Available online at www.sciencedirect.com
SCIENCE DIRECT®
Methods 34 (2004) 112–120
METHODS
www.elsevier.com/locate/ymeth

Methods for molecular dynamics simulations of protein folding/unfolding in solution

David A.C. Beck[a] and Valerie Daggett[b,*]

[a] *Biomolecular Structure and Design Program, University of Washington, Seattle, WA 98195-7610, USA*
[b] *Department of Medicinal Chemistry, University of Washington, Seattle, WA 98195-7610, USA*

Accepted 5 March 2004
Available online 2 June 2004

Abstract

All atom molecular dynamics simulations have become a standard method for mapping equilibrium protein dynamics and non-equilibrium events like folding and unfolding. Here, we present detailed methods for performing such simulations. Generic protocols for minimization, solvation, simulation, and analysis derived from previous studies are also presented. As a measure of validation, our water model is compared with experiment. An example of current applications of these methods, simulations of the ultrafast folding protein Engrailed Homeodomain are presented including the experimental evidence used to verify their results. Ultrafast folders are an invaluable tool for studying protein behavior as folding and unfolding events measured by experiment occur on timescales accessible with the high-resolution molecular dynamics methods we describe. Finally, to demonstrate the prospect of these methods for folding proteins, a temperature quench simulation of a thermal unfolding intermediate of the Engrailed Homeodomain is described.

Keywords: Protein folding; Protein unfolding; Molecular dynamics; Force field; Water model

1. Introduction

Molecular dynamics (MD) is a theoretical physics technique for the examination of molecular systems at atomic detail. It has a sound basis in statistical mechanics and classical physics [1–3]. MD has been used in areas as diverse as materials sciences [4], atmospherics [5], and in the biosciences for systems with lipids [6], nucleic acids [7–9], and proteins [10–13].

Accurate simulation of biomolecules in solution (i.e., the condensed phase) requires as much detail as possible in the internal representation of the system under study. For this reason, 'all atom' MD, where all of the atoms (including hydrogens) are treated explicitly during the calculations, is the most realistic approach, and generally prevails over 'united atom' (e.g., methyl groups treated as a single unit) [14], implicit solvent (e.g., distance dependent dielectric or other approximations to account for the lack of water molecules) [15], and other methods with reduced complexity. Another approach is to study very small proteins because their small size decreases the computational requirements [16]. More recently, increases in computer speed, and the proliferation of inexpensive multi-processor machines have enabled all-atom simulations of full proteins access to long simulation time scales [17].

All atom simulation techniques provide atomic resolution of equilibrium protein dynamic behavior and non-equilibrium events like protein folding and unfolding. When used in conjunction with experiment, simulations provide an enhanced view of the system under study. Recent work has examined aspects of protein folding and unfolding [18–23], and the mechanisms of chemical denaturants and co-solvents [18,24–26].

There are several well-known methods and implementations for molecular dynamics simulations of proteins and other biomolecules [14,27–30]. Here, we present our methods for all atom MD simulation of

* Corresponding author. Fax: 1-206-685-7420.
E-mail address: daggett@u.washington.edu (V. Daggett).

1046-2023/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2004.03.008

proteins in solution based on the force field and protocols described by Levitt et al. [27,28]. The known implementations of these methods include the ENCAD program [27] and *in lucem* Molecular Mechanics (known as *il*mm, our scalable parallel in-house program). The protocols presented are generic renditions of those used in a variety of protein studies from our laboratory.

2. Molecular dynamics simulation methods

Molecular dynamics is the time dependent integration of the classical equations of motion for molecular systems. The equations of motion, for all but the simplest systems, are of sufficient complexity that the integration must be done numerically over a large number of very small discrete timesteps rather than analytically in a continuous fashion. This treatment of time assumes that at any given discrete time step the atomic coordinates are fixed. This assumption holds if the magnitude of the time step is sufficiently small (e.g., approximately 2 fs, or less). At any given time step, these 'fixed' coordinates are used to calculate the potential energy and its first derivative, the force, using a molecular mechanics force field.

Generally, for any atom, evolution in time proceeds from step n to $n+1$ as described in Eq. (1), where subscripts denote the time step, Δt is the magnitude of the integration time step, a is the acceleration, f is the force on the atom, m is the atomic mass, v is the velocity, and x refers to the atomic coordinates:

$$\begin{aligned} a_n &= \frac{f_n}{m}, \\ v_{n+1} &= v_n + a_n \Delta t, \\ x_{n+1} &= x_n + v_n \Delta t + \frac{1}{2} a_n \Delta t^2. \end{aligned} \tag{1}$$

A long series of these steps generates a trajectory through phase space, the $6N$ dimensional space (where N is the number of atoms) defined by the three space vectors of the atoms' positions, and velocities. In general, post-simulation analysis is concerned with the atomic position (coordinates) subspace of phase space.

2.1. The microcanonical ensemble

The microcanoncial (NVE) ensemble fixes the number of atoms, the volume of the periodic box, and the total energy (potential and kinetic) of the system. Energy conservation is naturally satisfied for NVE when the classical equations of motion are used [1]. There are several other advantages to performing simulations in the NVE ensemble: there is no need to couple the microscopic system to macroscopic thermodynamic properties such as pressure and temperature on a step-to-step basis. As a result, the implementation is computationally efficient. An efficient computational approach implies fewer numerical operations, reducing the drift (from round-off errors) in the conserved property (energy), thereby maximizing the integrity of the simulation. In addition, attempting to control the properties of macroscopic variables, such as temperature and pressure, for distinctly microscopic systems is fundamentally flawed, and difficult to achieve.

2.2. Numerical integration

Stepwise numerical integration of the equations of motion can be performed in a variety of ways [1,2,14,27–31]; we use the Beeman algorithm as modified by Brooks (Eq. (2) [27,31]). Energy conservation with the Brooks–Beeman algorithm is better than that of the commonly used Verlet method. A range of integration timesteps was tested for stability (i.e., conservation of energy). For all but the most extreme cases, a Δt of 2 fs was found to be appropriate [28]. Larger values of Δt disrupt the continuity of the simulation and conservation of energy, while smaller values do not make efficient use of the computational resources

$$\begin{aligned} x_i(t+\Delta t) &= x_i(t) + v_i(t)\Delta t + [5a_i(t) - a_i(t-\Delta t)]\frac{\Delta t^2}{8}, \\ v_i(t+\Delta t) &= v_i(t) + [3a_i(t+\Delta t) + 6a_i(t) - a_i(t-\Delta t)]\frac{\Delta t}{8}. \end{aligned} \tag{2}$$

2.3. Molecular mechanics force field

An all atom molecular mechanics force field analytically describes the potential energy of a system in terms of the geometries of atomic centers. The energy calculation and dynamics (ENCAD) force field was originally described by Levitt [31] and was subsequently updated [27] and augmented to include the flexible three-center (F3C) water model [28]. As with other biomolecular force fields such as those in CHARMM [14,30] and AMBER [29], the potential energy parameters (e.g., ideal bond length, bond vibration energies) in the ENCAD force field are derived empirically from ab initio quantum mechanics, spectroscopy, and crystallography. Curious readers are directed to the history of the ENCAD force field and its genealogy [32] which includes a description of the original work from Lifson's group; the ECEPP force field; and protocols from Scheraga's lab [33–35]; the Kollman group force fields implemented within AMBER [36,37]; the GROMOS force field from van Gunsteren [38]; the hydrocarbon force fields (MM2-4) of Allinger et al. [39]; and a historical account of molecular dynamics and CHARMM by Karplus [40].

Eq. (3) contains the ENCAD potential function, V. It describes the potential energy as a function of internal coordinates which are calculated from the Cartesian coordinates. It enters from f in Eq. (1). V is expressed in two, three, and four body interaction terms. The first three terms represent the intramolecular interactions due to bond lengths, bond angles, and dihedral/torsion angles. The fourth term accounts for the 'non-bonded' energies attributed to the van der Waals and electrostatic interactions of atom pairs. Idealized plots of the constituent terms of the potential energy function are provided in Fig. 1

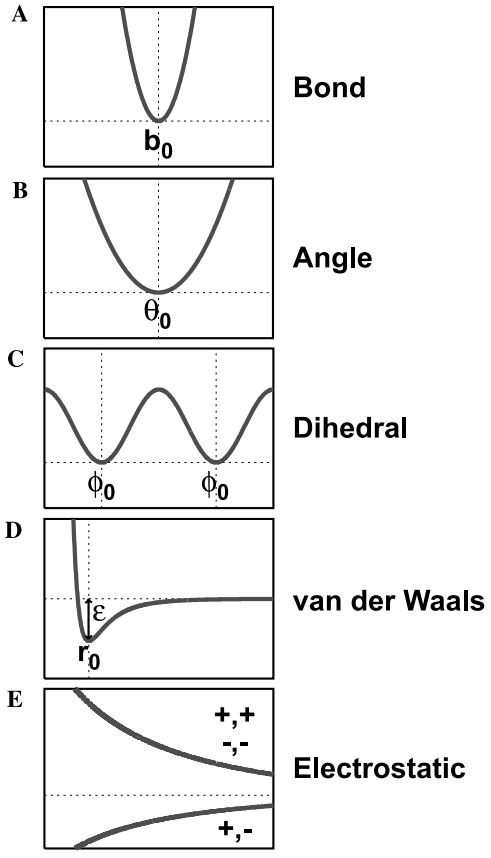

Fig. 1. Idealized plots of constituent terms of the ENCAD potential function, U. The potential energy of each term is the y-axis. (A) The harmonic term that describes the interaction energy of two bonded atoms as a function of the distance of their atomic centers with ideal distance b_0. (B) The harmonic term, similar in form to (A), but of lower energy, that describes the interaction of two atoms bonded to a third atom as a function of the angle between them with the ideal angle θ_0. (C) A typical periodic ($n = 2$) cosine term with a minimum at φ_0 used to describe both in- and out-of-plane (i.e., proper and improper) dihedral angle energies. Plots (A–C) share the same range for energy. (D) The van der Waals interaction energy of two atoms with ε and r_0 the geometric mean of their respective ε and r_0. (E) Three typical electrostatic interactions. The top line idealizes the interaction of charges with like signs while the bottom line idealizes the interaction of two charges with different signs. The sum of (D–E) constitutes the non-bonded interaction energy of two atomic centers.

$$f = -\left(\frac{\partial V}{\partial x}\right)$$

$$V = \sum_{i}^{\text{bonds}} K_{b,i}(b_i - b_{0,i})^2 + \sum_{i}^{\text{bond angles}} K_{\theta,i}(\theta_i - \theta_{0,i})^2 + \sum_{i}^{\text{torsion angles}} K_{\varphi,i}\{1 - \cos[n_i(\varphi_i - \varphi_{0,i})]\} + U_{\text{nb}}$$

$$V_{\text{nb}} = \sum_{\text{pairs}\cdot i,j}\left[\varepsilon_{ij}\left(\frac{r_{0,ij}}{r_{ij}}\right)^{12} - 2\varepsilon_{ij}\left(\frac{r_{0,ij}}{r_{ij}}\right)^{6}\right] + 332\sum_{\text{pairs}\cdot i,j}\left(\frac{q_i q_j}{r_{ij}}\right). \quad (3)$$

The energy of a covalent bond is treated as a harmonic oscillator with an energy minimum at b_0 (Fig. 1) and force constant of K_b. Bond angles are treated similarly with ideal angle θ_0 and force constant K_θ. The third term, used for dihedral and out-of-plane torsion angles, is represented by a cosine with n periods with a minimum energy at φ_0, and barrier to rotation force constant of $K\varphi$.

The van der Waals interaction energy of the atomic pair i and j is treated with a 12/6 Lennard-Jones function. When the pair distance r_{ij} is less than r_0, the geometric mean of r_i and r_j, the function is highly repulsive (Fig. 1). At r_{ij} greater than r_0, the interaction is attractive with a minimum value of ε_0, the geometric mean of ε_i, and ε_j. The interatomic attraction decreases as the separation distance approaches infinity. Pairs of atoms within the same molecule separated by fewer than four bonds are not included in this term.

The electrostatic interaction energy of the atomic pair i and j, with partial charges q_i and q_j, respectively, separated by distance r_{ij} is expressed with a Coulomb style potential. In this model the energy of interaction is favorable when the signs of the partial charges are different and unfavorable when they are the same. As with the Lennard-Jones potential, the energy of interaction gradually decreases to zero as r_{ij} approaches infinity.

The set of parameters for protein atoms including force constants, equilibrium values, r_i, ε_i, and partial charges q_i is available elsewhere [27]. The parameters for the flexible three-center (F3C) water model, an explicit solvent model designed for the ENCAD potential, are also available [28]. Additions for chemical denaturants [24–26] and other co-solvents and ions can be found in the references that describe their applications [18,19,25].

2.4. Non-bonded interaction cutoff

To mimic the solution state of a system, the simulation volume is treated as an infinitely repeating cell or 'periodic box.' Conceptually, this is similar to an orthorhombic unit cell in crystallography. The result is an infinite solution with a protein concentration approaching (but usually below) those in vivo. In practice, it is neither necessary nor computationally possible to

consider all of the non-bonded interactions arising from an infinite solution, as in Eq. (3). In fact, the dielectric constant becomes large at fairly short distances: $\varepsilon \sim 50$ at 10 Å separation of the charges and ~ 70 at 15 Å [41]. Therefore, it is common practice to use a non-bonded pair distance cutoff, r_c. Pairs separated by distances greater than r_c (e.g., 8, 10 Å, etc.) are not considered; that is, the energy of interaction beyond r_c is zero.

The NVE ensemble relies on the continuity of potential terms to conserve energy. Without further modification, such a scheme would be discontinuous at r_c. To maintain the integrity of the NVE ensemble and to preserve the energies and forces of interaction, the ENCAD force field uses a force-shifting cutoff, V_{fs}. This method smoothly and continuously shifts the energies and forces by subtracting from the original potential term, V_{nb}, its first order Taylor expansion about r_c, as in Eq. (4). The non-bonded potential terms and a more complete discussion can be found elsewhere [27]:

$$V_{fs}(r) = V_{nb}(r) - \left[V_{nb}(r_c) + (r - r_c)\left(\frac{dV_{nb}(r_c)}{dr} \right) \right]. \quad (4)$$

The choice of cutoff is not arbitrary. Clearly a very short cutoff ($\sim$4 Å) does not adequately model the electrostatic screening properties of systems. However, slightly longer cutoff ranges such as 8 and 10 Å have been shown in model peptide systems to behave very similarly to much longer cutoffs such as 12, 14, and 16 Å ([27,28] and [D.A.C. Beck, R.S. Armen, V. Daggett, 2003, manuscript in preparation]). In general, very long cutoffs ($\sim$20 Å) do not improve the fidelity of the calculations and take significantly more computational time (as r_c increases the number of pairs grows exponentially). Additional problems with very long cutoffs arise when they exceed half the periodic box dimensions. In this case, an atomic pair could have multiple degenerate interactions, which if evaluated would overestimate the energy of interaction and lead to perturbations in the atomic interactions.

3. Molecular dynamics simulation protocols

3.1. System preparation

The initial preparation of the system under study is vitally important. The molecular dynamics trajectory that is calculated can be highly dependent on the initial configuration. An ill-prepared system, e.g., one that contains atomic clashes, will begin a simulation with exceptional forces that could quickly disrupt the tertiary structure of the protein under examination. The resulting simulation could on all accounts be a correct biophysical interpretation, but one without any relevance to the intended topic of study. A standard set of procedures have been developed to reduce artifacts from inadequate preparation. The specific number of steps of minimization and dynamics may vary between applications, but the general protocol remains consistent.

For native state and unfolding simulations, an experimental structure (derived from crystallography or NMR experiments) is used. The crystal structures require that hydrogens be added. For refolding simulations, a starting structure can be taken from a thermal, or chemical unfolding MD trajectory. Pre- and post-transition state structures, folding intermediates, and structures from the denatured ensemble have all been used for this purpose [18–22]. The potential energy of the complete structure is minimized briefly (usually 200–1500 steps) with respect to the atomic coordinates, usually with a mix of steepest-descent and conjugate-gradient techniques [42]. The resulting 'minimized structure' is then ready to be solvated for simulation in solution.

The minimized structure is placed in an empty periodic box, the walls of which extend a specified distance (typically 8–12 Å) from the protein. This box is then filled with solvent. It is necessary to extend the box to or beyond r_c to eliminate any direct interactions between the protein's first and second solvation shells. Such interactions might alter the process under study. At distances past these shells, the water has been shown to behave as bulk [43].

Water molecules are added from periodic boxes pre-equilibrated to the appropriate density for the desired simulation temperature. Waters from the pre-equilibrated box are not added to the system if they are within a specified 'radius of exclusion' (typically 1.67–2.10 Å) from the protein. The waters (only) are minimized to smooth the solvent network before a short (typically 1–5 ps) MD simulation of the water (only) is performed. The protein is fixed during this process to encourage water to populate relevant hydration sites on the protein surface without causing disruption to the protein structure. In the final steps of preparation, the protein (only) is minimized followed by a minimization of the entire system (water and protein).

3.2. Modifications for higher order systems

The preparation of ternary and quaternary systems involving one or more co-solvents is more complex. Solvation of the protein with waters is performed as described above, after which water molecules are swapped out for randomly placed co-solvent molecules. Care must be taken to insert the correct number of co-solvent molecules and to adjust the box volume so as to match the experimental density at a given mole fraction. The water only is minimized to reorganize the network where it was disrupted by the insertion of co-solvent. Several successive rounds of isolated minimization and short MD simulations are performed on the water, co-solvent, and protein independently. When minimization

readily converges, the process is terminated, and the solution is ready for simulation.

3.3. Temperature

Studies performed with these methods are more concerned with behavior at a given temperature (e.g., 298 K) than at a given energy (e.g., −24,532 kcal/mol). However, in the NVE ensemble, the step-to-step kinetic energy (and thus the temperature) of the system may vary. At each step, the temperature T is calculated from the atomic velocities according to Eq. (5). In this expression, the sum is over all atoms, each with mass m_i, and instantaneous velocity of v_i. N is the number of atoms and K_b is the Boltzmann constant. Due to the step-to-step fluctuations, the mean of these instantaneous temperature samplings for a time interval (typically 100–500 steps) is a more appropriate measure of T.

$$T = \frac{\sum m_i v_i^2}{3NK_b}. \tag{5}$$

At the beginning of a simulation, small, equal, and opposite impulses are applied to randomly selected pairs of atoms. This process is continued until the Maxwellian velocity distribution for the system has a mean within a few Kelvin of the desired simulation temperature. The system must be brought to temperature slowly enough such that it is not shocked. Our current protocols heat the system by 0.05–0.1 K per step.

Simulation of protein native states frequently occurs at 298 K. For simulations of thermal unfolding, any temperature above the protein of interest's melting temperature, T_m, can be used. Our past studies have shown, however, that thermal unfolding is an activated process obeying the rules of Arrhenius behavior [21–23,44]. That is, increasing temperature does not alter the pathway of unfolding, only the rate. As a result, it is possible to simulate unfolding at temperatures significantly above a protein's T_m. The increased rate of unfolding allows short unfolding simulations to sample not only the transition state and early intermediates of unfolding but large regions of the denatured ensemble. With the Maxwellian temperature distribution, a 200 K increase in temperature corresponds to only a 30% increase in the mean atomic velocities.

In addition to the increased rate of unfolding, high temperature simulations benefit from reduced system density. As stated previously, the density of a system during preparation is set to the value obtained from experiment. At 498 K, the density of liquid water from experiment is 0.829 gm/ml [45]. Contrast this with 0.997 gm/ml for 298 K [45] and it is readily apparent that there will be significantly fewer non-bonded interaction partners at 498 K. The reduced number of partners translates into yet faster simulation run-times without disrupting the integrity of the study.

The energy drift arising with this method is primarily kinetic and due to numerical round-off. As a result, the mean system temperature over a large number of steps can be used to monitor energy conservation; when the mean temperature drifts, the velocities are rescaled. Using double precision (64 bit) operations, systems of modest complexity simulated at 298 K rescale once per 5.0×10^6 steps or every 10 ns.

4. Validation and results

As mentioned above, poorly prepared starting structures can introduce fictitious behavior into what is otherwise a correct biophysical simulation. Similarly, incorrect parameterization can cause improper dynamics. For these reasons, it is critical that rigorous comparisons with experiment be conducted to validate the simulation methodology. Here, we present a minimal set of experimental comparisons as means of validation, a brief synopsis of the most commonly used MD simulation analysis methods, and a glance at some current results.

4.1. Water

Explicit water models have a number of experimental observations against which they can be validated. The F3C model has been thoroughly tested and documented [28,43,46]. Here, we have chosen two of the most important bulk properties known from experiment: water self-diffusion and the radial distribution function; which reflect the dynamic behavior; and structure of the solvent, respectively. These properties are well reproduced with the methods described above and a commonly used non-bonded cutoff of 8 Å. The simulation used for these comparisons had 502 F3C water molecules at the experimental density of 0.997 gm/ml and was run for 11 ns. The first nanosecond was allocated to system equilibration.

The self-diffusion of water as a function of simulation time is presented in Fig. 2. The mean diffusion over the last nanosecond is 0.23 Å^2/ps, in agreement with experiment (0.23–0.25 Å^2/ps) [47]. The diffusion calculation converges with simulation time [2]. This convergence is common with much of MD analysis and reflects the need for averaging over long time scales to approximate sampling from ensembles. Another approach to long time scale sampling is to use numerous short simulations performed in parallel. Each simulation has a slight perturbation to its starting structure or a different random number seed during the heating stage. By the ergodic principle, the sampling of these multiple short simulations is equivalent to the sampling of a single long simulation [1].

Another commonly used property for validation of water models is the radial distribution function (RDF),

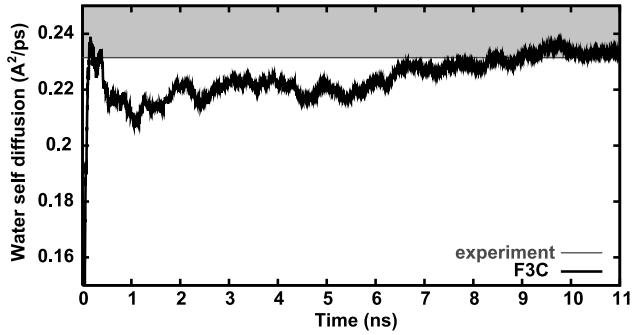

Fig. 2. Water self-diffusion from the F3C water model and experiment [47]. The F3C self-diffusion (black) converges to 0.23 Å^2/ps as the sampling interval increases. This value is in agreement with that from experiment, 0.23–0.25 Å^2/ps (shaded region).

also known as the pair distribution function, or $G(r)$ [1,43]. The RDF of oxygens in water, or $G_{OO}(r)$, represents the ensemble averaged number of oxygen–oxygen pairs found at a distance r. These two-dimensional functions can describe much of the three-dimensional structural quantities of homogeneous systems. Fig. 3 shows water RDFs (G_{OO}, G_{OH}, and G_{HH}) for the F3C water model and two calculated from Soper's neutron diffraction data (Soper A and B) [48]. The F3C model almost completely reproduces the height (number of pairs) and distance of peaks in the experimental RDFs. The height and distance of the first peak in the $G_{OO}(r)$ represent the first solvation shell of water. The height describes the coordination of the first shell, while the distance reflects the close tetrahedral arrangement of water's hydrogen bond network. The second and third

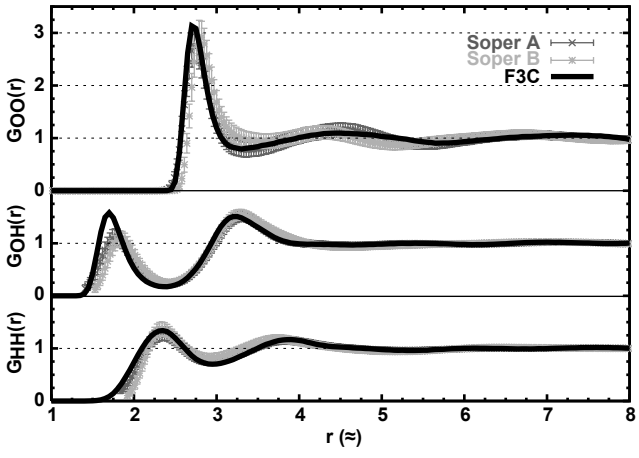

Fig. 3. Water radial distribution functions (RDF) from the F3C water model and neutron diffraction experiments [48]. Data for F3C calculated over the final 10 ns of an 8.0 Å non-bonded cutoff simulation of 502 waters at the experimental density of 0.997 mg/ml. G_{OO} refers to the RDF for oxygen–oxygen pairs, G_{OH} to oxygen–hydrogen, and G_{HH} to hydrogen–hydrogen. The F3C model (black) reproduces the peak heights and distances of the neutron diffraction data (grey). The experimental intra-molecular peaks have been removed for clarity.

peaks correspond to the second shell and bulk water. These peaks are less well defined from experiment as they are much more dynamic with respect to the origin than the first shell.

4.2. Protein folding/unfolding

The study of protein folding and unfolding pathways by molecular dynamics is aided by the choice of system under study. Ultrafast folding proteins are of particular interest because the folding and unfolding events measured by experiment occur on timescales accessible to MD. The Engrailed Homeodomain (En-HD) is a three helix bundle (61 residues) that refolds with a rate constant of 37,500 s^{-1} at 298 K and 51,000 at 315 K, as assessed by temperature jump relaxation experiments [21]. Fig. 4 presents data from three simulations of En-HD with an 8 Å cutoff range. The 298 K native state and 498 K thermal unfolding simulations have been described and compared in detail with experimental data from the laboratory of our collaborator Alan Fersht [21,22]. The folding simulation is a 298 K quench of the 5 ns thermal unfolding intermediate.

In its native state, En-HD's core consists of helices I, II, and III (colored in red, green, and blue, respectively, in Fig. 4). Helices I and II are connected by a 5 residue loop and form an anti-parallel scaffold against which helix III packs. The 7 residue N-terminal loop is highly mobile and is seen to come away from and return to the helical core multiple times in the native state simulation.

During unfolding, the core is weakened, and helix III begins to pull away from the I, II scaffold. This sequence of events, seen in multiple simulations at 348, 373, and 498 K, leads to the transition state ensemble, which occurs at approximately 0.26 ns in the 498 K simulation shown. The transition state identified from simulation is in good agreement with experiment [22]. As the unfolding proceeds, the helices separate, and in the process, lose tertiary contacts. The resulting denatured ensemble has a large amount of secondary structure, but few high order contacts. This is well described by the framework model of protein folding, where the slow step involves the correct tertiary packing of persistent local secondary structure.

The extent of secondary structure predicted for the denatured state [21] was recently confirmed with CD and NMR Hα chemical shift experiments [22]. NMR experiments, however, yielded no trace of tertiary interactions. These experiments were carried out on the L16A mutant, which shifts the equilibrium so that the denatured state (in this case the folding intermediate) is populated under physiological conditions.

Hydrogen exchange of En-HD (used to monitor spontaneous un- and refolding at physiological conditions by the exposure and protection of main chain amides) confirms that most of the residual helix in the

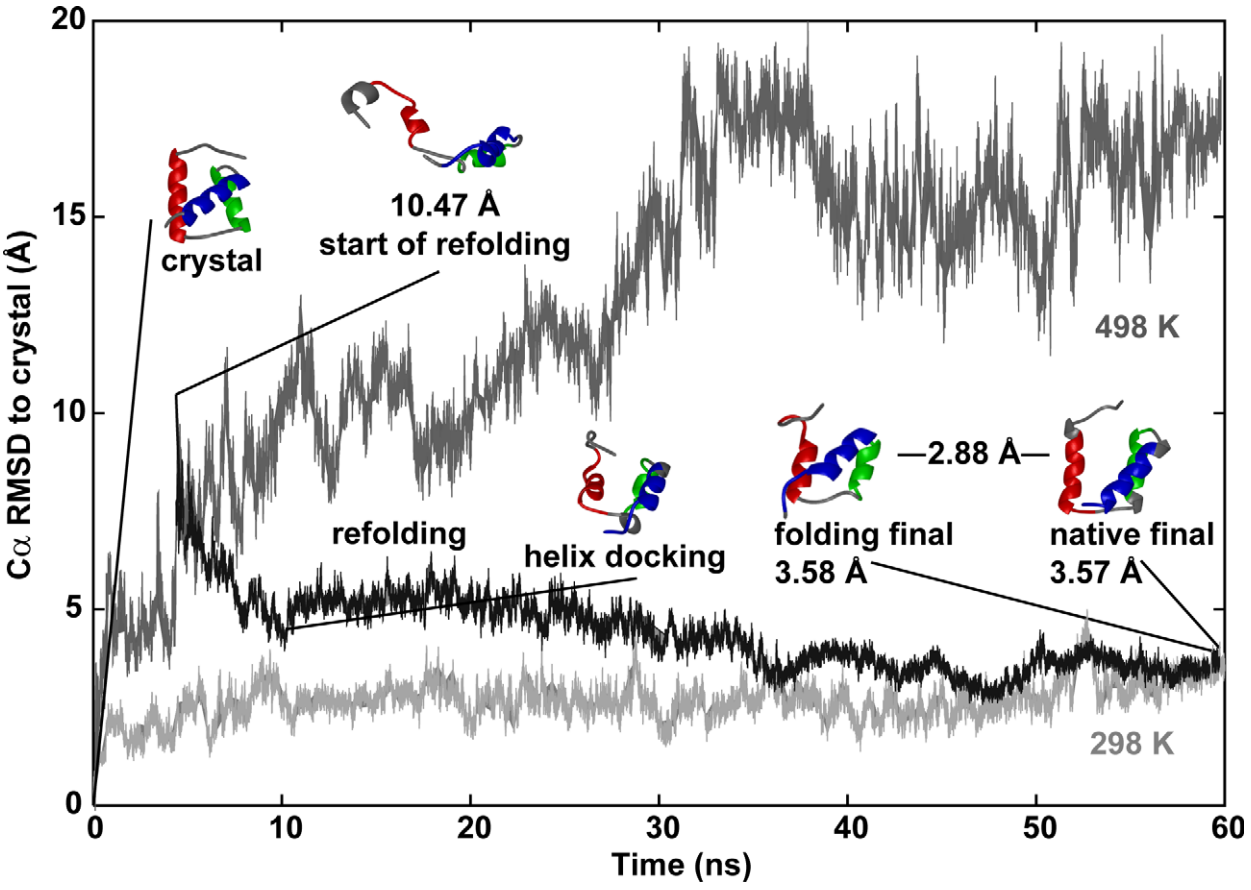

Fig. 4. Cα RMSD to crystal structure as a function of time for three MD simulations of the En-HD. En-HD is a three helix bundle (1enh [50]) with a fair amount of helical structure in its unfolding intermediate, and denatured ensemble. The native state and thermal unfolding simulations have been fully characterized and verified against experiment [21,22]. The 298 K native state simulation (light grey) is a reference against which to compare the 498 K thermal unfolding (dark grey) and 298 K folding/quench (black) simulations. The structures are colored according to the native state helices: helix I in red; helix II in green; and helix III in blue. In the native state simulation, the RMSD ranges from 2.0 to 3.5 Å. The fluctuations reflect the degree of mobility in the loops between helices. The transition state in the thermal unfolding simulation was seen at 0.26 ns. The 5 ns (10.47 Å Cα RMSD to crystal) structure from the unfolding simulation was used to seed the folding/temperature quench simulation. Within the first 5 ns, the folding system undergoes an initial collapse and refolds by the framework model to a final structure that is 3.58 Å from the crystal, and 2.88 Å from the final structure of the native state simulation. These simulations were performed with ENCAD and *il*mm, and used an 8 Å cutoff range [27]. The Cα RMSD calculation excludes the highly mobile N- and C-termini.

intermediate state is native, although transient non-native helices seen in the unfolding simulation are consistent with NMR chemical shift deviations of the L16A mutant. The extrapolated temperature dependent rates of unfolding from temperature jump experiments are in good agreement with those from simulation at high temperature, especially when considering the 'single molecule' aspect of simulation [22]. At 373 K, for example, the half-life of folding was ~2 ns from simulation, and about 5 ns by extrapolation of the experimental data.

A post-transition state starting structure from the thermal unfolding run was used for a temperature quench/refolding simulation. It is 10.47 Å Cα root-mean-square deviation (RMSD) from the crystal structure. This intermediate is non-native in that very few tertiary contacts are present, each helix lacks several turns, and the N-terminus contains a non-native helical segment. Protein refolding occurs very much as the reverse of denaturation: after quenching at 298 K, transient non-native helical segments are lost, and much of the native helical structure quickly returns (<5 ns). Subsequently, the I, II scaffold returns (see 'helix docking' in Fig. 4), and the swing arm of helix III begins to move toward the core (see 'refolding final' in Fig. 4). Although the final structure is similar to structures in the native state ensemble, the refolding simulation is ongoing in order to capture the complete atomic detail of the end-stages in helix docking.

For experimental comparisons, there are several other important computational analyses that must be performed. For example, one must demonstrate that the potential function and simulation protocols reproduce the structure, and dynamic behavior of the native state under folding conditions. The starting structure, or crystal structure in this case, is a useful reference against

which simulation can be compared. The native state RMSD to the crystal structure in Fig. 4 ranges from 2 to 4 Å. These fluctuations reflect the degree of mobility in the loops between the helices. In the thermal unfolding simulations, the RMSD rapidly diverges from the range of values experienced by the native ensemble to a value of 18.6 Å at 60 ns. The refolding simulation starts from the 5 ns, 10.47 Å, unfolding intermediate. The final structure of the folding simulation after a 55 ns simulation at 298 K has an RMSD of 3.58 Å. The final structure of the native state is similar, 3.57 Å. These two final structures are 2.88 Å from each other. The similarity in RMSD to the crystal structure of the final native and 'refolded' structures, in conjunction with their relatively low RMSD to each other, is an indication that the protein in the quenched, refolding simulation has become very native-like.

The RMSD alone is not a sufficient description of protein structure. Other relevant analyses include the calculation of solvent accessible surface area (SASA) and the number and persistence of residue–residue contacts. The mean and standard deviations of the total SASA (by the NACCESS method [49]) for the final nanosecond of the native (4753 ± 127 Å^2) and the refolding simulation (4803 ± 168 Å^2) overlap. The statistical similarity of these values further suggests the refolding run is adopting a very native like conformation. The SASA for the final nanosecond of the 498 K unfolding simulation (6335 ± 204 Å^2), however, is very different from the values for the native state and the folding run. Also of interest in studies of this type are the SASA breakdowns by residue, hydrophobicity, and side/main-chain (data not shown).

The total number of side-chain to side-chain contacts for the last nanosecond of these simulations was calculated. As with the SASA, the refolding simulation mean, and SD (138.8 ± 2.9) is within the fluctuations of the native state (143.6 ± 2.3), a further indication of refolding. In contrast, the unfolding simulation (83.0 ± 3.4) has about 60% of the contacts populated in the native state simulation. The denatured state of En-HD contains considerable residual helical structure in both the simulations and as assessed by experiment [22]. The high degree of contacts in the denatured state reflects intra-helical contacts, not contacts for docking of the helices. More detailed analysis of precisely which native contacts are preserved and which are lost is typical for such studies (data not shown).

5. Concluding remarks

Molecular dynamics is a useful tool for enhancing the information obtained from experiment about protein native states, thermal and chemical unfolding events, and folding pathways. These methods permit reliable unfolding of proteins in agreement with experiments probing both folding and unfolding. The discovery of ultrafast folding proteins bridge the gap between MD and experiment and illustrate the synergy between the two approaches: theorists get validation from experiment and experimentalists get atomic level detail from theory.

Acknowledgments

This work was supported by the National Institutes of Health (GM 50789 to V.D.). D.B. is supported by an NIH Molecular Biophysics Training Grant (National Research Service Award 5 T32 GM 08268). Some of the simulations presented were computed on hardware donated by Intel. University of California, San Francisco, MIDASPLUS, was used to prepare Fig. 4.

References

[1] M.P. Allen, D.J. Tildesley, Computer Simulation of Liquids, Oxford University Press, Oxford, 1987.
[2] J.M. Haile, Molecular Dynamics Simulation: Elementary Methods, Wiley, New York, 1992.
[3] J.A. McCammon, B.R. Gelin, M. Karplus, Nature 267 (1977) 585–590.
[4] K. Kremer, Macromol. Chem. Phys. 204 (2003) 257–264.
[5] P. Jungwirth, D. Tobias, J. Phys. Chem. B 106 (2002) 6361–6373.
[6] L. Saiz, S. Bandyopadhyay, M.L. Klein, Biosci. Rep. 22 (2002) 151–173.
[7] W. Wang, O. Donini, C.M. Reyes, P.A. Kollman, Annu. Rev. Biophys. Biomol. Struct. 30 (2001) 211–243.
[8] E. Giudice, R. Lavery, Acc. Chem. Res. 35 (2002) 350–357.
[9] J. Norberg, L. Nilsson, Acc. Chem. Res. 35 (2002) 465–472.
[10] M. Karplus, J.A. McCammon, Nat. Struct. Biol. 9 (2002) 646–652.
[11] T. Hansson, C. Oostenbrink, W. van Gunsteren, Curr. Opin. Struct. Biol. 12 (2002) 190–196.
[12] V. Daggett, Acc. Chem. Res. 35 (2002) 422–429.
[13] A. Warshel, Acc. Chem. Res. 35 (2002) 385–395.
[14] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, J. Comput. Chem. 4 (1983) 187–217.
[15] D. Bashford, D.A. Case, Annu. Rev. Phys. Chem. 51 (2000) 129–152.
[16] Y. Duan, P.A. Kollman, Science 282 (1998) 740–744.
[17] V. Daggett, Curr. Opin. Struct. Biol. 10 (2000) 160–164.
[18] D.O.V. Alonso, V. Daggett, J. Mol. Biol. 247 (1995) 501–520.
[19] D.O.V. Alonso, V. Daggett, Protein Sci. 7 (1998) 860–874.
[20] D. De Jong, R. Riley, D.O.V. Alonso, V. Daggett, J. Mol. Biol. 319 (2002) 229–342.
[21] U. Mayor, C.M. Johnson, V. Daggett, A.R. Fersht, Proc. Natl. Acad. Sci. USA 97 (2000) 13518–13522.
[22] U. Mayor, N.R. Guydosh, C.M. Johnson, J.G. Grossmann, S. Sato, G.S. Jas, S.M. Freund, D.O.V. Alonso, V. Daggett, A.R. Fersht, Nature 421 (2003) 863–867.
[23] R. Day, B.J. Bennion, S. Ham, V. Daggett, J. Mol. Biol. 322 (2002) 189–203.
[24] K.E. Laidig, V. Daggett, J. Phys. Chem. 100 (1996) 5616–5619.
[25] Q. Zou, B.J. Bennion, V. Daggett, K.P. Murphy, J. Am. Chem. Soc. 124 (2002) 1192–1202.

[26] B.J. Bennion, V. Daggett, Proc. Natl. Acad. Sci. USA 100 (2003) 5142–5147.
[27] M. Levitt, M. Hirshberg, R. Sharon, V. Daggett, Comput. Phys. Commun. 91 (1995) 215–231.
[28] M. Levitt, M. Hirshberg, R. Sharon, K.E. Laidig, V. Daggett, J. Phys. Chem. B 101 (1997) 5051–5061.
[29] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham, S. Debolt, D. Ferguson, G. Seibel, P. Kollman, Comput. Phys. Commun. 91 (1995) 1–41.
[30] A.D. Mackerrell, B. Brooks, C.L. Brooks, L. Nilsson, B. Roux, Y. Won, M. Karplus, in: P. Schleyer (Ed.), The Encyclopedia of Computational Chemistry, vol. 1, John Wiley, Chichester, 1998, pp. 271–77.
[31] M. Levitt, J. Mol. Biol. 168 (1983) 595–620.
[32] M. Levitt, Nat. Struct. Biol. 8 (2001) 392–393.
[33] F.A. Momany, R.F. Mcguire, A.W. Burgess, H.A. Scheraga, J. Phys. Chem. 79 (1975) 2361–2381.
[34] G. Nemethy, M.S. Pottle, H.A. Scheraga, J. Phys. Chem. 87 (1983) 1883–1887.
[35] M.J. Sippl, G. Nemethy, H.A. Scheraga, J. Phys. Chem. 88 (1984) 6231–6233.
[36] S.J. Weiner, P.A. Kollman, D.T. Nguyen, D.A. Case, J. Comput. Chem. 7 (1986) 230–252.
[37] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, P. Weiner, J. Am. Chem. Soc. 106 (1984) 765–784.
[38] W. van Gunsteren, X. Daura, A. Mark, in: P.V.R. Schleyer (Ed.), Encyclopedia of Computational Chemistry, John Wiley, New York, Chichester, 1998.
[39] N.L. Allinger, K.S. Chen, J.H. Lii, J. Comput. Chem. 17 (1996) 642–668.
[40] M. Karplus, Biopolymers 68 (2003) 350–358.
[41] V. Daggett, P.A. Kollman, I.D. Kuntz, Biopolymers 31 (1991) 285–304.
[42] W.H. Press, Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, Cambridge, New York, 1992.
[43] D.A. Beck, D.O. Alonso, V. Daggett, Biophys. Chem. 100 (2003) 221–237.
[44] N. Ferguson, J.R. Pires, F. Toepert, C.M. Johnson, Y.P. Pan, R. Volkmer-Engert, J. Schneider-Mergener, V. Daggett, H. Oschkinat, A. Fersht, Proc. Natl. Acad. Sci. USA 98 (2001) 13008–13013.
[45] G.S. Kell, J. Chem. Eng. Data 12 (1967) 66.
[46] K.E. Laidig, J.L. Gainer, V. Daggett, J. Am. Chem. Soc. 120 (1998) 9394–9395.
[47] K. Krynicki, C.D. Green, D.W. Sawyer, Discuss. Faraday Soc. 66 (1978) 199–208.
[48] A.K. Soper, Chem. Phys. 258 (2000) 121–137.
[49] S.J. Hubbard, J.M. Thornton, Department of Biochemistry and Molecular Biology, University College London, 1993.
[50] N.D. Clarke, C.R. Kissinger, J. Desjarlais, G.L. Gilliland, C.O. Pabo, Protein Sci. 3 (1994) 1779–1787.