Microscopic Reversibility of Protein Folding in Molecular Dynamics Simulations of the Engrailed Homeodomain[†]

Michelle E. McCully,[‡] David A. C. Beck,[§] and Valerie Daggett*[,‡,§]

Biomolecular Structure and Design Program and Department of Bioengineering, University of Washington, Box 355013, Seattle, Washington 98195-5013

Received January 21, 2008; Revised Manuscript Received April 16, 2008

ABSTRACT: The principle of microscopic reversibility states that at equilibrium the number of molecules entering a state by a given path must equal those exiting the state via the same path under identical conditions or, in structural terms, that the conformations along the two pathways are the same. There has been some indirect evidence indicating that protein folding is such a process, but there have been few conclusive findings. In this study, we performed molecular dynamics simulations of an ultrafast unfolding and folding protein at its melting temperature to observe, on an atom-by-atom basis, the pathways the protein followed as it unfolded and folded within a continuous trajectory. In a total of 0.67 μs of simulation in water, we found six transient denaturing events near the melting temperature (323 and 330 K) and an additional refolding event following a previously identified unfolding event at a high temperature (373 K). In each case, unfolding and refolding transition state ensembles were identified, and they agreed well with experiment on the basis of a comparison of *S* and Φ values. On the basis of several structural properties, these 13 transition state ensembles agreed very well with each other and with four previously identified transition states from high-temperature denaturing simulations. Thus, not only were the unfolding and refolding transition states part of the same ensemble, but in five of the seven cases, the pathway the protein took as it unfolded was nearly identical to the subsequent refolding pathway. These events provide compelling evidence that protein folding is a microscopically reversible process. In the other two cases, the folding and unfolding transition states were remarkably similar to each other but the paths deviated.

In 1925, Richard C. Tolman coined the term "microscopic reversibility" in reference to chemical reactions: "In recent years increasing use has been made of a new postulate which perhaps cannot yet be stated in its final form, but which requires in a general way in the case of a system in thermodynamic equilibrium not only that the total number of molecules leaving a given state in unit time shall on the average equal the number arriving in that state in unit time, but also that the number leaving by any particular path shall on the average be equal to the number arriving by the reverse of that particular path, thus excluding any cyclical maintenance of the equilibrium state. The writer has ventured to name this postulate *the principle of microscopic reversibility"* (*1*). This description was recast into structural terms in 1967 by Frank H. Westheimer (*2*) and was adopted by the IUPAC in 1999: "In the case of SN2 reactions at tetrahedral centers implying a formation of the trigonal bipyramid transition state (or intermediate) structure, the original formulation of the principle was extended in the following way: if a molecule or reactant enters a trigonal bipyramid at an apical position, this (or another) molecule or reactant must likewise leave the trigonal bipyramid from an apical position" (*3*).

The hypothesis that protein folding may, like chemical reactions, be microscopically reversible has since been offered. If this hypothesis is true, one would expect to observe identical transition states for folding and unfolding, and major events on the folding pathway would occur in reverse order in the unfolding pathway. Supporting evidence has been presented by Jackson et al., who showed through Φ value analysis on chymotrypsin inhibitor 2 (CI2)[1] that folding and unfolding transition states are the same, suggesting that the pathways are also the same (*4*). Additionally, molecular dynamics (MD) generated unfolding transition states (TS) of CI2 are in quantitative agreement with experimental data collected for both unfolding and refolding (*5, 6*). Furthermore, the unfolding and direct refolding pathways of CI2 were shown to be the same in a single continuous MD trajectory (*7*).

The latter MD study was conducted at the melting temperature (T_m) of the protein, at which point the folding and unfolding rates are equal and $\Delta G = 0$. For these reasons, exchange between the folded and unfolded states is dependent on the energy barrier ($\Delta G^{\ddagger} \approx 2.3$ kcal/mol) (*8*), and unfolding and refolding may occur in a single trajectory on

[†] This research was supported by National Institutes of Health Grant GM50789 (to V.D.).

* To whom correspondence should be addressed. E-mail: daggett@u.washington.edu. Phone: (206) 685-7420. Fax: (206) 685-3300.

[‡] Biomolecular Structure and Design Program.

[§] Department of Bioengineering.

[1] Abbreviations: CI2, chymotrypsin inhibitor 2; MD, molecular dynamics; TS, transition state; T_m, melting temperature; $\Delta G^{\ddagger}$, energy barrier/activation energy; N, native; N′, nearly native; D, denatured; rmsd, root-mean-square deviation; En-HD, engrailed homeodomain; TSE, transition state ensemble; *il*mm, *in lucem* molecular mechanics; MDS, multidimensional scaling.

10.1021/bi800118b CCC: $40.75
Published on Web 06/14/2008

time scales tractable by MD. The protein, CI2, passed through three different states, native (N), nearly native (N′), and denatured (D), and then returned to N′ over a time period of 60 ns. When the protein moved from N′ to D and back, unfolding and refolding transition states were identified, and they were the same. The Cα root-mean-square deviations (rmsds) from the native structure and internal contacts were analyzed to differentiate among the three different states. Day and Daggett (*7*) defined N′, an alternate, stable state for CI2 at elevated temperatures. The protein passed through N′ before it moved through its TS to D. N′ was characterized by many near-native interactions but elongated contact distances. In particular, Trp 5, a fluorescence unfolding probe, was buried in both N and N′, but not in D, which would be expected if both N and N′ were native but D was not. D had a disrupted hydrophobic core and loss of secondary structure. This CI2 study showed direct unfolding and refolding in a single continuous trajectory by the same structural pathways for the first time. Consequently, this behavior needs to be demonstrated in another system to ensure it is reproducible, which we describe here.

The engrailed homeodomain (En-HD) of *Drosophila melanogaster* is a 61-residue three-helix bundle. It is ultrafast folding ($k_F = 37500\ s^{-1}$ at 25 °C and $51000\ s^{-1}$ at 42 °C) and unfolding ($k_U = 1100\ s^{-1}$ at 25 °C and $205000\ s^{-1}$ at 63 °C), and its folding and unfolding pathways have been extensively characterized through combined experimental and MD studies (*9–12*). Folding for En-HD follows the framework model involving the docking of HI (residues 10−22), HII (residues 28−38), and HIII (residues 42−55) (*13*). These properties make En-HD especially well-suited for MD folding studies.

In this study, we performed MD simulations of En-HD near its T_m [52 °C = 325 K (*12*)] to compare unfolding and refolding under identical conditions. We analyzed five simulations, three at 323 K and two at 330 K. We compared them to four previously described thermal denaturation simulations, two each at 373 and 498 K and one native simulation at 298 K. The first of the 373 K simulations was found to contain a region of particular interest that had not been previously reported, so that simulation was also analyzed in detail. We identified and characterized three different states populated during unfolding and refolding: N, N′, and D. We also found six transient denaturing events in which En-HD partially unfolded and refolded in three of the five T_m simulations; from this, 12 unfolding and refolding transition state ensembles (TSE) were identified. En-HD unfolded in the 373K/1 simulation, and a TSE in agreement with experiment was reported previously (*9, 12, 13*). Further investigation of this simulation showed that En-HD later refolded, so we also describe this high-temperature refolding TSE. These 13 TSEs agree well with the four previously identified unfolding TSEs from the four high-temperature simulations. Besides defining TSEs, we analyzed the entire pathway En-HD followed as it unfolded and refolded. Five of the seven refolding pathways were nearly identical to the unfolding pathways that preceded them. These five examples are further evidence that the ensembles of folding and unfolding pathways are one and the same, and that protein folding is a microscopically reversible process. However, in the other two cases, En-HD passed through remarkably similar unfolding and refolding transition states, but only a portion of the actual refolding pathway was similar to the unfolding pathway.

METHODS

Molecular Dynamics Simulations. A total of nine MD simulations are addressed in this paper at the following temperatures with simulation times in parentheses: 298 K (100 ns), 323 K (100, 50, and 42 ns), 330 K (100 and 100 ns), 373 K (24 and 75 ns), and 498 K (20 and 60 ns), for a total of 0.67 μs. All four of the 373 and 498 K simulations have been described previously (*9, 12, 13*), as have the first two 323 K simulations and the 298 K simulation (*14*).

Both of the 330 K simulations were performed using our in-house molecular dynamics package, *in lucem* molecular mechanics (*il*mm) (*15*) with the Levitt et al. force field (*16*) using previously described protocols (*17*). The crystal structure (Protein Data Bank entry 1ENH) was minimized for 1000 steps and solvated in a box of F3C water molecules (*18*) such that there was at least 12 Å between the protein and the edge of the periodic box. The density was set to 0.985 g/mol, in agreement with the experimentally determined liquid−vapor coexistence curve for this temperature (*19*). One thousand steps of steepest descent minimization were performed on the water alone followed by 1 ps of dynamics. Next, the water and the protein were independently minimized for an additional 500 steps. Production simulations were performed for 100 ns, allowing all atoms to move, with structures written out every 1 ps. Long-range interactions were truncated after 8 Å using a force-shifted nonbonded cutoff. Our force-shifted cutoff method at this distance is an effective treatment of long-range interactions based on computational savings, energy conservation, and the ability to reproduce experimental results (*20*). The 323K/3 simulation followed the same protocol, except there was only 8 Å of padding between the protein and the edge of the box, the protein was minimized for 200 steps before adding water, and the simulation was run for 42 ns.

Cα rmsd Matrix and Three-Dimensional (3D) MDS. All-versus-all Cα rmsd matrices were calculated to identify clusters of structures with similar conformations. Granularities were chosen to give 1000−5000 time points over the period of interest. The Cα rmsd between each structure and every other structure was computed, resulting in a matrix with 1000^2−5000^2 data points. Low-Cα rmsd boxes on the diagonal represent a period of time during which the protein stayed in a particular conformation. When these boxes lie off the diagonal, they indicate conformations of similar structure visited discontinuously in time. As described previously (*13*), the "core" (residues 8−53) was usually used to calculate the Cα rmsd, rather than the whole protein (residues 3−56). Since the fluctuations of the terminal residues are not indicative of the overall motion of the protein and introduce noise, the five residues at the N-terminus and three at the C-terminus were not included where specified.

Using the program R (*21*), multidimensional scaling (MDS) was performed to project the matrix down to three dimensions. This scaling results in a 3D plot in which each point represents a structure, and the distance between any two points is proportional to the Cα rmsd between the respective structures. The points are connected in order of time for the period of interest. As with the matrix, a series

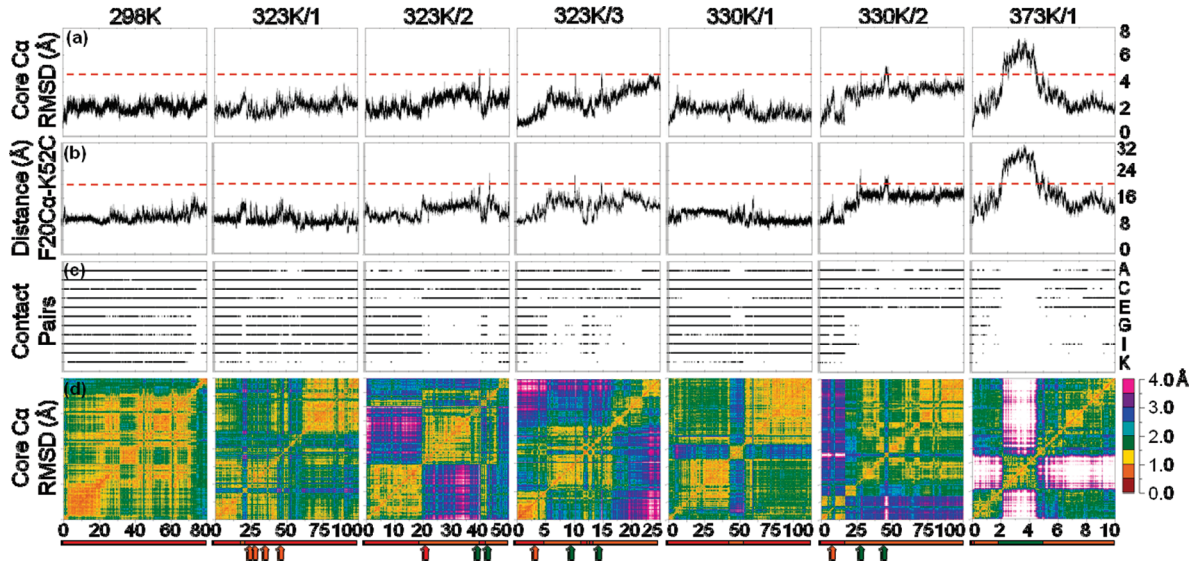


FIGURE 1: General properties for each simulation. (a) Cα rmsd of the core (residues 8−53) calculated over time for each simulation relative to the 0 ns structure. Values above ~4.5 Å are indicative of movement to D, as indicated by the dashed red line. (b) Distance between the Cα of Phe 20 and the backbone C of Lys 52 over time. N′ is characterized by values of ~15 Å, while distances of greater than 20 Å (dashed red line) appear when the protein moves to D. (c) Contacts made between residue pairs. Alternate pairs are labeled at the right from top to bottom: (A) Ile 45−Leu 38, (B) Ile 45−Leu 40, (C) Trp 48−Leu 16, (D) Phe 49−Leu 16, (E) Phe 49−Phe 20, (F) Phe 49−Arg 24, (G) Phe 49−Leu 26, (H) Lys 52−Phe 20, (I) Arg 53−Phe 20, (J) Arg 53−Arg 24, and (K) Arg 53−Leu 26. Contacts 49−24, 49−26, 52−20, and 53−24 are characteristically present in N but not N′, while additional contacts are lost during D. (d) All-vs-all core Cα rmsd matrix. Low-core Cα rmsd squares on the diagonal represent a period of time with similar structures, and when they are off the diagonal, they indicate that the structures from the two corresponding time periods are similar. Below each matrix is a timeline depicting the different states the protein assumes in each simulation: N (red), N′ (orange), and D (green). Arrows represent transiently occupied states.

of points close together indicates that the structures are similar. Using this plot, TSEs were chosen as the last structures leaving the extended native cluster for an unfolding event (*5*) or the first structures upon returning to the native cluster for a refolding event. The TSE was defined as the point of cluster exit and previous 5 ps for unfolding (*5*) and as the cluster re-entry and succeeding 5 ps for refolding.

HIII−Core Distance Calculation. The distance between the closest backbone atoms in the C-terminus of HIII and the HI−HII scaffold in the crystal structure was chosen to represent the movement of HIII. The atoms chosen were the Cα of Phe 20 and the backbone carbonyl C of Lys 52, and the distance was measured at 10 ps granularity. Since the 373K/1 simulation was so short, a granularity of 1 ps was used to give consistent sampling.

Average Structures. Average structures were calculated using 100 ps granularity for the long N and N′ time spans. For TS structures, all six structures in the TSE were included in the average. The Cα rmsd of the core residues (8−53) between the average structures was then calculated.

HIII−Core Contacts. A contact for a pair of residues was defined on the basis of whether any one of the heavy atoms in the first residue was below a set cutoff of any heavy atom in the second residue. This cutoff was defined as 5.4 Å for carbon−carbon distances and 4.6 Å for all other atom pairs. For Figure 4, the calculation was taken over the time period of interest with 1 ps granularity, and the percentage of structures in which the two specified residues were in contact was reported for the average measurements. For the whole-simulation graphs (Figure 1c), each of the contacts is listed as a separate horizontal line, and if the contact was present at the given time point along the *x*-axis, a cross (+) was plotted. Ten picosecond granularity was used.

To choose which contacts to report, we identified residue pairs in which one member of the pair was in HIII and the other was not. Of these, the only pairs that were selected were those that were in contact at least 25% of the time in the native (298 K) simulation.

Calculation of S Values. The *S* value is a semiquantitative structure index that provides an overall measure of the secondary and tertiary structure for each residue of the protein (*6*). *S* values agree well with experimental Φ values for a variety of proteins (*6, 9, 22–25*). The *S* value is the product of the extent of native secondary structure ($S_{2°}$) and native and nonnative tertiary contacts ($S_{3°}$) present in a given structure relative to the number of contacts in the crystal structure. A value of 1 for *S* corresponds to a nativelike extent of structure in the TS, while a value of 0 suggests the residue is unstructured. As previously described (*9*), $S_{3°}$ was used in place of *S* for residues Phe 8, Leu 26, and Leu 40. For these three residues, side chain interactions were maintained despite disorder in the main chain. Consequently, the product of $S_{2°}$ and $S_{3°}$ did not accurately represent the degree of structure retention.

Protein−DNA Interactions. To generate a semiquantitative measure of whether MD-generated En-HD conformers bind DNA, we measured distances between the DNA sugar−phosphate backbone and the HI−HII helical hairpin using the crystal structure of En-HD bound to DNA (PDB entry 3HDD). The crystal structure contains two nearly identical En-HD structures (core Cα rmsd = 0.27 Å), so we selected the one bound to the ideal TAATTA sequence for all measurements (*26*). En-HD binds DNA primarily through residues in HIII (major groove) and the N-terminus (minor groove) (*26, 27*). Since the N-terminus becomes structured only upon binding DNA (*26*), its conformation during our simulations should have no bearing on whether free En-HD

is structured enough to bind DNA. Using ProFit (*28*), the MD structures were first aligned with the DNA-bound structure based on a least-squares fit of the Cα atoms in HIII, the DNA-binding helix. We selected pairs of residues for the distance measurements, representative of one of two hydrogen bonds in the DNA-bound crystal structure that did not involve HIII or the N-terminus of En-HD. The atoms chosen for the measurement were the Cα of Tyr 25 from En-HD and the backbone P of thymine 28 from the DNA. For measurements over time, 10 ps granularity was used. A period of time beginning 1 ns after the TS and continuing for 1 ns was selected to represent D for all four high-temperature unfolding simulations. Since there was not a full nanosecond of denatured time for most of the four lower-temperature simulations, the most denatured structure based on 3D MDS of the Cα rmsd matrix was used.

Protein Images and Figures. All protein images were rendered using VMD (*29*), Cα rmsd matrices were made using R (*21*), and 3D MDS images were rendered with Chimera (*30*). Graphs were plotted and rendered in Gnuplot (*31*).

RESULTS

A total of 10 MD simulations were performed at five different temperatures (298, 323, 330, 373, and 498 K). We describe the major conformational states of En-HD in each of the 10 simulations: N, N′, TS, and D. When En-HD was in N, HIII was docked against the HI−HII scaffold. N′ was characterized by a slight movement of HIII toward the N-terminus without losing many contacts or solvating the hydrophobic core. When HIII moved out and away from the HI−HII scaffold, exposing the hydrophobic core, En-HD was deemed to be in D. The protein was not necessarily unfolded in D, but it was not native or, by definition, biologically active. Whenever the protein moved from N′ to D or from D back to N′, a TS was identified. These four states will be discussed further in the context of each simulation.

Overview of Simulations.

(i) 298 K. En-HD remained folded in the native 298 K simulation with a core Cα rmsd of 2.1 ± 0.3 Å (average ± one standard deviation) and an HIII−core distance (Phe20 Cα−Lys52 carbonyl C; see Methods) of 10.6 ± 1.4 Å for the first 80 ns of the 100 ns simulation (Figure 1a,b). During the final 20 ns of the simulation, the nine C-terminal residues formed a π-helix but HIII remained docked to the HI−HII scaffold. For this reason, the final 20 ns is not considered here.

(ii) 323K/1. In this 100 ns simulation, En-HD stayed mostly in N but briefly moved to N′ from 18 to 23 ns. It transiently moved to N′ approximately four more times over the next 30 ns and then stabilized in N for the remainder of the simulation (Figure 1). The protein did not populate D during this simulation, so there were no TSs identified.

(iii) 323K/2. En-HD was in N for the first ∼20 ns. At 19 ns, HIII moved ∼10 Å toward the N-terminus, entering N′ (Figure 3a). It remained in N′ until 39 ns when there was a large jump in the core Cα rmsd and HIII−core distance, reflecting the undocking of HIII and entrance into D. HIII moved sufficiently far away from the core (20 Å) for it to lose 10 of its 11 native core contacts (Figure 1c) and for the hydrophobic core to be solvated. This altered position was only transient, however, and HIII moved back to its position in N′ a short 0.28 ns later. Over the next 1 ns, HIII moved back to its N position where it stayed for ∼3 ns. HIII then returned to its N′ position transiently before the protein once again entered D at 43 ns. This transition was marked by another jump in the core Cα rmsd, a jump in the HIII−core distance, and loss of contacts (Figure 1). After 0.16 ns, the protein returned to N′, where it remained for the duration of the 50 ns simulation. The structures of En-HD in all four of its states are shown in Figure 2, and these structures are representative of those seen in the other six simulations.

(iv) 323K/3. The first 5 ns of this simulation was spent mostly in N, after which En-HD stabilized in N′ for an additional 5 ns. There was a transient movement to D at 10 ns, indicated by a spike in core Cα rmsd and HIII−core

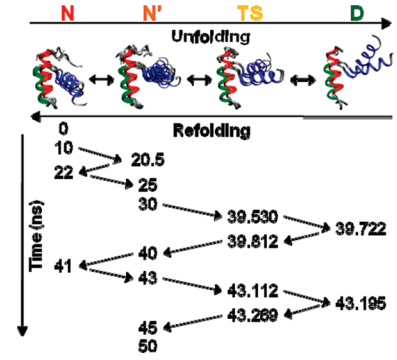


FIGURE 2: Structures and order of population of the four states in the 323K/2 simulation. Under each state is a set of representative structures from the course of the simulation overlaid and fit on HI−HII Cα atoms. HI (residues 8−20) is colored red, HII (residues 26−36) green, and HIII (residues 42−55) blue. The time point in nanoseconds of each structure is listed below, and arrows connect them in the order they occurred. Structures within each state are similar, while each state is distinct. In N and N′, HIII forms an ∼15° angle with the HI−HII scaffold. When En-HD reaches the TS, the angle is ∼30°, and it becomes even wider in D.

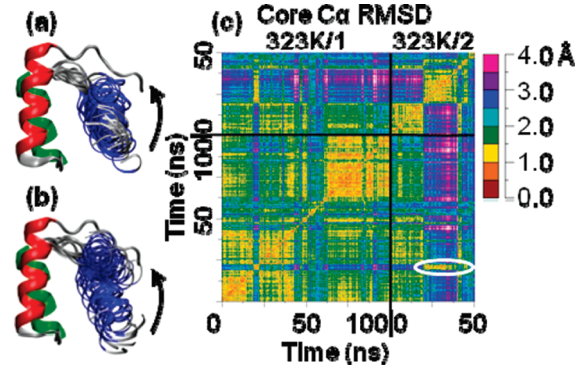


FIGURE 3: N and N′ structures and all-vs-all core Cα rmsd matrix for two 323 K simulations. (a) N and N′ from the first 38 ns of the 323K/2 simulation taken every 0.5 ns for HIII. (b) N and N′ from the first 100 ns of the 323K/1 simulation with structures taken every 1 ns for HIII. HI−HII is from the 0 ns structure. HIII moves toward the N-terminus but not out, so the hydrophobic core is not solvated. (c) All-vs-all core Cα rmsd matrix for the 323K/1 and 323K/2 simulations. 323K/1 and 323K/2 are in N′ for the longest time from 19 to 20 ns and from 20 to 39 ns, respectively. The color of the circled, low-core Cα rmsd box off the diagonal indicates that these two N′ states are as similar to each other as they are to themselves. Smaller boxes can be seen off the diagonal for the points in the 323K/1 simulation at which En-HD moves to N′ transiently indicating these transient N′ states are the same as the longer two.

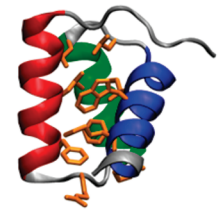
(a) Fraction of Time in Contact: N and N′

Res Pair	N 298K	N′ 323K 1	323K 2	323K 3	330K 1	330K 2	330K 2	330K 2	373K 1	373K 1	N′ Avg	N′ SD
	0-80	18-23	20-39	5-10	15-25	43-53	17-27	29-43	48-100	5-15	6-10	
45-38	0.87	0.92	0.79	0.44	0.32	0.00	0.74	0.42	0.55	0.10	0.27	0.46	0.30
45-40	0.84	0.97	0.99	1.00	0.99	0.99	1.00	1.00	0.99	1.00	0.96	0.99	0.01
48-16	0.58	0.00	0.44	0.69	0.37	0.00	0.12	0.01	0.06	0.68	0.40	0.28	0.27
49-16	0.53	0.62	0.86	0.06	0.49	0.00	0.88	0.95	0.95	0.01	0.51	0.53	0.39
49-20	0.80	0.96	0.95	0.80	0.91	0.25	0.95	0.48	0.59	0.62	0.91	0.74	0.25
49-24	0.46	0.04	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.10	0.01	0.02	0.03
49-26	0.80	0.29	0.01	0.05	0.00	0.12	0.01	0.00	0.00	0.40	0.01	0.09	0.14
52-20	0.31	0.20	0.01	0.04	0.01	0.69	0.00	0.00	0.00	0.17	0.00	0.11	0.22
53-20	0.75	0.66	0.52	0.02	0.01	0.82	0.47	0.00	0.00	0.04	0.20	0.27	0.31
53-24	0.52	0.09	0.01	0.00	0.00	0.28	0.00	0.00	0.00	0.02	0.01	0.04	0.09
53-26	0.60	0.03	0.00	0.00	0.00	0.28	0.00	0.00	0.00	0.01	0.00	0.03	0.09
Avg	0.64	0.43	0.42	0.28	0.28	0.31	0.38	0.26	0.29	0.29	0.30	0.32	0.06
% 298	1.00	0.62	0.58	0.39	0.39	0.54	0.52	0.37	0.40	0.42	0.41	0.47	0.09

(b) Fraction of Time in Contact: New and Known TSEs

Res Pair	Known TSs Avg	Known TSs SD	323K/2 39ns U	323K/2 39ns R	323K/2 43ns U	323K/2 43ns R	323K/3 10ns U	323K/3 10ns R	323K/3 14ns U	323K/3 14ns R	330K/2 27ns U	330K/2 28ns R	330K/2 44ns U	330K/2 47ns R	373K 5ns R	New TSs Avg	New TSs SD
45-38	0.15	0.22	0.00	0.00	0.00	0.00	0.33	0.00	0.83	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.24
45-40	0.91	0.18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00
48-16	0.46	0.35	0.00	0.00	0.00	0.00	0.17	0.50	0.83	1.00	0.00	0.00	0.00	0.00	0.00	0.19	0.35
49-16	0.00	0.00	0.12	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.33	1.00	0.00	0.20	0.37
49-20	0.60	0.27	0.88	0.33	0.00	0.00	0.83	1.00	0.50	0.83	1.00	0.00	0.00	1.00	0.00	0.49	0.45
49-24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
49-26	0.42	0.47	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05
52-20	0.13	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.09
53-20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
53-24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
53-26	0.06	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg	0.25	0.17	0.18	0.14	0.09	0.09	0.23	0.23	0.32	0.26	0.18	0.18	0.12	0.27	0.09	0.18	0.07
% 298	0.35	0.26	0.23	0.17	0.11	0.11	0.28	0.30	0.48	0.36	0.22	0.28	0.16	0.39	0.11	0.25	0.12

(c)



Leu 16 Ile 45
Phe 20 Trp 48
Arg 24 Phe 49
Leu 26 Lys 52
Leu 38 Arg 53
Leu 40

FIGURE 4: HIII−core residue pair fraction of time in contact for N, N′, and new TSEs in each simulation. Entries in panels a and b are colored pink for <25%, purple for 25−75%, and blue for >75%. (a) Contacts for N in the 298 K simulation and N′ in the 323, 330, and 373 K simulations. The temperature of the simulation, simulation number, and time span in nanoseconds of the N′ state are given. (b) Contacts for new and previously identified TSEs. The simulation temperature and number are given (373 K is 373K/1) as well as the time in nanoseconds during which the unfolding (U) or refolding (R) TSE occurred. The average fraction of time in contact was reported for each simulation as was the fraction of time in contact relative to the native (298 K) simulation. Additionally, the average and standard deviation of the fraction of time in contact were calculated for each contact pair across all compared time spans. The average fraction of contacts is quite different for each of the states, N, N′, and TS, with a low standard deviation, and those contacts that are lost are lost fairly consistently. (c) Each of the selected contact residues is shown on the En-HD structure colored orange.

distance. There was a transition back to N for 2 ns with a transient jump to N′ during that time. Then at 13 ns, there was a more stable transition to N′, interrupted only by a transient jump to D at 14 ns. En-HD remained in N′ until the end of the simulation. Only the first 25 ns of the 42 ns simulation is considered here to focus on the transitions of interest.

(v) 330K/1. En-HD was stable in N for the majority of this simulation, with the exception of the period from 43 to 53 ns (Figure 1). During this time, HIII unwound slightly at its N-terminus, causing it to move toward its N-terminus by a register, much like the movement previously described for N′. Because the movement was due to unwinding rather than a simple loop movement, N′ in this simulation had somewhat different properties (core Cα rmsd HIII−core and contact pattern) than N′ in the 323 K simulations (Figure 1).

(vi) 330K/2. This simulation began with 8 ns of N, moved to N′ for 1 ns, returned to N for 8 ns, and then shifted to N′ again. At 27 ns, the protein moved from N′ to D for 0.33 ns and then returned to N′ at 28 ns. Again, at 44 ns, En-HD moved from N′ to D and stayed there until 47 ns when it returned to N′. It remained in N′ for the duration of the 100 ns simulation (Figure 1). Each of the three nontransient N′ states here agrees very well with N′ in the 323 K simulations as shown in Figure 2.

(vii) 373K/1. En-HD moved from N to N′ within 0.5 ns and then proceeded to D. An unfolding TS was previously identified at 1.720 ns (*9, 12, 13*), and we identified a new refolding TS shortly after 5 ns. After En-HD returned to N′, it remained there for the remainder of the 24 ns simulation. We will focus only on the first 10 ns of the simulation here. After refolding, N′ did not agree very well with N′ in the other simulations near the T_m, due to the fluctuating π-helical structure adopted by the C-terminus of HIII around 7 ns (data not shown). The temperature of this simulation was nearly 40 K above the T_m of En-HD, so refolding is expected to occur only rarely and incompletely. We were lucky to observe such an event so that we can compare folding and unfolding within a continuous trajectory at high temperature.

Native′ State at Elevated Temperatures. Before HIII moved sufficiently far from the HI−HII scaffold for En-HD to become "denatured", it occupied two distinct positions. The conformation of En-HD with HIII positioned in a manner most similar to that of the crystal structure is N, and we define N′ as the state in which HIII slides ∼10 Å toward the N-terminus (Figure 3a,b).

In all three of the 323 K simulations and the second 330 K simulation, the same N′ structures were observed on the basis of core Cα rmsd and HIII−core contact similarity (Figures 3c, 4a, and 5a). On the basis of the core Cα rmsd matrix of the first two 323 K simulations (Figure 3c), the

(a) Core Cα RMSD for N and N' Average Structures

		N		N'									
		min	298K 0-80	323K 1 18-23	323K 2 20-39	323K 3 5-10	323K 3 15-25	330K 1 43-53	330K 2 17-27	330K 2 28-43	330K 2 48-100	373K 1 5-1.5	373K 1 6-10
N	min	0.00	1.82	2.22	2.72	2.41	3.10	2.20	2.68	3.10	3.22	2.05	2.22
N	298K 0-80ns		0.00	1.69	2.48	2.23	2.82	2.26	2.41	2.62	2.70	1.91	3.02
N'	323K/1 18-23ns			0.00	0.96	1.22	1.60	1.77	1.01	1.57	1.58	1.53	2.82
N'	323K/2 20-39ns				0.00	1.57	1.27	2.26	0.60	1.39	1.28	2.02	2.90
N'	323K/3 5-10ns					0.00	1.68	1.74	1.66	1.72	1.86	1.09	2.94
N'	323K/3 15-25ns						0.00	2.54	1.36	1.35	1.26	2.20	3.25
N'	330K/1 43-53ns							0.00	2.38	2.61	2.67	1.86	2.80
N'	330K/2 17-27ns								0.00	1.30	1.28	1.98	2.80
N'	330K/2 28-43ns									0.00	0.40	2.12	3.10
N'	330K/2 48-100ns										0.00	2.28	3.19
N'	373K/1 0.5-1.5ns											0.00	2.79
N'	373K/1 6-10ns												0.00

(b) Core Cα RMSD for New and Known TSE Average Structures

		N	New TS												Known TS			
		min	323K 2 39 U	323K 2 39 R	323K 2 43 U	323K 2 43 R	323K 3 10 U	323K 3 10 R	323K 3 14 U	323K 3 14 R	330K 2 27 U	330K 2 28 R	330K 2 44 U	330K 2 47 R	373K 1 5 R	373K 1 1 U	373K 2 1 U	498K 1 0 U	498K 2 0 U
N	min	0.00	3.46	3.35	3.17	3.50	2.58	2.60	3.02	2.86	3.30	3.62	4.10	3.01	3.39	2.53	2.68	2.86	2.93
New TS	323K/2 39ns Unfold		0.00	0.65	2.27	2.53	2.70	2.83	2.74	1.77	2.55	1.99	2.20	2.35	3.16	2.00	3.10	4.14	3.95
New TS	323K/2 39ns Refold			0.00	2.09	2.32	2.49	2.63	2.45	1.41	2.42	1.86	2.21	2.24	3.00	1.89	3.02	3.84	3.63
New TS	323K/2 43ns Unfold				0.00	0.65	1.31	1.35	1.74	1.84	1.11	1.74	2.14	2.01	1.91	1.70	2.22	2.72	2.90
New TS	323K/2 43ns Refold					0.00	1.39	1.36	1.74	1.99	1.35	2.01	2.30	2.34	2.23	1.87	2.54	2.72	3.01
New TS	323K/3 10ns Unfold						0.00	0.74	1.30	1.84	1.67	2.36	2.72	2.08	2.24	1.55	2.36	2.35	2.72
New TS	323K/3 10ns Refold							0.00	1.56	1.92	1.81	2.51	2.79	2.25	2.41	1.46	2.14	2.09	2.63
New TS	323K/3 14ns Unfold								0.00	1.57	2.18	2.41	2.71	2.23	2.70	1.92	2.50	2.65	2.97
New TS	323K/3 14ns Refold									0.00	2.34	2.00	2.56	2.29	2.85	1.39	2.52	3.12	3.07
New TS	330K/2 27ns Unfold										0.00	1.48	2.17	1.90	1.67	2.16	2.66	2.93	3.14
New TS	330K/2 28ns Refold											0.00	1.75	1.96	2.03	2.37	3.00	3.66	3.60
New TS	330K/2 44ns Unfold												0.00	1.77	2.65	2.60	3.14	3.93	3.95
New TS	330K/2 47ns Refold													0.00	2.15	2.28	2.63	3.21	3.29
New TS	373K/1 5ns Refold														0.00	2.71	2.94	3.08	3.06
Known TS	373K/1 1ns Unfold															0.00	2.24	2.84	2.92
Known TS	373K/2 1ns Unfold																0.00	2.45	2.80
Known TS	498K/1 0ns Unfold																	0.00	1.84
Known TS	498K/2 0ns Unfold																		0.00

FIGURE 5: Matrix of core Cα rmsds among average structures representative of N, N′, and TS. The rmsds reported are for the core Cα residues of the average structure over the time periods indicated and are colored to visualize trends 0−1 Å (red), 1−2 Å (orange), 2−3 Å (yellow), 3−4 Å (green), and 4−5 Å (blue). (a) Core Cα rmsds for average N′ structures. The simulation temperature, number, and time span in nanoseconds of N′ are given. With the exception of the later 373K/1 N′ structure, all of the N′ structures are within a 3 Å core Cα rmsd of each other, and with the exception of 330K/1, all of the low-temperature N′ structures are within 2 Å. (b) Core Cα rmsd for the 13 new average TSE structures and four previously identified. The simulation temperature, number, and period in nanoseconds during which the unfolding (U) or refolding (R) TSE occurred are given. The average TSE structures are most similar to other TSE structures at the same temperatures, and all are approximately the same core Cα rmsd from N (~2.5−3.5 Å).

conformations from 18 to 23 ns and the subsequent transient deviations from N in 323K/1 were the same as those from 20 to 40 ns and from 42 to 50 ns (with the exception of the unfolding events), respectively, in 323K/2. The core Cα rmsd between average structures of N′ in the 323K/1−3 and 330K/2 simulations also showed this similarity (Figure 5a). The 0.5−1.5 ns N′ in 373K/1 was in good agreement with N′ in 323K/1−3 and 330K/2 based on the core Cα rmsd, but the 6−10 ns N′ was less similar due to disruption of HIII.

The fraction of time the HIII−core contacts were made was remarkably similar for all 10 nontransient N′ conforma-

tions, with the exception of 330K/1 (Figure 4a). The average standard deviation for contact time over all 11 residue pairs was 19%, and if the 330K/1 simulation was excluded, it dropped to 15% (Figure 4). Contact versus time plots are useful for probing which pairs are in contact over the course of a simulation (Figure 1c). On the basis of this, contacts Phe 49−Arg 24 (F), Phe 49−Leu 26 (G), Lys 52−Phe 20 (H), and Arg 53−Arg 24 (K) were lost when En-HD moved from N to N′ and were mostly regained if it moved back to N from N′.

Properties of the Transition State Ensembles. A total of 13 new unfolding and refolding TSEs were identified in four simulations at 323, 330, and 373 K. The TSEs all had a low core Cα rmsd from each other (2.1 ± 1.0 Å), particularly the T_m TSEs [1.7 ± 0.9 Å (Figure 5b)]. The T_m TSEs were less similar to the high-temperature TSEs, with core Cα rmsds from the 373 K TSEs lower than those from the 498 K TSEs, but most were 2−4 Å core Cα rmsd between any two high-temperature and T_m TSE average structures. The core Cα rmsd from the native state was 3.1 ± 0.4 Å over all 17 TSEs, and the lowest core Cα rmsds were observed between the exit and re-entry TSEs at 39 and 43 ns in the 323K/2 simulation (0.65 Å in both cases).

The HIII−core contacts agreed very well between the new and previously identified TSEs (Figure 4b). The Ile 45−Leu 40 contact was consistently sustained in all of the TSEs, likely due to the residues' positions in the HII−HIII loop and at the N-terminal end of HIII, respectively. Where there were dissimilarities in contacts made between the known and new TSEs, it was usually the case that there were fewer contacts made in the new, lower-temperature TSEs.

There was a pattern of gain and loss of HIII−core contacts preceding and following the TSEs across the simulations. There were six residues involved in contacts characteristically lost in N′ (Phe 20, Arg 24, Leu 26, Phe 49, Lys 52, and Arg 53), and aside from the four pairs that lost contact in N′, these six residues were also involved in three more contacts: Phe 49−Phe 20, Arg 53−Phe 20, and Arg 53−Leu 26. Of these three pairs, the 53−26 contact was lost early in all six simulations at elevated temperatures (Figure 1c). In the case of the 12 new TSEs in the 323 and 330 K simulations, the 49−20 contact was lost within 1 ns before or immediately after the exit TS, and it was reformed within 0.1 ns following the re-entry TSs. The 53−20 contact was also lost at the same time (with one exception, 323K/3, 10 ns), and it was regained within 1 ns of all four re-entry TSs in the 323 K simulations; however, it never re-formed after the first re-entry TS in the 330K/1 simulation. In the 373K/1 simulation, the 49−20 pair was lost 0.5 ns after the exit TS and regained 0.1 ns after the re-entry TS, and the 53−20 pair was lost 1.5 ns before the exit TS and regained 1 ns after re-entry.

S values, which quantify local structure in TSEs, were calculated for the 13 TSEs and compared with experimentally determined Φ values. The *S* and Φ values agree very well for 10 of the 13 new TSEs, with linear correlation coefficients (*R*) ranging from 0.71 to 0.86 (Figure 6). The two 330K/2 44−47 ns TSEs did not agree well with experiment (*S* vs Φ, $R = 0.10$ and 0.03). Loss of secondary structure in the termini of HIII explains some of the disagreement. For example, Ala 43 from the N-terminal end of HIII completely lost its secondary structure, giving it an *S* value of 0 rather than its Φ value of 1.05. The 373K/1 5 ns re-entry TS had

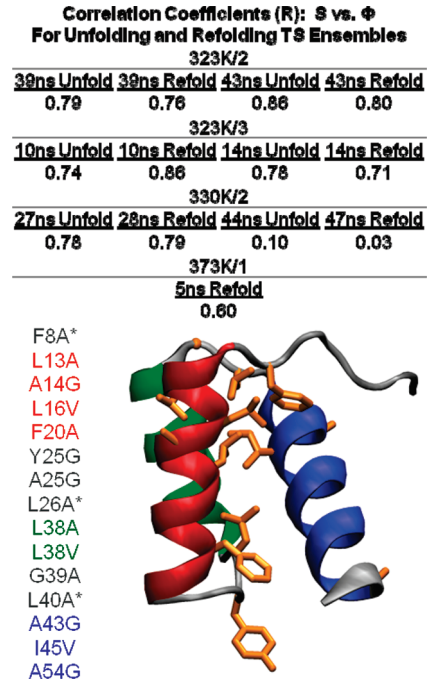


FIGURE 6: *S* values for the 13 new TSEs. The correlation between calculated *S* values and experimentally determined Φ values is quite good for the first 10 TSEs. The *S* and Φ values do not agree as well for the 330K/2 44−47 ns or 373K/1 TSEs, which is due to loss of secondary structure or altered packing. For the three residues denoted with asterisks, only S_{3_0} was reported (see Methods).

a slightly lower correlation coefficient ($R = 0.60$), with the largest disagreements seen for residues Leu 26, Leu 38, and Gly 39 which are located at the ends of HII. The secondary structure was as expected for these residues (turn, helix, and helix, respectively), so the discrepancy is due to altered packing.

Protein−DNA Interactions. When En-HD can no longer bind DNA, it loses its biological activity and is therefore denatured, although it may not be unfolded. Consequently, it is informative to determine whether the structures we identified as N, N′, and D can bind DNA. In an effort to quantify En-HD's DNA binding ability, we fit En-HD to DNA on the basis of the Cα atoms of the major binding helix (HIII) and assessed the extent to which the other binding interactions could be formed. Tyr 25 is not in HIII or the unstructured N-terminus, and it makes a hydrogen bond to the DNA backbone (*26, 27*). Since we forced HIII to bind, this residue was selected for distance calculations. Across the simulations, the Tyr 25−thymine 28 distances show a general trend of being longer for D than N, but the distance is also longer in N′ than N. Structures from the 323K/2 simulation and the Tyr 25−thymine 28 distance are given in Figure 9 and are representative of what we saw in the other simulations.

Unfolding and Refolding Pathways. There were seven unfolding and refolding events identified in four simulations during the following times: 39 and 43 ns in 323K/2, 10 and 14 ns in 323K/3, 27−28 and 44−47 ns in 330K/2, and 1−5 ns in 373K/1. Comparison of the structures indicates that the protein moved through the same conformations when it refolded as when it unfolded for a given unfolding−refolding event. This path retracing can be seen as an "X" on the core Cα rmsd matrix and as overlaid paths in the 3D projection of the matrix (Figure 7). As the protein moved from its most

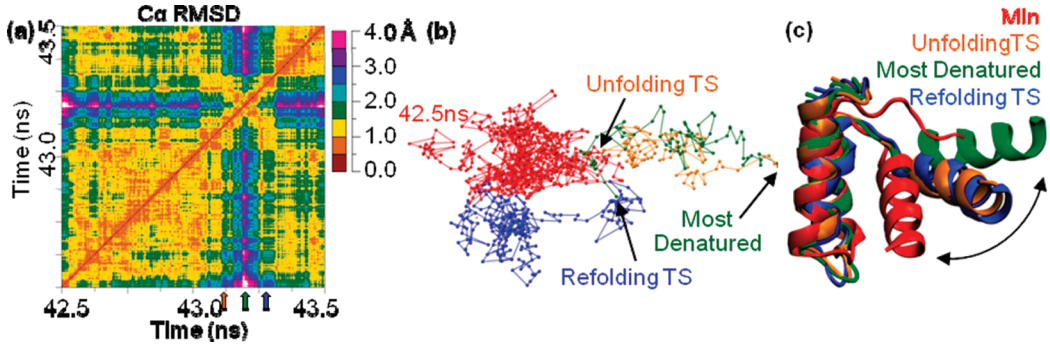


FIGURE 7: Comparison of 323K/2 43 ns unfolding and refolding TSEs and their pathway. (a) All-vs-all Cα rmsd matrix showing the unfolding TS (43.112 ns, orange arrow), the most denatured state (43.195 ns, green arrow), and the refolding TS (43.269 ns, blue arrow). The visible X is evidence that the conformations the protein takes as it leaves N′ to go to D are the same as those it takes when it returns to D, but in the reverse order. (b) 3D MDS projection of the matrix from panel a in which each of the points represents a structure and the distance between any two points is proportional to the Cα rmsd between the respective structures. The colors denote different periods in time: from 42.5 ns to the unfolding TS (red), from the unfolding TS to the most denatured conformation (orange), from the most denatured to the refolding TS (green), and from refolding to 43.5 ns (blue). That the paths that the protein followed as it moved from N′ to D and back are overlaid in the 3D projection indicates the conformations the protein took were very similar and in the reverse order. (c) Structures of the unfolding TS (orange), refolding TS (blue), most denatured conformation (green), and starting minimized structure (red). The structures were fit on the basis of the Cα atoms of the HI−HII scaffold. The two TS structures are nearly identical, while they are distinct from both N and D.

denatured point back to N′, a line perpendicular to the diagonal of the matrix is apparent. This line represents a series of conformations that is very similar to the series of conformations through which the protein passed just previously, but in the reverse order. In the 3D projection, this same phenomenon is seen as overlapping points along the path from N′ to the most denatured point, and then back to N′. This evidence was present to different extents for each of the seven unfolding−refolding events, but it is the most striking for 323K/2 at 39 and 43 ns, 323K/3 at 10 ns, 330K/2 at 27−28 ns, and 373K/1 at 1−5 ns.

While N′ is not in the denatured ensemble, it is distinct from N and, thus, a low-energy pathway to move between the two states must exist. As with the unfolding and refolding pathways, there was an X on the core Cα rmsd matrix when En-HD transiently moved from N′ to N and back. In the 323K/3 simulation, there were two transient N → N′ → N movements. There was an X visible on the core Cα rmsd matrix around both N → N′ → N transitions, and there was a third X off the diagonal, around the intersection of the times corresponding to both of the individual N → N′ → N transitions (Figure 8). This third X indicates not only that the N → N′ and N′ → N pathways were very similar for a single transition but also that the two N′ → N → N′ transitions were almost equally as similar.

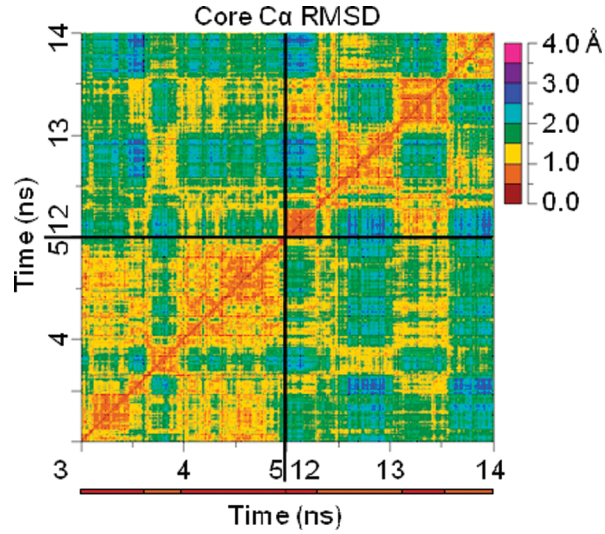


FIGURE 8: All-vs-all core Cα rmsd matrix of two transient N → N′ → N transitions from 323K/3. Much like for the N′ → D → N′ transition in the 323K/2 43 ns TS, an X is visible on the core Cα rmsd matrix for both N → N′ → N transitions. In this case, there is a third X that is apparent off the diagonal at the intersection of the times at which the other two X's occur. This suggests not only that the N ↔ N′ paths are very similar but also that the N → N′ → N path from the first transition is remarkably similar to that from the second.

DISCUSSION

Three different conformational states were populated by En-HD in our seven simulations: N, N′, and D. The protein's state was determined on the basis of a combination of measurements: core Cα rmsd, HIII−core distance, and HIII−core contact pattern (Figure 1). Measuring the core Cα rmsd to the minimized crystal structure was the foremost method in determining En-HD's state. The core Cα rmsd generally fluctuated between 1 and 3 Å when the protein was in N and between 2 and 4 Å for N′ (Figures 1a and 5). Values of >4.5 Å usually indicated a departure from N′ to D. The Phe 20 Cα−Lys 52 C distance, representative of the distance between HIII and the HI−HII scaffold, was also a good indicator of state. When the protein was in N, the distance fluctuated around 10 ± 2 Å, while an increase to 14 ± 2 Å indicated movement to N′ (Figure 1b). Distances of >20 Å occurred when En-HD moved to D. Movement between states was more clearly discerned on the basis of the HIII−core distance than on the basis of the core Cα rmsd. There was nearly always a clean jump in distance between the different states, which suggests that N and N′ are distinct states despite both being "native". The pattern of contacts between HIII and the rest of the protein also helped discriminate among different states. The contact pairs selected for analysis were deliberately chosen to be good representatives of N. Jumps in the core Cα rmsd to 2−4 Å and HIII−core distance to ~14 Å, which were characteristic of N′, coincided with the loss of four contact pairs: Phe

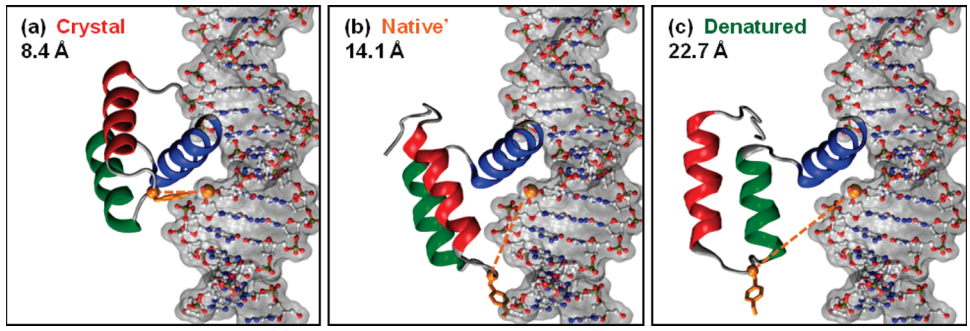


FIGURE 9: Structures from 323K/2 fit to DNA. En-HD colored with Tyr 25 in orange sticks. En-HD's Tyr 25 Cα and the DNA's thymine 28 phosphate are colored orange van der Waals spheres, and the distance between these two atoms is given and marked with an orange dashed line. (a) Crystal structure of En-HD bound to DNA (PDB entry 3HDD). (b) N′ at 30 ns in the 323K/2 simulation. (c) Most denatured state (D) from the 43 ns N′ → D → N′ transition, 43.195 ns. The structures show that En-HD moved slightly away from the DNA in N′ compared to N, and much farther in D.

49−Arg 24, Phe 49−Leu 26, Lys 52−Phe 20, and Arg 53−Arg 24 (Figure 1c).

To estimate the likelihood that the three different En-HD conformations bind DNA, we fit structures from our MD simulations onto the DNA-bound crystal structure and took a distance measurement that might discern between native (N or N′) and D. Representative structures from the 323K/2 simulation and distances are given in Figure 9. Even though both the N′ and D structures were distinct from the crystal structure, N′ conformations could more easily move back to N and be in a position to bind the DNA (and thus be biologically active) than those in D. For N′ → N movement to occur, HI and HII would have to slide along the DNA and HIII so that they could dock against HIII as in N. This movement was what we saw for all N → N′ transitions in our simulations without DNA. For a D structure to move to N, it would have to expel the water from its solvated hydrophobic core before HI and HII could dock back onto HIII and the DNA. Overall, D appears to be too distorted to function properly, and N′ falls somewhere between N and D. While N′ may be able to recover and clamp down on the DNA, it is also possible that it is too distorted and therefore inactive.

Having defined three different states for En-HD in seven independent simulations, we find it interesting to consider the variations within a state and how En-HD passes between them. N′ was observed in all six of the elevated-temperature simulations, but it was more similar in five of them (323K/1−3, 330K/2, and 373K/1) than it was in 330K/1 based on core Cα rmsd and HIII−core contacts (Figures 4a and 5a). Also, the first period of N′ in the 373K/1 simulation matched the first four simulations, while the second period was ambiguous. In all cases, HIII moved a register toward the N-terminus, away from the HI−HII turn, but N′ in 330K/1 and the end of 373K/1 was somewhat different from N′ in the four other simulations and in the beginning of 373K/1, despite having the same overall topology. This difference suggests that there may be multiple, subtly different N′ states.

In the four simulations where there was N′ ↔ D movement, a total of 13 new TSEs were identified. The TSEs within one temperature were most similar, and the best agreement was between the unfolding and refolding TSEs for a single transient unfolding−refolding event (Figure 5b). In these cases, only a small portion of D was sampled so it was likely that the protein would find a refolding path very similar to its unfolding path from the ensemble of paths available.

On the basis of the 11 HIII−core contact pairs selected for analysis, there was good agreement among all 17 TSEs [13 new and four previously identified (Figures 1c and 4b)]. In almost all of the cases where there was disagreement between the new and previously identified TSEs, it was the case that there were more contacts present in the high-temperature TSEs, which suggests they were more nativelike. It is expected that high-temperature TSEs will more closely resemble N than TSEs at the protein's T_m due to Hammond effects. This phenomenon causes the structure of the TSE to become more nativelike upon destabilization, in this case by heat (*22, 32, 33*).

Not only were the 13 new TSEs consistent with those previously identified, but 10 of them were in good agreement with experiment on the basis of comparison of calculated S values to experimental Φ values [$R = 0.71-0.86$ (Figure 6)]. The correlation was somewhat lower ($R = 0.60$) for the 373K/1 refolding TSE and significantly lower ($R = 0.10$ and 0.03) for the 330K/2 44−47 ns TSEs. The 330K/2 44−47 ns unfolding−refolding event followed a cyclical path as it folded and unfolded, yet its unfolding and refolding TSEs were very similar to each other [average structure core Cα rmsd = 1.77 Å (Figure 5b)]. The lack of agreement in this case may be illustrating discrepancies between single-molecule behavior and bulk measurements. That is, our aberrant TSE pair may be extreme members of the much broader ensemble probed experimentally.

The 13 new TSEs generally agreed well with each other, the four previously identified TSEs from high-temperature unfolding simulations, and experiment. The unfolding and refolding TSEs were equally similar, which is evidence that all 17 TSEs come from the same global ensemble of transition states for En-HD folding. Further, our data suggest that the ensemble of paths for unfolding and refolding is also the same, which supports our long-standing contention that protein folding is a microscopically reversible process.

The symmetrical order of contact pair loss and gain upon unfolding and refolding is further evidence that protein folding is microscopically reversible. The six residues involved in the four contact pairs that were characteristically lost in N′ made three additional contacts (Figures 1c and 4). One pair was lost early in the simulations, and the other two,

Phe 49−Phe 20 and Arg 53−Phe 20, were usually lost just before the unfolding TS and regained right after the refolding TS in the seven unfolding−refolding events. These six residues make up the half of the HIII−core contacts closest to the C-terminus (Figure 4c). Arg 24, Leu 26, and Lys 52 lost all of their HIII−core contacts in N′, but Phe 20, Phe 49, and Arg 53 maintained two of the seven original contact pairs. It was not until right after these two contact pairs were lost that En-HD reached its TS and became denatured, and they re-formed right after re-entering N′ from D in most cases. Thus, loss and gain of these six hydrophobic core contacts are critical steps on the N ↔ N′ and N′ ↔ D pathways, based on our 323 and 330 K simulations. This is evidence for microscopic reversibility in protein folding because there is a consistent order of loss of contacts in unfolding that is repeated in reverse order upon refolding. However, the pattern of loss by these six contact pairs was not consistently repeated in the four high-temperature unfolding simulations previously run at 373 and 498 K (data not shown).

The all-vs-all core Cα rmsd matrix and its 3D projection are arguably the best ways to observe the similarity of MD structures over time. Indeed, there was a visible X on the core Cα rmsd matrix for five of the seven unfolding−refolding events. Similarly, the structures from the unfolding TS to the most denatured point back to the refolding TS overlay on the 3D projection of the core Cα rmsd matrix for these five unfolding−refolding events (Figure 7). Both the X and the overlaid paths indicate that the conformations En-HD moved through from the unfolding TS to the most denatured point had a low core Cα rmsd from the conformations En-HD took as it moved back to the refolding TS, but in the reverse order.

X's were also visible for transient movements from N to N′. In 323K/3, not only was the N → N′ path similar to the N′ → N path, but also both N → N′ → N movements followed highly similar paths. There is an X on the diagonal of the all-vs-all core Cα rmsd matrix for each N → N′ → N movement, and there is also an off-diagonal X at the intersection of the times corresponding to each of the N → N′ → N events. While these four N ↔ N′ pathways are not complete folding or unfolding pathways, they are transitions between discrete states along the full folding pathway.

En-HD never moved from N to D without first passing through N′ each of the seven times it unfolded. In fact, in the 323K/2 and 323K/3 simulations, it moved back to N between the two unfolding−refolding events, passing through N′ on the way. On the basis of our simulations, the N ↔ N′ paths were part of the same ensemble as were the N′ ↔ D paths, and N′ is a necessary step between N and D. Together, these findings are evidence that the entire N → N′ → D and D → N′ → N pathways are mirror images of the same process, and thus, protein folding is a microscopically reversible process.

We identified and characterized four distinct states of En-HD (N, N′, TS, and D) which were consistent across simulations at 298, 323, 330, and 373 K. The core Cα rmsd, HIII−core contacts, the HIII−core distance, and predicted DNA binding ability were used to discriminate among the states and place them on the folding pathway. We identified seven transient denaturing events in six simulations and identified 13 new unfolding and refolding TSEs. The 13 new TSEs agreed well with four previously identified TSEs on the basis of the core Cα rmsd and HIII−core contacts as well as with experimental data based on Φ and *S* values. In five of the seven transient denaturing events, the unfolding pathway was nearly identical to the refolding pathway. We also found two N ↔ N′ transitions that followed the same pathway in the folding and unfolding directions for both transitions. These phenomena are evidence that the ensemble of folding and unfolding pathways are one and the same and that protein folding can be a microscopically reversible process.

ACKNOWLEDGMENT

We thank Darwin Alonso, Amanda Jonsson, and Dustin Schaeffer for helpful discussions and technical assistance.

REFERENCES

1. Tolman, R. C. (1925) The Principle of Microscopic Reversibility. *Proc. Natl. Acad. Sci. U.S.A. 11*, 436–439.
2. Westheimer, F. H. (1968) Pseudo-rotation in the hydrolysis of phosphate esters. *Acc. Chem. Res. 1*, 70–78.
3. Minkin, V. I. (1999) Glossary of Terms Used in Theoretical Organic Chemistry. *Pure Appl. Chem. 71*, 1919–1981.
4. Jackson, S. E., el Masry, N., and Fersht, A. R. (1993) Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: A critical test of the protein engineering method of analysis. *Biochemistry 32*, 11270–11278.
5. Li, A., and Daggett, V. (1994) Characterization of the transition state of protein unfolding by use of molecular dynamics: Chymotrypsin inhibitor 2. *Proc. Natl. Acad. Sci. U.S.A. 91*, 10430–10434.
6. Daggett, V., Li, A., Itzhaki, L. S., Otzen, D. E., and Fersht, A. R. (1996) Structure of the transition state for folding of a protein derived from experiment and simulation. *J. Mol. Biol. 257*, 430–440.
7. Day, R., and Daggett, V. (2007) Direct observation of microscopic reversibility in single-molecule protein folding. *J. Mol. Biol. 366*, 677–686.
8. Itzhaki, L. S., Otzen, D. E., and Fersht, A. R. (1995) The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: Evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol. 254*, 260–288.
9. Gianni, S., Guydosh, N. R., Khan, F., Caldas, T. D., Mayor, U., White, G. W., DeMarco, M. L., Daggett, V., and Fersht, A. R. (2003) Unifying features in protein-folding mechanisms. *Proc. Natl. Acad. Sci. U.S.A. 100*, 13286–13291.
10. Mayor, U., Grossmann, J. G., Foster, N. W., Freund, S. M., and Fersht, A. R. (2003) The denatured state of Engrailed Homeodomain under denaturing and native conditions. *J. Mol. Biol. 333*, 977–991.
11. Mayor, U., Guydosh, N. R., Johnson, C. M., Grossmann, J. G., Sato, S., Jas, G. S., Freund, S. M., Alonso, D. O. V., Daggett, V., and Fersht, A. R. (2003) The complete folding pathway of a protein from nanoseconds to microseconds. *Nature 421*, 863–867.
12. Mayor, U., Johnson, C. M., Daggett, V., and Fersht, A. R. (2000) Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc. Natl. Acad. Sci. U.S.A. 97*, 13518–13522.
13. DeMarco, M. L., Alonso, D. O. V., and Daggett, V. (2004) Diffusing and colliding: The atomic level folding/unfolding pathway of a small helical protein. *J. Mol. Biol. 341*, 1109–1124.
14. Beck, D. A., and Daggett, V. (2007) A One-Dimensional Reaction Coordinate for Identification of Transition States from Explicit Solvent Pfold-Like Calculations. *Biophys. J. 93*, 3382–3391.
15. Beck, D. A. C., Alonso, D. O. V., and Daggett, V. (2000−2008) *il*mm, *in lucem* molecular mechanics, University of Washington, Seattle.
16. Levitt, M., Hirshberg, M., Sharon, R., and Daggett, V. (1995) Potential-Energy Function and Parameters for Simulations of the Molecular-Dynamics of Proteins and Nucleic-Acids in Solution. *Comput. Phys. Commun. 91*, 215–231.
17. Beck, D. A. C., and Daggett, V. (2004) Methods for molecular dynamics simulations of protein folding/unfolding in solution. *Methods 34*, 112–120.

18. Levitt, M., Hirshberg, M., Sharon, R., Laidig, K. E., and Daggett, V. (1997) Calibration and testing of a water model for simulation of the molecular dynamics of proteins and nucleic acids in solution. *J. Phys. Chem. B 101*, 5051–5061.
19. Haar, L., Gallagher, J. S., and Kell, G. S. (1984) *NBS/NRC Steam Tables: Thermodynamic and Transport Properties and Computer Programs for Vapor and Liquid States of Water in SI Units*, Hemisphere, Washington, DC.
20. Beck, D. A., Armen, R. S., and Daggett, V. (2005) Cutoff size need not strongly influence molecular dynamics results for solvated polypeptides. *Biochemistry 44*, 609–616.
21. Team, R. C. (2004) R: A language and environment for statistical computing, Royal Foundation for Statistical Computing, Vienna.
22. Daggett, V., Li, A., and Fersht, A. R. (1998) Combined molecular dynamics and φ-value analysis of structure-reactivity relationships in the transition state and unfolding pathway of barnase: Structural basis of Hammond and anti-Hammond effects. *J. Am. Chem. Soc. 120*, 12740–12754.
23. Li, A., and Daggett, V. (1998) Molecular dynamics simulation of the unfolding of barnase: Characterization of the major intermediate. *J. Mol. Biol. 275*, 677–694.
24. Fulton, K. F., Main, E. R., Daggett, V., and Jackson, S. E. (1999) Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J. Mol. Biol. 291*, 445–461.
25. Day, R., and Daggett, V. (2005) Sensitivity of the folding/unfolding transition state ensemble of chymotrypsin inhibitor 2 to changes in temperature and solvent. *Protein Sci. 14*, 1242–1252.
26. Fraenkel, E., Rould, M. A., Chambers, K. A., and Pabo, C. O. (1998) Engrailed homeodomain-DNA complex at 2.2 Å resolution: A detailed view of the interface and comparison with other engrailed structures. *J. Mol. Biol. 284*, 351–361.
27. Kissinger, C. R., Liu, B. S., Martin-Blanco, E., Kornberg, T. B., and Pabo, C. O. (1990) Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: A framework for understanding homeodomain-DNA interactions. *Cell 63*, 579–590.
28. Martin, A. C. R. (1992−2001) *ProFit: Protein Least Squares Fitting*, University College London, London.
29. Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD: Visual Molecular Dynamics. *J. Mol. Graphics 14*, 33–38.
30. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera: A visualization system for exploratory research and analysis. *J. Comput. Chem. 25*, 1605–1612.
31. Williams, T., and Kelley, C. (1986−1993,1998,2004) *Gnuplot* (www.gnuplot.info).
32. Matthews, J. M., and Fersht, A. R. (1995) Exploring the energy surface of protein folding by structure-reactivity relationships and engineered proteins: Observation of Hammond behavior for the gross structure of the transition state and anti-Hammond behavior for structural elements for unfolding/folding of barnase. *Biochemistry 34*, 6805–6814.
33. Day, R., Bennion, B. J., Ham, S., and Daggett, V. (2002) Increasing temperature accelerates protein unfolding without changing the pathway of unfolding. *J. Mol. Biol. 322*, 189–203.

BI800118B