Multimolecule test-tube simulations of protein unfolding and aggregation

Michelle E. McCully*a, David A. C. Beckb, and Valerie Daggettc

*Biomedical Structure and Design Program and Department of Bioengineering and eScience Institute and Department of Chemical Engineering, University of Washington, Seattle, WA 98195

Edited by William A. Eaton, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, and approved September 25, 2012 (received for review April 28, 2012)

Molecular dynamics simulations of protein folding or unfolding, unlike most in vitro experimental methods, are performed on a single molecule. The effects of neighboring molecules on the unfolding/folding pathway are largely ignored experimentally and simply not modeled computationally. Here, we present two all-atom, explicit solvent molecular dynamics simulations of 32 copies of the Engrailed homeodomain (EnHD), an ultrafast-folding and -unfolding protein for which the folding/unfolding pathway is well-characterized. These multimolecule simulations, in comparison with single-molecule simulations and experimental data, show that intermolecular interactions have little effect on the folding/unfolding pathway. EnHD unfolded by the same mechanism whether it was simulated in only water or also in the presence of other EnHD molecules. It populated the same native state, transition state, and folding intermediate in both simulation systems, and was in good agreement with experimental data available for each of the three states. Unfolding was slowed slightly by interactions with neighboring proteins, which were mostly hydrophobic in nature and ultimately caused the proteins to aggregate. Protein–water hydrogen bonds were also replaced with protein–protein hydrogen bonds, additionally contributing to aggregation. Despite the increase in protein–protein interactions, the protein aggregates formed in simulation did not do so at the total exclusion of water. These simulations support the use of single-molecule techniques to study protein unfolding and also provide insight into the types of interactions that occur as proteins aggregate at high temperature at an atomic level.

protein folding | protein dynamics

The folding pathway of the Engrailed homeodomain (EnHD) has been extensively characterized through a combination of experimental and computational techniques. EnHD is a three-helix bundle protein, with helices I and II packing antiparallel and helix III docking across them (Fig. 1). EnHD folds and unfolds on ultrafast timescales, which makes its folding pathway a good candidate to be studied by simulation. The structure of the transition state (TS) was first predicted by molecular dynamics (MD) simulations (1) and later validated by experimental techniques (2). A folding intermediate was identified by experiment (3) and structurally characterized by MD (4), and the MD-predicted structure was later validated by NMR (5). In addition, the protein adheres to the principle of microscopic reversibility in simulations at its melting temperature where unfolding and refolding occur in a single continuous trajectory (6). Similarly, when high-temperature unfolding simulations are quenched to folding permissive temperatures, the protein refolds by the reverse of unfolding (7). Despite the fact that the experimental techniques used provide ensemble measurements and the MD simulations are single-molecule (SM) in structure, the agreement has been very good, allowing for a much richer description of the folding/unfolding process of EnHD than would be possible through either approach alone.

SM vs. ensemble measurements aside, protein concentration is a variable that differs greatly between simulation, in vitro experiments, and protein folding in vivo. MD simulations of protein folding/unfolding are effectively at infinite dilution. Structural and kinetic measurements of EnHD have used protein concentrations on the order of 10 μM to 10 mM. On the other end of the scale, proteins fold in vivo in a crowded cellular environment with predicted concentrations of ~300 μg/mL (8). For EnHD, a 7-kDa protein, this cellular concentration would equate to ~40 mM. The effect of neighboring molecules on the folding pathway has been largely ignored, particularly in computational studies, without regard to whether the low concentrations are realistic. Given the wealth of information for EnHD, it is a good system for investigating the effect of neighbors on protein behavior.

Here, we present MD simulations of protein unfolding in a multimolecular system, which we refer to as test-tube (TT) simulations. Our system consisted of 32 copies of EnHD solvated with explicit water, resulting in a concentration of 18 mM. This system was heated to 25 °C to probe the dynamics and conformational properties of the native state of EnHD in the presence of neighboring molecules. In addition, the system was heated to 225 °C to investigate the effect of neighbors on the thermal unfolding pathway. Previous studies have shown that the unfolding of EnHD is an activated process and the pathway is independent of temperature from 75 °C to 225 °C, such that the process is just faster at higher temperature (1–3, 9).

The TT simulations were compared with multiple independent SM simulations (7 simulations at 25 °C and 10 at 225 °C). The native dynamics and unfolding pathway were largely unaffected by the presence of neighboring molecules, although unfolding progressed somewhat more slowly in the TT simulation. The native (N), TS, intermediate (I) state, and denatured (D) state populated during unfolding in the SM and TT simulations agree equally well with existing experimental data. Molecules in the TT systems aggregated, with the high-temperature simulation showing one main cluster and the low-temperature simulation showing many smaller, dynamic clusters. Most of the contacts between proteins were hydrophobic in nature, in contrast to those in the SM simulations. At high temperature, nonpolar packing interactions (or hydrophobic interactions) that were lost upon unfolding were replaced with hydrophobic interactions between neighboring proteins. Hydrogen bonds also formed between protein molecules, many at the exclusion of water, further promoting aggregation. Thus, these simulations provide a molecular picture of protein aggregation at elevated temperature.

Results and Discussion

To create the multimolecule TT system, 32 copies of EnHD were placed on a lattice and solvated by water, giving a concentration of ~18 mM (Fig. 1). The distance between the centers of mass of any molecule and that of its closest neighbor in the initial lattice was a minimum of 51 Å, and the closest atoms were 23 Å apart. After constructing the system, the temperature was brought to 25 °C or 225 °C. Fig. 1 shows the evolution of the two TT systems over time.

Author contributions: M.E.M., D.A.C.B., and V.D. designed research; M.E.M. and D.A.C.B. performed research; M.E.M. and V.D. analyzed data; and M.E.M. and V.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1To whom correspondence may be addressed. E-mail: daggett@uw.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1201809109/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1201809109

PNAS | October 30, 2012 | vol. 109 | no. 44 | 17851–17856
At high temperature, the molecules quickly began to interact, aggregate, and form a large number of intermolecular contacts. In contrast, at 25 °C, the proteins moved more freely through solution, transiently interacting with neighboring molecules.

**Nature of Intermolecular Interactions.** The types of contacts that occurred within a molecule differed in proportion and number from the types of contacts that occurred between molecules. Fig. 24 shows the proportion of hydrogen bonds, hydrophobic interactions, and nonspecific interactions that occur as intramolecular and intermolecular contacts. In the SM and TT simulations, intramolecular contacts were primarily nonspecific (65–68%), followed by hydrophobic (28–31%). However, the contacts between protein molecules in the 25 °C and 225 °C simulations were primarily hydrophobic (58% and 52%), with only 38% and 40% nonspecific interactions.

The total number of interactions present within molecules, between molecules, and with water was the same for SM vs. TT simulations at a given temperature (Fig. 2B). For each of the four systems, contacts were approximately equally split between intramolecular and intermolecular. Overall, there were fewer contacts at high temperature than at low temperature because the density of water is lower at 225 °C. Additionally, there were fewer intramolecular contacts at 225 °C because EnHD unfolded. Contacts with water were lost in the two TT simulations and replaced with contacts between proteins. Despite the aggregation, EnHD still had twice as many contacts with water than with other protein molecules at 225 °C and 10 times as many at 25 °C (Fig. 2B).

Hydrogen bonds were the only type of intramolecular interaction to remain constant between the native and high-temperature simulations (Fig. 2B). EnHD made ~45 hydrogen bonds with itself, primarily within the helical backbone, in all four systems. The native hydrogen bonds that were lost during unfolding were replaced with nonnative hydrogen bonds, primarily between side chains, consistent with previous studies (7). In the high-temperature TT simulation, EnHD made fewer hydrogen bonds with water than in the SM simulations, replacing them with hydrogen bonds to neighboring protein molecules.

The most dramatic increase in contacts in the TT simulation reflected hydrophobic interactions (Fig. 2B). Proteins in the SM and TT simulations had the same number of intramolecular hydrophobic interactions, but in the TT simulations at both temperatures, there was an increase in the number of intermolecular protein–protein interactions. EnHD gained hydrophobic contacts with neighboring molecules without a net loss of intramolecular hydrophobic contacts (Fig. 2B).

Although many hydrophobic residues made contacts with neighboring molecules, many more were exposed to solvent and did not make favorable intermolecular interactions. It is geometrically impossible for all hydrophobic residues on EnHD to be buried, even in N, so solvent exposure of some residues is expected. Hydrophobic clusters also frequently formed within a molecule and consisted of native and non-native interactions. At the end of the high-temperature TT simulation, molecule 23 had a hydrophobic patch on helix III that was buried in a hydrophobic pocket created by two other copies of EnHD (Fig. 3A–D). Fig. 3A, C, and E show progressively closer views of molecule 23, for example, within the context of the other molecules colored by hydrophobic (green) and polar (blue) groups. A slice into the binding surface (Fig. 3F) shows the interactions between hydrophobic groups of molecule 23 and the other proteins that formed the binding pocket.

Protein folding is driven by release of water from exposed nonpolar groups and the burial of hydrophobic groups in the core, as reflected in the decrease in the total number of hydrophobic contacts in the SM native vs. unfolding simulations (Fig. 2B). In the low- and high-temperature TT simulations, there was an increase in total hydrophobic interactions relative to the native SM simulations. Just as folding is driven by the need to bury hydrophobic groups, protein–protein association was also dominated by hydrophobic interactions. Indeed, the hydrophobic contacts lost upon unfolding were replaced with intermolecular hydrophobic
In terms of burial of hydrophobic surface area, aggregation is as effective as folding.

**Behavior in N for SM vs. TT Simulations.**

The TT simulation at 25 °C is in good agreement with previously reported SM simulations (1, 7) as shown in Table S1 based on core (residues 8–53) Cα rmsd (2.2 ± 0.6 for SM vs. 2.58 ± 1.52 Å for the TT simulations), fraction α-helix (0.71 ± 0.05 vs. 0.69 ± 0.03), fraction of native contacts satisfied (0.77 ± 0.04 vs. 0.78 ± 0.05), and fraction of nuclear Overhauser effect cross-peaks (NOEs) satisfied (87 ± 3 vs. 85 ± 3%). Overall, the properties of the molecules in the SM and TT simulations indicated that the presence of neighboring molecules had no effect on the structural properties of N for EnHD.

**Effect of Intermolecular Interactions on Unfolding Pathway of EnHD.**

The extent of unfolding was comparable between the SM and TT simulations as shown in the plots of the average core Cα rmsd, percentage α-helix, and fraction of native contacts across the 10 independent SM simulations and all 32 molecules in the 225 °C TT simulation (Fig. S1). Instead of considering many properties independently, as earlier, these and other properties can be combined to create a property space that captures conformational states and unfolding pathways and allows for rigorous comparison of different simulations. Structures that are close together when plotted in property space, or a projection thereof, have similar structural properties. Such an approach was taken to develop a general, protein-independent property space to compare folding pathways of different proteins (10, 11), and a tailored, protein-specific property space was used previously for EnHD (7, 12). Here, 39 properties were included, and a principal component (PC) analysis was performed on the property space. The properties with the highest weights in the first PC (PC1) were core Cα rmsd, native contacts, and distances between residues in HI and HII (Fig. S2A). The first and second components accounted for 64% of the total variance (Fig. S2B).

Property space projections (PC1 vs. PC2) for all structures from the simulations are shown in Fig. 4. The SM and TT simulations occupied much of the same space in this representation, with the majority of structures contributing to two peaks, representing N and I (Fig. 4). These states, as well as the TSs (Fig. 4, orange dots) and final structures (Fig. 4, green dots), are in the same respective regions in the SM and TT simulations, indicating contacts. In terms of burial of hydrophobic surface area, aggregation is as effective as folding.

**Behavior in N for SM vs. TT Simulations.**

The TT simulation at 25 °C is in good agreement with previously reported SM simulations (1, 7) as shown in Table S1 based on core (residues 8–53) Cα rmsd (2.2 ± 0.6 for SM vs. 2.58 ± 1.52 Å for the TT simulations), fraction α-helix (0.71 ± 0.05 vs. 0.69 ± 0.03), fraction of native contacts satisfied (0.77 ± 0.04 vs. 0.78 ± 0.05), and fraction of nuclear Overhauser effect cross-peaks (NOEs) satisfied (87 ± 3 vs. 85 ± 3%). Overall, the properties of the molecules in the SM and TT simulations indicated that the presence of neighboring molecules had no effect on the structural properties of N for EnHD.
The properties of the 25 structures from the NMR-derived ensemble of the L16A folding intermediate (5) are overlaid on the plot in cyan (Fig. 4). This intermediate was engineered to be highly populated under physiological conditions through mutation of Leu16 to Ala so that structural studies could be performed (3). The L16A intermediate fell in a highly populated region of the property space characterized by high α-helical content and short distances between the 16 key HI–HII and HIII–core residue pairs. Thus, our simulations, in combination with the property space analysis, correctly identified I. Further analysis of the distributions and types of contacts, as well as general properties of the SM vs. TT N and I, is shown in Fig. S4 and Table S2. The contacts closely mirrored what was observed for the respective simulation sets as a whole, and the properties of N and I were the same between the SM and TT structures (Fig. 2). Moreover, N overall had fewer intermolecular contacts than D (Fig. S5), in agreement with Fig. 2B. However, within D, the number of contacts made between proteins was independent of how unfolded EnHD was.

TS ensembles for the high-temperature unfolding simulations were identified for the eight new SM simulations [in addition to the two published simulations (1)] and each molecule of the TT simulation, as described previously (13) and in Methods. A total of 40 TS ensembles were identified for the 32 molecules in the TT simulation (every protein unfolded once, and four molecules transiently refolded and unfolded again). Past work has demonstrated that the TS for EnHD is characterized by HI-III pulling away from the HI–HII scaffold and exposing the hydrophobic core (1, 2, 4), and this is indeed the case for the TSs here (Fig. 5A). Distributions of the core Ca rmbsd, time when the TS occurred, and the number of native contacts present in the TS were very similar for EnHD in the SM and TT simulations (Fig. 5B, D, and E). There were a few late TSs in the TT simulations, which is consistent with the slower kinetics observed for the three properties illustrated in Fig. 5. Using the average final time to reach the first TS during unfolding (SM, 351 ± 172 ps; TT, 462 ± 138 ps) as the half-life t1/2, the unfolding rate kU may be estimated for both ensembles, as described previously (4), as 2.6 ± 1.7 × 10^9 s^{-1} for the SM simulations and 2.4 ± 1.6 × 10^9 s^{-1} for the molecules in the TT simulation.

A semiquantitative structure index (S-value) was calculated for each of the TS ensembles as the ratio of the fraction of native structures from the unfolding TT simulation are provided in the property space projection in Fig. S3 to better visualize the space. The properties of the 25 structures from the NMR-derived ensemble of the L16A folding intermediate (5) are overlaid on the plot in cyan (Fig. 4). This intermediate was engineered to be highly populated under physiological conditions through mutation of Leu16 to Ala so that structural studies could be performed (3). The L16A intermediate fell in a highly populated region of the property space characterized by high α-helical content and short distances between the 16 key HI–HII and HI–II–III–core residue pairs. Thus, our simulations, in combination with the property space analysis, correctly identified I. Further analysis of the distributions and types of contacts, as well as general properties of the SM vs. TT N and I, is shown in Fig. S4 and Table S2. The contacts closely mirrored what was observed for the respective simulation sets as a whole, and the properties of N and I were the same between the SM and TT structures (Fig. 2). Moreover, N overall had fewer intermolecular contacts than D (Fig. S5), in agreement with Fig. 2B. However, within D, the number of contacts made between proteins was independent of how unfolded EnHD was.

TS ensembles for the high-temperature unfolding simulations were identified for the eight new SM simulations [in addition to the two published simulations (1)] and each molecule of the TT simulation, as described previously (13) and in Methods. A total of 40 TS ensembles were identified for the 32 molecules in the TT simulation (every protein unfolded once, and four molecules transiently refolded and unfolded again). Past work has demonstrated that the TS for EnHD is characterized by HI-III pulling away from the HI–HII scaffold and exposing the hydrophobic core (1, 2, 4), and this is indeed the case for the TSs here (Fig. 5A). Distributions of the core Ca rmbsd, time when the TS occurred, and the number of native contacts present in the TS were very similar for EnHD in the SM and TT simulations (Fig. 5B, D, and E). There were a few late TSs in the TT simulations, which is consistent with the slower kinetics observed for the three properties illustrated in Fig. 5. Using the average final time to reach the first TS during unfolding (SM, 351 ± 172 ps; TT, 462 ± 138 ps) as the half-life t1/2, the unfolding rate kU may be estimated for both ensembles, as described previously (4), as 2.6 ± 1.7 × 10^9 s^{-1} for the SM simulations and 2.4 ± 1.6 × 10^9 s^{-1} for the molecules in the TT simulation.

A semiquantitative structure index (S-value) was calculated for each of the TS ensembles as the ratio of the fraction of native
secondary structure and the fraction of contacts present in the TS relative to N (14). S-values can be compared with the corresponding experimental Φ-values for EnHD (2). Both S- and Φ-values reflect the extent of structure in a given residue in the TS ensemble, whereby a value of 0 suggests D-like and a value of 1 suggests N-like extent of structure. Correlations between S- and Φ-values were good for all 50 TSs, with the majority falling within an R range of 0.60 to 0.85 (Fig. 5C).

The order that the five key contacts between HI and HII form upon refolding is generally consistent in temperature quenched simulations of EnHD (7). In refolding, the contacts form in the following order: Arg30-Glu19, Leu34-Leu16, Glu37–Arg15 and Leu34–Arg15, then Leu38–Gln12. In both the SM and TT unfolding simulations presented here, the five contacts were usually lost in the same order: Glu37–Arg15, Leu34–Arg15, Leu38–Gln12, Leu34–Leu16, then Arg30–Glu19 (Fig. 6). The last two contacts lost upon unfolding (Leu34–Leu16 and Arg30–Glu19) were the same two that were gained first in refolding, in the opposite order. Notably, the Arg30–Glu19 contact was present in the starting structure for the refolding simulations, so it was always the first contact to form. The first three contacts lost in unfolding (Glu37–Arg15, Leu34–Arg15, Leu38–Gln12) were gained last in refolding. However, they were gained in the same order they were lost, rather than the opposite. Leu38–Gln12 was consistently the last and least likely contact to reform in the refolding simulations, but it was usually lost third in these unfolding simulations. Curiously, there was not a single simulation or individual molecule in which the five contacts were lost in the exact reverse order in which they were gained in the quenched refolding simulations.

Conclusions

Here we present two TT simulations that probe the interactions between molecules in the N state and during thermal unfolding. EnHD formed clusters in the multimolecule TT simulations at low and high temperature, and it formed fewer intermolecular interactions in N than when it was in D. Hydrophobic packing interactions lost upon unfolding were replaced with intermolecular hydrophobic contacts in the high-temperature TT simulation. EnHD gained protein–protein hydrogen bonds in the TT simulations while losing such interactions with water. However, overall, there were many fewer contacts made between protein molecules than within, and most of the intermolecular interactions were with water rather than with other proteins. Although the molecules were largely interacting with each other in the TT simulations, it was not at the total exclusion of water.

The unfolding pathway was largely unaffected by the presence of neighboring protein molecules, although it was moderately slowed down. The structures from the SM and TT simulations occupied the same regions of property space with similar distributions. TS ensembles agreed well with each other, as well as with experimental data, based on several individual properties. The 39-dimensional property space correctly identified the folding intermediate, and the correlation with structures from the NMR ensemble of the L16A intermediate was excellent. The order of contact loss between HI and HII was consistent between the SM and TT simulations.

Despite the fact that MD is typically an SM technique, it consistently reproduces ensemble experimental measurements. Here, we created a small ensemble—although still many orders of magnitude smaller than the number of molecules probed by experimental methods—and we obtained good agreement with experimental NOEs in N, correlation between S- and Φ-values for TSs, and overlap in property space with the experimentally derived folding intermediate. The behavior in N and unfolding pathways were remarkably similar in the SM and TT simulations despite the high degree of aggregation observed, particularly at high temperature. Although neighboring molecules did not perturb the unfolding pathway, they did alter the kinetics, slowing down the process by 32% (comparing the average times vs. the first unfolding TS). These TT simulations provide insight into the nature of interactions in protein aggregates, showing hydrophobic aggregation through folded and unfolded segments of the structure.

Methods

MD Simulation Parameters. The MD simulations were performed using the in luceum molecular mechanics modeling package (15) with the Levitt et al. force field (16). Several of the SM simulations (all runs at 25 °C, and runs 1 and 2 at 225 °C) were reported previously (1, 2, 4, 7). The starting structure for the simulations was the crystal structure (Protein Data Bank (PDB) ID code 1enh (17); Fig. 1). To create the multimolecule TT system, 32 of these structures were arranged in a face-centered cubic lattice with sides of length ~144 Å, giving concentrations of ~18 mM. A temperature of 225 °C was selected for the unfolding studies because SM unfolding simulations at this temperature for EnHD and other proteins reproduce data from experiments and lower-temperature simulations (1, 4, 9, 18). Use of a high temperature allows us to unfold EnHD faster and therefore better sample D. The system was solvated with flexible F3C water (19), and the water density was set based on the simulation temperature according to the experimentally determined liquid-vapor coexistence curve [25 °C, 0.997 g/mL (20); 225 °C, 0.829 g/mL (21)]. The resulting systems had 85,230 and 71,148 water molecules for the simulations was the crystal structure [Protein Data Bank (PDB) ID code 1enh (17); Fig. 1]. To create the multimolecule TT system, 32 of these structures were arranged in a face-centered cubic lattice with sides of length ~144 Å, giving concentrations of ~18 mM. A temperature of 225 °C was selected for the unfolding studies because SM unfolding simulations at this temperature for EnHD and other proteins reproduce data from experiments and lower-temperature simulations (1, 4, 9, 18). Use of a high temperature allows us to unfold EnHD faster and therefore better sample D. The system was solvated with flexible F3C water (19), and the water density was set based on the simulation temperature according to the experimentally determined liquid-vapor coexistence curve [25 °C, 0.997 g/mL (20); 225 °C, 0.829 g/mL (21)]. The resulting systems had 85,230 and 71,148 water molecules for a total of 285,994 and 243,748 atoms at 25 °C and 225 °C, respectively. The NVE microcanonical ensemble (constant number of particles, volume, and energy) was used with 2-fs time steps, and structures were saved every 1 ps for analysis. An 8-Å force-shifted nonbonded cutoff was used (22), and the nonbonded list was updated every two steps. For the SM simulations, there

Fig. 6. HI–HII contact loss. (A) Five contact pairs between HI–HII, circled by color: Glu37–Arg15 (red), Leu34–Arg15 (orange), Leu38–Gln12 (green), Leu34–Leu16 (cyan), and Arg30–Glu19 (purple). (B) Number of times each contact pair was lost, first, second, third, fourth, and fifth in the nine SM simulations in which all five contact pairs were lost. (C) Number of times each contact pair was lost in the 27 molecules in the TT simulations in which all five contact pairs were lost.
were a total of seven performed at 25 °C (2 × 80 ns, 2 × 50 ns, 3 × 20 ns) and 10 at 225 °C (1 × 39 ns, 1 × 60 ns, 8 × 50 ns), totaling 819 ns. The TT system was simulated for 50 ns at 25 °C and 175 ns at 225 °C for a total of 225 ns, giving an equivalent of 7.2 μs of SM data. In total, there was more than 8 μs of simulation data in this study.

The NMR structure of the L16A mutant of EnHD (a surrogate for I) was previously solved (3, 5). Each of the 25 models in the NMR structure (PDBe ID code 1ztr) were truncated to residues 3 to 56, and the same properties were calculated for each model as for the MD-generated structures.

**MD Simulation Analysis.** A total of 39 physical properties were monitored for all simulations to create an alternate description of the trajectory in property space, which can be very helpful for comparing different trajectories (7, 10–12) (Fig. S2). Cα rmsd to the minimized crystal structure was calculated for the core residues (i.e., 8–53), which excludes the floppy N- and C-termini. The percentage of residues forming a-helix was calculated by our in-house implementation of the DSSP algorithm, which bases secondary structure assessments on hydrogen bonding patterns (23). Center-of-mass distances were calculated between 16 residue pairs previously found to be indicative of the folded N state (6, 7). The number of native residue–residue contacts were counted and classified as occurring between main chain and side chain atoms and whether they were present in the starting structure (native-nonautonomous). Nonnative noncovalent interactions contained carbon atoms that were ≤5.4 Å apart or any other nonhydrogen atoms that did not meet the aforementioned criteria but were ≤4.6 Å apart were considered in contact.

Contacts were also classified by type. Hydrogen bonds were defined as when the donor atom’s hydrogen was ≤2.6 Å from the acceptor atom and the angle between the three atoms was within 45° of linearity. Hydrophobic interactions were for aliphatic carbons atoms separated by ≤5.4 Å. Any other pair of non-hydrogen atoms that did not meet the aforementioned criteria but were ≤4.6 Å apart were classified as an “other,” nonspecific contact. Only contacts between non-neighboring residues were considered.

Solvent accessible surface area (SASA) was calculated for the core residues using our in-house implementation of the NACCESS algorithm (24) and a probe radius of 1.4 Å. The resulting SASA was classified as main chain or side chain and as polar or nonpolar. SASA of Trp48, the fluorescence probe for folding, was included as well, as were the radius of gyration and end-to-end distance. PC analysis was carried out on the resulting normalized 39-dimensional property space as described previously (11). N and I were defined for further analysis as the structures in the bins in the PCT vs. PC histograms that had >40% of the maximum bin count for the four simulation sets (SM and TT at 25 °C and 225 °C).

The 25 °C simulations were compared with experiment via NOE satisfaction. A total of 654 NOEs are available for our construct, residues 3 to 56 (25). An NOE was considered satisfied if the r²-weighted distance between the closest protons in the NOE was less than 5.5 Å, which was the longest cutoff published with the EnHD experimental set.

In the SM and TT 225 °C simulations, TS ensembles for the unfolding and refolding events were determined. Four molecules in the TT simulation were found to transiently refold, so a refolding and additional unfolding TS were identified in these four cases. Unfolding TSs were considered the point of no return from the N-like cluster of the 3D multidimensional scaling of the all-against-all Cα rmsd matrix (13). To create a TS ensemble, the TS was taken as the final point in the N-like cluster and the previous 5 ps. For refolding, the TS was the first point in the N cluster and the subsequent 5 ps (6). The S-value was calculated over the ensemble as the product of S2° and S3° for each residue (14), with S2° the fraction of N secondary structure and S3° the fraction of native and non-native contacts in the TSE relative to the number of contacts in the crystal structure. S-values are a semiquantitative reflection of structure in the TS, as the Cα experimental φ-values are based on energetics but are used to infer structural attributes of the TS (2). Both take values between 0 and 1 and can be compared and used for validation of the MD-generated TS structures (2, 4).

**ACKNOWLEDGMENTS.** This research was supported by National Institutes of Health (NIH) Grant GM50789 (to V.D.) and by the Department of Defense through the National Defense Science and Engineering Graduate Fellowship Program (M.E.M.). Computer time for the test-tube simulations was provided through the Department of Energy (DOE) Office of Biological Research by the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US DOE under contract DE-AC02-05CH11231. Additional computer time for these simulations was provided by NIH National Center for Research Resources Grant 5T1RR023044-01 through the Multi-Tiered Proteomic Compute Cluster. Protein images were made with University of California, San Francisco, Chimera, and plots were made in Gnuplot and Microsoft Excel.

Fig. S1. Properties of thermal unfolding simulations over time. (A) Cα rmsd of the core residues (i.e., 8–53) to the crystal structure is plotted over time. (B) Percentage of residues in α-helix, as defined by DSSP, is plotted over time. (C) Fraction of native residue–residue contacts (present in the minimized crystal structure) is plotted over time. On average, EnHD unfolded more and faster in the single-molecule (SM) simulations. However, by the ends of the respective simulations, EnHD unfolded about the same amount on average (SM, n = 10; TT, n = 32).
Fig. S2. Property space. (A) The 39 properties making up the property space are plotted along with their loadings in the first (red) and second (blue) principal component (PC). (B) The variance (red) and total variance (blue) captured is plotted for each PC.
Fig. S3. Final structures plotted in PC space. The structure of each of the 32 molecules from the 225 °C test-tube (TT) simulation at 175 ns is plotted at its value in the first two dimensions of PC space. Proteins are colored by helix, and residues whose center of mass distances contribute to the property space are shown in sticks and colored by atom (Gln12, Arg15, Leu16, Glu19, Phe20, Arg24, Leu26, Leu-34, Glu37, Leu38, Leu40, Arg30, Ile45, Trp48, Phe49, Lys52, Arg53).
Fig. S4. Types of intermolecular vs. intramolecular interactions in the native (N) vs. intermediate (I) state. (A) The distributions of structures in the first and second PCs of the 39-dimensional property space are plotted for the pooled simulations at 25 °C and 225 °C for the SM (n = 81,672) and TT (n = 720,000) systems. The 25 °C and 225 °C distributions were independently normalized to the bin, with the maximum number of structures and plotted from white (i.e., 0.0) to gray (i.e., 0.4) to red/blue (i.e., 1.0). Structures from the bins with >0.4 maximum structures were defined as the N and I peaks (SM N, blue, n = 18,901; SM I, red, n = 12,098; TT N, light blue, n = 70,684; and TT I, light red, n = 283,035). (B) The fraction of contacts, as classified as hydrogen bonds (Hbond, cyan), hydrophobic interactions (Hphob, purple), and others. Legend continued on following page.
Table S1. Properties of N simulations

<table>
<thead>
<tr>
<th>Property</th>
<th>SM simulations</th>
<th>TT simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core Cα rmsd*, Å</td>
<td>2.22 ± 0.61</td>
<td>2.58 ± 1.52</td>
</tr>
<tr>
<td>Fraction α-helix†</td>
<td>0.71 ± 0.05</td>
<td>0.69 ± 0.03</td>
</tr>
<tr>
<td>Fraction native contacts‡</td>
<td>0.77 ± 0.04</td>
<td>0.78 ± 0.05</td>
</tr>
<tr>
<td>Fraction NOEs satisfied§</td>
<td>0.87 ± 0.03</td>
<td>0.85 ± 0.03</td>
</tr>
</tbody>
</table>

*rmsd was calculated over the Cα atoms of residues 8 to 53. All properties are given as the average ± 1 SD (n = 7 runs for the SM simulations and n = 32 molecules for the multimolecule simulation).
†DSSP was used to determine what fraction of the 54 residues was in α-helix.
‡Two residues were considered in contact if the distance between at least one nonhydrogen atom from each was less than 5.4 Å for carbon/carbon pairs and 4.6 Å for all other pairs. A contact was considered native if it was present in the minimized crystal structure.
§A total of 654 reported nuclear Overhauser effect cross-peaks (NOEs) were used for comparison. An NOE was considered satisfied if the r−6 weighted distance between closest equivalent protons was ≤5.5 Å.
Table S2. Properties of SM vs. TT N and I states

<table>
<thead>
<tr>
<th>Property</th>
<th>N state</th>
<th>I state</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SM simulations</td>
<td>TT simulation</td>
</tr>
<tr>
<td>Core Cα rmsd*, Å</td>
<td>2.18 ± 0.42</td>
<td>1.75 ± 0.44</td>
</tr>
<tr>
<td>Fraction α-helix†</td>
<td>0.73 ± 0.06</td>
<td>0.69 ± 0.02</td>
</tr>
<tr>
<td>Fraction native contacts‡</td>
<td>0.78 ± 0.02</td>
<td>0.80 ± 0.03</td>
</tr>
</tbody>
</table>

* rmsd was calculated over the Cα atoms of residues 8 to 53. All properties are given as the average ± 1 SD (n = 18,901 structures N SM; 12,098 I SM; 70,684 N TT; 283,035 I TT).
† DSSP was used to determine what fraction of the 54 residues was in α-helix.
‡ Two residues were considered to be in contact if the distance between at least one nonhydrogen atom from each was less than 5.4 Å for carbon/carbon pairs and 4.6 Å for all other pairs. A contact was considered N if it was present in the minimized crystal structure. There were 196 native contacts in the starting structure.