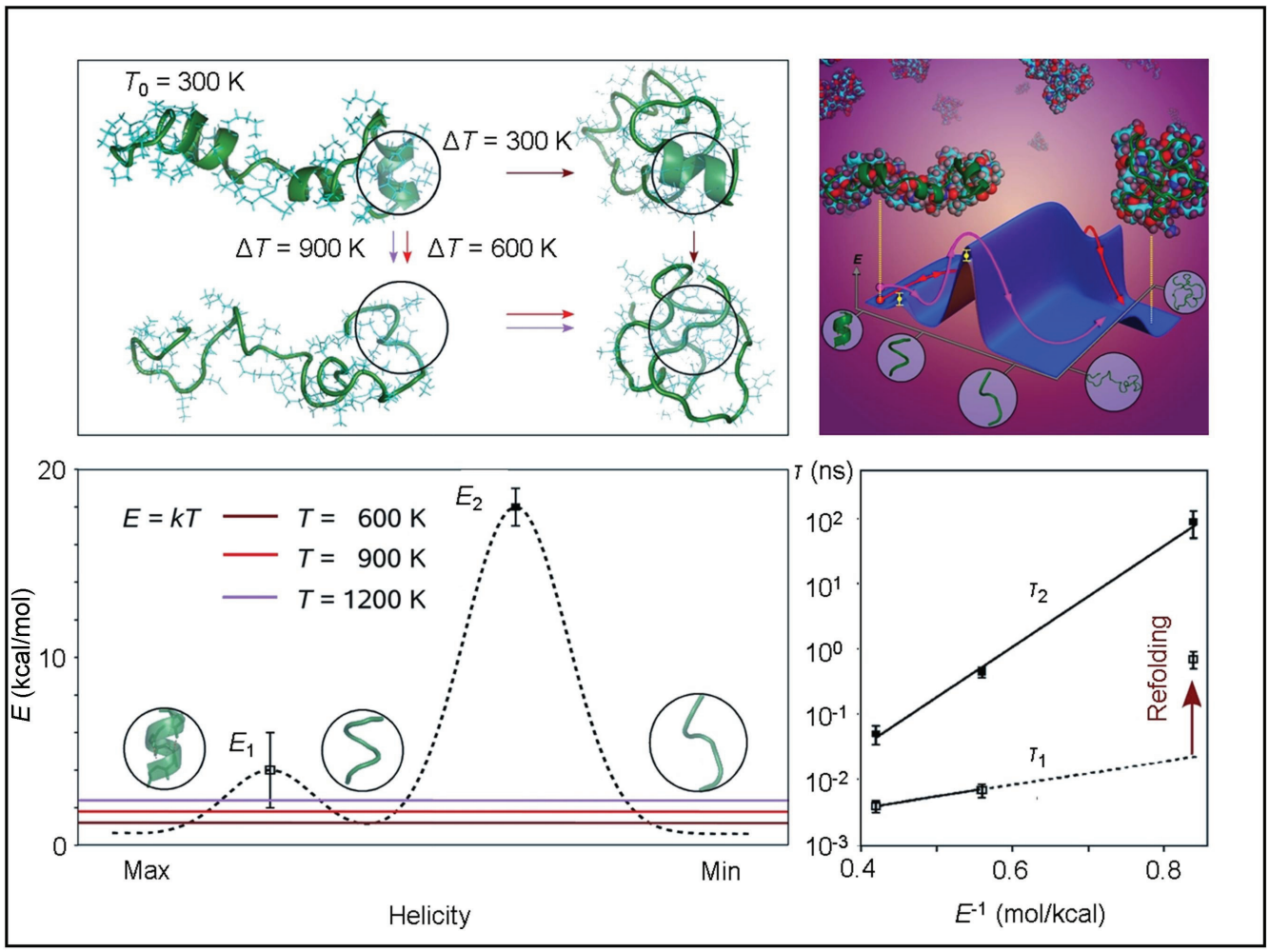

Fig. 4.15. Local and global structural dynamics during unfolding. Following the temperature jump, the initial ensemble undergoes both local denaturation of helical motifs (blue circular highlight) as well as global coiling (top left). For ΔT = 300 K, the unfolding pathway (dark red arrows) involves global coiling and enthalpy-driven contraction followed by loss of local helicity, whereas for ΔT = 600 and 900 K (light red and magenta arrows) the order of these two processes is reversed. The loss of local helical structure is further separated into two distinct time scales corresponding to fast bond disruption (τ_1) and slower loss of structural helicity (τ_2) which are separated by up to 2 orders of magnitude in their time scales (bottom right). Note that at ΔT = 300 K, bond disruption is reversible, which significantly slows the dynamics (red arrow in inset). The unraveling of the helix therefore involves surmounting two barriers, E_1 and E_2 which are separated by the kinetic intermediate structure (bottom left). The unfolding dynamics can be summarized on a schematic energy landscape defined by local and global order parameters (top right).

involves crossing a 4 ± 2 kcal/mol barrier as evidenced from the ΔT = 600 and 900 K time scales (Figure 4.15, bottom). We note that τ_1 at ΔT = 300 K is significantly longer than that predicted for barrier crossing because, unlike the irreversible bond breaking process at higher temperatures, the hydrogen-bond disruption at this temperature is a dynamical interplay of bond breaking and reformation on the nanosecond time scale (Figure 4.15, bottom right).

The magnitude of the energy barrier to structural helicity loss, which is much higher than that characteristic of the native hydrogen bond disruption, indicates that the helical structural motif is much more persistent than its bonding energetics would suggest. As a result, α-helix unfolding appears to be (at least) a three-state process with the helical-but-unbound state being a well-defined population of structures which, depending on the magnitude of the *T*-jump applied, may give rise to a long-lived kinetic intermediate dominating the unfolding dynamics. In fact, at ΔT = 300 K, this local structural motif is much longer-lived than the overall global structural relaxation.

Global (dis)order: coherent vs. incoherent dynamics evidenced in the radii of gyration. Another important characteristic of the macromolecular ensemble under study that can be readily obtained from the simulated UED data is the temporal evolution of the RMS radius of gyration, $<R_g^2(t)>_n^{1/2}$, which measures the *compactness* of structures across the ensemble and is calculated from the ensemble-averaged radial distribution function: $< R_g^2(t) >_n = 1/2 \int_0^\infty < f(r,t) >_n \, r^3 dr$ (see Section 3.3 for methodological details). Shown in Figure 4.14, right, are the temporal-evolution profiles of $<R_g^2(t)>_n^{1/2}$ as obtained from MD/UED data averaged over *n* = 200 independent unfolding trajectories of $T\beta_9$. A

striking feature characteristic of the higher *T*-jumps (ΔT = 600, 900 K) is the initial *expansion* of macromolecular structures across the ensemble taking place on ultrafast time scales, which is followed by a (much slower) contraction processes ensuing at longer time scales (Figure 4.14, bottom right; ΔT = 600 K: data not shown). For these higher *T*-jumps, the formerly α-helical moieties of $T\beta_9$ expand *coherently* in the gas phase because of the ultrafast (coherent) cleavage of the α-helical hydrogen bonds, which is followed by the diffusion-driven conformational smearing of the α-helical structural motif and the subsequent formation of compact globular structures. Notably, the ensemble-averaged behavior characteristic of ΔT = 300 K appears to be radically different (Figure 4.14, top right): the (faster) contraction process that is accomplished within a few nanoseconds leads to the (slower) long-lived contraction dynamics at longer times. The former process is due to (non-native) long-range hydrogen bond formation, whereas the latter one is associated with the (diffusion-driven) conformational dynamics leading to a more compact globular structure. Last but not least, we note that in contrast to MD/UED simulation results reported here, the empirical background correction implicit in the data refinement procedure, the finite active *s*-range of the UED diffractometer, and the spatial resolution of the CCD detector are factors that affect the experimental determination of $<R_g^2(t)>_n^{1/2}$ in practice.

As a side note, the above non-equilibrium behavior is strikingly reminiscent of that modeled theoretically and observed experimentally in this laboratory for the two-component (substrate-adsorbate) assemblies subjected to a laser-pulse irradiation (183, 184, 252, 253). The transient anisotropic change in c_0 of the adsorbates (fatty-acid and

phospholipid layers) described in such studies is vastly different from that observed in the steady state. At equilibrium, the observed changes are in a_0 and b_0 (not in c_0), and the diffraction intensity monotonically decreases, reflecting the thermal, incoherent motions (Debye–Waller effect) and phase transitions. On the ultrashort time scale, the expansion is along c_0, unlike in the thermal case, and the amplitude of the ensuing change is much larger than that predicted for incoherent thermal expansion. The changes in Bragg-spot intensity and width are very different from those observed during the course of equilibrium heating as well. Following an ultrafast *T*-jump, the structure coherently expands within ~10 ps (because of atomic displacements) along the *c*-direction. On the nanosecond and longer time scale, the structure shrinks as it reaches the equilibrium state (incoherent movement of atoms), and the original configuration is recovered by heat diffusion on the millisecond time scale between pulses. This behavior is in contrast with that observed at steady state, as mentioned above.

Summary. In a series of studies presented here, the (ensemble-averaged) temporal evolution of the fractional helical content as well as the time scales characteristic of helicity loss were obtained for the helical protein $T\beta_9$ following a variety of temperature jumps in vacuo. Importantly, the local (hydrogen-bond-specific) dynamics were explored together with the global structural evolution by examining the small-*r* and large-*r* regimes of $\langle f(r, t)\rangle_n$ and $\langle f_B(r, t)\rangle_n$ and calculating the corresponding radii of gyration. In doing so we demonstrated that the above rudiments of UED data analysis constitute a universal coarse-graining approach that simultaneously captures local and global structural fingerprints characteristic of biological macromolecules in real time.

As with the DNA double helix, the interplay of enthalpic and entropic forces in $T\beta_9$ can be seen in the different dependences of these two processes on temperature. For lower *T*-jumps (ΔT = 300 K), conformational diffusion was shown to *precede* any significant hydrogen bond cleavage. However, with the increasing ΔT, the Arrhenius bond breaking process accelerates exponentially whereas the rate characteristic of the diffusive motion is roughly proportional to the square root of the final temperature of the ensemble. At a certain critical final temperature the time scales for the two processes cross, and, for higher *T*-jumps (ΔT = 600, 900 K), the native contacts appear to break *before* a significant global conformational change can occur. The above behavior is evident in both the ensemble-averaged radial distribution functions and the radii of gyration, but calculation of their backbone-related temporal profiles, $\langle f_B(r, t)\rangle_n$ and $(\langle R_g^2(t)\rangle_n)_B^{1/2}$, is *required* to obtain a clear-cut picture of the structural change. In addition, unfolding of the helix was found to follow a three-state process in which the intermediate ensemble is represented by a population lacking the canonical helical structure and hydrogen-bonding, but nevertheless possessing residual helicity. The barriers of this three-state mechanism were calculated from the ensemble-averaged kinetic data and the transition from the intermediate state ensemble to the coil state was found to be the rate-limiting step, being slower than the global structural relaxation for lower *T*-jumps.

In Section 4.2.1 we pointed out that, to properly account for the actual structural changes induced by external perturbations, the correct balance between comprehensiveness and structural specificity must be attained. The temperature dependence of the global and local dynamics of $T\beta_9$ in vacuo is summarized in Figure 4.15, top. The elucidation of the three-state unfolding kinetics characteristic of the local structure and the persistence of the

intermediate ensemble state on time scales longer than that associated with the global structural reorganization implies that the use of global order parameters such as the radius of gyration or RMSD with respect to the native-state structure to monitor the progress of (un)folding may underestimate the unfolding time. Therefore, analyses lacking a local-structure-specific metric such as $\langle f(r, t)\rangle_n$ or preferably $\langle f_B(r, t)\rangle_n$ may turn out to be insufficient. In thinking about macromolecular dynamics characteristic of both nucleic acids and proteins, the results of Figure 4.15 challenge the intuitive notion that the largest length scales are necessarily associated with the longest time scales.

4.2.4 Effect of Solvent on Protein Mobility

In the above Section 4.2.2, the equilibrium structure and time-dependent unfolding dynamics of DNA were investigated both in the presence and absence of water. The effect of the solvent was mainly examined in terms of its influence on structural properties such as the overall (ensemble-averaged) morphology of DNA macromolecules and disruption of local helical ordering associated, e.g., with nucleation of an internal bubble during the course of unfolding. In the present Section, we consider the effect of water on the dynamics of both secondary and tertiary structure of the 20-residue Trp-cage mini-protein depicted in Figure 4.14, top right (115).

Explicit-atom ensemble-convergent MD simulations were performed for 38 independent trajectories of Trp-cage in aqueous solution, each lasting 60 ns. To enable comparison with free macromolecules, 76 independent trajectories of Trp-cage, each lasting 2 μs, were generated in vacuo as well. All simulations were carried out using the CHARMM suite of programs and force field (118) and coupled to a Nose thermostat to

obtain a canonical room temperature ensemble. The ensemble convergence was assessed by ensuring that doubling of the simulation time window did not significantly affect the results. The solution phase simulations were performed on the peptide, 3914 TIP3P (220) water molecules, and one chlorine atom for neutrality. The system was restricted to a cubic box with initial sides of 50 Å and equilibrated at constant temperature (298 K) and pressure (1 atm) with periodic boundary conditions. The trajectories were seeded from the 38 NMR structural variants. For gas phase simulations, 1 ns long solution phase trajectories were generated for each of the 38 NMR structural variants and the final conformations from these trajectories were used to double the number of initial structures. The gas phase simulations were also performed without coupling to a thermal bath (microcanonical ensemble) to confirm that the free energy landscape was not significantly affected.

Shown in Figure 4.16 are temporal RMSD profiles with respect to the experimentally measured native structure of the polypeptide as obtained for four independent trajectories in aqueous solution (top left) and in vacuo (bottom left). Although vibrational and rotational fluctuations do occur in the gas phase on the smallest length-scales, larger-scale fluctuations are virtually frozen out. This may be attributed to the high energetic penalty associated with breaking inter-residue hydrogen bonds in the absence of compensating hydrogen bonding provided by the hydration shell. For all trajectories, the expected time to witness RMSD change of 0.5 Å is increased by up to two orders of magnitude in the absence of water. This translates to a reduction of up to 1.6 kcal/mol in the free energy landscape roughness (Figure 4.16, bottom right), which is defined to be the logarithm of the wait time multiplied by kT. These findings are consistent with electrospray mass spectrometry experiments indicating that

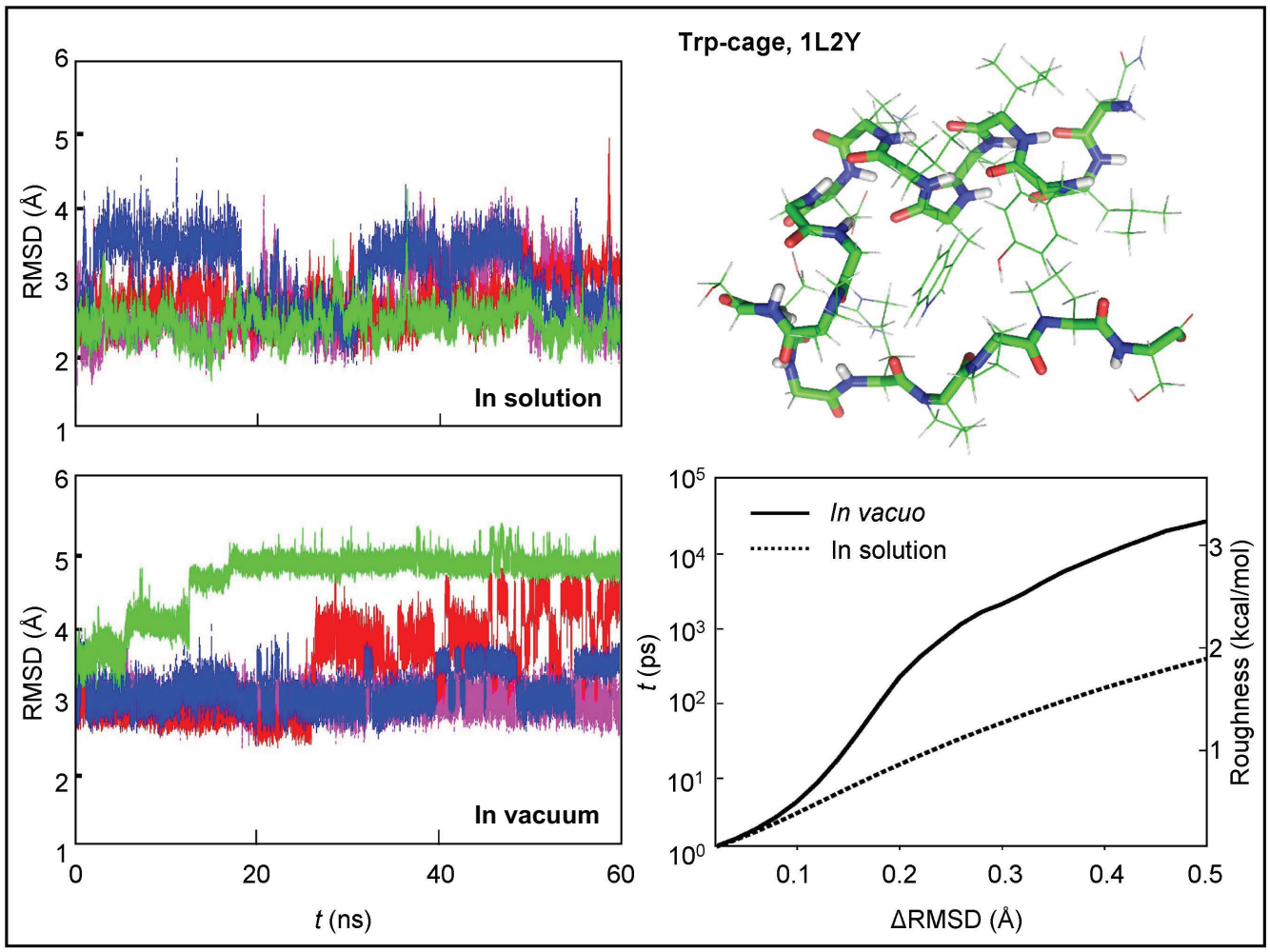

Fig. 4.16 Roughness of the trp-cage energy landscape. The *rmsd* as a function of time is shown for four representative trajectories of the solvated (top left) and vacuum (bottom left) MD simulations of the trp-cage mini-protein (top right). Whereas atomic-scale fluctuations are still present in vacuum, dynamical transitions on larger length scales are suppressed as compared with the solvated case. The change in the *rmsd* as a function of time, and averaged over all trajectories, is also shown for the solvated (dotted line) and vacuum (solid line) states (bottom right). The roughness of the energy landscape, which by definition starts from zero and increases by the logarithm of the wait time, is shown on the right axis. Note the "freezing out" of larger-scale conformational fluctuations upon removal of water.

conformational interconversions in the protein cytochrome c are up to five orders of magnitude slower in vacuo than in solution (254). We conclude that the radically different protein dynamics observed in vacuo is associated with the “glassy” behavior induced by the conformational frustration. Therefore, water is not only crucial for proper biological structure formation, but it also facilitates large-scale conformational motions that would be disabled in gas phase macromolecules.

Chapter 5

RESULTS: PROTEIN FOLDING

Although proteins typically possess numerous degrees of mechanical freedom, this alone is not sufficient to account for their complexity. Rather, it is the cooperative behavior of their constituents relevant at all levels of protein folding and function that renders biomolecular complexity a unique and fascinating phenomenon. For the purpose of exploring the thermodynamics and kinetics of the folding process, this complexity can be condensed to a (multi-dimensional) free energy landscape through which the protein traverses a multitude of folding pathways funneling to the native state (109, 255-258). Depending on the (system-specific) interplay of entropic and enthalpic factors, and the presence of kinetically stable intermediates, the folding process may occur on various time scales ranging from microseconds to minutes (257). In the present Chapter, simple yet predictive analytic methods with which to elucidate the overarching mechanisms and calculate the actual rates of protein folding are outlined and tested against the results of ensemble-convergent MD simulations and experimental evidence. Beginning with α-helix nucleation and propagation and ending with tertiary structure formation, the fundamental limits of length and time scales are obtained at each level of the protein structural hierarchy. Importantly, the results obtained are further used to assess the validity of a number of now-prevalent paradigms concerning the fundamental mechanisms of protein folding. In so doing, we demonstrate that the existing consensus in regard to such mechanisms needs to be drastically revisited.

5.1 SECONDARY STRUCTURE KINETICS AND THE SPEED LIMIT

5.1.1 Helix Nucleation

Introduction. The requirement that three adjacent sets of backbone torsional angles be specifically defined (Section 2.2.2) for α-helix nucleation to occur, renders this the most basic cooperative process ubiquitous in protein folding. This cooperative behavior separates α-helix nucleation from more elementary protein dynamics such as, e.g., single-bond rotation or non-specific contact formation. From the experimental perspective, although the associated kinetics have been measured by a number of workers and found to occur in hundreds of nanoseconds (259-261), the multiple possible nucleation sites along the peptide, in conjunction with other kinetics processes such as helix propagation and non-native contact formation, appear to determine the overall time scale of helix formation which is dependent on the protein sequence, temperature, and length of the helical moiety. Hence, for such studies the elementary event of helix nucleation is convoluted by other processes that may dominate the dynamics observed. Early MD simulations yielded a variety of time scales ranging from hundreds of picoseconds to a few nanoseconds for α-helix nucleation in short polypeptides (262-264). Prior to the ultrafast measurements presented in this Section, such time scales were not accessible to (fast) *T*-jump experiments, which were limited to 10–20 ns temporal resolution (256, 259-261, 265-270). With the ultrafast *T*-jump methodology developed in this laboratory (271, 272), we were able to isolate, for the first time, the fundamental processes pertinent to helix nucleation (90), which is the speed limit of protein folding (273).

In the following, ultrafast *T*-jump spectroscopy, analytical modeling, and ensemble-convergent MD simulations are employed to glean a comprehensive picture of the folding of an α-helical nucleus. For example, the free energy landscape constructed using the MD data revealed that, despite its relative simplicity, helix nucleation exhibits a complex dynamical behavior similar to that typically associated with the folding of the entire protein (255). Perhaps the most significant finding, both experimentally and theoretically, is that the nucleation process can be decomposed into (fast) cooperative annealing and (rate-determining) conformational diffusion stages, the latter of which can be described analytically.

To isolate the helix nucleation step from other processes such as, e.g., helix propagation (or growth), which are known to convolute the associated kinetics, the α-helix nucleation dynamics in polypeptides composed of only five residues (Wgh5 and Wh5; Figure 3.2, bottom center) was investigated. For Wgh5 and Wh5, the dependence of the nucleation rate on both the polypeptide sequence and temperature was explored. To enable comparison with longer polypeptides for which the helix growth beyond a single turn is not precluded by the small number of residues, the refolding kinetics of a 21 residue polypeptide (Wh21) was measured. The experimental results obtained were found to be in agreement with the predictions made using an analytic theoretical model as well as ensemble-convergent MD simulations. The joint theoretical, experimental, and computational efforts undertaken during the course of the study reported below provided a holistic and intuitive picture for the elementary steps of helix nucleation, and yielded predictive insight into the associated structural dynamics taking place on a feature-rich free energy landscape.

Experimental. Helix nucleation rates were measured using time-resolved fluorescence spectroscopy for the following macromolecules: alanine-based pentapeptide, Ac-W-A_3-H-NH_2 (Wh5), glycine-based pentapeptide, Ac-W-G_3-H-NH_2 (Wgh5), and alanine-based twenty-one-residue-long polypeptide, Ac-W-$(A)_3$H-$(A_3RA)_3$A-NH_2 (Wh21) (Figure 5.1). To ensure that the side chain of the histidine residue was protonated, the polypeptides studied were prepared in acetate buffer (pH = 4.8). Due to the higher helix propensity of alanine compared to glycine (274), Wh5 and Wgh5 display different helix content and stability. In addition, their minimal size ensures that helix propagation is precluded. The tryptophan residue in each pentapeptide serves as a sensitive probe of the conformational change induced by the ultrafast *T*-jump because its fluorescence is quenched by the histidine residue when the two residues come to close proximity. The polypeptide samples obtained from California Peptide Research were characterized by some 98% purity. N-acetyl-L-tryptophanamide (NATA) purchased from Sigma was more than 99% pure. The polypeptide concentrations were assessed from the optical absorbance at 280 nm, using the molar extinction coefficients of 5690 $M^{-1}cm^{-1}$. Solutions were buffered with 20 mM sodium acetate (pH = 4.8). The near-IR signal and idler pulses were generated by two optical amplifier systems pumped by a Ti:sapphire amplifier laser system operating at 800 nm with a repetition rate of 200 Hz. The (initiating) *T*-jump pulse was set to 1.45 μm with a sufficient energy, typically 15–20 μJ at the location of the sample. The experimental *T*-jump procedure utilizing ultrafast-laser setup has been described in detail elsewhere (271,275).

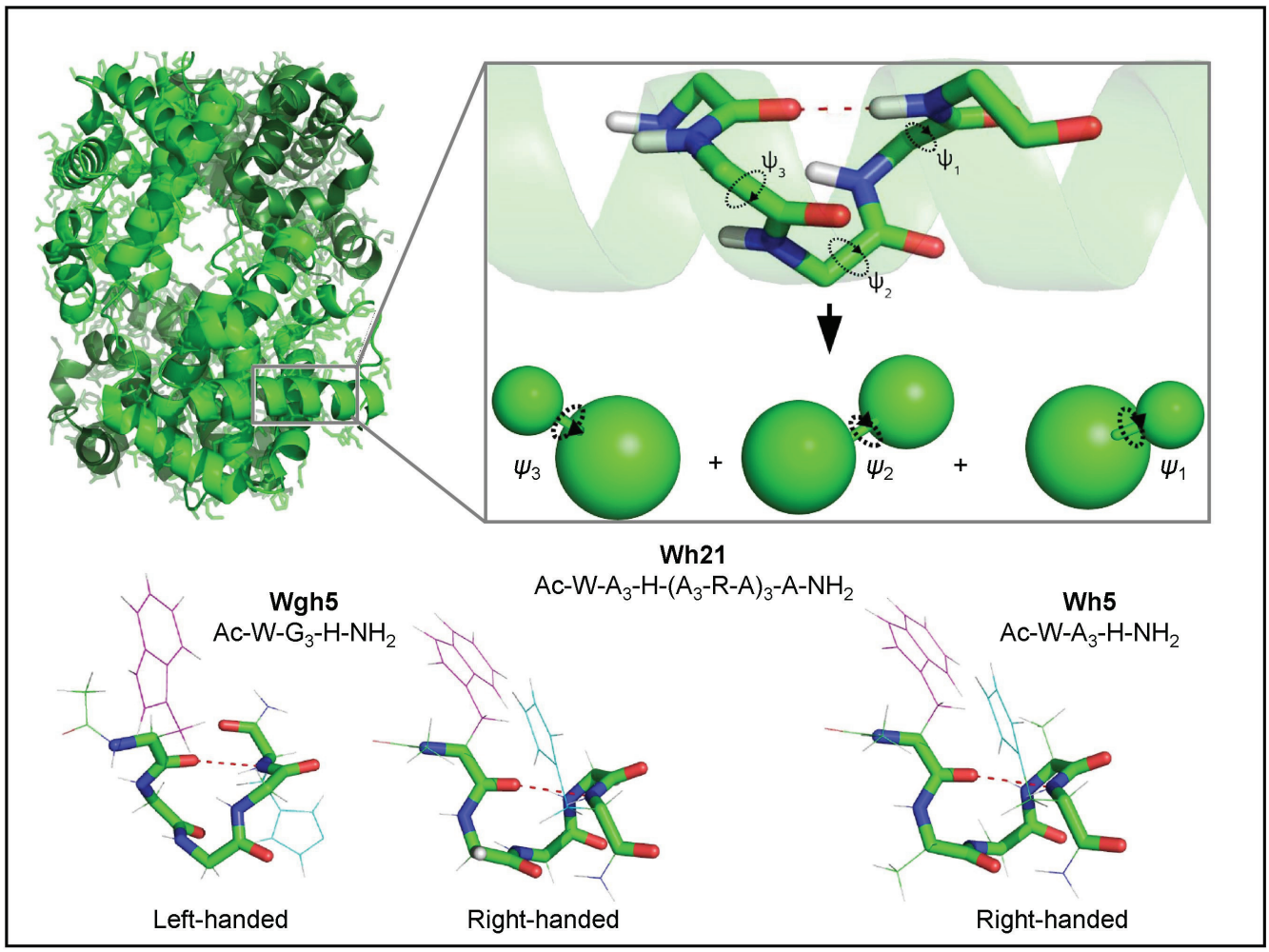

Fig. 5.1. The α-helix nucleus. Three torsion angles are the cooperative degrees of freedom that define the backbone conformational state of the five-residue segment responsible for α-helix nucleation (top inset). The memory-loss diffusion model maps the molecular content (both backbone and side chain atoms) into effective van der Waals spheres on each side of a torsion angle. The helix initiation time is the characteristic time for all three torsion angles to diffuse to the helical domain, subject to thermal excitation and viscous drag on the effective spheres (see section 5.1.1 in Text). Structures of the alanine-rich Wh5 and both right- and left-handed structures of the glycine-rich Wgh5 are given (bottom). In addition, the 21-residue alanine-rich polypeptide Wh21 was also studied. Tryptophan and histidine residues constitute residues 1 and 5 of Wh5, Wgh5, and Wh21, enabling fluorescence probing of the conformational state as a function of the tryptophan-histidine distance.

Shown in Figure 5.2, top left, are representative (equilibrium) far-UV CD spectra of Wgh5 and Wh5. Remarkably, the results of Figure 5.2 indicate that such short polypeptides do exhibit measurable helix formation in aqueous solutions. The (alanine-based) Wh5 is characterized by a right-handed α-helical structure and its CD spectrum is found to be in good agreement with those reported for short polypeptides and protein helices (276, 277). The (glycine-based) Wgh5, on the other hand, produces a dominant left-handed helix signature, similar to the left-handed-helix spectrum reported for D-amino acid polypeptides (278); however, although the left-handed helical population is dominant, the structures we probe for Wgh5 are those of right-handed helices because the strong tryptophan-histidine quenching mechanism is sterically suppressed in the left-handed helix due to the backbone geometry (Figure 5.1). Using the CD spectra obtained at T = 310 K, the ensemble-wide percent helicity was found to be 20 ± 5 and 5 ± 2 for Wh5 and Wgh5, respectively, and the fluorescence signal measured during the course of the T-jump experiments revealed the decrease in right-handed helix concentration for Wgh5 when compared with Wh5. Helices of varying lengths were observed for Wh21 by varying the concentration of the denaturing agent guanidine hydrochloride (Gdmcl), and the spectrum of partially denatured Wh21 was shown to approach that of Wh5 when sufficient denaturant was added to reduce the longer helices to a single helical turn (Figure 5.2, bottom left).

During the course of the T-jump experiments, polypeptide solutions were heated using an ultrafast IR pulse centered at 1.45 μm, and the transition to the new equilibrium over time was monitored through the quenching of tryptophan fluorescence by histidine. Figure 5.2, top right, depicts the T-jump-induced kinetics for the final temperature of T = 300 K as obtained for Wgh5 (blue line) and Wh5 (green line) over a time window ranging

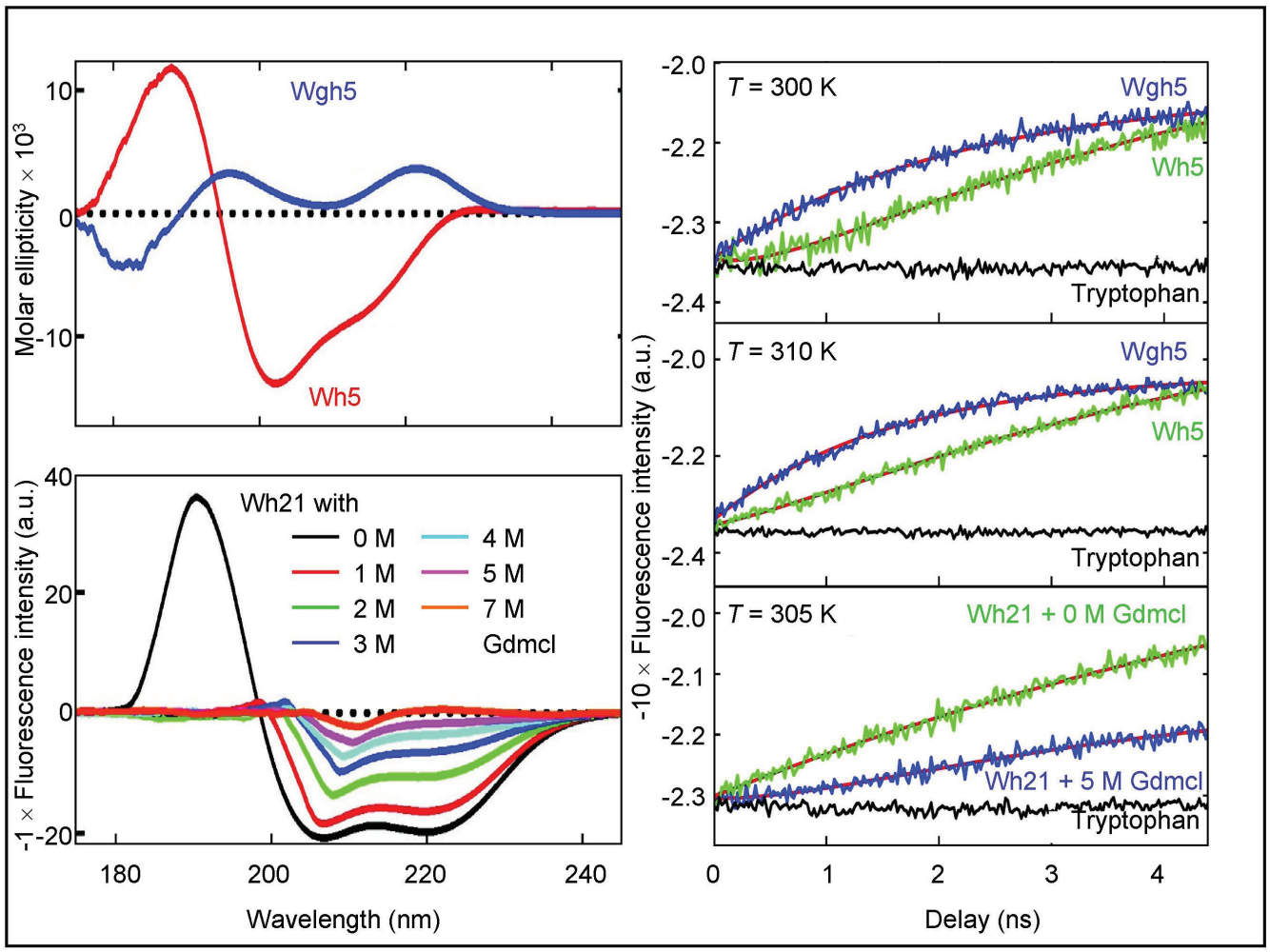

Fig. 5.2. CD spectra and ultrafast *T*-jump transients of α-helix nucleation. Far UV CD spectra of Wgh5 (blue) and Wh5 (red) at 266 K in aqueous solution, showing the signatures of left-handed, and right-handed α-helix content, respectively (top left). CD spectra of Wh21 as a function of guanidine hydrochloride (Gdmcl) concentration at room temperature shows the progressive denaturation of helical content (bottom left). Transient evolution of the tryptophan fluorescence was monitored for Wgh5 (blue) and Wh5 (green) after ultrafast *T*-jump to final temperatures of 300 K and 310 K over a time window ranging from t = -50 ps to t = 4.4 ns (top and center right). The fluorescence signals are shown with negative amplitude in arbitrary units. The initial heating is through the excitation of the overtone of the OH stretching vibration of water, inducing a 12° C temperature jump. The transient evolution of the tryptophan fluorescence of Wh21 in acetate buffer at pH = 4.8 with 0 M (green) and 5 M (blue) Gdmcl, following the *T*-jump to 305 K final temperature, shows the helix refolding following the excitation (bottom right). The flat black curve is the fluorescence of the free tryptophan in water under identical conditions, following the *T*-jump, which is used as a baseline. All experiments were performed in acetate buffer at pH = 4.8.

from t = 0 to t = 4.4 ns. We note that within the first 50 ps the tryptophan fluorescence decreases dramatically due to the temperature rise resulting from the water thermalization (90). The unfolding and solvent relaxation (solvation) events that occur during this time interval cannot be resolved. It is not until the completion of the thermalization process that the conformational diffusion rate equilibrates at the final temperature leading to an additional quenching (decrease) in the tryptophan fluorescence caused by the proximity of the histidine residue when the peptide conformation nears that of the α-helix. Therefore, to isolate the refolding relaxation dynamics, the "time zero" is set to be 50 ps following the *T*-jump throughout the *T*-jump experiments reported here.

Notably, the relaxation profile of Wgh5 can be described by a single exponential function, whereas for Wh5, a double exponential function has to be invoked. The resulting time constants are: τ = 2.2 ± 0.3 ns and τ_1 = 0.85 ± 0.3 ns, τ_2 = 5.3 ± 1.9 ns for Wgh5 and Wh5, respectively. Remarkably, the above observations indicate that the protein sequence determines the folding behavior even at the smallest scale of length. Thus, for Wh5, the steric hindrance arising from the presence of the methyl groups of alanine side chains separates the torsional diffusion from local annealing, whereas the absence of such hindrance in Wgh5 results in single exponential behavior during the course of the helix nucleation process.

By varying the magnitude of the *T*-jump, the temperature dependence of the above time scales was studied as well. Thus, at elevated temperatures, the rate of helix formation was found to increase. At T = 310 K, the time constant measured were: τ = 1.4 ± 0.2 ns and τ_1 = 0.65 ± 0.25 ns, τ_2 = 4.7 ± 0.6 ns for Wgh5 and Wh5, respectively (Figure 5.2, center

right). This trend is consistent with the conformational diffusion being the rate-limiting step in the process of folding (see below). The faster component (τ_1) contributes ca. 10 to 20% of the measured signal amplitude depending on the final temperature of Wh5. We note that, in the case of Wh5, the error bar for the slower component (τ_2) is large because of the short experimental time window. For the latter reason, we repeated the experiment at a higher temperature to identify the asymptotic level of the recovery, and confirmed that the quoted experimental errors were correctly estimated. From the observed dynamics, we infered that the studied process was (at least) three-state, and, as was evidenced from the measured disparity between signal amplitudes of the faster and slower components, the rate-determining step was associated with the transition from the unfolded state to the intermediate state (or ensemble); importantly this interpretation was borne out by the theoretical analysis (see below). We emphasize that, for Wh5, the asymptotic level of the recovery was found to be a factor of three lower at T = 330 K than the maximum value measured at T = 310 K. This observation is consistent with the relative helix content obtained from the (equilibrium) CD data representative of those temperatures.

To investigate the impact of the polypeptide length on the structural dynamics of α-helix formation, *T*-jump relaxation measurements were also performed on the twenty-one-residue-long α-helical polypeptide Wh21 in the presence and absence of a denaturant. Figure 5.2, bottom right (green line) displays the faster relaxation signal component as obtained for Wh21 at a final temperature of T = 305 K. Although 21-residue-long alanine-based polypeptides appear to be well-studied systems (260), to the best of our knowledge this is the first evidence of the existence of a few-nanoseconds-long transient behavior exhibited by the macromolecule. Given the identical locations of the tryptophan and

histidine residues within the two structures under study, by comparison with the faster (sub-nanosecond) process measured for Wh5 we conclude that the rate obtained for Wh21 is associated with the backbone conformational dynamics.

The effect of denaturing agent on the above results is dramatic. Figure 5.2, bottom right (blue line) shows the ultrafast kinetics of Wh21 following the *T*-jump to the final temperature of T = 305 K in 5 M guanidine hydrochloride. The double exponential behavior observed is characterized by the $\tau_1 = 0.8 \pm 0.3$ ns and $\tau_2 = 2.8 \pm 1.0$ ns components which contribute 15 and 85 percent of the measured signal amplitude, respectively, thus mirroring the temporal behavior of Wh5, the sequence of which is identical to that of the first five residues of Wh21. It follows that, in the absence of denaturant, the refolding kinetics measured for Wh21 was convoluted by the effect of helix propagation from distant parts of the polypeptide sequence.

Coarse-grained analytic model. The dependence of experimentally measured time scales discussed above on the sequence, temperature, and length of the polypeptides under study elucidates the nature of the elementary processes involved in α-helix folding. Notably, the fast relaxation dynamics, which occurs within a few hundred picoseconds, takes place on the same time scale as the (ultrafast) formation or breaking of a hydrogen bond between tryptophan residue 1 and histidine residue 5 (279), which requires backbone annealing. The observed double exponential behavior of Wh5 with well-separated time constants indicates that α-helix nucleation for Wh5 cannot be described as a two-state process. Early MD simulations on short α-helical polypeptides have, in fact, identified locally stable intermediates corresponding to collapsed but not folded structures (280). On the other

hand, the single exponential behavior we observed for (glycine-based) Wgh5 does not necessarily imply that the folding process is two-state because multiple experimental signal components with similar time scales may not be easily separable. To aid in resolving the above uncertainty, a coarse-grained diffusion model of α-helix nucleation was constructed. The findings made using the analytical model were in good agreement with those of the *T*-jump experiments as well as the ensemble-convergent MD simulations.

Importantly, the model is designed to discriminate between two regimes, each of which is associated with its own characteristic time scale: (i) slow non-cooperative diffusive search over the collective degrees of freedom for the helical basin on the free energy landscape, and (ii) fast cooperative annealing to the native state within the found helical basin. It is assumed that the characteristic time associated with surmounting the steric barrier, which is required to escape the helical basin, is longer than the annealing time but shorter than that of torsional diffusion. This assumption was supported a posteriori by ensemble-convergent MD simulations (see below). The presence of the helical basin on the free energy landscape transforms the concept of a *continuous* rate-limiting diffusion involving three sets of backbone torsional angles into that of the *discrete* number of trials necessary to observe three independent events. In what follows, we demonstrate that the conformational diffusion process implies a well-defined time constant τ which is associated with the time interval between individual trial attempts.

Within the framework of our model, the studied polypeptide is represented as a chain, the conformation of which is determined by a set of backbone torsional angles [two torsional angles for each of the three residues that define a single turn of the helix; (Figure

5.1, top right)]. Thus, the process of α-helix nucleation which involves formation of one full turn of the helix can be parameterized using three pairs of backbone torsional angles (φ_1,ψ_1), (φ_2,ψ_2), and (φ_3,ψ_3). For the native (helical) hydrogen bond contact to form, the three sets of torsional angles must attain the helical configuration. A conformation of the studied polypeptide is defined to be helical when all the torsional angles are within the right-handed α-helix domain of the Ramachandran diagram; for such conformations, fast cooperative annealing to the native state can occur. Because Ramachandran angles φ are always within the helix domain, the above conformational diffusion problem is reduced to the probability of finding all three values of ψ to be in the helical domain. Therefore, the helix nucleation time, t_{init}, can be expressed as:

$$t_{\text{init}} = \frac{\tau}{p^3}, \qquad [5.1]$$

where τ is the characteristic time required for each degree of conformational freedom to thermally diffuse to an uncorrelated state (which is equivalent to the conformational memory loss) and p is the probability of finding each backbone torsional angle ψ to be in the α-helical domain. We note that the conformational diffusion relevant for α-helix nucleation involves the superposition of three rotational motions (or modes), and the masses of rotating bodies are shared between the modes (Figure 5.1, top right).

Each Ramachandran angle ψ is then defined by the relative rotation of the bodies on either side of the bond (modeled to first order as spheres) in a viscous Brownian temperature bath. The rotational auto-correlation time τ can be calculated as the temporal decay time of $<\cos\psi>$, where angular brackets denote ensemble averaging. Because the

diffusion process is time-translation invariant, $\langle\cos\psi\rangle$ decays by the same factor after every time interval of fixed duration, i.e., $\langle\cos\psi\rangle = e^{-t/\tau}$. This exponential decay naturally associates τ with the unique randomization time characteristic of the polypeptide conformation. By expanding both sides of this relation and matching the first-order term for early times, we obtain: $\langle\psi^2\rangle = 2t/\tau$. Unrestricted rotation with respect to the dihedral angle ψ is equivalent to independent rotational motion of two spheres on each side of the bond, denoted by θ_a and θ_b. Accordingly, $\langle\psi^2\rangle = \langle\theta_a^2\rangle + \langle\theta_b^2\rangle$. From Einstein's (1D) diffusion equation, which is valid for rotational diffusion at early times (281), $\langle\theta^2\rangle = 2kTt/\gamma$, where k, T, and γ are the Boltzmann constant, absolute temperature, and friction coefficient, respectively. Substituting the latter expression for $\langle\theta^2\rangle$ and then $\langle\psi^2\rangle$, one finds:

$$\tau = \frac{\gamma_a \gamma_b}{kT(\gamma_a + \gamma_b)}. \qquad [5.2]$$

Assuming that the rotating bodies are spheres of volume V immersed in a fluid with viscosity η, one obtains (282) $\gamma = 6\eta V$. Therefore, τ can be expressed in terms of the temperature T, viscosity of water η, and volumes of the rotating bodies on either side of each bond ($V_{a/b}$, which are similar for all three bonds): $\tau = 6\eta V_a V_b / kT(V_a + V_b)$. We note that Equation 5.2 is similar to the one obtained by Debye for the dielectric relaxation time of water using a geometric method (283), except for water dipole being a single body with two effective degrees of rotational freedom, rather than the superposition of three rotational modes, each of which is associated with a single degree of conformational freedom of the studied polypeptide.

Recently, the classical Ramachandran diagram of sterically allowed (and prohibited) structural domains in polypeptides (79,284) was updated by Ho *et. al* (285) to reflect the empirically observed dihedral angle distributions. The (non-glycine/proline) diagram thus obtained is reproduced in Figure 5.3, top left inset, with the most sterically favorable backbone conformations represented by the dark blue areas of the diagram. According to the diagram, there exists a weak steric barrier separating the domain populated by right-handed α-helices from that of the β-strands (the regions within these domains corresponding to the α-helix and β-strand conformations are colored in red and yellow, respectively). We note that the right handed α-helical domain spans the entire range of the most sterically favorable φ-values and about 50% of the most sterically favorable ψ-values represented by the upper left quadrant of the diagram, neglecting the (much smaller) left-handed α-helical domain, which is isolated by a significant steric barrier (285); therefore, $p \approx 1/2$. By substituting the value of p and Equation 5.2 into Equation 5.1, the nucleation time can be expressed as:

$$t_{init} = \frac{48\eta V_a V_b}{kT\left(V_a + V_b\right)}. \qquad [5.3]$$

For α-helix nucleation taking place in the interior of a long polypeptide, V_a and V_b should represent the volume associated with the Kuhn length of the chain, which, in proteins, is typically ca. 3 to 4 residues long (286). Thus, the spherical approximation is justified even for helices nucleating in the interior of long proteins, where the nucleation rate is approximately three times as slow as that of a similar process taking place near the ends of the chain (we note that, for helices nucleating in the interior of a protein, the increased

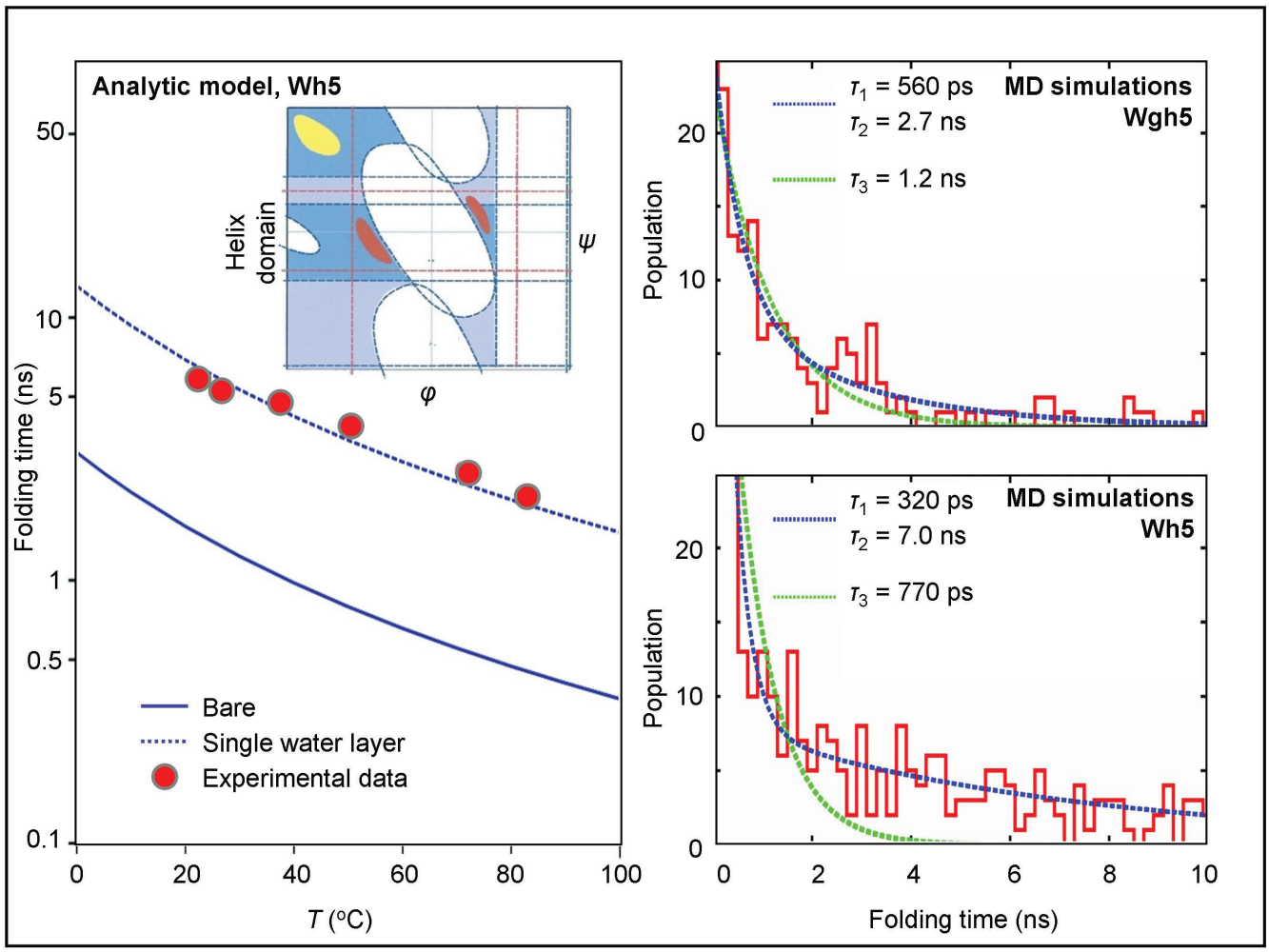

Fig. 5.3. Theoretical and computational results for α-helix nucleation. The helical and non-helical domains used in the model are taken from the corresponding regions of the (non-glycine) Ramachandran plot (left inset). The α-helix regions (both left- and right-handed) are shown in red whereas the β-strand region is shown in yellow. Note that each region is within a larger conformational domain (dark blue). Whereas crossing between right-handed helix and β-strand domains requires traversing a region of minimal steric strain (light blue), the left-handed helical domain is separated from the other domains by a region of substantial steric strain (white). The predictions of the model are compared with the experimental results for Wh5 as a function of temperature, both for the bare rotation of the residues and for the more realistic case of dragging a single layer of water during the rotation (left). The predicted rate agrees with the experimental results, both in absolute terms and as a function of temperature. MD results showing the histogram distribution of folding times for Wgh5 (top right) and Wh5 (bottom right) are presented. The refolding statistics for Wh5 display two well-separated time scales while those for Wgh5 indicate a single time scale. The values of all fitted time scales are in good agreement with those measured experimentally, and the single-versus-double timescale behavior is explained by the free energy landscapes of the two sequences (see Figure 5.4).

values of $V_{a/b}$ are associated with an increased friction between the solvent and solute). In the following, Equation 5.3 will be employed to calculate α-helix nucleation rates for real polypeptides, which will be further compared to those obtained using the ultrafast *T*-jump experiments as well as ensemble convergent MD simulations.

For Wh5, the (slowest) internal rotational motion with respect to the torsional angle ψ_2 is characterized by $V_a = 323$ Å^3 and $V_b = 232$ Å^3 in the absence of the hydration shell. The effective volumes assuming one or two layers of water bound with varying tightness to the surface of the rotating bodies, which is typical for proteins, can be estimated using the van der Waals diameter ($d = 2.82$ Å) of a water molecule (287). At $T = 293$ K, given that the viscosity of water is $\eta = 0.001$ Ns/m^2 (288), we obtain $t_{init} \approx 5$ ns in the assumption of a single-layer hydration shell (289), which is consistent with experimental observations. For Wh5, the characteristic time scale of α-helix nucleation as obtained using Equation 5.3 and ultrafast *T*-jump measurements is plotted in Figure 5.3, left, as a function of temperature. We note that the effect of the temperature on polypeptide folding rate is mainly a consequence of the temperature dependence of solvent viscosity. The order-of-magnitude estimate provided by our simple model is surprisingly good, given that both the amino acid residue shapes and intramolecular interactions are completely neglected. Remarkably, the analytic description of cooperative diffusion can accurately account for the processes involved in α-helix nucleation as well as the overall folding times. It is also noteworthy that the slope of the temperature dependence of the Wh5 folding time scale as obtained using Equation 5.3 is in excellent agreement with the experimental results.

Computational. To validate the assumptions and predictions of the above analytic model, and to obtain an insight into higher-order details of the experimentally obtained time-dependent data such as, e.g., the double and single-exponential refolding behavior exhibited by Wh5 and Wgh5, respectively, ensemble-convergent MD simulations were carried out using the CHARMM (118) suite of programs and the CHARMM22/CMAP force field. For Wh5 (Wgh5), the polypeptide structure was centered in the cubic primary-simulation cell with initial box length of 30.0 (28.5) Å. To mimic a 13° *T*-jump, periodic boundary conditions at $T = 311$ K were employed. The starting-point structures of both polypeptides were assumed to have random backbone configurations and the end-capping identical to that characteristic of the *T*-jump measurements. In addition to the polypeptides, 872 (747) TIP3P water molecules and one chloride ion were added as a 61.5 (71.7) mM salinity solvent yielding an electrically neutral system which comprised 2,698 (2,314) individual atoms. To calculate α-helix refolding times, an unfolding event was defined as accomplished when the RMSD with respect to the (canonical) α-helical structure was found to exceed 2 Å. In contrast, a refolding event was tabulated when the RMSD lower than 0.5 Å was detected for four consecutive (1 ps) MD frames following an unfolding event. The latter assumption is justified because the mean amplitude of the conformational fluctuations within a helical polypeptide equals 0.5 Å. One hundred independent MD trajectories, each lasting 100 nanoseconds, were obtained, yielding a total simulation time of 10 μs for each polypeptide.

For every (1 ps) MD frame, the RMSD of the polypeptide conformation with respect to that of a canonical α-helix was calculated. All conformations characterized by RMSD ≤ 0.7 Å were considered to be α-helical because the outer envelope of the (well-

defined) RMSD basin representing α-helical structures throughout the MD simulations corresponded to 0.7 Å. At T = 311 K, using the above criterion, the population of right-handed α-helices was estimated to constitute 20 and 3 percent of the studied macromolecular ensemble for Wh5 and Wgh5, respectively. The fraction of left-handed α-helices for Wgh5 was 6%, in agreement with the CD results dominated by the left-handed structure fingerptint.

For every instance of helix nucleation, the folding time is defined to be the elapsed time between an unfolding event and a refolding event. For the right-handed helices, the total number of independently obtained folding times was 595 and 244 for Wh5 and Wgh5, respectively, allowing for an ensemble-level assessment of the time scale characteristic of the folding process. As shown in the histogram (frequency distribution) plots of Figure 5.3, right, both polypeptides display a variety of refolding time scales ranging from hundreds of picoseconds to nanoseconds. We note that, for Wh5, the distribution pattern as obtained from the ensemble-convergent MD simulations fits closely to a double-exponential function and poorly to a single exponential function. The two time constants associated with the above biexponential behavior, τ_1 = 320 ps and τ_2 = 7.0 ns, appear to be in good agreement with those found experimentally at the same temperature. For Wgh5, the quality of a single-exponential fit is comparable to that of the double exponential fit, yielding a single time constant of τ = 1.2 ns, which is in good agreement with the single-exponential behavior found experimentally. The analysis of representative MD trajectories reveals that experimentally measured rates corresponding to shorter time scales represent hydrogen-bond formation and local structural perturbations whereas those corresponding to longer time scales represent conformational diffusion starting from a random-coil structure. The

conformational diffusion process is faster in Wgh5 than in Wh5 due to the decrease in both viscous drag and steric strain characteristic of the (smaller) glycine side chains. For Wgh5, the temporal overlap of (local) annealing and (global) diffusion processes results in a single exponential time dependence, which is in agreement with the experimental findings.

Using the data obtained from the ensemble-convergent MD simulations summarized above, the free energy landscapes of (un)folding were constructed for both Wgh5 and Wh5, as shown in Figure 5.4. Each point on the landscapes is parameterized by the polar coordinates (r, ϕ), where r is the oxygen-to-nitrogen distance between residue 1 and residue 5 (1O-5N), and ϕ is the torsional angle defined as 1O-2Cα-4Cα-5N. This choice of order parameters allows for an intuitive visualization of the entire conformational space, with r being the hydrogen bond distance and ϕ representing the overall backbone twist. The landscapes of Figure 5.4 are constructed by plotting the natural logarithm (multiplied by kT) of the fraction of time the polypeptide was found to spend at each point of the (r, ϕ)-space throughout the MD simulation window (we note that the total MD simulation time was sufficient to construct an ensemble-convergent landscape). Importantly, for Wgh5 (Figure 5.4, left), the landscape is characterized by two minima with opposite twisting directions populated by the left-handed and right-handed helical structures. Besides the greater stability of the left-handed conformers, the Wgh5 landscape reveals the lack of free energy barriers separating the helical-structure domains from the unfolded-structure domains. In contrast, for Wh5, the steric hindrance induced by the presence of methyl group side chains results in the vanishing of the minimum populated by left-handed helical structures in the case of Wgh5, which is accompanied by the formation

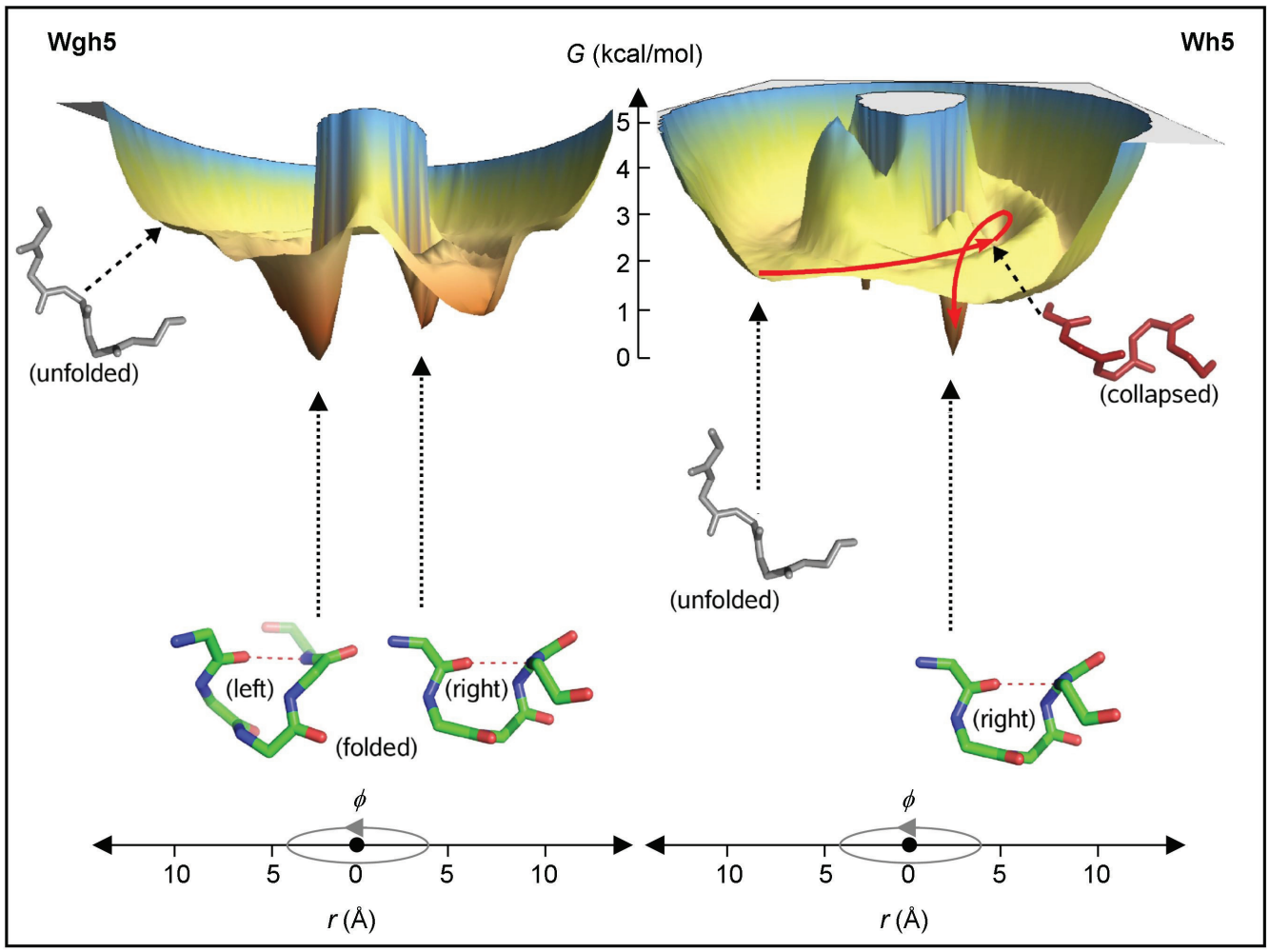

Fig. 5.4. Helix nucleation free energy landscapes for different peptide sequences. The landscapes are parameterized using the α-helix hydrogen bond distance and overall backbone twisting angle in polar coordinates (see section 5.1.1 in Text). The landscape for Wgh5 shows the existence of both left- and right-handed helices (green structures), with the left-handed helix more stable (left). In contrast, the landscape of Wh5 shows only the right-handed helix with a steric wall separating the helical domain from the unfolded state (right). The dominant pathway is defined (red arrow) by a slow diffusion from the unfolded state (gray structure) to the helical domain (red structure), followed by fast annealing to the helical (green) structure, justifying the analytical model (see Text).

of a steric wall separating the only remaining (right-handed) helical structure domain from the unfolded state (Figure 5.4, right).

The (conformational-memory-loss) analytic model therefore captures the dominant folding mechanism which implies that the peptide undertakes the (rate-limiting) diffusive search to find the helical basin. This is then followed by a (much faster) search for the native helical fold (i.e., annealing), which is facilitated by the presence of the steric wall limiting conformational diffusion out of the helical basin during the course of the annealing process (Figure 5.4, right). Importantly, the model appears to properly account for the role played by the steric wall, which separates the (rate-limiting) non-cooperative conformational diffusion from the cooperative annealing to the native state. As evident in the free energy landscape of Wgh5, in the absence of the steric wall, the peptide can diffuse out of the helical basin, and the process of α-helix nucleation is no longer split into two distinguishable temporal regimes. The results shown in Figure 5.4 validate the assumptions made in the analytical diffusion model, and completely reproduce the experimental findings.

Summary. Using the experimental ultrafast *T*-jump methodology characterized by an atomic-scale temporal resolution, we observed the folding of the shortest secondary structures possible, representing the *speed limit* of the protein folding dynamics. With the aid of experimental, theoretical, and computational studies of the sequence, temperature and length dependencies of the folding rates, a clear-cut quantitative picture of the α-helix nucleation process was constructed. The rate-limiting step of the process corresponds to the (non-cooperative) conformational diffusion search for the helical basin, and the helix

nucleation time is accurately described by the closed-form solution of Equation 5.3. There also exists a faster component of the folding process, which corresponds to (cooperative) annealing to the α-helical state. The steric (side-chain) hindrance is responsible for the separation of the time scales of these two processes. The above findings demonstrate that the interplay of conformational entropy and collective energetics occurs on even the smallest length and time scales of protein folding.

5.1.2 Helix Propagation

The formation and decay of α-helices play a pivotal role in protein folding (and misfolding). Indeed, even for proteins containing little native helical structure, helix formation, concomitant with hydrophobic collapse, seems to be the universal precursor of the folding process (290), much as the non-native conversion of helices to β-strands is the precursor to protein aggregation diseases such as Alzheimer's (83). Though the thermodynamics of such transformations was well understood beginning with the 1950s, relatively little was known about their kinetics until the end of the 1990s (41, 89). With the advent of fast *T*-jump experimental techniques (un)folding time scales became readily measurable, but the underlying structural dynamics was still inaccessible and, until very recently, the fundamental processes involved remained in the dark. In Section 5.1.1, the helix nucleation step was studied in detail, both theoretically and experimentally. In the present Section the analysis is extended to the process of *propagation*, or elongation, of the α-helical nucleus following the formation of the first turn of the helix.

The now prevalent *kinetic zipper* (KZ, or nucleation–elongation) model (41, 42, 291) of helix formation was suggested by Thompson *et al*. in the late 1990s, and it has been

widely used to interpret experimental observations ever since (292). Within the framework of the model, the nucleation step, which involves the entropy-driven conformational search for the correct helical alignment of three sets of consecutive backbone torsional angles (Section 5.1.1), is postulated to be the rate-limiting step of the process. The subsequent "zipping" (or elongation) of the helical fragment, which requires the helical alignment of only one set of such torsional angles, is viewed to proceed at a much faster rate. However, in light of helix folding experiments, in particular the recent ultrafast investigations into the helix nucleation process (90, 91), the kinetic zipper interpretation needs to be reevaluated.

First and foremost, if the concept summarized above were correct, the folding time scales measured experimentally, being primarily entropic in nature, would not differ by several orders of magnitude for α-helical polypeptides of different sequences and similar lengths (293). Second, neglecting any type of interactions besides native helical contact formation, e.g., nonnative structures stabilized by long-range hydrogen bonding contacts (294) and β-hairpin-like structures (141), as well as ignoring the fact that the so-called "random-coil" (nonhelical) state of the protein is neither random nor completely denatured (Section 4.1), can hardly be considered valid approximations. Third, in a series of studies from this laboratory described in the previous Section, we embarked on ultrafast (un)folding experiments, analytic modeling, and ensemble-convergent MD simulations that yielded, for the first time, a comprehensive atomic-scale picture of the elementary processes pertinent to the α-helix nucleation event (90, 91). These studies unambiguously demonstrated that the α-helix nucleation, when observed in isolation, occurs on a few-nanosecond time scale, which is consistent with analytical and computational assessments carried out in this (90, 91) and other laboratories (264, 295). In contrast, the overall helix

folding time scale at room temperature is on the order of hundreds of nanoseconds (42, 89); therefore, the *helix propagation* must be the rate-determining step. Finally, we note that the enthalpic barrier to the (diffusion-limited) helix formation is at most 4 kcal/mol (due to the temperature-dependence of solvent viscosity), whereas the measured barrier is 8 kcal/mol. The discrepancy of ~4 kcal/mol between the kinetic zipper interpretation and experimental evidence cannot be accounted for by mechanisms like, e.g., dipole-dipole interactions, as readily acknowledged by Thompson *et al.* (42). In the following we demonstrate that the above limitations of the KZ model stem from the neglect of misfolded intermediates characterized by non-helical hydrogen bonding contacts, which significantly retard the α-helix propagation and increase the (effective) enthalpic barrier.

Given that carrying out massively distributed ensemble-convergent MD simulations constitutes a challenge even for relatively small proteins that consist of hundreds of atoms, developing a simple yet predictive analytical model of protein (un)folding would greatly aid the detailed exploration of the physicochemical processes involved. More importantly, analytical methods, even those limited to low structural resolution, provide the crucial causal connection between fundamental driving forces and the resulting collective behavior that cannot be obtained from experimental or computational observations. In the present Section, we develop the KIS model of α-helix propagation, and demonstrate that it naturally accounts for the relevant kinetics and structural dynamics of the process under study. Importantly, the above approach can serve as a model framework for analyzing order–disorder transitions in systems for which non-native interactions are important.

Computational. To illustrate the dominant effects of misfolding, which cannot be accounted for within the now-prevalent paradigm, we begin by examining polyalanine, which is the standard testbed of α-helix formation. MD folding simulations were performed using the CHARMM program suite (118) and the CHARMM22/CMAP force field on an ensemble of 25-residue-long polyalanine chains (A_{25}), each of which was solvated by 7583 TIP3P water molecules in a cubic box with periodic boundary conditions; the sides of the box equilibrated to 60 Å at 278 K. Long-range electrostatic interactions were computed using the PME method (121, 122). Initially, the system was heated to T = 500 K during 500 ps and 100 conformational snapshots were randomly selected from the last 250 ps of the resulting trajectory to construct a macromolecular ensemble. To study the folding of A_{25} in a regime strongly favoring α-helix formation, the final temperature was chosen to be T = 278 K ($\Delta T < 0$); for all trajectories, the MD simulation time window was 14 ns long.

A careful examination of representative structures that dominate the MD ensemble during the initial stages of folding reveals that α-helical nuclei appear to form almost instantaneously (within several nanoseconds) across the entire ensemble, whereas water-mediated structures resembling β-hairpins that are, in fact, more abundant than α-helices (Figure 5.5, bottom) tend to impose restrictions on the helix elongation (propagation) rate. Indeed, because growth of helical islands requires unraveling of misfolded structures that hinder further zipping, the helix propagation process in polyalanine appears to be significantly *slower* than the helix nucleation process. It is also noteworthy that, because the (longer) misfolded sequences associated with close-to-an-end α-helix nucleation should take longer time to unravel than the (shorter) ones arising upon α-helix nucleation somewhere in the middle of the polypeptide chain, a helical island that nucleates in the

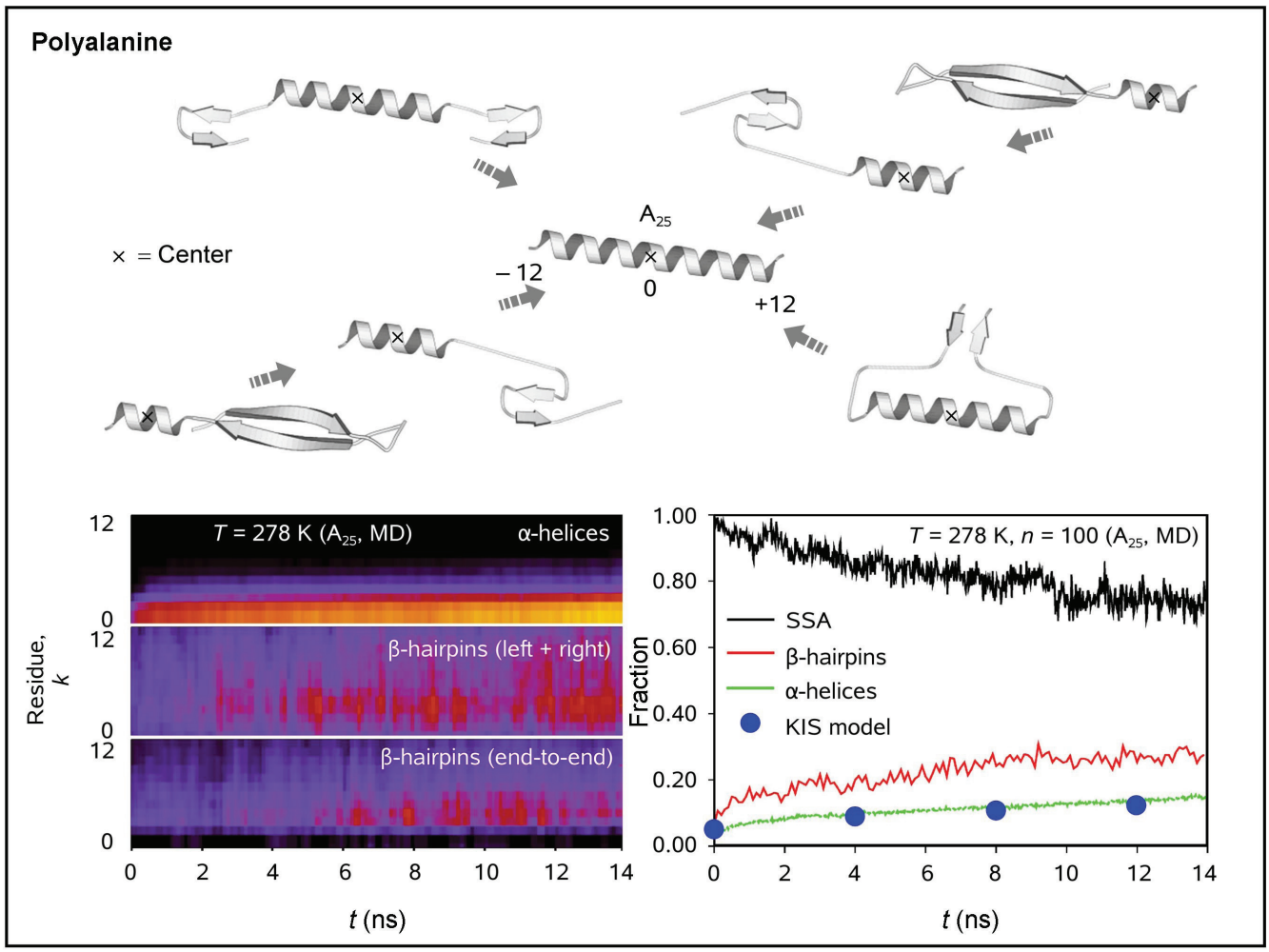

Fig. 5.5. Coil-to-helix transformation of polyalanine (A_{25}) observed with molecular dynamics. Distributed ensemble-convergent MD simulations carried out for n = 100 trajectories of A_{25} at T = 278 K indicate that the growth of α-helices (black: 0%; yellow: 100%) is hindered by formation of β-hairpins (black: 0%; yellow: 20%). In each of the three panels, individual residues k are numbered from –24 to +24 with the center of the helical segment of each member of the ensemble aligned at the origin by definition (top). Helical-island and non-native hairpin formation processes are depicted as a function of $|k|$ (bottom left). The single sequence approximation (SSA) is a reasonable description of the dynamics, and the presence of hairpin-type contacts suppresses helix formation rates (bottom right), a result predicted by the KIS model (see Figures 5.6-7).

vicinity of the center of the chain should grow faster than the one that nucleates closer to an end (Figure 5.5, top).

Another important observation we made during the course of the ensemble-convergent MD simulations was the validity of the SSA, which assumes a single helix nucleation site within the polypeptide. As evidenced by the results shown in Figure 5.5, bottom right, the SSA holds for ca. 80% of the studied macromolecular ensemble of A_{25}, which may be rationalized in terms of a significant barrier to nucleating multiple helices. For A_{25}, this barrier is induced by the β-hairpin-like misfolded structures that prevent formation of new nucleation sites following the initial nucleation event. Importantly, the SSA allows all relevant partially folded states to be described using two variables *i* and *j* (defined below and in Figure 5.6, left), thereby providing a comprehensive coverage of the entire configurational state space in two dimensions. In what follows, we take advantage of the SSA to formulate a quantitative pictorial model of helix formation. It is demonstrated that, by capturing the effect of misfolding, the KIS model provides a much more accurate description of the observed secondary structure kinetics than the KZ model of α-helix folding.

KIS model. In the present Section, the KIS model employed in our studies of temperature-induced *unfolding* of DNA/RNA hairpins (Section 4.2.1) is further extended to describe α-helix *folding* kinetics using a parameterization scheme specific to the α-helical hydrogen-bond contacts between amino acids. For a polypeptide chain capable of forming *L* α-helical hydrogen bonds, the reaction coordinates *i* and *j* are chosen to be the number of such bonds that are broken ("unzipped") counting from the C- and N-termini of the polypeptide chain,

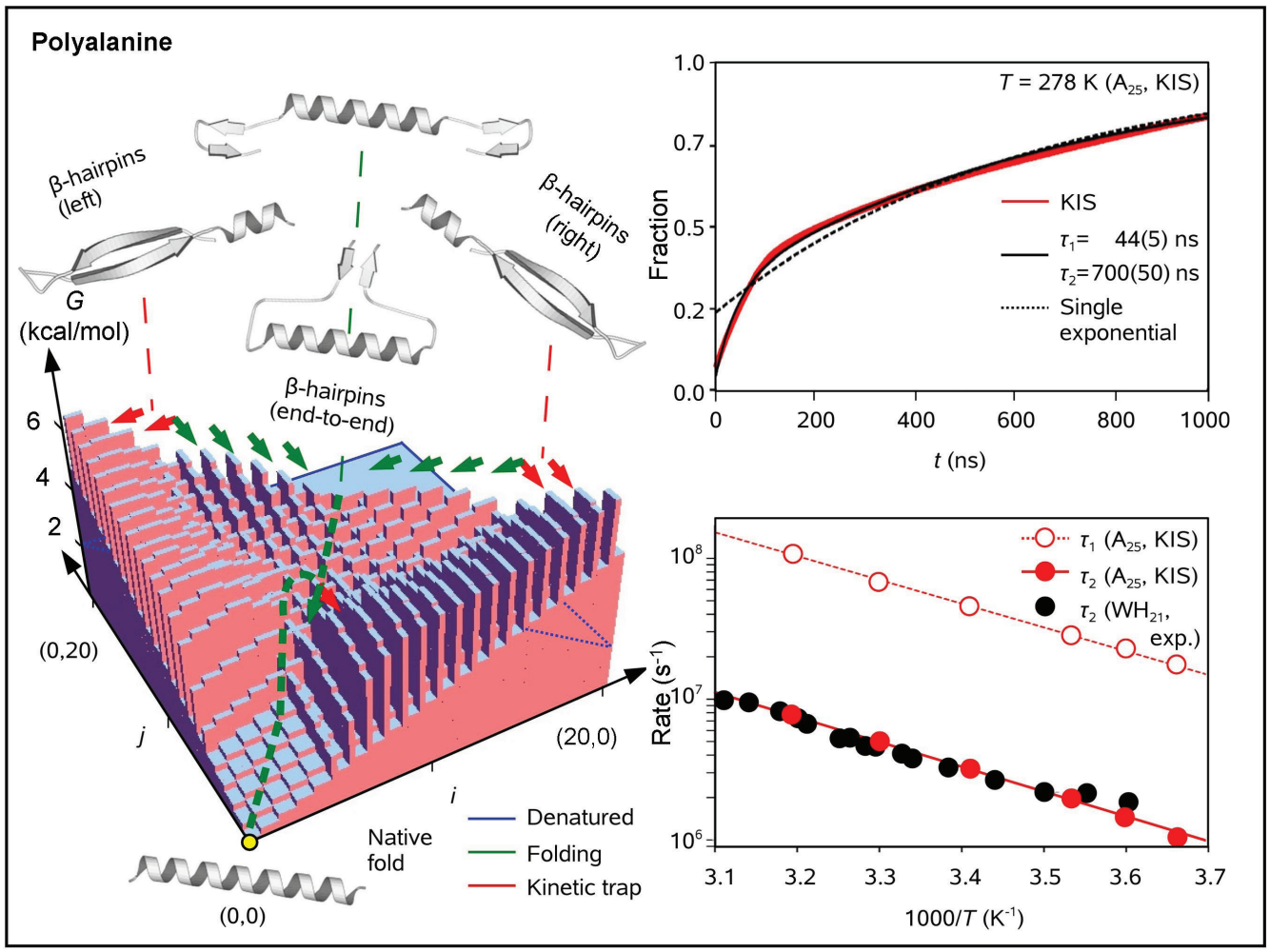

Fig. 5.6. Free energy KIS landscape of polyalanine (A_{25}) helix formation in aqueous solution. The KIS model computes the landscape using a small set of experimentally obtained enthalpy and entropy data for any temperature (see section 5.1.2 in Text); the landscape shown corresponds to T = 278 K, for which full folding occurs at equilibrium (left). The walls in the maze-like landscape are the enthalpic barriers associated with breaking misfolding backbone contacts in order to propagate the helix. The representative structures of indicated subdomains ($i >> j$, $j >> i$, and $i \approx j$) of the misfolded basin, as well as the native-fold structure ($i = 0$, $j = 0$) that dominates the semi-native basin ($0 \le i, j \le 4$), are plotted in gray. Green dashed line denotes the locus of folding pathways. Reaction coordinates i and j are the number of broken ("unzipped") native contacts from the C- and N-termini of the chain, respectively. The KIS landscape elucidates the folding mechanism and bottlenecks, and is used to predict folding time constants via Monte Carlo simulations. The biexponential rise seen in experimental folding studies is reproduced by the ensemble-averaged helicity fraction as obtained from Monte Carlo simulations performed on the KIS landscape (top right), and arises from the pathway bifurcation of the landscape (left). The obtained rate agrees with the measured value over all relevant temperatures, thereby reproducing the observed effective energy barrier which is the (negative) slope of the Arrhenius plot (bottom right).

respectively (Figure 5.6, left). We note that the choice of the coordinates implicitly constrains the model to the SSA. All states are then represented by the unique set of coordinates $\{(i, j)\}$ on the 2D reaction coordinate grid, with the native-fold (α-helical) structure of the protein located at (0, 0). The only state that is not associated with a unique point on the grid is the "coil" (or completely unfolded ensemble), which is represented by the points on the diagonal boundary of the coordinate space (i, $L - i$). We note that each state (i, j) corresponds to an ensemble of structures that share the same set of intact α-helical hydrogen bonds but may differ in their detailed atomic coordinates. The free-energy landscape $\Delta G(i, j)$ which defines the folding behavior of the polypeptide is then obtained by calculating the free energy for each (i, j)-state with respect to the unfolded state of the macromolecular ensemble:

$$\Delta G(i,j) = \left(L - i - j - 1\right)\Delta G_{\text{prop}}(T) + \Delta G_{\text{nuc}}(T)\,. \qquad [5.4]$$

Here, $\Delta G_{\text{prop}}(T)$ is the free energy change associated with forming a single backbone hydrogen bond to propagate the helix, and $\Delta G_{\text{nuc}}(T)$ is the free energy change corresponding to the formation of the helix nucleus. Similarly to Section 4.2.1, the (temperature-dependent) free energy differences can be obtained from the corresponding experimentally measured enthalpy and entropy changes: $\Delta G(T) = \Delta H - T\Delta \text{S}$. The interstate barriers crossed during the course of $(i, j - 1) \leftarrow (i, j) \rightarrow (i - 1, j)$ transitions between individual microstates are given by:

$$\Delta G_{(i,j)\to(i-1,j)} = \Delta H_{\beta} P_{\beta}\left(i; i, j\right), \qquad [5.5a]$$

$$\Delta G_{(i,j)\to(i,j-1)} = \Delta H_{\beta} P_{\beta}\left(L - j; i, j\right). \qquad [5.5b]$$

Here, ΔH_β is the enthalpic barrier associated with breaking a backbone–backbone (β-hairpin type) hydrogen bond and $P_\beta(m; i, j)$ is the probability of having to break such a non-native hydrogen bond at position $1 \le m \le L$ of the state (i, j) in order to complete the helix propagation step. We note that $P_\beta(m; i, j)$ is given by the statistical weight of the sub-population of the state (i, j) which is characterized by a non-native hydrogen bond at position m, divided by the statistical weight of the state (i, j):

$$P_\beta(m;i,j) = \frac{\exp\left(-G(m;i,j)/kT\right)}{\exp\left(-G(i,j)/kT\right)}. \qquad [5.6]$$

For a given state (i, j), let N be the number of microstates in which the β-hairpin type non-native contacts are formed. The denominator of Equation 5.6 can then be expressed as the sum of statistical weights of the N microstates:

$$\exp\left(-G(i,j)/kT\right) = \sum_{n=1}^{N} \exp\left(\frac{-d(n)\Delta G_\beta(T) - \Delta G_{\text{loop}}(n)}{kT}\right), \qquad [5.7]$$

where $d(n)$ is the number of β-hairpin type contacts, $\Delta G_\beta(T)$ is the free energy change upon formation of a β-hairpin type contact, and $\Delta G_{\text{loop}}(n)$ is mainly due to the entropy correction associated with loop formation in microstate n. We note that all N states in the sum can be computationally generated. The numerator of Equation 5.6 is calculated in the same way as the denominator except for the constraint that there is a non-native contact at position m.

The experimentally-obtained thermodynamics parameters employed in the KIS model of folding of A_{25} (in units of kcal/mol) are given by: $\Delta G_{\text{prop}}(T) = -1.3 + 0.004T$ (296-298), $\Delta G_{\text{nuc}}(T) = -1.3 + 0.014T$ (299), $\Delta H_\beta = 4$ (300), and $\Delta G_\beta(T) = -1 + 0.002T$

(300). We note that the barrier height ΔH_β is estimated from the electrostatic stabilization associated with formation of a β-hairpin and that its magnitude is reduced to account for solvation effects when the enthalpic component of $\Delta G_\beta(T)$ is calculated. Finally, for a given misfolded microstate n, $\Delta G_{loop}(n)$ is obtained from the entropy of a polymer loop of length l, which can be expressed as $A\ln(l/l_{ref})$ (301), with the reference loop length $l_{ref} = 10$, and the excluded volume pre-factor $A = -2.4$ kcal/mol (302). Loops characterized by $l < 3$ are disregarded within the framework of the model.

Crucially, although the KIS landscape of A_{25} is only parameterized by the number of broken native (hydrogen bond) contacts on either end of the α-helix, the information about misfolded (β-hairpin type) states which retard the folding kinetics is nevertheless contained in the interstate barriers of the landscape (Equations 5.5 and 5.6). As demonstrated below, in contrast to the behavior exhibited by simple chemical systems, the (misfolded-structure-dominated) interstate barriers in systems with numerous degrees of conformational freedom can overwhelm the underlying free energy landscape defined by the (i, j)-states, thereby determining the overall folding behavior and time scale.

Despite being feature-rich, the free energy landscape of α-helix folding as obtained from the KIS model for A_{25} (Figure 5.6, left) can be characterized as consisting of two major domains occupied by structurally diverse populations: (i) misfolded and (ii) semi-native. The latter, much smaller in size, represents ideal helices with up to four broken terminal contacts that rapidly zip toward the native fold, whereas the former constitutes a maze that funnels the helices nucleating close to the center of the polypeptide chain into the fast-folding pathway and, simultaneously, retards those nucleating in the vicinity of the

termini. The structures populating the larger basin range from β-hairpin-dominated (and kinetically-trapped in the rough $i >> j$ and $j >> i$ sub-domains) to long α-helix stretches capped by small misfolded loops ($i \approx j$). The former gradually evolve toward the native fold as the β-hairpin stretches are slowly unraveled, which gives rise to a severely misfolded population, whereas the latter need to surmount a (relatively low) free-energy barrier in order to proceed with the end segment zipping. Given the apparent roughness of the landscape, the interconversions between individual microstates are driven by unraveling of misfolded-structure contacts rather than conformational diffusion.

For the coil-to-helix transformation of A_{25} in aqueous solution, ensemble-convergent Monte Carlo simulations, starting with a single random nucleation site, were performed on a number of KIS landscapes as obtained for a range of temperatures. Throughout the simulations, the duration of the elementary temporal step was set equal to the characteristic helix propagation time, which can be obtained by multiplying the denominator in Equation 5.3 by $2^{(3 - 1)} = 4$. (We note that, in contrast to the α-helix nucleation mechanism which involves the alignment of three consecutive sets of backbone torsional angles, the elongation of a helical nucleus requires the alignment of a single set of such angles.) At $T = 278$ K, for which the ensemble-averaged fraction of α-helix content approaches 100% at equilibrium, the simulations predict a temporal helicity profile characterized by $\tau_1 = 44(5)$ ns and $\tau_2 = 700(50)$ ns (Figure 5.6, top right). The clear-cut biexponential behavior characteristic of A_{25} as obtained using the KIS model may be attributed to the bifurcating nature of the free energy landscape. Indeed, the faster (tens of nanoseconds) component which contributes ~30% of the signal amplitude appears to stem

from the rapid zipping of inner segments of the polypeptide chain, whereas the elongation of terminal helices which occurs on a much slower time scale, and accounts for ~70% of the signal amplitude, determines the rate-limiting step. Notably, the above picture is consistent with the findings made during the course of early experimental (42, 89, 303) and computational (MD) studies (304, 305). As shown in Figure 5.6, bottom right, the folding rates for A_{25} as obtained from the KIS model are in excellent agreement with the results of experimental measurements of polyalanine (un)folding kinetics. The effective barrier assessed using the (Arrhenius) dependence of folding rates on temperature, 7.9(2) kcal/mol, which exhibits a weak variation with the polypeptide chain length (data not shown) and remains consistent over a wide temperature range, is also in excellent agreement with an early experimental estimate of 8 kcal/mol (42). Therefore, the extra 4 kcal/mol barrier height, which was previously unaccounted for by the KZ model, is associated with the barrier to collective cleavage of (non-native, hairpin-forming) backbone–backbone hydrogen bonds that impede the coil-to-helix transition. The effect of misfolding barriers on α-helix formation can be seen in the dramatic dependence of folding time scales on the location of the seeding helical nucleus within the chain (Figure 5.7). Thus, the helicity profile representative of α-helix folding in A_{25} (Figure 5.7, left) is dominated by toward-the-ends, toward-the-center, and past-the-center zipping processes which take place on short, intermediate, and long time scales, respectively (Figure 5.7, right). The helical nuclei that form close to the center of the polypeptide chain tend to elongate (propagate) rapidly, whereas misfolded structures associated with nucleation sites located near the ends of the chain severely hinder the process of folding; hence the biexponential behavior evidenced in the results of Figure 5.6, top right. Because of the dominant role played by the interstate

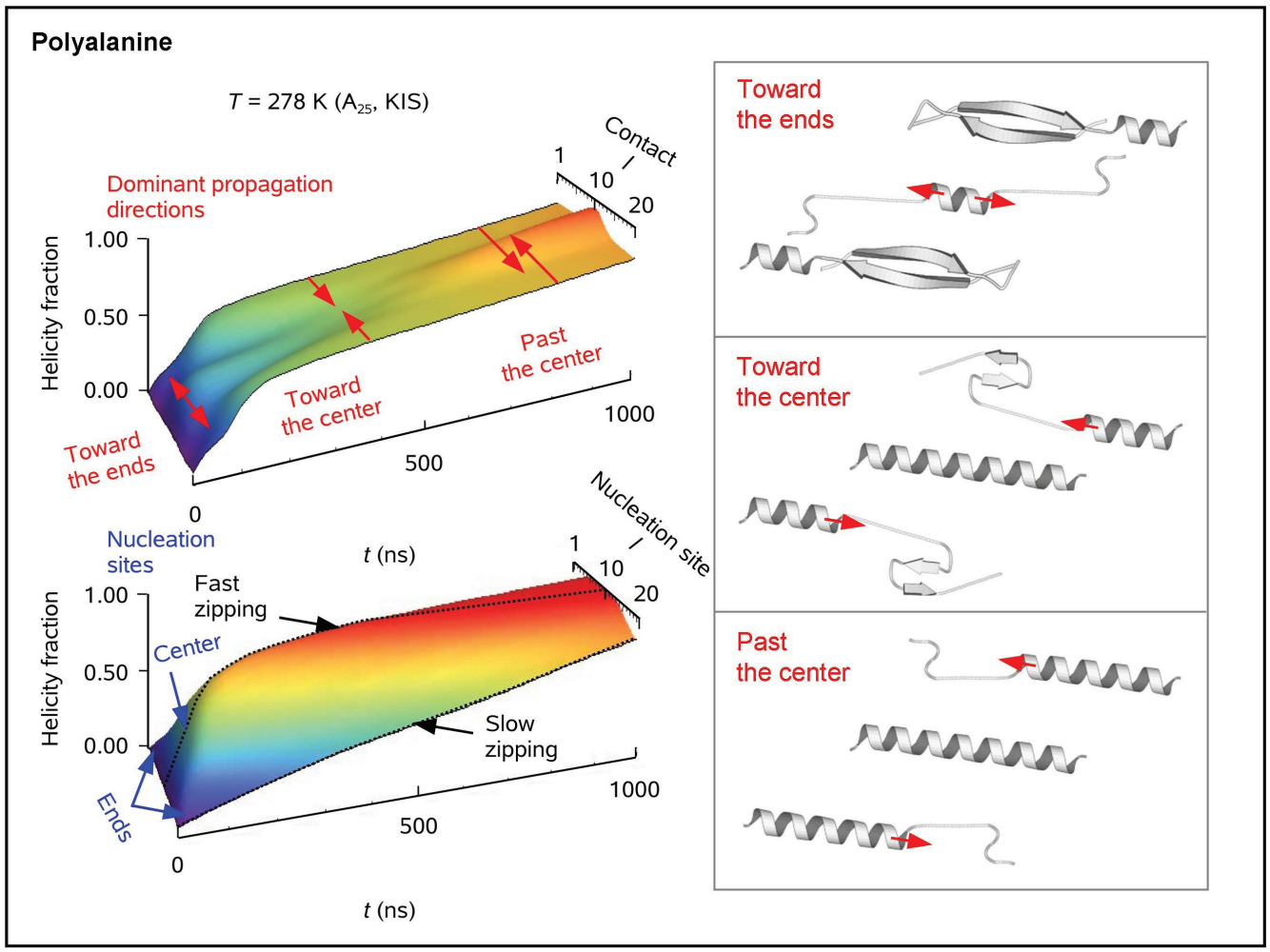

Fig. 5.7. Monte Carlo simulations on the KIS landscape. The process of α-helix formation is dominated by toward-the-ends, toward-the-center, and past-the-center zipping at short, intermediate, and long times, respectively (top left and right insets). Helical nuclei that form close to the middle of the sequence tend to elongate rapidly, whereas misfolded structures associated with nucleation sites located in the vicinity of the termini severely hinder the folding process (bottom left). This bifurcation in timescales is a direct consequence of the bifurcation of landscape pathways (see Figure 5.6).

barriers (free energy landscape roughness associated with formation of β-hairpin type structures) during the course of α-helix formation in A_{25}, the α-helix folding rates of heteropolypeptides are predicted by the KIS model to be very sensitive to the β-hairpin-forming propensities characteristic of the constituent amino acids (see below).

α-helix folding in heteropolypeptides and proteins. To elucidate implications of protein sequence heterogeneity for α-helix folding in biologically relevant systems, we examined formation of the two (major and minor) α-helical moieties of protein thymosin-β_9 ($T\beta_9$, PDB ID: 1HJ0 (113); Figure 5.8, top left) in aqueous solution with the aid of ensemble-convergent MD simulations. The MD setup, including the pre-heating step to initialize the unfolded state, was identical to that of A_{25}, except that 23848 TIP3P waters were used to solvated the bigger $T\beta_9$ structure, corresponding to box lengths of 89 Å at $T = 278$ K. The MD simulation window was 9 ns long.

Naively, one would expect A_{25} to fold somewhat faster than, e.g., the major α-helical sequence of $T\beta_9$ which is of the same length. However, in polyalanines, the (highest) helix-forming propensity characteristic of Ala residues competes with the formation of misfolded structures (β-hairpins) across the entire ensemble (141). The latter effect may be attributed to the (spatially compact) Ala side chains that leave the C, N, O atoms of the protein backbone exposed and, therefore, prone to hydrogen-bond formation. Paradoxically, larger side chains, regardless of their characteristic helix-forming propensities, appear to facilitate α-helix formation at early times by hindering misfolded structure formation. Indeed, the thicker side-chain "coating" of the protein backbone sequence limits β-hairpin nucleation across the ensemble of $T\beta_9$ (Figure 5.8, top right).

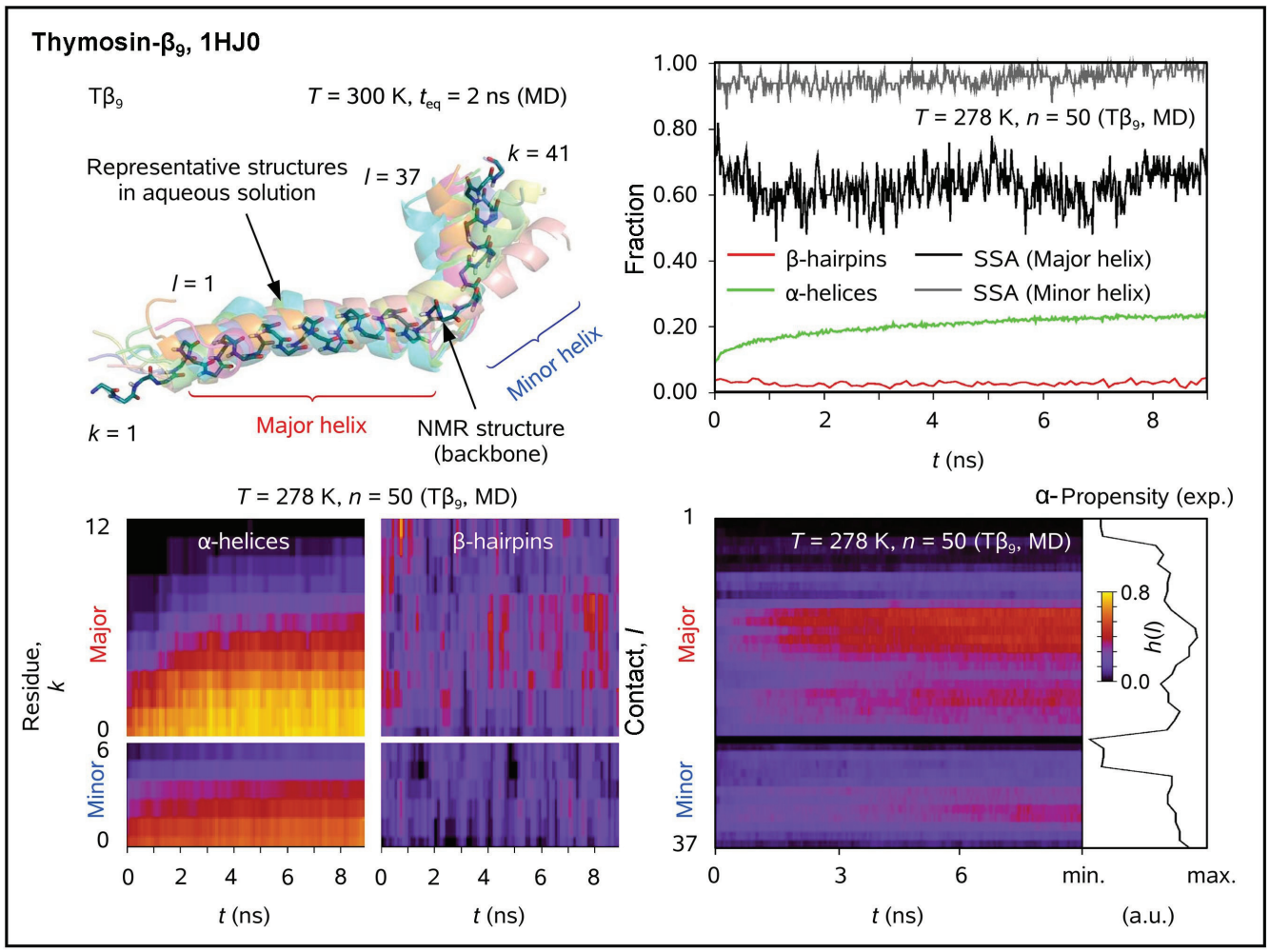

Fig. 5.8. Coil-to-helix transformation of thymosin-β_9 observed with molecular dynamics. Shown are a number of equilibrium structures (t_{eq} = 2 ns, cartoon) as obtained from our ensemble-convergent MD simulations in pure water superimposed on the experimentally determined structure (backbone skeleton) at T = 298 K. 41 residues of Tβ_9 can engage in up to 37 $l \rightarrow l + 4$ α-helical contacts. The ordered conformation of Tβ_9 includes two α-helical sequences, "major" and "minor", which extend from residues 4 and 31 to 23 and 37, respectively. The two helices are joined together by a disordered loop sequence (residues 28 to 30). Another loop is located at the N-terminus of the protein (top left). Distributed ensemble-convergent MD simulations carried out for n = 50 trajectories of Tβ_9 at T = 278 K indicate that the growth of α-helices (black 0%, yellow 100%) is minimally hindered by formation of β-hairpins (black 0%, yellow 20%), which may be due to steric repulsion of larger side chains that disfavor loop formation (bottom left). Note the disappearance of end-to-end misfolded structures prevalent in the kinetics of A_{25} (see Figure 5.5) (bottom left), as well as the faster helix formation, lower strand content compared with A_{25}, and the breakdown of the SSA for the major helical segment of Tβ_9 at early times (top right). The α-helix formation rates within the protein correspond well to experimental helix-forming propensities of the residues in the protein sequence (bottom right).

Nevertheless, non-helical contacts still exist, and they seem to accumulate at the boundaries separating ordered (helical) and disordered moieties of the protein (Figure 5.8, bottom left).

It is noteworthy that extended helical islands nucleating at the initial stages of folding of $T\beta_9$, as evident in the results of MD simulations, correspond to the areas of maximum experimentally-determined helix-forming propensities (Figure 5.8, bottom right). Given that the SSA breaks down for the major helical sequence of $T\beta_9$ right from the beginning of the folding process (Figure 5.8, top right), there must exist a statistically significant population of macromolecules that feature an internal loop (or "bubble") located between the two helical nuclei of the sequence at early times. We note that formation of the bubble can be attributed to a pronounced dip in the helix-forming propensity profile of $T\beta_9$ (Figure 5.8, bottom right). Notably, along with the structures characterized by internal bubbles, there also exist those containing single major helical nuclei as well as those containing single minor helical nuclei (or both). Validity of the SSA for the minor helical sequence of $T\beta_9$ may be attributed to its smaller size and more uniform helix forming propensity as determined by its sequence.

Summary. For proteins in general, an interplay between non-helical contacts (misfolded structures) and helix-forming propensities of individual residues is expected to determine time scales of helix formation, with the role of misfolded intermediates varying according to the protein sequence. However, for polyalanine, which is often considered to be the "hydrogen atom" of helix formation studies, the mechanism appears to be much clearer cut. For α-helix formation in A_{25}, the rate limiting step corresponds to the elongation (or "zipping") process, rather than that of helix nucleation. Contrary to the common wisdom

(223), α-helix formation is, in general, a three-step event: it requires formation of a helical island, unraveling of misfolded terminus (or termini) which often implies existence of long-lived kinetic intermediates, and, finally, zipping of the termini. The above behavior is dictated by the topology of the free energy landscape that guides protein (un)folding: for A_{25} it consists of three extensive domains occupied by kinetically trapped ($i << j$ and $i >> j$), misfolded ($i \approx j$, $i, j > 4$) and semi-native ($i, j \leq 4$) macromolecular structures (Figure 5.6, left). The apparent roughness of the landscape which hinders the diffusion search on its surface can be thought of as a maze that funnels unfolded structures into a variety of routes; the helical islands that form close to the middle of the sequence grow faster than those nucleating in the vicinity of its termini.

Interestingly, the above results are reminiscent of an early experimental study of a simple chemical reaction from this laboratory (306), in which two-body and three-body dissociation pathways originating from the same starting point on the potential energy landscape of HgI_2 were found to be associated with totally different time scales. Similarly to (inorganic) AB_2 molecules, fluxional biological macromolecules often undergo non-equilibrium transformations that are guided by *bifurcations* on the free energy landscape. Importantly, for α-helical segments that satisfy the SSA, these landscapes can be *comprehensively* parameterized on the KIS free energy landscape, with the key step in the coarse-graining analysis being the compaction of the numerous (non-native) degrees of structural freedom into the barriers of the (native-contact-specific) KIS landscape. For A_{25}, the folding pathways naturally emerge on the kinetic maze of the landscape, bundling themselves into two bifurcating ensembles separated by one order of magnitude in their

characteristic α-helix folding rates. Similarly to the oversimplified yet predictive valence-shell-electron-pair-repulsion (VSEPR) model of G. N. Lewis (307), the KIS model as applied to protein folding quantitatively explains both the kinetics and complex structural dynamics of α-helix formation from physical principles, a crucial function missing from experiments, whether done in the laboratory or on the computer.

5.2 TERTIARY STRUCTURE KINETICS AND THE LENGTH LIMIT

Ever since the discovery that proteins can spontaneously self-assemble into unique 3D shapes called *native folds* (308), the mechanism of this process has been a focus of biological research. It has been demonstrated that random protein sequences can attain unique ground states which are separated from other possible folds by significant energy gaps (309). However, assuming the existence of a unique native fold, there is no assurance that the protein can efficiently parse the fold space to find it. In particular, how nature is able to search the exponentially increasing number of folds accessible to proteins of non-trivial length has not been explicitly elucidated. Thus, Levinthal famously estimated that for a protein consisting of 150 amino acids, in which every degree of freedom is discretized into only ten possible values, it would take much longer than the age of the universe to sample all the folds even at the limit of molecular motion (6). In Section 5.1, we addressed the issue of folding kinetics of a (locally) periodic structure, the α-helix, which is a ubiquitous motif in proteins. The problem of the overwhelmingly large search space arises from the global and asymmetric nature of the tertiary structure. In what follows, we address the latter issue.

The *qualitative* resolution of the Levinthal paradox has been provided by the concept of the folding funnel, whereby a global bias in the (multi-dimensional) energy landscape channels the protein toward the sub-space comprising the native fold (102). Given that, by introducing an artificial search bias in favor of native contacts (310) or designing sequences favoring specific secondary structures (311), accelerated folding was computationally confirmed, it has long been a common belief that the funnel arises from evolutionary tuning of the intramolecular interactions via sequence mutation and that feasible protein folding times are the result of natural selection.

In contrast to the above picture, in the present Section we demonstrate that a fundamental effect, namely the hydrophobic force, is *sufficient* to account for (fast) protein folding kinetics without invoking any selection bias. The global tendency for hydrophobic residues to segregate in the interior of proteins has long been recognized as a major force that drives protein folding (100, 312). Indeed, it has been shown experimentally (313, 314) and computationally (129) that proteins undergo hydrophobic collapse in the earliest stages of folding. By considering the degeneracy of all folds, we demonstrate that hydrophobic collapse, and hydrophobic/hydrophilic residue segregation, lead to realistic folding time scales for globular proteins and protein domains, which are independently folding subunits that constitute larger proteins (315), thus *quantitatively* resolving the paradox. We also find an upper limit imposed by nature on the length of protein domains (~200 amino acids), for which such hydrophobic packing constraints would allow the native state to be identified within a biologically reasonable time scale through a hypothetical exhaustive search. By comparing the above result to the experimentally obtained distribution of protein domain lengths, we find that most protein structures do fall below this “hydrophobic length limit,”

although it can be exceeded due to the influence of other processes, besides hydrophobic collapse, that affect protein folding.

To date, many attempts have been made to estimate the reduction of the effective search space due to the hydrophobic force. For a self-avoiding chain (SAC) composed of L residues on a 3D cubic lattice, the number of unique conformational folds (degeneracy) was found to be $N_{\mathrm{SAC}} \sim L^{0.16}4.68^{L}$ (316). If we further restrict the chain to adopt maximally compact folds, as defined in the mean-field theory treatment (101), the resulting degeneracy may be expressed as $N_{\mathrm{Compact}} = (6/e)^{L}$, where e is the base of the natural logarithm. We note that, although compaction significantly reduces the search space, and is a driving factor for secondary structure formation (317), the degeneracy is still astronomically large even for the smallest proteins.

The crucial step is the (further) reduction of the conformational space by choosing only those compact folds with hydrophobic residues (H) maximally buried, in the sense of maximizing the number of H–H contacts, into the interior of the protein, and polar residues (P) on the outside. This minimalist HP representation of protein structures has long been a mainstay of analytic investigations of protein folding (66). We define $N_{\mathrm{HP}}(s)$ to be the degeneracy of self-avoiding compact folds with maximum H/P segregation, which is a function of s, the sequence of H and P residues along the chain. Because hydrophobic residues are randomly distributed along the protein sequence (318), the average conformational degeneracy of a collapsed H/P-segregated protein, $\langle N_{\mathrm{HP}} \rangle$, is equal to $N_{\mathrm{HP}}(s)$ averaged over *all* possible sequences s of length L, where the angular brackets denote averaging over the protein sequence space. Therefore, $\langle N_{\mathrm{HP}} \rangle$ represents the size of

the effective conformational space that may be sampled when a randomly-generated polypeptide sequence folds in a polar solvent like water.

The chief difficulty in obtaining $\langle N_{HP} \rangle$ is associated with the constraint that all monomers must be connected in a chain. This constraint had made $N_{HP}(s)$ impossible to analytically compute for any given sequence s, much less $\langle N_{HP} \rangle$ averaged over all values of s. Estimates that neglect the linear sequence of the chain, e.g. by assigning independent probabilities for each hydrophobic residue to be in the interior of the protein, overlook the crucial role of the above constraint in reducing the degeneracy. Thus, $\langle N_{HP} \rangle$ has been found to grow almost as quickly as $N_{Compact}$ as a function of L (66), an estimate nine orders of magnitude too large for $L = 100$ (see below).

Alternatively, computational efforts have been underway to explicitly account for chain connectivity. For 2D chains of $L < 28$, Camacho *et al.* computed $\langle N_{HP} \rangle$ by enumerating all such folds for all possible sequence permutations, assuming that half the residues were hydrophobic and half polar (319). On the 2D lattice, $\langle N_{HP} \rangle$ was found to increase on a (sub)logarithmic rate and plateau at $L \sim 10$. However, due to the exponentially growing number of both conformers and sequences with increasing L, the above approach becomes computationally intractable for longer sequences. Because the computational cost is even more severe for a 3D lattice, the protein conformational degeneracy in 3D space was extrapolated from the 2D lattice results (320). However, for a random sampling of HP sequences of $L = 48$, it was found via explicit calculations that $N_{HP}(s)$ ranges from "thousands to millions" on a cubic lattice (99), which is incompatible with the logarithmic growth rate expected for 2D lattices. In the following, we analytically

calculate $\langle N_{HP} \rangle$ in 3D space, taking into account, unlike the previous theoretical work, the fact that the individual residues on the 3D lattice must be interconnected to form a chain. Importantly, in the present study, the protein chain length L is the only parameter of the solution obtained.

Lattice Model. We define a connectivity map to be a spatial arrangement of (interacting) H and P residues (Figure 5.9). Within the framework of the HP model, each H–H contact in the map decreases the total energy by a fixed amount, whereas all other contacts do not contribute energetically. All calculations presented below ignore pre-factors of order unity. For a particular sequence s of length L, $N_{HP}(s)$ technically describes the number of conformations of sequence s that attain the lowest-energy spatial arrangement (i.e., the "optimal map") for that sequence. We note that this may not be the same as the optimal map for *all* sequences of length L, which is termed the "global optimal map". Further, we define $N^*_{HP}(s)$ as the number of conformations that attain the global optimal map. If the protein sequence s can attain the global optimal map, then $N^*_{HP}(s) = N_{HP}(s)$, otherwise $N^*_{HP}(s) = 0$. Hence, $\langle N^*_{HP} \rangle$ is a lower bound imposed on $\langle N_{HP} \rangle$. If $N^*_{HP}(s) = 0$, then $N_{HP}(s)$ is bounded by the probability of the event that none of the optimal folds of the sequence s can be locally perturbed to attain lower energy maps. Consequently, on a 3D lattice, the sequences that cannot attain the global optimal map do not contribute significantly to $\langle N_{HP} \rangle$, hence $\langle N_{HP} \rangle \approx \langle N^*_{HP} \rangle$ (see details below and Figure 5.10).

To calculate $\langle N^*_{HP} \rangle$, consider the sequence vs. fold space representative of an L-residue-long protein (Figure 5.9). Each column in the Figure corresponds to a unique protein sequence, and each row corresponds to a unique compact conformation. Because

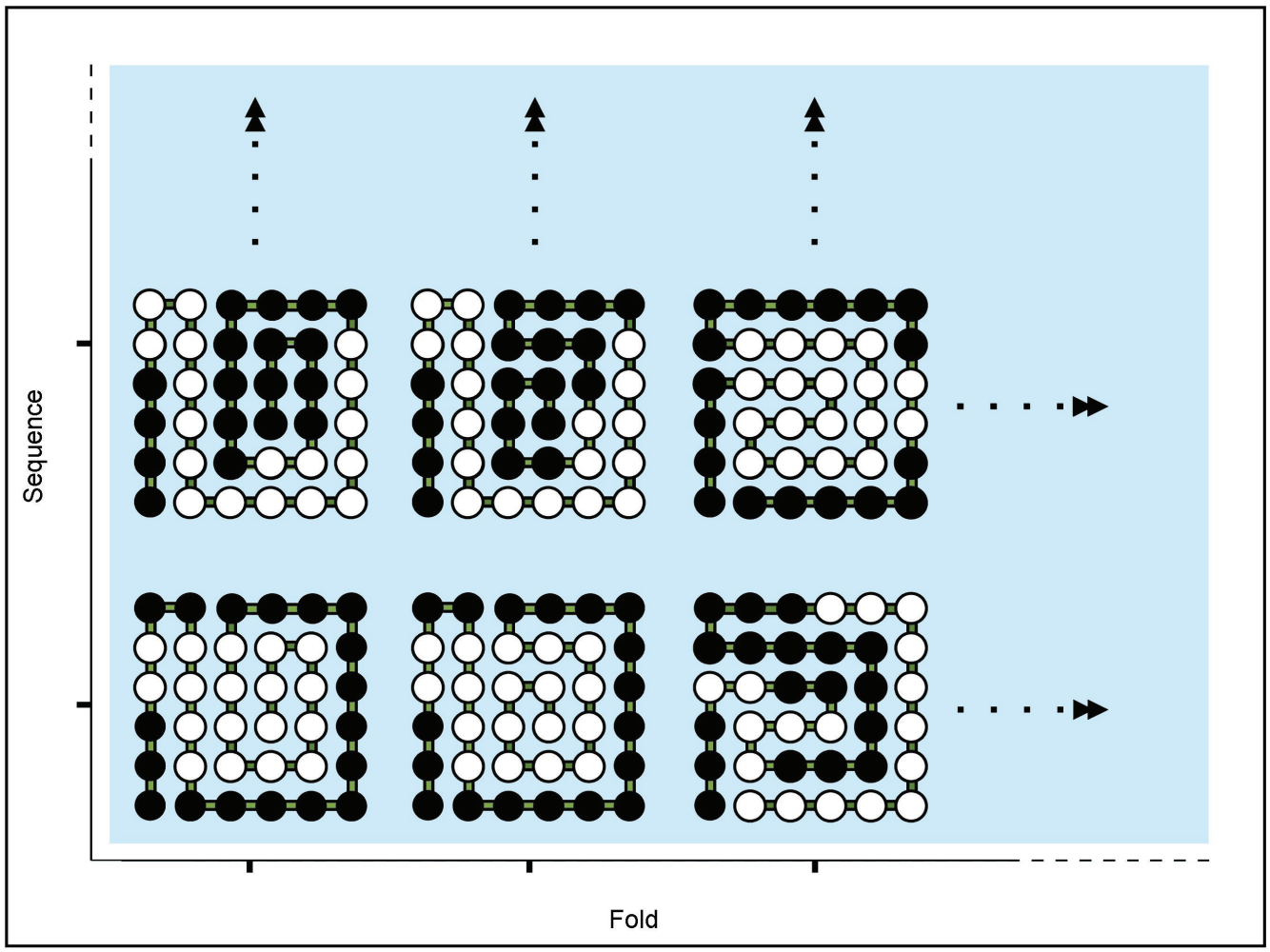

Fig. 5.9. Schematic table of folds versus sequences. White and black circles denote hydrophobic (H) and polar (P) residues, respectively. Each pattern of H and P circles constitutes a map. For any given map, a fixed sequence (e.g. row 2) can have multiple (or no) conformation folds which produce the same map (e.g. columns 1 and 2). On the other hand, each fold (column) has one and only one sequence (row) that can produce a given map.

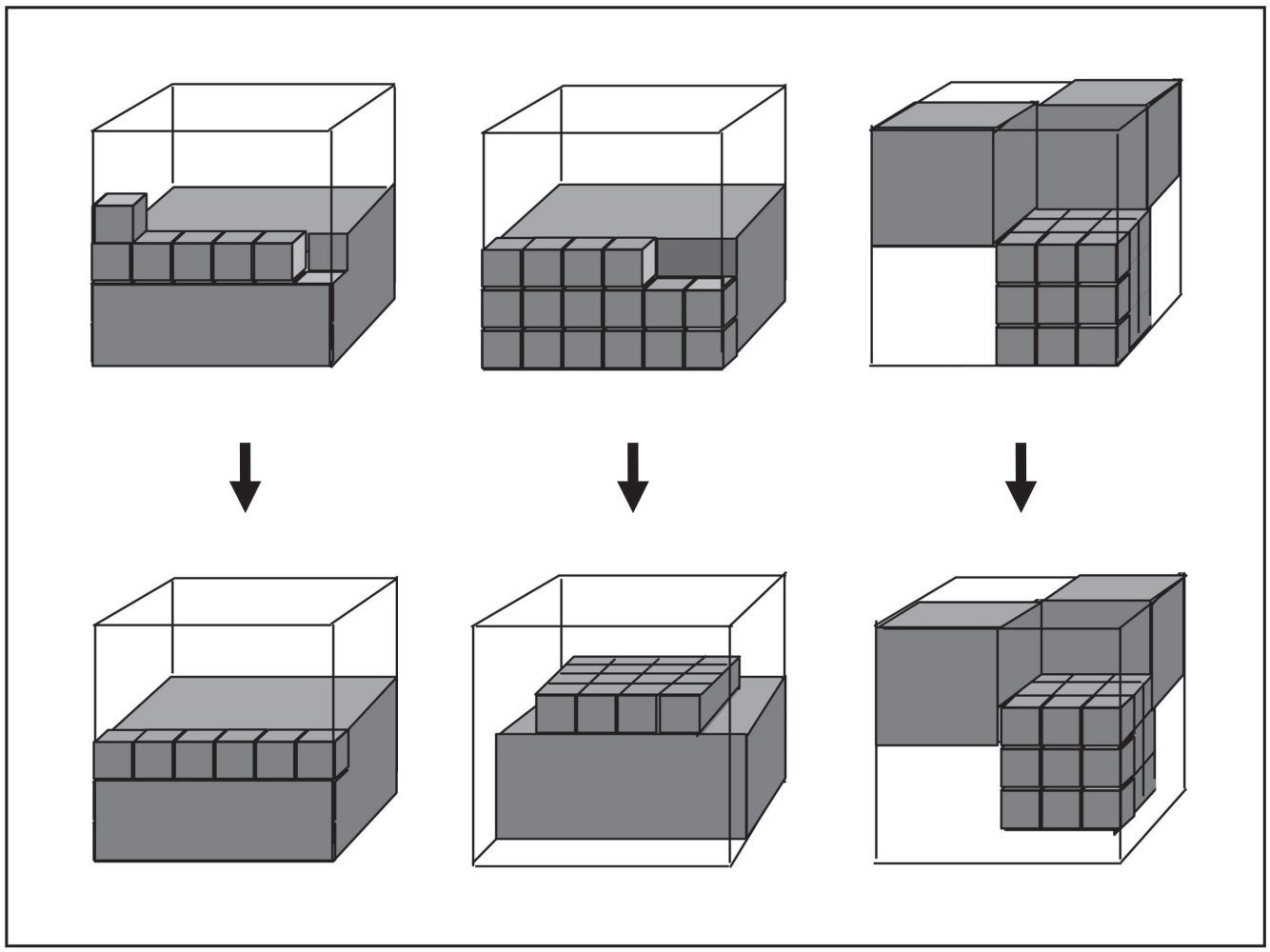

Fig. 5.10. Minimal sub-volume required for H-P swapping leading to lower energy. The schematic illustrates the three types of swapping for a cubic lattice of length 6. The hydrophobic residues (H) are shown in gray and the hydrophilic residues (P) are transparent for clarity purposes. The individual gray cubes denote the hydrophobic residues selected to be in the minimal sub-volume *n*. The three types are (top) swapping on one face, (center) swapping involving multiple faces, and (bottom) merging distinct hydrophobic regions. Note that the schematic illustrates extreme cases; most configurations require smaller distortions.

proteins are, on average, characterized by equal numbers of hydrophobic and polar residues (321), there exist $\binom{L}{L/2}$ columns and $(6/e)^L$ rows (cf. N_{Compact}). For clarity, shown in Figure 5.9 is the space representative of a protein chain of length $L = 36$ on a 2D lattice, although the concept is identical in 3D. Thus,

$$\left\langle N_{\text{HP}}^{*} \right\rangle = L^{2/3} \binom{L}{L/2}^{-1} \sum_{s=1}^{\binom{L}{L/2}} d(s), \qquad [5.8]$$

where $d(s)$ is the number of conformations for which the chosen sequence s can attain the global optimal map. We note that the pre-factor $L^{2/3}$ in Equation 5.8 accounts for the number of distinct global optimal maps due to (possible) hydrophobic pockets located along the H-surface which can accommodate leftover H residues (unless the number of H residues is a perfect cube). Importantly, all relevant constraints, including the chain connectivity, are captured by $d(s)$. Because the degeneracy calculated by summing over all columns of the global optimal map is equivalent to that calculated by summing over all rows of the map, we obtain:

$$\left\langle N_{\text{HP}}^{*} \right\rangle = L^{2/3} \binom{L}{L/2}^{-1} \sum_{f=1}^{(6/e)^L} D(f), \qquad [5.9]$$

where $D(f)$ is the (sequence) degeneracy of the conformation f. For any map, a given sequence can attain the map with multiple (or none) of its conformations; on the contrary, any conformation can attain any map with exactly one sequence: $D(f) = 1$ for all f. Therefore, Equation 5.9 transforms into: $\left\langle N_{HP}^{*} \right\rangle = L^{2/3} \binom{L}{L/2}^{-1} (6/e)^L$, and, applying Stirling's

approximation, we arrive at the following main results. The degeneracy can be calculated as:

$$\left\langle N_{\mathrm{HP}}^{*} \right\rangle = L^{7/6}\left(\frac{3}{e}\right)^{L} \approx L\left(\frac{3}{e}\right)^{L}, \qquad [5.10]$$

and the protein folding time associated with exhaustive conformational search is given by:

$$\tau_{\mathrm{folding}} = \left\langle N_{\mathrm{HP}}^{*} \right\rangle \cdot \tau_{\mathrm{sampling}}, \qquad [5.11]$$

where τ_{sampling} is the time characteristic of an elementary conformational interconversion. We note that the above results were obtained for a 3D lattice. In 2D space, the $(3/e)^{L}$ term of Equation 5.10 transforms into $(2/e)^{L}$, which is an exponentially decreasing function. In the 2D case, longer chains are characterized by lower probabilities of folding into the global optimal map; as L increases, the chain continuously "runs out" of ways to fold into the optimal map so that the maps of higher energies become optimal. This turnover of the optimal map causes $\langle N_{\mathrm{HP}} \rangle$, the optimal map degeneracy, to plateau as a function of L, in agreement with the explicit 2D calculations (319).

Importantly, the results obtained using Equation 5.10 are also in agreement with 3D lattice simulations. Consistent with the "thousands to millions" of degenerate optimal conformations estimated from an explicit enumeration, $\langle N_{\mathrm{HP}} \rangle$ approaches 10^4 for a chain of $L = 48$. At a higher level of chemical accuracy, for a lattice chain of $L = 80$ that consists of the 20 known types of biological amino acids and possesses a unique native fold, it was

found that at most 10^6 Monte Carlo steps were required to attain the native-fold structure (322). This is in agreement with the estimate of Equation 5.10: $\langle N_{HP} \rangle = 5 \times 10^5$ for $L = 80$.

Through multiplication by the experimentally determined $\tau_{sampling} = 10$ ns (323), Equation 5.11 relates the degeneracy to the protein folding time $\tau_{folding}$. Presented in Figure 5.11, top, as functions of L in 3D are the folding (or conformational-space sampling) times of the self-avoiding chain, compact self-avoiding chain, and the (average) collapsed H/P-segregated chain with characteristic degeneracies of N_{SAC}, $N_{Compact}$, and $\langle N_{HP} \rangle \approx \langle N^*_{HP} \rangle$, respectively. We note that, in accordance with Levinthal's findings, in the cases of N_{SAC} and $N_{Compact}$ the number of available conformations becomes astronomically large even for very short chains. Nevertheless, if confined to $\langle N_{HP} \rangle$ by the hydrophobic force, the exhaustive search of the same sub-space can be accomplished on biological time scales (nanoseconds to minutes) for $L < 200$. Because $\langle N_{HP} \rangle$ increases exponentially with L, proteins cannot complete an exhaustive search of the hydrophobic sub-space beyond the above length; this is, therefore, the exhaustive hydrophobic search length limit imposed by nature on protein domains.

Also shown in Figure 5.11, top, are the experimentally measured folding times for 65 single-domain proteins (106, 324). For $L < 100$, the folding times agree with the exhaustive sequence-averaged folding time $\tau_{folding}$, with the variance arising from the particular protein sequence. For $L > 100$, the average folding time falls below $\tau_{folding}$, which is indicative of the onset of other factors such as, e.g., evolutionary sequence selection favoring faster kinetics, despite the overall folding time scale for $L < 200$ being dominated by the H/P collapse.

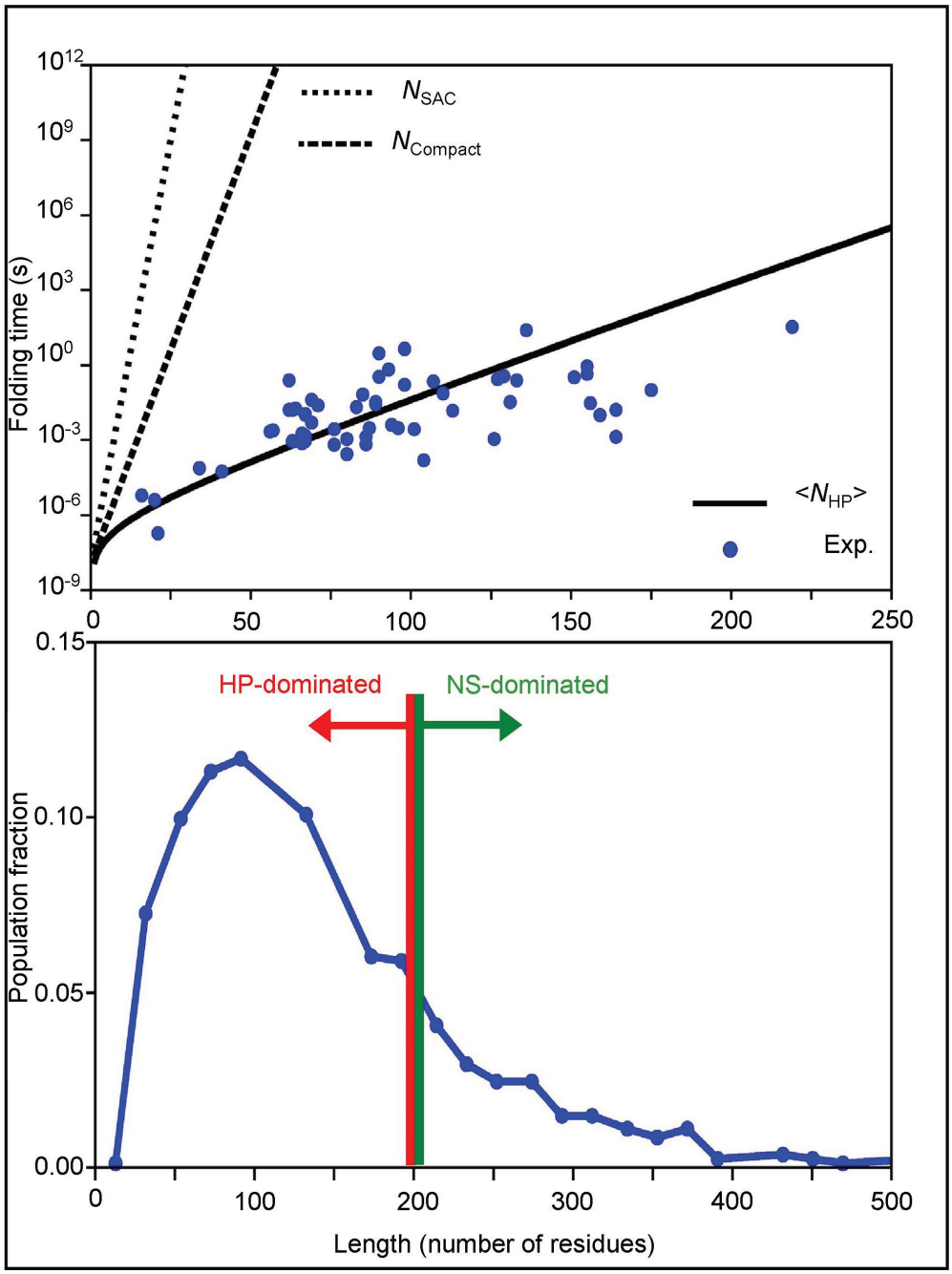

Fig. 5.11. Length and time limits of folding. Degeneracy of conformations on a cubic lattice is plotted as a function of chain length (top). Conformational degeneracies of self-avoiding chain (SAC), self-avoiding compact chain, and the sequence-averaged lowest energy HP chain are shown with dotted, dashed, and solid lines, respectively. The degeneracies are multiplied by the 10 nanosecond residue reorganization time to obtain the folding times (Equation 5.11). The predicted limit (above which exhaustive search cannot lead to sufficiently fast folding), is $L \sim 200$ amino acids. Experimentally measured folding times are also shown, indicating that faster-than-exhaustive-search folding occurs for $L > 100$. The experimental domain length distribution of a representative set of 1236 proteins shows that for $L > 100$, the population fraction begins to decay (bottom). The protein populations are divided into the HP-dominated regime for L below the exhaustive search length limit (red arrow), and the natural selection (NS)-dominated regime for L above the length limit (green arrow). All experimental data are shown in blue. See Text for further descriptions of the regimes.

Although proteins often consist of more than 1000 amino acids, protein domains are on average 100 amino acids long, typically ranging from 50 to 200 (26, 325), with 90% of them being shorter than 200 (326). The above observation is in agreement with what we have identified as the domain length regime for which the hydrophobic–polar interactions are sufficient to ensure the fast (HP-dominated) folding. Beyond the above regime, accelerated protein folding rates are facilitated by mechanisms other than the hydrophobic force such as, e.g., sequences characterized by smaller degeneracies than that typical of the average sequence, evolutionary funneled free energy landscapes, periodic local structures arising from repeated insertion mutations (327), and molecular chaperones assisting in folding (328). These types of auxiliary mechanisms allow for the existence of domains that exceed the hydrophobic length limit; the fast folding of proteins in this second regime is therefore consistent with the effect of *natural selection* (NS-dominated). Shown in Figure 5.11, bottom, is the domain length distribution of a representative sample of 1236 proteins (321) and its partitioning into the two regimes. Consistent with the data of Figure 5.11, top, the abundance of proteins of length L begins to decrease near $L = 100$, which is in agreement with the onset of evolutionary pressure to select for sequences that fold faster than the exhaustive search at this length scale would allow. However, because the folding degeneracy increases exponentially with L, the fraction of protein domains exceeding the (L= 200) hydrophobic length limit is small. We note that the fundamental length limit pertinent to protein folding may have forced most proteins with $L > 200$ to evolve as modular combinations of smaller domains.

Lattice model details. In the following we justify a key step made during the course of the above analytic derivation, namely: $\langle N_{HP} \rangle \approx \langle N^*_{HP} \rangle$. We denote a sequence “optimal” if it

can attain the global optimal map, and "sub-optimal" otherwise. Further, we define p to be the fraction of sequences that can only fold into sub-optimal maps, $\langle N_s \rangle$ to be the average ground state degeneracy over all such sub-optimal sequences, and $\langle N_o \rangle$ to be the average ground state degeneracy over all optimal sequences. To prove that $\langle N_{HP} \rangle \approx \langle N^*_{HP} \rangle$, we note that $\langle N_{HP} \rangle = p\langle N_s \rangle + (1 - p)\langle N_o \rangle = p\langle N_s \rangle + \langle N^*_{HP} \rangle$; it is therefore sufficient to show that $p\langle N_s \rangle < \langle N^*_{HP} \rangle$. To this end, note that if a sequence is sub-optimal, there exist lower energy states that it cannot attain by locally perturbing any of its ground state conformations. Being a globally sub-optimal sequence, each ground state fold f of the sequence s contains a minimal-sized sub-volume $n(f, s)$ of the lattice in which the number of H–H contacts can be increased (and thus the energy decreased) by changing the positions of H and P residues to form a new map with lower energy in the sub-volume. For each ground state fold f, assuming that there exists at least one other fold besides the starting fold which preserves the chain connectivity to the outside of the sub-volume, we define P_f to be the probability of the event that at least one such fold can attain a lower energy. Then, the probability of finding that each residue of $n(f, s)$ matches that of the lower-energy map is $(1/2)^{n(f, s)}$. Therefore, $P_f > (1/2)^{n(f, s)}$, where the greater-than sign is due to the possibility of attaining multiple conformations, multiple lower energy maps, and unequal numbers of H and P residues within the sub-volume which can all increase the latter probability. Then, the probability that none of the ground state folds can be locally perturbed to achieve a lower energy is given by:

$$\prod_{i=1}^{N_s}(1-P_f) \approx 1-\sum_{f=1}^{N_s}P_i < 1-\sum_{f=1}^{N_s}(1/2)^{n(f,s)} < 1-\langle N_s \rangle (1/2)^{\langle n \rangle}, \qquad [5.12]$$

where $\langle n \rangle$ is the average size of the minimal volume over all sequences and over all ground state folds of each sequence. The approximation in Equation 5.12 is justified because $P_f N_S < 1/2$ for the locally optimal fold. Since probabilities must be non-negative, we obtain from Equation 5.12 that $\langle N_s \rangle < 2^{\langle n \rangle}$.

An estimate for $\langle n \rangle$ can be obtained by noting that there exist three generic types of H–P swapping to achieve lower energy. First, we consider the (most likely) case in which at least one extra H–H contact can be created by rearranging one surface of the hydrophobic region. The volume $n(f, s)$ is therefore associated with a path on the H-surface such that the H residue(s) at one end of the path swap with the P residue(s) at the other end; the volume is thus at most twice the length of the cubic lattice: $n(f, s) < 2L^{1/3}$ (Figure 5.10, left) for any given fold f and sequence s.

The second case to consider implies that the rearrangement may require multiple H residues on one face of an H-region to swap with P residues on a different face. In this case (Figure 5.10, center) the maximum n is limited to the area of a face: $n(f, s) < L^{2/3}$ for any given fold f and sequence s.

Finally, we consider the case which requires two separate H-regions to connect. In this case, all H-regions must be cubic, otherwise the above case 1 or case 2 would apply. The maximum $n(f, s)$ corresponds to the case in which the lattice is divided into a 3D checkerboard with each H-region being a cube of sides at most $L^{1/3}/2$. Thus, the maximum $n(f, s)$ corresponds to that of this cube plus one face of the adjacent cube so that the cube may be shifted by one notch and thereby make contact with another cube. Therefore, we

obtain $n(f, s) < (L^{1/3}/2)^2 \cdot (L^{1/3}/2+1) = L/8 + L^{2/3}/4$ (Figure 5.10, right) for any given fold f and sequence s.

Importantly, in all three cases, $2^{n(f, s)} < \langle N^*_{HP} \rangle$ for any given fold f and sequence s. Because the above inequality is true for any f and s, it should also hold when $n(f, s)$ is averaged over all f and s: $2^{\langle n \rangle} < \langle N^*_{HP} \rangle$. For example, for a protein chain of length L = 200, we obtain $\langle N^*_{HP} \rangle \sim 10^{12}$, whereas $2^n < 10^5$, 10^{10} , and 10^{10} for (the worst-case scenarios of) the three cases, respectively. Therefore, $p\langle N_s \rangle < 2^{\langle n \rangle} < \langle N^*_{HP} \rangle$, and thus $\langle N_{HP} \rangle \approx \langle N^*_{HP} \rangle$.

Computational. To complement the general, yet coarse grained, results above, we also performed ensemble-convergent MD simulations using the CHARMM suite of programs and force field (118) on a single polypeptide to gain insight at the atomistic level. For these studies, we chose the 20-residue Trp-cage mini-protein (115), which despite its small size possesses both secondary and tertiary structure, the latter being represented, e.g., by the burial of the large hydrophobic tryptophan side-chain in the protein interior. Computational details pertinent to these simulations have been provided in Section 4.2.4.

Presented in Figure 5.12 are the free energy landscapes of Trp-cage as obtained both in vacuo (top) and in solution (bottom) as functions of the RMSD calculated with respect to the original (experimentally determined) macromolecular structure; also shown in the Figure are the solvent-accessible surface areas of the hydrophobic residues (exposed H-areas) in both environments. The free energy landscapes are binned into 100 × 100 grids of distinct microstates, as defined by these two order parameters, and the corresponding

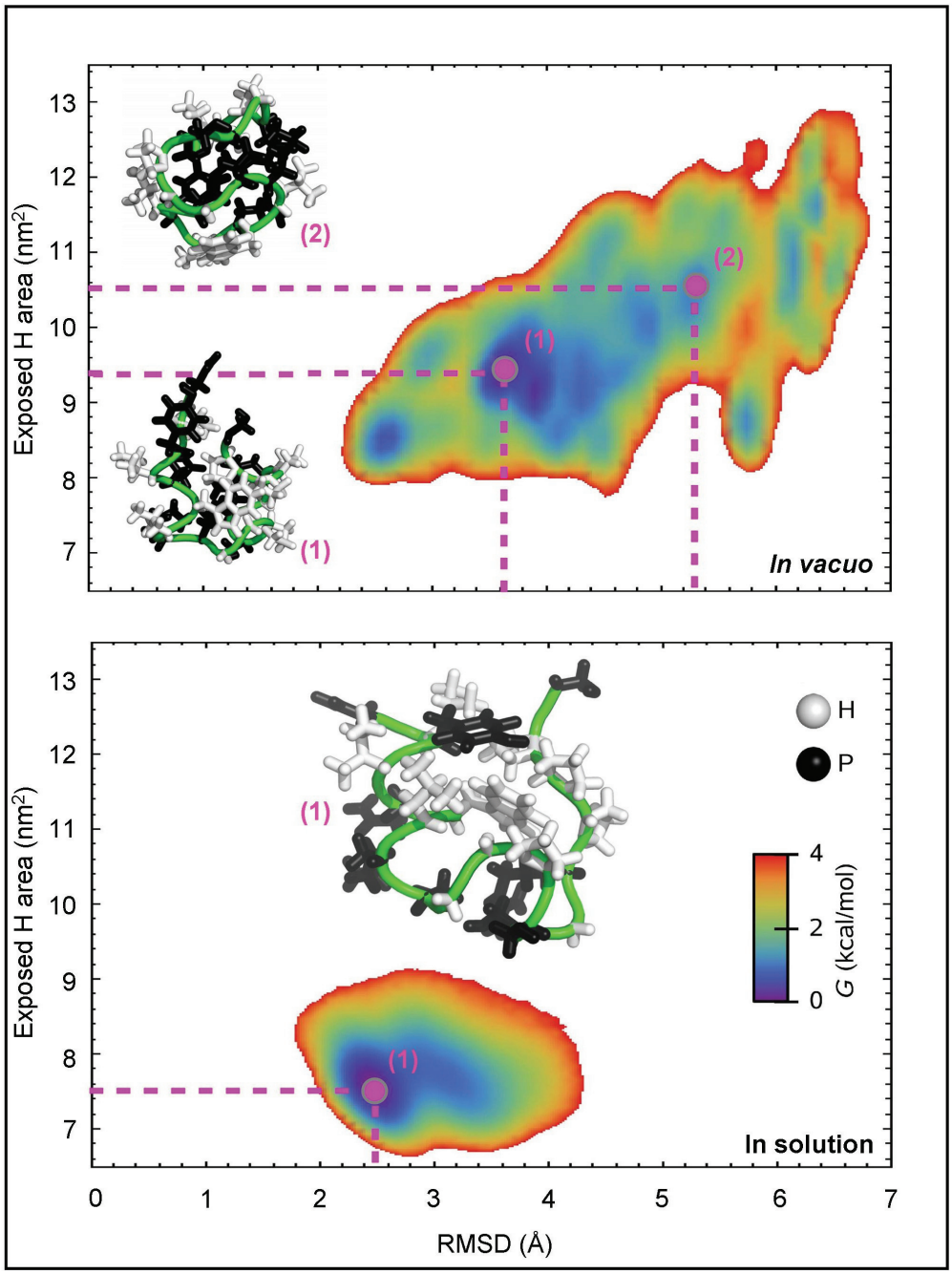

Fig. 5.12. Free energy landscapes of trp-cage in the presence and absence of water from MD simulations. White and black denote hydrophobic (H), and polar (P) residues respectively. The two order parameters are root-mean-squared deviation (*rmsd*) from the experimentally determined starting structure, and the solvent accessible surface area of the hydrophobic side chains (exposed H area). When solvated (bottom), the landscape is restricted to the funnel containing the native state; in vacuum (top), there are a multitude of minima, all with similar free energies, that are no longer constrained to minimize the exposed H area. Some representative structures are also shown, including an “inside-out” conformation sampled during the vacuum simulations.

free energies are calculated as $\Delta G_i = kT\ln[P(i)]$, where $P(i)$ is the fraction of total simulation time spent in microstate i. We note that, in water, the peptide is confined to a free energy basin with burial of hydrophobic residues, including tryptophan (small exposed H-area) and a macromolecular-structure ensemble similar to that found experimentally (low RMSD). In the absence of water, on the other hand, the hydrophobic residues are more solvent-exposed and there exist multiple conformational basins distributed throughout the entire free energy landscape. Importantly, a number of minima thus obtained correspond to predicted "inside out" conformational ensembles (329), in which the hydrophobic residues are on the outside and the polar residues are buried. In the gas phase, the polypeptide no longer spends the majority of its time in the lowest free energy basin. Crucially, in accordance with the findings made using the above lattice model, the conformational space available for sampling is greatly reduced in aqueous solution because the peptide is restricted to folds with buried hydrophobic residues.

Because Trp-cage is unique, with its hydrophobic "core" primarily consisting of a single residue, care must be made when extrapolating specific dynamical behaviors to proteins in general. However, as a minimal-size peptide with tertiary structure for which comprehensive and statistically significant information can be obtained with atomic resolution, Trp-cage has been used to confirm that the results obtained from the lattice model do extend to the physical world.

Summary. In the present Section, we addressed the apparent paradox of overwhelming fold degeneracy in protein folding, a problem that is analogous to the question of how proteins (and therefore genes) evolve by natural selection within the immense space of possible

sequences (330). Smith argued that the latter paradox vanishes if incremental evolutionary steps confine the protein sequence within the exponentially smaller sequence sub-space that corresponds to good fitness (331). Just like evolutionary fitness is a global order parameter that keeps the sequence search within the fruitful sub-space of all sequences, hydrophobic–hydrophobic contact area is a global order parameter that keeps the fold search within the fruitful sub-space of folds. Here, we *quantitatively* demonstrated that the size of the hydrophobically collapsed sub-space is indeed small enough to be realistically sampled during the course of protein folding.

The coarse grained lattice model has been used to derive the hydrophobic length limit imposed on protein domains ($L < 200$). Below this limit, proteins could in principle randomly sample the entire folding sub-space consistent with hydrophobic collapse and hydrophobic/hydrophilic segregation. Above this limit, the hydrophobically-constrained fold space increases exponentially beyond what is accessible by random search. Consequently, the evolution of larger proteins is consistent with the model of modular growth, involving the aggregation, swapping, and duplication of stable domains (332). In this latter regime, natural selection may be necessary to enhance the folding rate using sequence-specificity and/or chaperones. The all-atom ensemble-convergent MD simulations explicitly demonstrated the role of the hydrophobic effect in drastically reducing the search space on the free energy landscape.

In addition to providing a mechanistic insight into the role of physical forces in shaping the length-scale and evolution of proteins, the results presented here may be useful in protein characterization and engineering. For example, a useful metric that quantifies the

effect of protein sequence on folding rate is the ratio of a protein's folding time to $\tau_{folding}$, the exhaustive search time of an average HP chain of the same length. For de novo protein design, we predict that for $L < 200$ it is not necessary to engineer a kinetic pathway which leads to the desired native state; as long as the native state is thermodynamically stable and the roughness of the conformational free energy landscape is sufficiently low, the protein will fold in a reasonable time.

Chapter 6

CONCLUSIONS

The first part of this Thesis, covered in Chapter 4, aimed at elucidating the general issues of macromolecular dynamics, which governs a wide range of biologically-relevant phenomena, including protein folding. Such issues as the dynamics at different length scales, the role of water, the effect of temperature, and the nature of the unfolded ensemble, are all critical pieces of the puzzle of macromolecular folding and function. It was found that even in the unfolded state, macromolecules can possess a significant geometric bias, as measured by the persistence length, due mainly to the cumulative effect of the (weak) backbone torsional potentials rather than excluded volume effects. For the order–disorder transition of key biological motifs, the identity and nature of kinetic unfolding intermediates were found to be predictable using experimentally measured thermodynamic parameters of the constituent interactions. The unfolding dynamics of both proteins and nucleic acids were found to occur at a variety of length scales; depending on the temperature, local structural changes were demonstrated to take place either before or after global dynamics. Furthermore, the unfolding of local secondary structure was shown to involve surmounting a "hidden" energy barrier due to persistent structural rigidity which, in the case of α-helices, was much higher than the bond breaking barrier. Throughout, it was found that the solvent has profound effects on the kinetic behavior, such as the "unfreezing" of large-scale conformational motions. The ensemble-averaged radial distribution function not only proved to be a powerful coarse-graining methodology, but

also suggested the possibility of observing the (un)folding dynamics of gas phase biopolymers via the structural resonance in prospective ultrafast electron diffraction experiments.

In light of the general macromolecular motions encountered in Chapter 4, the second part of this Thesis, covered in Chapter 5, is concerned with understanding the protein folding process and its rate. It is found that different mechanisms dominate the folding at different length scales ranging from the local dynamical steps of secondary structure to the global tertiary structure. At all length scales, we have derived folding solutions based on the fundamental underlying forces. The most basic initial step of folding, α-helix nucleation, is rate-limited by *cooperative* Brownian diffusion of multiple degrees of freedom occurring in a few nanoseconds, and is the **speed limit** of protein structure formation. This folding time was independently obtained from a novel Langevin dynamics model, ensemble-convergent all-atom simulations, as well as the first experimental studies to isolate helix nucleation with the necessary ultrafast temporal resolution. The propagation (growth) of the helical nucleus to form extended α-helices was found, in some cases, to be rate-limited by the need to break non-native contacts such as β-hairpins that impede helix growth, and therefore explained why measured rates were an order of magnitude slower than predicted by conformational diffusion. Therefore, although α-helix, β-strand, and tertiary structure formation are completed at different time scales, all three processes are nevertheless dynamically coupled; contrary to the prevailing paradigm of the kinetic zipping mechanism, the formation of structures consisting solely of local contacts can nevertheless be dominated by the interference of long-range contacts. The kinetic intermediate structure model provides a general way to map the relevant local and

long-range interactions onto a free energy landscape that describes the kinetics of such systems.

As for tertiary structure, during the course of finding the native conformation (fold), proteins sample a structural sub-space that is typically hundreds of orders of magnitude smaller than their full conformational space. At this length scale, the overwhelming fold degeneracy is the rate-limiting factor. Whether proteins' ability to quickly parse this vast space of possible folds is due to the intrinsic physical constraints or the result of natural selection, and what is the largest foldable protein, were open questions. We derived the sequence-averaged degeneracy of a lattice polymer in which the hydrophobic residues are maximally clustered in the center as a result of the hydrophobic force. The exhaustive search time is therefore proportional to this degeneracy. From the analytical solutions, we identify two regimes. The first, corresponding to protein length $L \leq 200$ amino acids, is the regime for which the hydrophobic force allows for an exhaustive conformational search in a biologically feasible time scale. This regime dictates the experimentally observed length and time scales of protein folding, as well as explains why larger proteins are modularly constructed, and accounts for the fast folding rates of most single domain proteins. The second regime is characteristic of $L > 200$, for which natural selection must be involved to enable sufficiently fast folding. Ensemble-convergent MD simulations of the Trp-cage mini-protein confirm this picture on the free energy landscape. The finding that the hydrophobic force is, by itself, sufficient to enable the folding of most single domain proteins, as well as the experimental fact that few protein domains exceed the predicted hydrophobic **length limit** of $L \sim 200$, advances our understanding of the protein folding process at a fundamental level.

The above results, which combine to form a predictive theoretical framework of protein folding kinetics spanning the range of relevant lengthscales for the folding of single-domain proteins, are summarized in Figure 6.1. Helix nucleation occurs via the cooperative diffusive mechanism, helix formation is hindered by misfolding, and tertiary structure formation is dominated by the conformational degeneracy (entropy). The theories presented, which model these phenomena, correctly predict the folding rates and other experimental observables, and are parameterized by well-defined experimentally determined constants with no adjustable parameters. This framework clarifies a few longstanding issues of protein folding such as widely varying helix folding rates, the quantitative resolution of the Levinthal paradox, and the separation of the role of intrinsic physical forces from that of fine-tuning, for example by natural selection. We find that even for highly heterogeneous and complex biological systems, physics exerts predictable limits on speed and size.

The line of research reported in this Thesis naturally suggests the investigation of new issues. These include (i) the physical properties intrinsic to β-strand formation (ii) the role of protein sequence in determining the folding pathways within the conformational subspace dictated by the hydrophobic force, whose size was found in Section 5.2, and (iii) the influence of solvent structure on protein folding dynamics beyond the capacity of water to reduce the roughness of the free energy landscape, as reported in Section 4.2.4. In regards to the last issue, throughout this work, both for protein folding in particular and macromolecular dynamics in general, the striking sensitivity of the structural dynamics and folding feasibility to solvation state (and temperature) indicates that the full role of water in biological function has yet to be elucidated, and serves as a reminder that Earth possesses

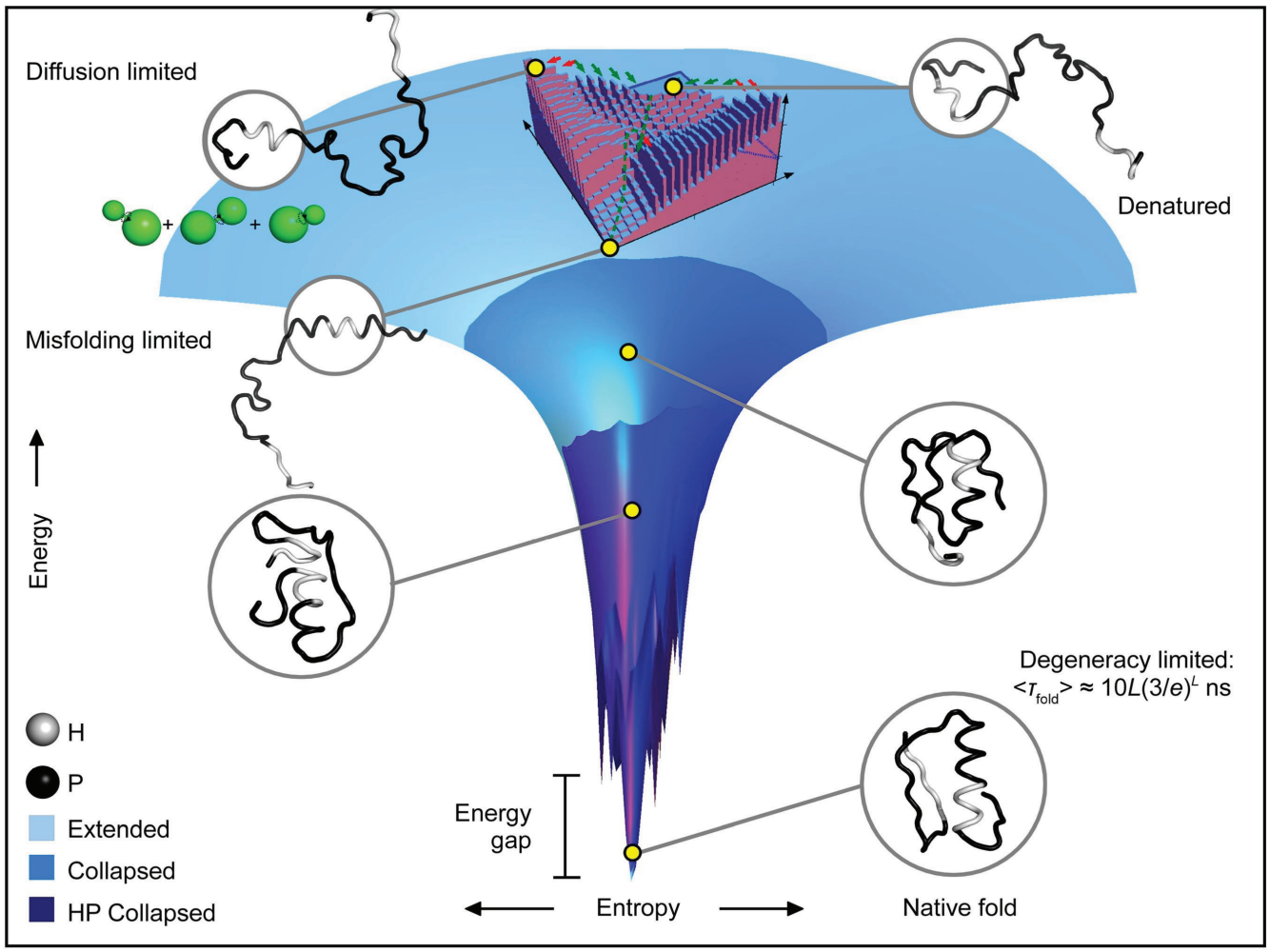

Fig. 6.1. A picture of protein folding kinetics. The results of Chapter 5 are schematically summarized on the energy landscape. The fastest event is helix nucleation, followed by helix propagation, hydrophobic collapse, and segregation of hydrophobic residues (H) into the interior of the protein. The rate limiting step is the final step of folding to the native state, whose structure can be exhaustively found for an average protein sequence of length L, in about $10L(3/e)^L$ ns (see section 5.2). At these different length and time scales, different issues dominate the folding. These mechanisms are cooperative diffusion, misfolding, and conformational degeneracy for helix nucleation, helix propagation, and tertiary structure formation, respectively.

the relatively narrow range of conditions suited to life. Indeed, it is often by witnessing conditions elsewhere that we become grateful for what we have.

BIBLIOGRAPHY

1. Horton HR, Moran LA, Ochs RS, Rawn DJ, & Scrimgeour KG (2002) *Principles of biochemistry* (Pearson Prentice Hall, Upper Saddle River, NJ) 3rd Ed.
2. Finkelstein AV & Ptitsyn OB (2002) *Protein physics: A course of lectures* (Academic Press, New York, NY).
3. Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181(4096):223-230.
4. Dill KA, Ozkan SB, Weikl TR, Chodera JD, & Voelz VA (2007) The protein folding problem: When will it be solved? *Curr Opin Struct Biol* 17(3):342-346.
5. Lodish HF, Berk A, Zipursky SL, Matsudaira P, Baltimore D, & Darnell J (2000) *Molecular cell biology* (W.H. Freeman, New York, NY) 4th Ed.
6. Levinthal C (1969) in *Mossbauer spectroscopy in biological systems*, ed Debrunner JTP, Tsibris JCM, & Munck E (University of Illinois Press, Urbana, IL), pp 22-24.
7. Watson JD & Crick FHC (1953) A structure for deoxyribose nucleic acid. *Nature* 171(4356):737-738.
8. Wilkins MHF, Stokes AR, & Wilson HR (1953) Molecular structure of deoxypentose nucleic acids. *Nature* 171(4356):738-740.
9. Franklin RE & Gosling RG (1953) Molecular configuration in sodium thymonucleate. *Nature* 171(4356):740-741.
10. Wang AHJ, Quigley GJ, Kolpak FJ, Crawford JL, van Boom JH, van der Marel G, & Rich A (1979) Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* 282(5740):680-686.
11. Wing R, Drew H, Takano T, Broka C, Tanaka S, Itakura K, & Dickerson RE (1980) Crystal structure analysis of a complete turn of B-DNA. *Nature* 287(5784):755-758.
12. Knorr D & Sinskey AJ (1985) Biotechnology in food production and processing. *Science* 229(4719):1224-1229.
13. McKeage K & Goa KL (2001) Insulin glargine: A review of its therapeutic use as a long-acting agent for the management of type 1 and 2 diabetes mellitus. *Drugs* 61(11):1599-1624.
14. Roweis S, Winfree E, Burgoyne R, Chelyapov NV, Goodman MF, Rothemund PW, & Adleman LM (1998) A sticker-based model for DNA computation. *J Comput Biol* 5(4):615-629.
15. Adleman LM, Rothemund PWK, Roweis S, & Winfree E (1999) On applying molecular computation to the data encryption standard. *J Comput Biol* 6(1):53-63.
16. Rothemund PWK, Papadakis N, & Winfree E (2004) Algorithmic self-assembly of DNA Sierpinski triangles. *PLoS Biol* 2(12):e424.
17. Andersen ES, Dong M, Nielsen MM, Jahn K, Subramani R, Mamdouh W, Golas MM, Sander B, Stark H, Oliveira CL, Pedersen JS, Birkedal V, Besenbacher F, Gothelf KV, & Kjems J (2009) Self-assembly of a nanoscale DNA box with a controllable lid. *Nature* 459(7243):73-76.
18. Rothemund PWK (2006) Folding DNA to create nanoscale shapes and patterns. *Nature* 440(7082):297-302.
19. Dietz H, Douglas SM, & Shih WM (2009) Folding DNA into twisted and curved nanoscale shapes. *Science* 325(5941):725-730.
20. Strachan T & Read AP (2004) *Human molecular genetics* (Garland Science, Independence, KY) 3rd Ed.

21. Campbell NA, Reece JB, Taylor MR, Simon EJ, & Dickey JL (2008) *Biology : Concepts and connections* (Benjamin Cummings, San Francisco, CA) 6th Ed.
22. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, & Dekker J (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289-293.
23. Hermann T & Patel DJ (2000) RNA bulges as architectural and recognition motifs. *Structure* 8(3):R47-R54.
24. Salazar M, Fedoroff OY, Miller JM, Ribeiro NS, & Reid BR (1993) The DNA strand in DNA·RNA hybrid duplexes is neither B-form nor A-form in solution. *Biochemistry* 32(16):4207-4215.
25. Tinoco I & Bustamante C (1999) How RNA folds. *J Mol Biol* 293(2):271-281.
26. Doolittle RF (1995) The multiplicity of domains in proteins. *Annu Rev Biochem* 64:287-314.
27. Rashin AA, Iofin M, & Honig B (1986) Internal cavities and buried waters in globular proteins. *Biochemistry* 25(12):3619-3625.
28. Richards FM (1974) The interpretation of protein structures: Total volume, group volume distributions and packing density. *J Mol Biol* 82(1):1-14.
29. Pauling L (1960) *The nature of the chemical bond and the structure of molecules and crystals: An introduction to modern structural chemistry* (Cornell University Press, Ithaca, NY) 3d Ed.
30. Pauling L & Corey RB (1951) Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proc Natl Acad Sci USA* 37(5):235-240.
31. Pauling L & Corey RB (1951) The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci USA* 37(5):251-256.
32. Huang K (2005) *Lectures on statistical physics and protein folding* (World Scientific, Hackensack, NJ).
33. Selkoe DJ (2003) Folding proteins in fatal ways. *Nature* 426(6968):900-904.
34. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, & Bourne PE (2000) The Protein Data Bank. *Nucl Acids Res* 28(1):235-242.
35. Whittle PJ & Blundell TL (1994) Protein Structure-Based Drug Design. *Annu Rev Biophys Biomol Struct* 23:349-375.
36. Schames JR, Henchman RH, Siegel JS, Sotriffer CA, Ni H, & McCammon JA (2004) Discovery of a novel binding trench in HIV integrase. *J Med Chem* 47(8):1879-1881.
37. Huston JS, Levinson D, Mudgett-Hunter M, Tai MS, Novotny J, Margolies MN, Ridge RJ, Bruccoleri RE, Haber E, Crea R, & Oppermann H (1988) Protein engineering of antibody binding sites: Recovery of specific activity in an anti-digoxin single-chain Fv analogue produced in Escherichia coli. *Proc Natl Acad Sci USA* 85(16):5879-5883.
38. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, & Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302(5649):1364-1368.
39. Looger LL, Dwyer MA, Smith JJ, & Hellinga HW (2003) Computational design of receptor and sensor proteins with novel functions. *Nature* 423(6936):185-190.
40. Andersen ES (2010) Prediction and design of DNA and RNA structures. *New Biotechnol* 27(3):184-193.

41. 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(30):9200-9210.
42. Thompson PA, Munoz V, Jas GS, Henry ER, Eaton WA, & Hofrichter J (2000) The helix–coil kinetics of a heteropeptide. *J Phys Chem B* 104(2):378-389.
43. Decatur SM & Antonic J (1999) Isotope-edited infrared spectroscopy of helical peptides. *J Am Chem Soc* 121(50):11914-11915.
44. Barth A (2007) Infrared spectroscopy of proteins. *Biochim Biophys Acta* 1767(9):1073-1101.
45. Huang CY, Getahun Z, Zhu Y, Klemke JW, DeGrado WF, & Gai F (2002) Helix formation via conformation diffusion search. *Proc Natl Acad Sci USA* 99(5):2788-2793.
46. Lednev IK, Karnoup AS, Sparrow MC, & Asher SA (1999) α-helix peptide folding and unfolding activation barriers: A nanosecond UV resonance Raman study. *J Am Chem Soc* 121(35):8074-8086.
47. Lednev IK, Karnoup AS, Sparrow MC, & Asher SA (2001) Transient UV Raman spectroscopy finds no crossing barrier between the peptide α-helix and fully random coil conformation. *J Am Chem Soc* 123(10):2388-2392.
48. Clarke DT, Doig AJ, Stapley BJ, & Jones GR (1999) The α-helix folds on the millisecond time scale. *Proc Natl Acad Sci USA* 96(13):7232-7237.
49. Nesmelova I, Krushelnitsky A, Idiyatullin D, Blanco F, Ramirez-Alvarado M, Daragan VA, Serrano L, & Mayo KH (2001) Conformational exchange on the microsecond time scale in α-helix and β-hairpin peptides measured by ^{13}C NMR transverse relaxation. *Biochemistry* 40(9):2844-2853.
50. Zhou J, Ha KS, La Porta A, Landick R, & Block SM (2011) Applied force provides insight into transcriptional pausing and its modulation by transcription factor NusA. *Mol Cell* 44(4):635-646.
51. Block SM, Blair DF, & Berg HC (1989) Compliance of bacterial flagella measured with optical tweezers. *Nature* 338(6215):514-518.
52. Comstock MJ, Ha T, & Chemla YR (2011) Ultrahigh resolution optical trap with single-fluorophore sensitivity. *Nat Methods* 8(4):335-340.
53. Kubelka J (2009) Time-resolved methods in biophysics 9. Laser temperature-jump methods for investigating biomolecular dynamics. *Photochem Photobiol Sci* 8(4):499-512.
54. Ladd MFC & Palmer RA (1977) *Structure determination by X-ray crystallography* (Plenum Press, New York, NY).
55. von Laue M (1913) X-radiation interferences. *Phys Z* 14:1075-1079.
56. Bragg WH (1912) On the direct or indirect nature of the ionization by X-rays. *Philos Mag* 23(133-138):647-650.
57. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, & Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* 181(4610):662-666.
58. Drenth J (2007) *Principles of protein X-ray crystallography* (Springer, New York, NY) 3rd Ed.
59. Zewail AH (2006) 4D ultrafast electron diffraction, crystallography, and microscopy. *Annu Rev Phys Chem* 57:65-103.

60. Massover WH (2007) Radiation damage to protein specimens from electron beam imaging and diffraction: A mini-review of anti-damage approaches, with special reference to synchrotron X-ray crystallography. *J Synchrotron Rad* 14(1):116-127.
61. Fujiyoshi Y & Unwin N (2008) Electron crystallography of proteins in membranes. *Curr Opin Struct Biol* 18(5):587-592.
62. Mitra K & Frank J (2006) Ribosome dynamics: Insights from atomic structure modeling into cryo-electron microscopy maps. *Annu Rev Biophys Biomol Struct* 35:299-317.
63. Gahlmann A, Park ST, & Zewail AH (2009) Structure of isolated biomolecules by electron diffraction–laser desorption: Uracil and guanine. *J Am Chem Soc* 131(8):2806-2808.
64. Gahlmann A, Lee IR, & Zewail AH (2010) Direct structural determination of conformations of photoswitchable molecules by laser desorption–electron diffraction. *Angew Chem Int Ed Engl* 49(37):6524-6527.
65. Lee IR, Gahlmann A, & Zewail AH (2012) Structural dynamics of free amino acids in diffraction. *Angew Chem Int Ed Engl* 51(1):99-102.
66. Dill KA (1985) Theory for the folding and stability of globular proteins. *Biochemistry* 24(6):1501-1509.
67. Jarzynski C (1997) Nonequilibrium equality for free energy differences. *Phys Rev Lett* 78(14):2690-2693.
68. Crooks GE (1999) Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys Rev E* 60(3):2721-2726.
69. Flory PJ (1953) *Principles of polymer chemistry* (Cornell University Press, Ithaca, NY).
70. Dill KA & Shortle D (1991) Denatured states of proteins. *Annu Rev Biochem* 60:795-825.
71. Lopez CF, Darst RK, & Rossky PJ (2008) Mechanistic elements of protein cold denaturation. *J Phys Chem B* 112(19):5961-5967.
72. Kolmogorov AN (1991) The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. *Proc R Soc Lond A* 434(1890):9-13.
73. Ringe D & Petsko GA (2003) The 'glass transition' in protein dynamics: What it is, why it occurs, and how to exploit it. *Biophys Chem* 105(2-3):667-680.
74. Mount DW (2004) *Bioinformatics: Sequence and genome analysis* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY) 2nd Ed.
75. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, & Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6(12):e28766.
76. Simons KT, Bonneau R, Ruczinski I, & Baker D (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* 37(S3):171-176.
77. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim BH, Das R, Grishin NV, & Baker D (2009) Structure prediction for CASP VIII with all-atom refinement using Rosetta. *Proteins* 77(S9):89-99.
78. Zewail AH (2008) in *Physical biology: From atoms to medicine*, ed Zewail AH (Imperial College Press, London, UK).
79. Ramachandran GN, Ramakrishnan C, & Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7(1):95-99.

80. Kawashima Y, Usami T, Ohashi N, Suenram RD, Hougen JT, & Hirota E (2006) Dynamical structure of peptide molecules. *Acc Chem Res* 39(3):216-220.
81. Wales DJ (2003) *Energy landscapes with application to clusters, biomolecules and gl;asses* (Cambridge University Press, Cambridge, UK).
82. Malmström BG & Andersson B (2001) in *The Nobel prize: The first 100 years*, ed Wallin Levinovitz A & Ringertz N (World Scientific, Singapore), pp 73-108.
83. Dobson CM (2003) Protein folding and misfolding. *Nature* 426(6968):884-890.
84. Zimm BH & Bragg JK (1959) Theory of the phase transition between helix and random coil in polypeptide chains. *J Chem Phys* 31(2):526-535.
85. Mermin ND & Wagner H (1966) Absence of ferromagnetism or antiferromagnetism in one- or two-dimensional isotropic Heisenberg models. *Phys Rev Lett* 17(22):1133-1136.
86. Hamada D, Segawa S, & Goto Y (1996) Non-native α-helical intermediate in the refolding of β-lactoglobulin, a predominantly β-sheet protein. *Nature Struct Biol* 3(10):868-873.
87. Li J, Shinjo M, Matsumura Y, Morita M, Baker D, Ikeguchi M, & Kihara H (2007) An α-helical burst in the src SH3 folding pathway. *Biochemistry* 46(17):5072-5082.
88. Yamamoto M, Nakagawa K, Fujiwara K, Shimizu A, Ikeguchi M & Ikeguchi M (2011) A native disulfide stabilizes non-native helical structures in partially folded states of equine β-lactoglobulin. *Biochemistry* 50(49):10590-10597.
89. Williams S, Causgrove TP, Gilmanshin R, Fang KS, Callender RH, Woodruff WH, & Dyer RB (1996) Fast events in protein folding: Helix melting and formation in a small peptide. *Biochemistry* 35(3):691-697.
90. Mohammed OF, Jas GS, Lin MM, & Zewail AH (2009) Primary peptide folding dynamics observed with ultrafast temperature jump. *Angew Chem Int Ed Engl* 48(31):5628-5632.
91. Lin MM, Mohammed OF, Jas GS, & Zewail AH (2011) Speed limit of protein folding evidenced in secondary structure dynamics. *Proc Natl Acad Sci USA* 108(40):16622-16627.
92. Cooper A (1976) Thermodynamic fluctuations in protein molecules. *Proc Natl Acad Sci USA* 73(8):2740-2741.
93. Derrida B (1981) Random energy model: An exactly solvable model of disordered systems. *Phys Rev B* 24(5):2613-2626.
94. Lau KF & Dill KA (1990) Theory for protein mutability and biogenesis. *Proc Natl Acad Sci USA* 87(2):638-642.
95. Shakhnovich EI & Gutin AM (1989) Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. *Biophys Chem* 34(3):187-199.
96. Gutin AM & Shakhnovich EI (1993) Ground state of random copolymers and the discrete random energy model. *J Chem Phys* 98(10):8174-8177.
97. Pande VS, Grosberg AY, Joerg C, & Tanaka T (1996) Is heteropolymer freezing well described by the random energy model? *Phys Rev Lett* 76(21):3987-3990.
98. Luthey-Schulten Z, Ramirez BE, & Wolynes PG (1995) Helix–coil, liquid–crystal, and spin–glass transitions of a collapsed heteropolymer. *J Phys Chem* 99(7):2177-2185.
99. Yue K & Dill KA (1995) Forces of tertiary structural organization in globular proteins. *Proc Natl Acad Sci USA* 92(1):146-150.

100. Kauzmann W (1959) Some factors in the interpretation of protein denaturation. *Adv Protein Chem* 14:1-63.
101. Orland H, Itzykson C & de Dominicis C (1985) An evaluation of the number of hamiltonian paths. *J Physique Lett* 46(8):353-357.
102. Frauenfelder H, Sligar SG, & Wolynes PG (1991) The energy landscapes and motions of proteins. *Science* 254(5038):1598-1603.
103. Karplus M & Weaver DL (1976) Protein folding dynamics. *Nature* 260(5550):404-406.
104. Karplus M & Weaver DL (1994) Protein folding dynamics: The diffusion–collision model and experimental data. *Protein Sci* 3(4):650-668.
105. Nolting B & Agard DA (2008) How general is the nucleation–condensation mechanism? *Proteins* 73(3):754-764.
106. Plaxco KW, Simons KT, & Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 277(4):985-994.
107. Daggett V & Fersht A (2003) The present view of the mechanism of protein folding. *Nature Rev Mol Cell Biol* 4(6):497-502.
108. Mccammon JA, Gelin BR, & Karplus M (1977) Dynamics of folded proteins. *Nature* 267(5612):585-590.
109. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, & Wriggers W (2010) Atomic-level characterization of the structural dynamics of proteins. *Science* 330(6002):341-346.
110. Bowman GR, Huang X, & Pande VS (2009) Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods* 49(2):197-201.
111. Bowman GR & Pande VS (2010) Protein folded states are kinetic hubs. *Proc Natl Acad Sci USA* 107(24):10890-10895.
112. Lane TJ, Bowman GR, Beauchamp K, Voelz VA, & Pande VS (2011) Markov state model reveals folding and functional dynamics in ultra-long MD trajectories. *J Am Chem Soc* 133(45):18413-18419.
113. Stoll R, Voelter W, & Holak TA (1997) Conformation of thymosin β_9 in water/fluoroalcohol solution determined by NMR spectroscopy. *Biopolymers* 41(6):623-634.
114. Heintz D, Reichert A, Mihelic M, Voelter W, & Faulstich H (1993) Use of bimanyl actin derivative (TMB-actin) for studying complexation of β-thymosins: Inhibition of actin polymerization by thymosin β_9. *FEBS Lett* 329(1-2):9-12.
115. 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(44):12952-12953.
116. Alder BJ & Wainwright TE (1959) Studies in molecular dynamics 1. General method. *J Chem Phys* 31(2):459-466.
117. Levitt M & Warshel A (1975) Computer simulation of protein folding. *Nature* 253(5494):694-698.
118. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, & Karplus M (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* 4(2):187-217.
119. Frenkel D & Smit B (2002) *Understanding molecular simulation: From algorithms to applications* (Academic Press, San Diego, CA) 2nd Ed.
120. Ewald PP (1921) The calculation of optical and electrostatic grid potential. *Ann Phys* 64(3):253-287.

121. Darden T, York D, & Pedersen L (1993) Particle mesh Ewald: an *N*Log(*N*) method for Ewald sums in large systems. *J Chem Phys* 98(12):10089-10092.
122. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, & Pedersen LG (1995) A smooth particle mesh Ewald method. *J Chem Phys* 103(19):8577-8593.
123. Levitt M & Sharon R (1988) Accurate simulation of protein dynamics in solution. *Proc Natl Acad Sci USA* 85(20):7557-7561.
124. Israelachvili J & Wennerstrom H (1996) Role of hydration and water structure in biological and colloidal interactions. *Nature* 379(6562):219-225.
125. Voelz VA, Bowman GR, Beauchamp K, & Pande VS (2010) Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1–39). *J Am Chem Soc* 132(5):1526-1528.
126. Bowman GR, Voelz VA, & Pande VS (2011) Atomistic folding simulations of the five-helix bundle protein $\lambda_{6\text{-}85}$. *J Am Chem Soc* 133(4):664-667.
127. Northrup SH, Pear MR, Lee CY, McCammon JA, & Karplus M (1982) Dynamical theory of activated processes in globular proteins. *Proc Natl Acad Sci USA* 79(13):4035-4039.
128. Sugita Y & Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* 314(1-2):141-151.
129. Duan Y & Kollman PA (1998) Pathways to a protein folding intermediate observed in a 1 microsecond simulation in aqueous solution. *Science* 282(5389):740-744.
130. Case DA & Karplus M (1979) Dynamics of ligand binding to heme proteins. *J Mol Biol* 132(3):343-368.
131. Lu H & Schulten K (1998) Steered molecular dynamics study of force induced domain unfolding. *J Mol Graph Model* 16(4-6):290-290.
132. Grubmuller H, Heymann B, & Tavan P (1996) Ligand binding: Molecular mechanics calculation of the streptavidin biotin rupture force. *Science* 271(5251):997-999.
133. Berneche S & Roux B (2001) Energetics of ion conduction through the K+ channel. *Nature* 414(6859):73-77.
134. Nose S (1984) A molecular dynamics method for simulations in the canonical ensemble. *Mol Phys* 52(2):255-268.
135. Hoover WG (1985) Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A* 31(3):1695-1697.
136. Parrinello M & Rahman A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* 52(12):7182-7190.
137. Nose S & Klein ML (1983) Constant pressure molecular dynamics for molecular systems. *Mol Phys* 50(5):1055-1076.
138. Hess B, Bekker H, Berendsen HJC, & Fraaije JGEM (1997) LINCS: A linear constraint solver for molecular simulations. *J Comput Chem* 18(12):1463-1472.
139. Miyamoto S & Kollman PA (1992) SETTLE: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J Comput Chem* 13(8):952-962.
140. Mortenson PN & Wales DJ (2001) Energy landscapes, global optimization and dynamics of the polyalanine $Ac(ala)_8NHMe$. *J Chem Phys* 114(14):6443-6454.
141. Mortenson PN, Evans DA, & Wales DJ (2002) Energy landscapes of model polyalanines. *J Chem Phys* 117(3):1363-1376.
142. Zewail AH (2000) in *Les prix Nobel: The Nobel prizes 1999*, ed Barany A (Almqvist and Wiksell, Stockholm, Sweden), pp 110-203.

143. Thomas JM (2008) in *Physical biology: From atoms to medicine*, ed Zewail AH (Imperial College Press, London, UK).
144. Shorokhov D & Zewail AH (2008) 4D electron imaging: Principles and perspectives. *Phys Chem Chem Phys* 10(20):2879-2893.
145. Barran PE (2011) (Re)solution of a protein fold without solution. *Angew Chem Int Ed Engl* 50(14):3120-3122.
146. Florance HV, Stopford AP, Kalapothakis JM, McCullough BJ, Bretherick A, & Barran PE (2011) Evidence for α-helices in the gas phase: A case study using melittin from honey bee venom. *Analyst* 136(17):3446-3452.
147. Benesch JLP & Robinson CV (2009) Biological chemistry dehydrated but unharmed. *Nature* 462(7273):576-577.
148. Srinivasan R, Lobastov VA, Ruan CY, & Zewail AH (2003) Ultrafast electron diffraction (UED): A new development for the 4D determination of transient molecular structures. *Helv Chim Acta* 86(6):1763-1838.
149. Ihee H, Lobastov VA, Gomez UM, Goodson BM, Srinivasan R, Ruan CY, & Zewail AH (2001) Direct imaging of transient molecular structures with ultrafast diffraction. *Science* 291(5503):458-462.
150. Zewail AH (1991) Femtosecond transition-state dynamics. *Faraday Discuss* 91:207-237.
151. Williamson JC & Zewail AH (1991) Structural femtochemistry: Experimental methodology. *Proc Natl Acad Sci USA* 88(11):5021-5025.
152. Zewail AH (2006) 4D ultrafast electron diffraction, crystallography, and microscopy. *Annu Rev Phys Chem* 57:65-103.
153. Srinivasan R, Feenstra JS, Park ST, Xu SJ, & Zewail AH (2005) Dark structures in molecular radiationless transitions determined by ultrafast diffraction. *Science* 307(5709):558-563.
154. Xu SJ, Park ST, Feenstra JS, Srinivasan R, & Zewail AH (2004) Ultrafast electron diffraction: Structural dynamics of the elimination reaction of acetylacetone. *J Phys Chem A* 108(32):6650-6655.
155. Park ST, Feenstra JS, & Zewail AH (2006) Ultrafast electron diffraction: Excited state structures and chemistries of aromatic carbonyls. *J Chem Phys* 124(17):174707.
156. Park ST, Gahlmann A, He Y, Feenstra JS, & Zewail AH (2008) Ultrafast electron diffraction reveals dark structures of the biological chromophore indole. *Angew Chem Int Ed Engl* 47(49):9496-9499.
157. Kuchitsu K ed (1998) *Structure data of free polyatomic molecules* (Springer Verlag, Berlin-Heidelberg, Germany).
158. Shorokhov D, Park ST, & Zewail AH (2005) Ultrafast electron diffraction: Dynamical structures on complex energy landscapes. *Chem Phys Chem* 6(11):2228-2250.
159. Hargittai I & Hargittai M ed (1988) *Stereochemical applications of gas-phase electron diffraction, part. A: The electron diffraction technique* (VCH, New York, NY).
160. Ihee H, Cao J, & Zewail AH (1997) Ultrafast electron diffraction: Structures in dissociation dynamics of $Fe(CO)_5$. *Chem Phys Lett* 281(1-3):10-19.
161. Novikov VP, Sipachev VA, Kulikova EI, & Vilkov LV (1993) A comparison of amplitudes and shrinkage corrections for $C_6Cl_3(NO_2)_3$ calculated using conventional and new procedures. *J Mol Struct* 301:29-36.
162. Sipachev VA (2004) The use of quantum-mechanical third-order force constants in structural studies. *J Mol Struct* 693(1-3):235-240.

163. Osina EL, Mastryukov VS, Vilkov LV, & Cyvin SJ (1975) Relations between mean amplitudes of vibration and corresponding internuclear distances 3. Amplitudes for CC, CH distances and Badgers rule. *J Struct Chem* 16(6):977-978.
164. Mastryukov VS (1976) Relations between mean amplitudes of vibration and corresponding internuclear distances 6. Bonded CX distances. *J Struct Chem* 17(1):69-73.
165. Mastryukov VS & Osina EL (1976) Relationship between vibrational amplitudes and internuclear distances 4. Amplitudes for element-hydrogen distances. *J Struct Chem* 17(1):147-148.
166. Woenckhaus J, Kohling R, Thiyagarajan P, Littrell KC, Seifert S, Royer CA, & Winter R (2001) Pressure-jump small-angle X-ray scattering detected kinetics of staphylococcal nuclease folding. *Biophys J* 80(3):1518-1523.
167. Lin MM, Shorokhov D, & Zewail AH (2011) Structural dynamics of free proteins in diffraction. *J Am Chem Soc* 133(42):17072-17086.
168. Yang DS, Gedik N, & Zewail AH (2007) Ultrafast electron crystallography 1. Nonequilibrium dynamics of nanometer-scale structures. *J Phys Chem C* 111(13):4889-4919.
169. Frisch M J, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA, Stratmann RE, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN, Strain MC, Farkas O, Tomasi J, Barone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Ochterski J, Petersson GA, Ayala PY, Cui Q, Morokuma K, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Gonzalez C, Challacombe M, Gill PMW, Johnson BG, Chen W, Wong MW, Andres JL, Head-Gordon M, Replogle ES, & Pople JA (1998) GAUSSIAN 98, (Gaussian Inc., Pittsburgh, PA), A7.
170. Berendsen HJC, van der Spoel D, & van Drunen R (1995) GROMACS: A message-passing parallel molecular dynamics implementation. *Comput Phys Commun* 91(1-3):43-56.
171. Humphrey W, Dalke A, & Schulten K (1996) VMD: Visual molecular dynamics. *J Mol Graph* 14(1):33-38.
172. Delano WL The PyMOL molecular graphics system (Schrödinger LLC, Cambridge, MA), 1.2r3pre.
173. Wolfram S (2010) MATHEMATICA (Wolfram Research, Champaign, IL), 8.
174. Williams T & Kelly C (2011) GNUPLOT, 4.5.
175. Berendsen HJC & Hayward S (2000) Collective protein dynamics in relation to function. *Curr Opin Struct Biol* 10(2):165-169.
176. Guay J, Lambert H, Gingras-Breton G, Lavoie JN, Huot J, & Landry J (1997) Regulation of actin filament dynamics by p38 map kinase-mediated phosphorylation of heat shock protein 27. *J Cell Sci* 110 (P3):357-368.
177. Kopito RR (2000) Aggresomes, inclusion bodies and protein aggregation. *Trends Cell Biol* 10(12):524-530.
178. Burghardt TP, Hu JY, & Ajtai K (2007) Myosin dynamics on the millisecond time scale. *Biophys Chem* 131(1-3):15-28.
179. Lockless SW & Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286(5438):295-299.

180. Lee J, Natarajan M, Nashine VC, Socolich M, Vo T, Russ WP, Benkovic SJ, & Ranganathan R (2008) Surface sites for engineering allosteric control in proteins. *Science* 322(5900):438-442.
181. Feige MJ, Hendershot LM, & Buchner J (2010) How antibodies fold. *Trends Biochem Sci* 35(4):189-198.
182. Glaeser RM (2008) Macromolecular structures without crystals. *Proc Natl Acad Sci USA* 105(6):1779-1780.
183. Tang J, Yang DS, & Zewail AH (2007) Ultrafast electron crystallography 3. Theoretical modeling of structural dynamics. *J Phys Chem C* 111(25):8957-8970.
184. Chen S, Seidel MT, & Zewail AH (2006) Ultrafast electron crystallography of phospholipids. *Angew Chem Int Ed Engl* 45(31):5154-5158.
185. Yamakawa H (1971) *Modern theory of polymer solutions* (Harper & Row, New York, NY).
186. Brion P & Westhof E (1997) Hierarchy and dynamics of RNA folding. *Annu Rev Biophys Biomol Struct* 26:113-137.
187. Varani G (1995) Exceptionally stable nucleic acid hairpins. *Annu Rev Biophys Biomol Struct* 24:379-404.
188. Glucksmann-Kuis MA, Dai X, Markiewicz P, & Rothman-Denes LB (1996) E. coli SSB activates N4 virion RNA polymerase promoters by stabilizing a DNA hairpin required for promoter recognition. *Cell* 84(1):147-154.
189. Uhlenbeck OC (1990) Tetraloops and RNA folding. *Nature* 346(6285):613-614.
190. Zhang W & Chen SJ (2002) RNA hairpin-folding kinetics. *Proc Natl Acad Sci USA* 99(4):1931-1936.
191. Sorin EJ, Rhee YM, Nakatani BJ, & Pande VS (2003) Insights into nucleic acid conformational dynamics from massively parallel stochastic simulations. *Biophys J* 85(2):790-803.
192. Kannan S & Zacharias M (2007) Folding of a DNA hairpin loop structure in explicit solvent using replica-exchange molecular dynamics simulations. *Biophys J* 93(9):3218-3228.
193. Bonnet G, Krichevsky O, & Libchaber A (1998) Kinetics of conformational fluctuations in DNA hairpin-loops. *Proc Natl Acad Sci USA* 95(15):8602-8606.
194. Jung J & Van Orden A (2006) A three-state mechanism for DNA hairpin folding characterized by multiparameter fluorescence fluctuation spectroscopy. *J Am Chem Soc* 128(4):1240-1249.
195. Goddard NL, Bonnet G, Krichevsky O, & Libchaber A (2000) Sequence dependent rigidity of single stranded DNA. *Phys Rev Lett* 85(11):2400-2403.
196. Ma H, Proctor DJ, Kierzek E, Kierzek R, Bevilacqua PC, & Gruebele M (2006) Exploring the energy landscape of a small RNA hairpin. *J Am Chem Soc* 128(5):1523-1530.
197. Ansari A, Kuznetsov SV, & Shen Y (2001) Configurational diffusion down a folding funnel describes the dynamics of DNA hairpins. *Proc Natl Acad Sci USA* 98(14):7771-7776.
198. Ma H, Wan C, Wu A, & Zewail AH (2007) DNA folding and melting observed in real time redefine the energy landscape. *Proc Natl Acad Sci USA* 104(3):712-716.
199. Enderlein J (2007) Collapsed but not folded: Looking with advanced optical spectroscopy at protein folding. *Chem Phys Chem* 8(11):1607-1609.

200. Miller TF, 3rd, Vanden-Eijnden E, & Chandler D (2007) Solvent coarse-graining and the string method applied to the hydrophobic collapse of a hydrated chain. *Proc Natl Acad Sci USA* 104(37):14559-14564.
201. Lin MM, Meinhold L, Shorokhov D, & Zewail AH (2008) Unfolding and melting of DNA (RNA) hairpins: The concept of structure-specific 2D dynamic landscapes. *Phys Chem Chem Phys* 10(29):4227-4239.
202. Zhang W & Chen SJ (2001) Predicting free energy landscapes for complexes of double-stranded chain molecules. *J Chem Phys* 114(9):4253-4266.
203. Ivanov V, Zeng Y, & Zocchi G (2004) Statistical mechanics of base stacking and pairing in DNA melting. *Phys Rev E* 70(5):051907.
204. Ares S, Voulgarakis NK, Rasmussen KO, & Bishop AR (2005) Bubble nucleation and cooperativity in DNA melting. *Phys Rev Lett* 94(3):035504.
205. Kuznetsov SV, Shen Y, Benight AS, & Ansari A (2001) A semiflexible polymer model applied to loop formation in DNA hairpins. *Biophys J* 81(5):2864-2875.
206. Poland D & Scheraga HA (1970) *Theory of helix-coil transitions in biopolymers: Statistical mechanical theory of order-disorder transitions in biological macromolecules* (Academic Press, New York, NY).
207. Wartell RM & Benight AS (1985) Thermal denaturation of DNA molecules: A comparison of theory with experiment. *Phys Rep* 126(2):67-107.
208. Paner TM, Amaratunga M, Doktycz MJ, & Benight AS (1990) Analysis of melting transitions of the DNA hairpins formed from the oligomer sequences d[GGATAC(X)$_4$GTATCC] (X = A, T, G, C). *Biopolymers* 29(14):1715-1734.
209. Klump H & Ackermann T (1971) Experimental thermodynamics of the helix–random coil transition 4. Influence of the base composition of DNA on the transition enthalpy. *Biopolymers* 10(3):513-522.
210. Frank-Kamenetskii MD (1971) Simplification of the empirical relationship between melting temperature of DNA, its GC content and concentration of sodium ions in solution. *Biopolymers* 10(12):2623-2624.
211. SantaLucia J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA* 95(4):1460-1465.
212. Petersheim M & Turner DH (1983) Base-stacking and base-pairing contributions to helix stability: Thermodynamics of double-helix formation with CCGG, CCGGp, CCGGAp, ACCGGp, CCGGUp, and ACCGGUp. *Biochemistry* 22(2):256-263.
213. Yamakawa H & Stockmayer WH (1972) Statistical mechanics of wormlike chains 2. Excluded volume effects. *J Chem Phys* 57(7):2843-2854.
214. Mills JB, Vacano E, & Hagerman PJ (1999) Flexibility of single-stranded DNA: Use of gapped duplex helices to determine the persistence lengths of poly(dT) and poly(dA). *J Mol Biol* 285(1):245-257.
215. SantaLucia J & Hicks D (2004) The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* 33:415-440.
216. Hilbers CW, Haasnoot CA, de Bruin SH, Joordens JJ, van der Marel GA, & van Boom JH (1985) Hairpin formation in synthetic oligonucleotides. *Biochimie* 67(7-8):685-695.
217. Haasnoot CAG, Hilbers CW, van der Marel GA, van Boom JH, Singh UC, Pattabiraman N, & Kollman PA (1986) On loop folding in nucleic-acid hairpin-type structures. *J Biomol Struct Dyn* 3(5):843-857.

218. Paner TM, Amaratunga M, & Benight AS (1992) Studies of DNA dumbbells 3. Theoretical analysis of optical melting curves of dumbbells with a 16-base-pair duplex stem and T*n* end loops (*n* = 2, 3, 4, 6, 8, 10, 14). *Biopolymers* 32(7):881-892.
219. van Dongen MJP, Mooren MMW, Willems EFA, van der Marel GA, van Boom JH, Wijmenga SS, & Hilbers CW (1997) Structural features of the DNA hairpin d(ATCCTA-GTTA-TAGGAT): Formation of a G-A base pair in the loop. *Nucl Acid Res* 25(8):1537-1547.
220. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, & Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79(2):926-935.
221. 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(28):6474-6487.
222. Lindahl E, Hess B, & van der Spoel D (2001) GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J Mol Model* 7(8):306-317.
223. Sorin EJ & Pande VS (2005) Exploring the helix–coil transition via all-atom equilibrium ensemble simulations. *Biophys J* 88(4):2472-2493.
224. Rueda M, Kalko SG, Luque FJ, & Orozco M (2003) The structure and dynamics of DNA in the gas phase. *J Am Chem Soc* 125(26):8007-8014.
225. Levine L, Gordon JA, & Jencks WP (1963) Relationship of structure to effectiveness of denaturing agents for deoxyribonucleic acid. *Biochemistry* 2(1):168-175.
226. Herskovits TT & Harrington JP (1972) Solution studies of nucleic-acid bases and related model compounds - solubility in aqueous alcohol and glycol solutions. *Biochemistry* 11(25):4800-4811.
227. Langridge R, Wilson HR, Hooper CW, Wilkins MHF, & Hamilton LD (1960) Molecular configuration of deoxyribonucleic acid 1. X-ray diffraction study of a crystalline form of the lithium salt. *J Mol Biol* 2(1):19-37.
228. Langridge R, Marvin DA, Seeds WE, Wilson HR, Hooper CW, Wilkins MHF, & Hamilton LD (1960) Molecular configuration of deoxyribonucleic Acid 2. Molecular models and their Fourier transforms. *J Mol Biol* 2(1):38-62.
229. Fuller W, Wilkins MHF, Wilson HR, & Hamilton LD (1965) Molecular configuration of deoxyribonucleic acid 4. X-ray diffraction study of a form. *J Mol Biol* 12(1):60-76.
230. Arnott S, Chandrasekaran R, Birdsall DL, Leslie AGW, & Ratliff RL (1980) Left-handed DNA helices. *Nature* 283(5749):743-745.
231. Richmond TJ & Davey CA (2003) The structure of DNA in the nucleosome core. *Nature* 423(6936):145-150.
232. Fuller W, Forsyth T, & Mahendrasingam A (2004) Water–DNA interactions as studied by X-ray and neutron fibre diffraction. *Phil Trans R Soc Lond B* 359(1448):1237-1247.
233. Gale DC & Smith RD (1995) Characterization of noncovalent complexes formed between minor groove binding molecules and duplex DNA by electrospray ionization mass spectrometry. *J Am Soc Mass Spectr* 6(12):1154-1164.
234. Schnier PD, Klassen JS, Strittmatter EF, & Williams ER (1998) Activation energies for dissociation of double strand oligonucleotide anions: Evidence for Watson-Crick base pairing in vacuo. *J Am Chem Soc* 120(37):9605-9613.

235. Wan KX, Gross ML, & Shibue T (2000) Gas-phase stability of double-stranded oligodeoxynucleotides and their noncovalent complexes with DNA-binding drugs as revealed by collisional activation in an ion trap. *J Am Soc Mass Spectr* 11(5):450-457.
236. Hofstadler SA & Griffey RH (2001) Analysis of noncovalent complexes of DNA and RNA by mass spectrometry. *Chem Rev* 101(2):377-390.
237. Lucas AA & Lambin P (2005) Diffraction by DNA, carbon nanotubes and other helical nanostructures. *Rep Prog Phys* 68(5):1181-1249.
238. Cheatham TE, 3rd & Kollman PA (2000) Molecular dynamics simulation of nucleic acids. *Annu Rev Phys Chem* 51:435-471.
239. Norberg J & Nilsson L (2002) Molecular dynamics applied to nucleic acids. *Acc Chem Res* 35(6):465-472.
240. Tidor B, Irikura KK, Brooks BR, & Karplus M (1983) Dynamics of DNA oligomers. *J Biomol Struct Dyn* 1(1):231-252.
241. Prabhakaran M, Harvey SC, Mao B, & McCammon JA (1983) Molecular dynamics of phenylalanine transfer-RNA. *J Biomol Struct Dyn* 1(2):357-369.
242. Nordlund TM, Andersson S, Nilsson L, Rigler R, Gräslund A, & McLaughlin LW (1989) Structure and dynamics of a fluorescent DNA oligomer containing the EcoRI recognition sequence: Fluorescence, molecular dynamics, and NMR studies. *Biochemistry* 28(23):9095-9103.
243. Kosikov KM, Gorin AA, Zhurkin VB, & Olson WK (1999) DNA stretching and compression: Large-scale simulations of double helical structures. *J Mol Biol* 289(5):1301-1326.
244. Weisenseel JP, Reddy GR, Marnett LJ, & Stone MP (2002) Structure of an oligodeoxynucleotide containing a 1,N^2-propanodeoxyguanosine adduct positioned in a palindrome derived from the Salmonella typhimurium hisD3052 gene: Hoogsteen pairing at pH 5.2. *Chem Res Toxicol* 15(2):127-139.
245. Foloppe N & MacKerell AD (2000) All-atom empirical force field for nucleic acids 1. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J Comput Chem* 21(2):86-104.
246. MacKerell AD & Banavali NK (2000) All-atom empirical force field for nucleic acids 2. Application to molecular dynamics simulations of DNA and RNA in solution. *J Comput Chem* 21(2):105-120.
247. Torshin IY, Weber IT, & Harrison RW (2002) Geometric criteria of hydrogen bonds in proteins and identification of 'bifurcated' hydrogen bonds. *Protein Eng* 15(5):359-363.
248. Pal SK & Zewail AH (2004) Dynamics of water in biological recognition. *Chem Rev* 104(4):2099-2123.
249. Ball P (2008) Water as an active constituent in cell biology. *Chem Rev* 108(1):74-108.
250. Lin MM, Shorokhov D, & Zewail AH (2006) Helix-to-coil transitions in proteins: Helicity resonance in ultrafast electron diffraction. *Chem Phys Lett* 420(1-3):1-7.
251. Idiris A, Alam MT, & Ikai A (2000) Spring mechanics of α-helical polypeptide. *Prot Eng* 13(11):763-770.
252. Chen S, Seidel MT, & Zewail AH (2005) Atomic-scale dynamical structures of fatty acid bilayers observed by ultrafast electron crystallography. *Proc Natl Acad Sci USA* 102(25):8854-8859.

253. Seidel MT, Chen S, & Zewail AH (2007) Ultrafast electron crystallography 2. Surface adsorbates of crystalline fatty acids and phospholipids. *J Phys Chem C* 111(13):4920-4938.
254. Horn DM, Breuker K, Frank AJ, & McLafferty FW (2001) Kinetic intermediates in the folding of gaseous protein ions characterized by electron capture dissociation mass spectrometry. *J Am Chem Soc* 123(40):9792-9799.
255. Onuchic JN, Luthey-Schulten Z, & Wolynes PG (1997) Theory of protein folding: The energy landscape perspective. *Annu Rev Phys Chem* 48:545-600.
256. Callender RH, Dyer RB, Gilmanshin R, & Woodruff WH (1998) Fast events in protein folding: The time evolution of primary processes. *Annu Rev Phys Chem* 49:173-202.
257. Cavalli A, Haberthur U, Paci E, & Caflisch A (2003) Fast protein folding on downhill energy landscape. *Prot Sci* 12(8):1801-1803.
258. Liu F & Gruebele M (2008) Downhill dynamics and the molecular rate of protein folding. *Chem Phys Lett* 461(1-3):1-8.
259. Huang CY, Klemke JW, Getahun Z, DeGrado WF, & Gai F (2001) Temperature-dependent helix–coil transition of an alanine based peptide. *J Am Chem Soc* 123(38):9235-9238.
260. Kim W, Yamato I, & Ando T (2011) Alanine-based peptides containing polar residues presumably favour α-helical structure entropically. *Mol Simulat* 37(5):379-385.
261. Gooding EA, Ramajo AP, Wang,JW, Palmer C, Fouts E, & Volk M (2005) The effects of individual amino acids on the fast folding dynamics of α-helical peptides. *Chem Commun* (48):5985-5987.
262. Buchete NV & Hummer G (2008) Coarse master equations for peptide folding dynamics. *J Phys Chem B* 112(19):6057-6069.
263. Duan Y, Wang L, & Kollman PA (1998) The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. *Proc Natl Acad Sci USA* 95(17):9897-9902.
264. Tobias DJ, Mertz JE, & Brooks CL, 3rd (1991) Nanosecond time scale folding dynamics of a pentapeptide in water. *Biochemistry* 30(24):6054-6058.
265. Huang CY, Getahun Z, Wang T, DeGrado WF, & Gai F (2001) Time-resolved infrared study of the helix–coil transition using ^{13}C-labeled helical peptides. *J Am Chem Soc* 123(48):12111-12112.
266. Wang T, Du D, & Gai F (2003) Helix–coil kinetics of two 14-residue peptides. *Chem Phys Lett* 370(5-6):842-848.
267. Gilmanshin R, Williams S, Callender RH, Woodruff WH, & Dyer RB (1997) Fast events in protein folding: Relaxation dynamics of secondary and tertiary structure in native apomyoglobin. *Proc Natl Acad Sci USA* 94(8):3709-3713.
268. Maness SJ, Franzen S, Gibbs AC, Causgrove TP, & Dyer RB (2003) Nanosecond temperature jump relaxation dynamics of cyclic β-hairpin peptides. *Biophys J* 84(6):3874-3882.
269. Dyer RB, Gai F, & Woodruff WH (1998) Infrared studies of fast events in protein folding. *Acc Chem Res* 31(11):709-716.
270. Bernard V (2002) *Molecular fluorescence: Principles and applications* (Wiley-VCH, Weinheim Germany).

271. Ma H, Wan C, & Zewail AH (2006) Ultrafast *T*-jump in water: Studies of conformation and reaction dynamics at the thermal limit. *J Am Chem Soc* 128(19):6338-6340.
272. Mohammed OF, Samartzis PC, & Zewail AH (2009) Heating and cooling dynamics of carbon nanotubes observed by temperature-jump spectroscopy and electron microscopy. *J Am Chem Soc* 131(44):16010-16011.
273. Kubelka J, Hofrichter J, Eaton WA (2004) The protein folding 'speed limit'. *Curr Opin Struct Biol* 14(1)76-88.
274. Chakrabartty A, Schellman JA, & Baldwin RL (1991) Large differences in the helix propensities of alanine and glycine. *Nature* 351(6327):586-588.
275. Schmeisser M, Thaller A, Iglev H, & Laubereau A (2006) Picosecond temperature and pressure jumps in ice. *New J Phys* 8:104.
276. Chin DH, Woody RW, Rohl CA, & Baldwin RL (2002) Circular dichroism spectra of short, fixed-nucleus alanine helices. *Proc Natl Acad Sci USA* 99(24):15416-15421.
277. Sadqi M, de Alba E, Perez-Jimenez R, Sanchez-Ruiz JM, & Munoz V (2009) A designed protein as experimental model of primordial folding. *Proc Natl Acad Sci USA* 106(11):4127-4132.
278. Mortishire-Smith RJ, Drake AF, Nutkins JC, & Williams DH (1991) Left-handed α-helix formation by a bacterial peptide. *FEBS Lett* 278(2):244-246.
279. Margulis CJ, Stern HA, & Berne BJ (2002) Helix unfolding and intramolecular hydrogen bond dynamics in small α-helices in explicit solvent. *J Phys Chem B* 106(41):10748-10752.
280. Soman KV, Karimi A, & Case DA (1993) Molecular dynamics analysis of a ribonuclease C-peptide analog. *Biopolymers* 33(10):1567-1580.
281. Islam MA (2004) Einstein-Smoluchowski diffusion equation: A discussion. *Phys Scripta* 70(2-3):120-125.
282. Bird RB, Stewart WE, & Lightfoot EN (2007) *Transport phenomena* (John Wiley, New York, NY) 2nd Ed.
283. Debye PJW (1960) *Polar molecules* (Dover Publications, New York, NY).
284. Ramachandran GN & Sasisekharan V (1968) Conformation of polypeptides and proteins. *Adv Protein Chem* 23:283-437.
285. Ho BK, Thomas A, & Brasseur R (2003) Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the α-helix. *Protein Sci* 12(11):2508-2522.
286. Toan NM, Marenduzzo D, & Micheletti C (2005) Inferring the diameter of a biopolymer from its stretching response. *Biophys J* 89(1):80-86.
287. Franks F (2000) *Water: A Matrix of Life* (RSC, Cambridge, UK) 2nd Ed.
288. Sengers JV & Watson JTR (1986) Improved international formulations for the viscosity and thermal conductivity of water substance. *J Phys Chem Ref Data* 15(4):1291-1314.
289. Svergun DI, Richard S, Koch MHJ, Sayers Z, Kuprin S, & Zaccai G (1998) Protein hydration in solution: Experimental observation by X-ray and neutron scattering. *Proc Natl Acad Sci USA* 95(5):2267-2272.
290. Roder H (2004) Stepwise helix formation and chain compaction during protein folding. *Proc Natl Acad Sci USA* 101(7):1793-1794.

291. Doshi UR & Munoz V (2004) The principles of α-helix formation: Explaining complex kinetics with nucleation–elongation theory. *J Phys Chem B* 108(24):8497-8506.
292. Ferguson N & Fersht AR (2003) Early events in protein folding. *Curr Opin Struct Biol* 13(1):75-81.
293. Doshi U (2008) in *Protein folding, misfolding and aggregation: Classical themes and novel approaches*, ed Munoz V (RSC Publishing, Cambridge, UK), pp 35-36.
294. Bertsch RA, Vaidehi N, Chan SI, & Goddard WA, 3rd (1998) Kinetic steps for α-helix formation. *Proteins* 33(3):343-357.
295. Hummer G, Garcia AE, & Garde S (2001) Helix nucleation kinetics from molecular simulations in explicit solvent. *Proteins* 42(1):77-84.
296. Scholtz JM, Qian H, Robbins VH, & Baldwin RL (1993) The energetics of ion-pair and hydrogen-bonding interactions in a helical peptide. *Biochemistry* 32(37):9668-9676.
297. Schellman JA (1955) The stability of hydrogen-bonded peptide structures in aqueous solution. *C R Trav Lab Carlsberg Ser Chim* 29(14-15):230-259.
298. Yang AS & Honig B (1995) Free-energy determinants of secondary structure formation 1. α-helices. *J Mol Biol* 252(3):351-365.
299. Wang L, O'Connell T, Tropsha A, & Hermans J (1995) Thermodynamic parameters for the helix–coil transition of oligopeptides: Molecular dynamics simulation with the peptide growth method. *Proc Natl Acad Sci USA* 92(24):10924-10928.
300. Yang AS & Honig B (1995) Free-energy determinants of secondary structure formation 2. Antiparallel β-sheets. *J Mol Biol* 252(3):366-376.
301. Chan HS & Dill KA (1989) Intrachain loops in polymers: Effects of excluded volume. *J Chem Phys* 90(1):492-509.
302. Nagi AD & Regan L (1997) An inverse correlation between loop length and stability in a four-helix-bundle protein. *Fold Des* 2(1):67-75.
303. Huang JJT, Larsen RW, & Chan SI (2012) The interplay of turn formation and hydrophobic interactions on the early kinetic events in protein folding. *Chem Commun* 48(4)487-497.
304. Smith AV & Hall CK (2001) α-helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model. *Proteins* 44(3):344-360.
305. Chowdhury S, Zhang W, Wu C, Xiong G, & Duan Y (2003) Breaking non-native hydrophobic clusters is the rate-limiting step in the folding of an alanine-based peptide. *Biopolymers* 68(1):63-75.
306. Zhong D & Zewail AH (1998) Femtosecond real-time probing of reactions 23. Studies of temporal, velocity, angular, and state dynamics from transition states to final products by femtosecond-resolved mass spectrometry. *J Phys Chem A* 102(23):4031-4058.
307. Lewis GN (1916) The atom and the molecule. *J Am Chem Soc* 38(4):762-785.
308. Dill KA, Ozkan SB, Shell MS, & Weikl TR (2008) The Protein Folding Problem. *Annu Rev Biophys* 37:289-316.
309. Shakhnovich EI & Gutin AM (1990) Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature* 346(6286):773-775.
310. Zwanzig R, Szabo A, & Bagchi B (1992) Levinthal's paradox. *Proc Natl Acad Sci USA* 89(1):20-22.
311. Dinner AR, Sali A, & Karplus M (1996) The folding mechanism of larger model proteins: Role of native structure. *Proc Natl Acad Sci USA* 93(16):8356-8361.

312. Tanford C (1968) Protein denaturation. *Adv Protein Chem* 23:121-282.
313. Agashe VR, Shastry MCR, & Udgaonkar JB (1995) Initial hydrophobic collapse in the folding of barstar. *Nature* 377(6551):754-757.
314. Dasgupta A & Udgaonkar JB (2010) Evidence for initial non-specific polypeptide chain collapse during the refolding of the SH3 domain of PI3 kinase. *J Mol Biol* 403(3):430-445.
315. Richardson JS (1981) The anatomy and taxonomy of protein structure. *Adv Protein Chem* 34:167-339.
316. Fisher ME & Hiley BJ (1961) Configuration and free energy of a polymer molecule with solvent interaction. *J Chem Phys* 34(4):1253-1267.
317. Chan HS & Dill KA (1989) Compact polymers. *Macromolecules* 22(12):4559-4573.
318. White SH & Jacobs RE (1990) Statistical distribution of hydrophobic residues along the length of protein chains: Implications for protein folding and evolution. *Biophys J* 57(4):911-921.
319. Camacho CJ & Thirumalai D (1993) Minimum energy compact structures of random sequences of heteropolymers. *Phys Rev Lett* 71(15):2505-2508.
320. Thirumalai D, O'Brien EP, Morrison G, & Hyeon C (2010) Theoretical perspectives on protein folding. *Annu Rev Biophys Biomol Struct* 39:159-183.
321. White SH & Jacobs RE (1993) The evolution of proteins from random amino acid sequences 1. Evidence from the lengthwise distribution of amino acids in modern protein sequences. *J Mol Evol* 36(1):79-95.
322. Shakhnovich EI (1994) Proteins with selected sequences fold into unique native conformation. *Phys Rev Lett* 72(24):3907-3910.
323. Zana R (1975) On the rate determining step for helix propagation in the helix–coil transition of polypeptides in solution. *Biopolymers* 14(11):2425-2428.
324. Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, & Finkelstein AV (2003) Contact order revisited: Influence of protein size on the folding rate. *Protein Sci* 12(9):2057-2062.
325. Wheelan SJ, Marchler-Bauer A, & Bryant SH (2000) Domain size distributions can predict domain boundaries. *Bioinformatics* 16(7):613-618.
326. Islam SA, Luo J, & Sternberg MJE (1995) Identification and analysis of domains in proteins. *Protein Eng* 8(6):513-526.
327. Sandhya S, Rani SS, Pankaj B, Govind MK, Offmann B, Srinivasan N, & Sowdhamini R (2009) Length variations amongst protein domain superfamilies and consequences on structure and function. *PLoS ONE* 4(3):e4981.
328. Ellis RJ & van der Vies SM (1991) Molecular chaperones. *Annu Rev Biochem* 60:321-347.
329. Wolynes PG (1995) Biomolecular folding in vacuo!!!(?). *Proc Natl Acad Sci USA* 92(7):2426-2427.
330. Salisbury FB (1969) Natural selection and the complexity of the gene. *Nature* 224(5217):342-343.
331. Smith JM (1970) Natural selection and the concept of a protein space. *Nature* 225(5232):563-564.
332. Heringa J & Taylor WR (1997) Three-dimensional domain duplication, swapping and stealing. *Curr Opin Struct Biol* 7(3):416-421.