A One-Dimensional Reaction Coordinate for Identification of Transition States from Explicit Solvent P_{fold}-Like Calculations

David A. C. Beck and Valerie Daggett
Department of Bioengineering, University of Washington, Seattle, Washington 98195-5061

ABSTRACT A properly identified transition state ensemble (TSE) in a molecular dynamics (MD) simulation can reveal a tremendous amount about how a protein folds and offer a point of comparison to experimentally derived Φ_F values, which reflect the degree of structure in these transient states. In one such method of TSE identification, dubbed P_{fold}, MD simulations of individual protein structures taken from an unfolding trajectory are used to directly assess an input structure's probability of folding before unfolding, and P_{fold} is, by definition, 0.5 for the TSE. Other, less computationally intensive methods, such as multidimensional scaling (MDS) of the pairwise root mean-squared deviation (RMSD) matrix of the conformations sampled in a thermal unfolding trajectory, have also been used to identify the TSE. Identification of the TSE is made from the original MD simulation without the need to run further simulations. Here we present a P_{fold}-like study and describe methods for identification of the TSE through the derivation of a high fidelity, bounded, one-dimensional reaction coordinate for protein folding. These methods are applied to the engrailed homeodomain. The TSE identified by this approach is essentially identical to the TSE identified previously by MDS of the pairwise RMSD matrix. However, the cost of performing P_{fold}, or even our reduced P_{fold}-like calculations, is at least 36,000 times greater than the MDS method.

INTRODUCTION

Protein folding/unfolding in cooperative systems is typically described as a reaction. As with chemical reactions, the system proceeds along a pathway or pathways that is divided into states such as denatured, intermediate, and native states. The progression from one state to the next requires the system to surmount a free energy barrier that limits the forward and reverse rates. At any given transition state (TS), the probability of a forward or reverse transition is 0.5. Expressed another way, at these critical states along the protein folding/unfolding reaction coordinates, the system exhibits an equivalent propensity to fold and unfold. Because of their vital role in protein folding, TSs and their location on a protein-folding reaction coordinate are of particular interest. As an example of the utility of this information, when a transition state ensemble (TSE) can be accurately identified, it is possible to predict sequence substitutions that increase the folding rate (1), which is also a powerful validation method.

Given the need to structurally characterize TSs of protein folding, several in silico methods have been developed to identify members of the TSE. We developed the first such method (hereafter referred to as our method) for identification of the TS of folding/unfolding from thermal unfolding molecular dynamics (MD) simulations using explicit solvent (2,3). Our approach focuses on the information contained in the pairwise Cα root mean-squared deviation (RMSD) matrix, or the $N \times N$ matrix of RMSD resulting from the best fit of Cα coordinates (4) for N protein structures against all N structures. The N structures are ordered by ascending simulation time. The resulting $N \times N$ matrix is symmetric with a zero diagonal. For plots of a representative pairwise distance matrix, see Fig. 3 in Kazmirski and Daggett (5). The pairwise distance matrix is then reduced to a lower dimensional space, typically three dimensions, by classical multidimensional scaling (MDS) and plotted in the reduced subspace. For depictions of a three-dimensional MDS of the pairwise distance matrix, see Fig. 6 in DeMarco et al. (6). By visual inspection and other clustering methods, e.g., GDBSCAN (7), it is possible to identify the TS of folding/unfolding by locating the first point after the exit from the native-like cluster (2).

Our original work on identification of TSEs and enhanced analyses of property space was done with chymotrypsin inhibitor 2 (CI2) and barnase (2,3,5,8). In fact, the CI2 TS assignment was verified through a simplified P_{fold}-like study (9). More recently, these methods have been applied to a family of three-helix bundles that act as transcription factors —including the ultrafast folding and unfolding, DNA-binding protein, the engrailed homeodomain (EnHD) (6,10–13). EnHD is an ideal system for theoretical studies of protein folding as it exhibits unfolding and folding rates on the timescale of MD simulation, i.e., nanoseconds (ns) to microseconds (μs). EnHD has been well characterized by experiment and theory (6,10–15). TSE predictions of EnHD done in 2000 (10) were shown to be in quantitative agreement with experiment some years later (11). EnHD is also known to fold via an intermediate (10), thereby exhibiting three-state kinetics. The MD-generated intermediate state structures were confirmed by NMR 5 years after their publication (10,16). Early unfolding (i.e., from native to intermediate) is essentially a two-state process with a k_f of 37,500 $\pm$ 1600 s^{-1} at 298 K

Submitted October 31, 2006, and accepted for publication July 16, 2007.

Address reprint requests to Valerie Daggett, Tel.: 206-685-7420; Fax: 206-685-3252; E-mail: daggett@u.washington.edu.

Editor: Kathleen B. Hall.

0006-3495/07/11/3382/10 $2.00

doi: 10.1529/biophysj.106.100149

and a maximum k_f of 51,000 ± 1500 s^{-1} at 315 K (12). The unfolding process is independent of temperature (10,12) and experiments probing both the folding and unfolding directions demonstrate that the pathway is robust (10–12).

A pairwise RMSD matrix for EnHD is shown after MDS reduction to three dimensions in Fig. 1. These data are from the first 500 ps of a previous thermal (498 K) unfolding simulation of EnHD that has been extensively validated against experiment (10–12). Time is indicated by connectivity of successive states and by the coloring of the points from black to white, where black is the beginning of the simulation. The TSE identified by our method is shown with the spheres of larger radius. There are two distinct lobes that correspond to pre-TS and post-TS structures.

In another approach to identify the TS, some have opted to utilize the aforementioned relationship between the probability of folding before unfolding, a quantity dubbed P_{fold}. Structures that fold half the time and unfold half the time have a $P_{fold} = 0.5$ and are therefore members of the TSE. An individual structure's P_{fold} value is calculated by using the structure as input to a number of Bernoulli trials (where the result is either 0 or 1) involving in silico calculations meant to mimic protein folding (17–22). For each structure, the number of trials often ranges from tens to hundreds, e.g., 20 in Shimada and Shakhnovich (22) and 400 in Du et al. (21). The large number of trials is required because the error in Bernoulli sampling is $1/\sqrt{N}$ where N is the number of tests. The input structures for this approach are typically derived from thermal denaturation simulations, as was done in this study. It is not always clear if the calculations meant to mimic protein folding can reliably reproduce realistic pathways for protein folding because they often use implicit solvent methods, discontinuous sampling strategies, or nonphysical dynamics.

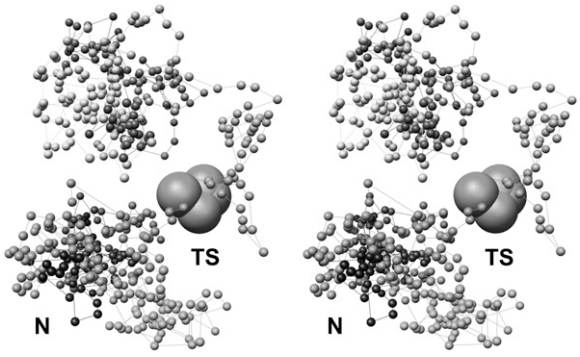

FIGURE 1 MDS of the first 500 ps of the pairwise Cα RMSD matrix for a thermal unfolding simulation of EnHD. Each point represents a structure, with the distance between structures approximating the Cα RMSD between structures. Points are gray-scale coded with ascending time starting from black and finishing at white. The previously identified TS (255–260 ps) is denoted with the large spheres and the designation TS. The native cluster is at the bottom of the figure and bears the designation ''N''.

The intention of this study is to compare the TS from a thermal unfolding simulation of EnHD identified in Mayor et al. (10) by our method with the TS from the same thermal unfolding simulation, identified through P_{fold}-like calculations with explicit solvent. As we believe that realistic simulation of proteins necessitates the inclusion of explicit solvent, we have applied our standard MD simulation methods and protocols (23) to an adapted P_{fold} methodology to reflect this.

Our previous combined theoretical and experimental studies of EnHD validated the thermal unfolding simulations at a variety of temperatures against experiment (10–12). As they have been extensively documented, they provide a reliable base from which to draw structures for further investigations, such as this study. In another previous study of EnHD, the 5 ns structure (10.47 Å Cα RMSD to native) from one of the high temperature (498 K) trajectories was refolded to a structure with analytical properties (e.g., solvent-accessible surface area (SASA), Cα RMSD to crystal, etc.) bounded by the native cluster's property space (23). The successful refolding demonstrates an essential point for a P_{fold}-like study that relies on the ability of software and methods to refold a protein: that our methods are capable of refolding EnHD.

In P_{fold} studies a one-dimensional reaction coordinate is used to evaluate a trajectory and assign to it a measure of how native-like the protein has become. Often, in P_{fold} and other protein folding/unfolding studies, one or two properties such as the fraction of native contacts (Q) or radius of gyration (Rg) are used either directly or together to construct a reaction coordinate (18,24–26). These can be poor choices for some proteins because such properties are neither monotonic nor one-to-one. This problem is well understood and documented (27,28). As an example, consider a protein with a loop that can open and close and acts as a lid for the active site. These opening and closing motions are the mechanism of regulation of access to the active site and are expected in native-state dynamics. In such a protein, a conformation with the lid open may have substantially fewer native contacts when compared to a closed conformation. If Q is used as a reaction coordinate, the protein in the open conformation would appear more denatured than it really is. Conversely, for a protein like EnHD that retains a significant amount of helical structure in its post-TS conformations, Q may overestimate the ''foldedness'' of the protein.

Another common one-dimensional folding metric is an atomic position RMSD (or a derivative quantity) relative to a single crystal or NMR structure. The idea is simple and attractive: the more native-like a protein is, the lower its RMSD. Again, a loop acting as an active-site lid can have a drastic effect on the RMSD—artificially raising it. In some circumstances, it may be possible to remove the lid residues from the RMSD, but this may result in overestimating the ''foldedness'' of post-TS structures. It may also be possible to use alternate structural similarity metrics like the CONGENEAL dissimilarity measure (29). This metric is the sum of

normalized differences between two weighted internal distance matrices each derived from a given protein structure. Each pairwise distance (d_{ij}) is weighted by raising it to the negative second power (i.e., d_{ij}^{-2}), which has the effect of having short distances contribute more than long distances. With such a metric, the hinge motion of a loop would contribute little to the structural dissimilarity. However, for this very reason, these metrics lack a signal/noise ratio appropriate for detecting the subtleties of conformations near the TS.

We believe that any attempt to quantify foldedness by a small number of structural or property space comparisons to a single native microstate conformation will be insufficient for most proteins, particularly those with low free energy barriers to unfolding. The native states of all proteins are made up of microstates that exchange relatively freely between themselves. Therefore, it is unlikely that a single conformation (or property) will be sufficient to describe the entire folded ensemble. Others have recognized this limitation and had success using density-based clustering techniques (30) and nonlinear dimensional reduction techniques such as self-organizing maps and other machine-learning techniques (31–33).

Here we present a property space (5) composed of the most informative 32 properties that we, as a matter of course, use in first-pass analysis of any protein system. The property space is then used to construct a one-dimensional reaction coordinate that is derived not from a single structure, but rather from a reference data set containing nearly 10 million conformations (and associated microstates) and their analytical properties. Unlike the previously mentioned density-based and machine-learning techniques, this approach does not involve conformational clustering. The process outlined for this study is relatively free from the structural biases arising from different protein topologies and can be applied to any given protein to generate a one-dimensional folding metric or reaction coordinate given enough reference data for the native state ensemble.

MATERIALS AND METHODS

MD simulations

Starting structures for the P_{fold}-like simulations were taken from a thermal unfolding simulation that has been described previously (6,10,11). The MDS-assigned TSE for this simulation occurs from 255 to 260 ps. The early folding/unfolding conformations were sampled by taking 22 structures at 20 ps intervals from 0 to 420 ps. More denatured conformations were sampled by using the 5 and 60 ns structures. The 24 structures used for our P_{fold}-like study were simulated using in lucem molecular mechanics (*il*mm) (D. A. C. Beck, D. Alonso, and V. Daggett, University of Washington, 2006) with our standard methods (23) and protocols consistent with those used for the thermal unfolding simulation (6,10,11). For each structure, 20.2 ns simulations were carried out at 298, 310, and 323 K with explicit solvent at experimental densities, 0.997, 0.993, and 0.988 g/ml, respectively (34,35). These temperatures are below the T_m of EnHD, 325.15 K (10), although in the case of 323 K, just barely. This resulted in a total of 72 simulations: three trials for each of the 24 structures. The five simulations started from the 340, 360, 380, 400, 420 ps and the two simulations started from the 5 and 60 ns snapshots were extended an additional 10 ns to improve sampling. The MD timescale of 20.2–30.2 ns was derived from a previous study demonstrating that early refolding events (i.e., those confined to the two-state time regime) of EnHD can occur on this timescale (23).

All but two of the reference simulations used in the construction of the one-dimensional reaction coordinate were conducted in *il*mm with our standard methods and protocols (23). The exceptions were two of the 498 K simulations, which were conducted in ENCAD (36) and described previously (6). The native reference set consisted of 12 simulations totaling 0.8 μs of simulation time at 298, 310, and 323 K. The nonnative reference set consisted of 180 ns of simulation time at 323, 423, and 498 K. In total, the reference data set consisted of 9.8×10^6 samples. The complete list of reference trajectories, their lengths, and combined percentage of NMR nuclear Overhauser effects (NOEs) satisfied can be found in Table 1. In total, this study contains ~1.8 μs of explicit solvent MD.

Property space

The raw coordinate data from the simulations were reduced in dimensionality by the construction of a property space description of the trajectories (5). Individual dimensions in this reduced space are composed of analytical or physical properties of the protein derived from the three-dimensional coordinates, such as Q, secondary structure content of various types, and SASA. Initially, 32 properties were included in the property space: Rg, end-to-end distance, CONGENEAL dissimilarity score to crystal, Cα-RMSD to crystal, Cα-RMSD of residues 6–51 of crystal, main-chain (MC) SASA, side-chain (SC) SASA, polar SASA, nonpolar SASA, MC polar SASA, SC polar SASA, MC nonpolar SASA, SC nonpolar SASA, total SASA, total native contacts, native MC to MC contacts, native MC to SC contacts, native SC to SC contacts, nonnative MC to MC contacts, nonnative MC to SC contacts, nonnative SC to SC contacts, total nonnative contacts, intramolecular polar contacts, intramolecular hydrophobic contacts, intramolecular polar to nonpolar atom contacts, intermolecular polar contacts, intermolecular polar to nonpolar contacts, helical content, β content, extended content, other content (i.e., not in standard ψ/φ definitions), and Pardi et al. helical content (37).

SASA was calculated according to the algorithm of Lee and Richards (38) with a probe radius of 1.4 Å, as implemented in *il*mm. Contacts are considered only between heavy atoms. Native contacts are those present in the crystal structure. Unless polar or nonpolar is specified, a contact is

TABLE 1 Reference data used for P_{fold}-like calculations

Temperature (K)	Total length (ns)	Component lengths (ns)	Cα RMSD* (Å)	NOEs satisfied† (%)
298	266	101,51,51,21,21,21	2.0 ± 0.5	92.8
310	396	101,101,101,51,21,21	2.4 ± 0.6	94.4
323	266	101,51,51,21,21,21	2.1 ± 0.6	93.7
498	274	60‡,51,51,51,40§,21	9.0 ± 4.3	72.3

*The N-terminal five residues and C-terminal three residues were not included in the RMSD as they are highly mobile and do not alter the EnHD major groove-binding interface (13,44–52).

†603 NOEs from NMR experiments conducted in the Fersht lab (T. Rutherford, S. Freund, and A. R. Fersht, personal communication). An NOE was considered satisfied if the r^{-6} weighted distance between closest protons was less than the inferred experimental value or 5.0 Å, whichever is smaller.

‡Previously published thermal unfolding trajectory (6,10–12,53).

§Previously published thermal unfolding trajectory (6).

defined as two carbon atoms whose separation distance is <5.4 Å or two noncarbon or one carbon and one noncarbon atoms whose separation distance is <4.6 Å. A polar contact is between two atoms whose absolute charges are <0.3 q and whose separation distance is <4.6 Å. A nonpolar contact is between two atoms whose absolute charges are <0.3 q and separation distance is <5.4 Å. A polar to nonpolar contact is one between two atoms, one with an absolute charge <0.3 q, and one whose absolute charge is <0.3 q and whose separation distance is <4.6 Å.

Each of these properties was normalized by its variance so that all contributed equally. Upon examination of the native and nonnative reference data set property spaces, it was observed that 10 properties contribute significant information for the distinction of native and nonnative configurations. The properties are total intramolecular contacts, total native contacts, total polar to nonpolar intramolecular contacts, total SASA, SC nonpolar SASA, MC polar SASA, helical content, all Cα RMSD to crystal, residues 6–51 Cα RMSD to crystal (dynamic N-terminal removed), and the CONGENEAL dissimilarity score to crystal. Several of the properties appear to be redundant, e.g., native intramolecular contacts and total intramolecular contacts. Indeed there is a certain amount of colinearity; however, these properties do speak to different aspects of protein structure. In the example of native and total intramolecular contacts, proteins tend to shed native contacts and gain nonspecific intramolecular contacts as they unfold, but this enthalpic exchange need not be one-to-one gained, nor need it occur simultaneously. Therefore, each property provides unique information about specific aspects of the underlying processes.

The mean distance between two structures, P_i and P_s, represented in their property space of N_p properties, can be calculated as

$$\langle |P_i - P_s| \rangle = \frac{\sum_{p}^{N_p} |x_{p,i} - x_{p,s}|}{N_p}, \quad (1)$$

where $x_{p,i}$ is the value of property p for structure i. For a cluster of structures F, with cardinality N_s, the mean distance in Eq. 1 can be used to calculate an average of the mean distance to the cluster of structures F:

$$\langle P_i - F \rangle = \frac{\sum_{s}^{N_s} \langle |P_i - P_s| \rangle}{N_s}. \quad (2)$$

One-dimensional reaction coordinate

The property space data in the reference set were used to build a one-dimensional folding/unfolding reaction coordinate with a range of 0 (unfolded) to 1 (folded). We note that ''unfolded'' in this reaction coordinate is any member of the intermediate ensemble or beyond in the denaturing direction such that the system is treated as two-state, commensurate with the early stages of the unfolding process. The mean distance in property space for every reference structure to the native reference cluster was calculated (Eq. 2). The native cluster consisted of all structures from the 298, 310, and 323 K simulations listed in Table 1. No structures from a given simulation were compared to any other structures in the same simulation. A histogram of these values was used to determine a sigmoidal function with the desired range and midpoint (0.5) value. The midpoint was centered on the valley (i.e., the lowest populated value) between the first mode (native cluster) and second mode (unfolding intermediate/denatured state) in the histogram. The final reaction coordinate related the unbounded one-dimensional mean distance in property space to the native cluster, $\langle P_i - F \rangle$, to a bounded (0–1) one-dimensional function, $\chi(\langle P_i - F \rangle)$ and was expressed by the equation

$$\chi(\langle P_i - F \rangle) = 1 - \left(\frac{1}{1 + e^{-(\langle P_i - F \rangle - 0.12)}} \right). \quad (3)$$

Computing a P_{fold}-like quantity from simulations

For each of the three P_{fold}-like assessment conditions (at 298, 310, and 323 K), the final location along the folding reaction coordinate was determined by the following process: if any structure's $\chi(\langle P_i - F \rangle)$ was within 0.05 of folded (i.e., 0.95) or unfolded (i.e., 0.05) the simulation was determined to have folded or unfolded, respectively. Otherwise $\chi(\langle P_i - F \rangle)$was averaged over the final nanosecond. Finally, the resultant values from the three simulations of each target were averaged for the starting structures' final P_{fold}-like quantity. This is in contrast to the well-known implicit solvent P_{fold} simulations in the literature where a large set of Bernoulli trials are used. For N Bernoulli trials, the error is a well-studied quantity, $1/\sqrt{N}$. In our case, due to the computing requirements of all atom MD in water, we are forced to use a much smaller number of simulations and take their mean reaction coordinate location. To improve our limited sampling, a set of temperatures (298, 310, and 323 K) was used, all under the T_m of EnHD. The higher temperatures also correspond to faster refolding temperatures of EnHD (10). In Fig. 2, we present a flowchart of the P_{fold}-like quantity assignment process.

RESULTS

Property space

Fig. 3 depicts the histograms of two normalized properties from the full 32-dimensional property space: MC nonpolar SASA (*A*) and CONGENEAL dissimilarity score (*B*). Twenty-two of the properties in the full space were similar to those depicted in Fig. 3 *A* in that they provided little or no ability to distinguish between native and nonnative conformations. However, the 10 properties selected for the construction of the one-dimensional reaction coordinate possessed a mode in their distributions that contained primarily native conformations, such as CONGENEAL in Fig. 3 *B*. These 10 properties also represent the most heavily weighted components in principal components analysis (PCA) of a subset of the 9.8×10^6 samples (see Table 2 for the final loadings). From Table 2, it is evident that they are redundant with all property loadings for the first three principal components (PC) being roughly equal.

Histograms of Cα-RMSD, *Rg*, and native contacts, often used as reaction coordinates for protein folding, are presented in Fig. 4, *A–C*, respectively. The histograms are broken down into native and nonnative groups—with the native group comprising the reference data sets for 298, 310, and 323 K. The nonnative data are from the 498 K reference set members. It is evident by the overlap between native and nonnative data sets in these histograms that no one of these properties alone is sufficient for a one-dimensional reaction coordinate. Also in Fig. 4, *D–F*, are the histograms of the two-dimensional reaction coordinates derived from the combinations of these simple properties for the reference folding and unfolding simulations. The TSE identified by MDS in the 498 K data set is plotted with yellow circles to emphasize that no one of these two-dimensional spaces is sufficient for identification of the TS.

One-dimensional reaction coordinate

A histogram of the values resulting from application of Eq. 2 to the 10 aforementioned properties of the reference data set

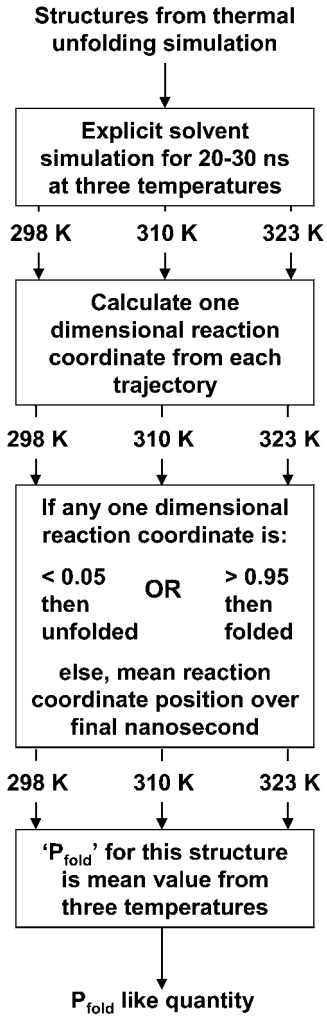

FIGURE 2 Process diagram of P_{fold}-like calculations in this study. For each candidate TS structure from the thermal unfolding simulation, three simulations are performed, one each at 298, 310, and 323 K. These simulations contain all atom and explicit solvent. The resulting trajectory is analyzed, and for each structure in the trajectory its location along the one-dimensional reaction coordinate is determined. Where a trajectory had a structure that was <5% folded by the reaction coordinate the trajectory was labeled as unfolded, and conversely where a trajectory reached a structure that was <95% folded the trajectory was assigned as folded. In those cases where neither of these binary states was achieved, the mean location of the trajectory along the one-dimensional reaction coordinate was assigned. Finally, the mean value of these quantities for all three temperatures was averaged, resulting in the final ''P_{fold}-like'' quantity for a given starting structure.

is presented in Fig. 5. It was used to determine the sigmoidal function with the desired range and midpoint (0.5) value. The midpoint was centered on the valley (i.e., the lowest populated value) between the first mode (native cluster) and second mode (unfolding intermediate/denatured state) in the histogram: 0.12 (arbitrary) units.

The P_{fold}-like quantity from simulations

The mean one-dimensional reaction coordinate results from 298, 310, and 323 K simulations for each starting structure from the validated thermal unfolding simulation with error as standard deviation are presented in Fig. 6. Based on the data, a sigmoidal function was fit to the P_{fold}-like values with weights on the individual data points corresponding to the inverse of the standard deviation. The TSE identified from the 0.5 value of the fitted function was 250 ± 2 ps. The final location of the TSE was relatively independent of fit (±2 ps), individual weights, and software used for the fit. The final fit was performed using the GNU Scientific Library (39). The average Cα RMSD between the structures within the TSE derived from P_{fold} is 1.0 ± 0.24 Å. The ensemble had 119 ± 7 hydrogen bonds to water, 42 ± 2 intramolecular hydrogen bonds, and 548 intramolecular nonpolar contacts. The TSE identified using our P_{fold}-like method is very similar to that derived using our clustering approach. The Cα RMSD between the two sets of structures in Fig. 6 *B* is 1.28 ± 0.22 Å.

Visualization of representative simulations

Three-dimensional projections of the full 32-dimensional property space for various simulations (native, nonnative, and P_{fold}) provide a way to visually inspect and compare multiple MD trajectories. However, these three-dimensional projections account for only ~71% of the variance in the underlying property space. A nine-dimensional projection is required to account for >90% of the variance, but this is difficult to display. Fig. 7, *A* and *B*, depicts the projection of the reference native state onto the first three PC resulting from a diagonalization of the property space correlation matrix. In *A*, only data from 298 K simulations are shown. Simulation time is denoted by gray scale shading from black (start) to white (end) of the underlying color representing each trajectory. Of note are the two distinct native substates. In *B*, data from all native state simulations are shown. The barrier between the native subensembles seems to be less pronounced, if present at all, indicating free exchange between these substates when the temperature is increased to 310 and 323 K.

Fig. 7 *C* depicts the property-space projections of native (*green*, *blue*, and *cyan* for 298, 310, and 323 K, respectively) and thermal unfolding simulations (*red* for 498 K) used as reference. Fig. 7 *C* contains the same data as 4 *B* except that the 498 K data are included, which required a change in the scale. Two lobes in the thermal unfolding simulations are evident (*right upper* and *lower* lobes), corresponding to the unfolding intermediates of EnHD (see DeMarco et al. (6) for a more complete discussion). The manifold encapsulating the unfolding trajectories is substantially larger in volume than the native state simulation manifold. The three lower panels (*D–F*) of Fig. 7 depict the property space projections in *C* with the inclusion of three different sample P_{fold}-like simulations (in *gray scale*). From left to right, they represent simulations that resulted in (*D*) a folded conformation, (*E*) a TS-like conformation, and (*F*) an unfolding intermediate conformation. It should be noted again that these projections account for only 71% of the variance in the

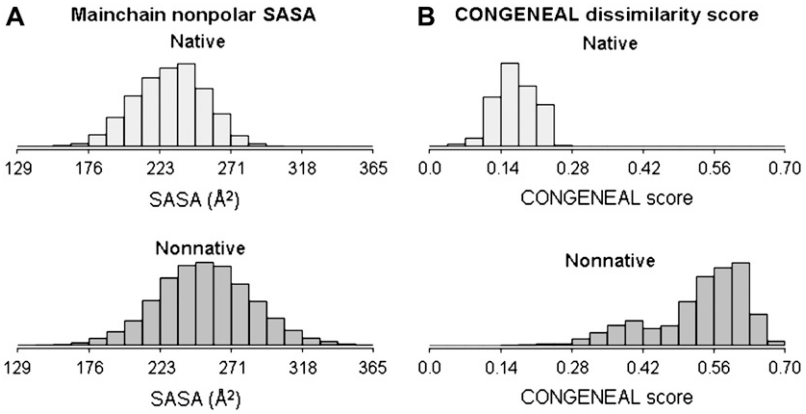

FIGURE 3 Histograms of two properties from native and nonnative reference simulation sets for EnHD. In both panels, the *y* axis is probability of occurrence. (*A*) The MC nonpolar SASA for (*top*) native reference data in Table 1 and (*bottom*) nonnative reference data in Table 1. Note that the histograms overlap such that native and nonnative configurations are not readily identifiable by these data alone. (*B*) The CONGENEAL dissimilarity score for (*top*) native reference data in Table 1 and (*bottom*) nonnative reference data in Table 1. Note that the histograms do not substantially overlap such that this property provides significant discriminatory power between native and nonnative configurations.

underlying property space. Even with this limitation, the spatial regions corresponding to native, transition, and nonnative states are evident. As the nonnative manifold was larger than the native manifold, the unfolded P_{fold}-like simulation in *F* has a larger *Rg* than that of the frustrated TS-like P_{fold}-like simulation in panel *E*, which is larger than that of the stably native-like P_{fold}-like simulation in panel *D*.

DISCUSSION

The length of the individual P_{fold} simulations in the study (i.e., not the reference simulations in Table 1) is short (20.2–30.2 ns) when compared to implicit solvent simulations (which can be up to 100 ns) (40). However, as we are treating only the early events in EnHD unfolding/late events in folding, this simulation length virtually assures that we are in the two-state behavior time region. Most of our simulations committed to folding or unfolding within the first 10 ns. That is, by the one-dimensional reaction coordinate, the simulations exhibited a strong tendency to immediately collapse, resulting in a lower *Rg*. This collapse, a result of the quenching process, typically resulting in one of two possibilities: a native-like collapsed configuration or a nonnative collapsed configuration. In these cases, the reaction coordinate location was easily identified. As expected, those starting structures around and just after the TS required more time to clearly commit by the one-dimensional reaction coordinate.

In addition, due to the computational requirements involved in doing all-atom explicit solvent simulations, we performed only three trial-folding simulations for each input structure. This is substantially fewer than others doing P_{fold} calculations, e.g., Rao and Caflisch (19) performed 100 per input structure. Due to the aforementioned Bernoulli nature of these trials in true P_{fold} studies, 400 trials per input structure are required to obtain a P_{fold} value accurate to within 5%. Thus we have developed an analogous method to compute quantities similar to those from P_{fold} given the inherent and very real limitations on all-atom explicit solvent simulation time.

The reaction coordinate we present here can be generalized and applied to any protein. Properties that do not contribute significant information for a given system, e.g., helical content for a β protein, will fall out of the calculation. In this instance, when we added the β content to our analysis, the quantity was virtually zero for all structures and did not alter the values resulting from Eq. 1. We are investigating the

TABLE 2 Weights derived by PCA of property spaces of reference data

	Eigenvector (decreasing by eigenvalue)			
Analytical property of protein structure	1	2	3	4
CONGENEAL (28) dissimilarity score	−0.33935	−0.25109	−0.25306	0.332823
Cα RMSD (all residues)	−0.3111	−0.48533	−0.02658	−0.30236
Cα RMSD (residues 6–51)	−0.32622	−0.39662	−0.05337	−0.328
Helical content	0.300256	−0.1179	0.453613	0.151191
MC polar SASA	−0.30024	0.462168	−0.15005	−0.48973
SC nonpolar SASA	−0.30421	−0.14675	0.51218	0.271036
Total SASA	−0.3261	0.154481	0.435532	0.069418
Total contacts	0.308812	−0.26417	−0.40982	0.230881
Native contacts	0.345916	0.076993	0.204635	−0.38375
Other contacts	0.295693	−0.44214	0.205626	−0.38559
Eigenvalues	7.103064	1.009966	0.928239	0.283860
% of variance captured	71.03	81.13	90.41	93.25

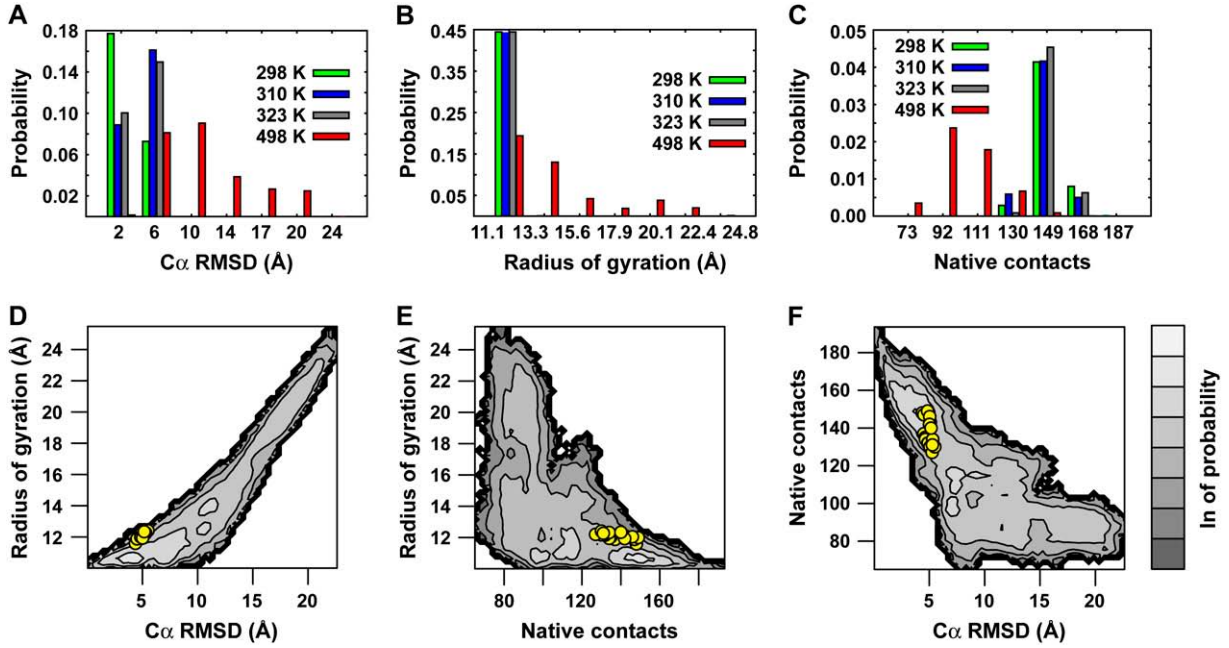

FIGURE 4 Histograms of analytical properties from reference data sets used for P_{fold}. Histograms of (*A*) Cα-RMSD, (*B*) *Rg*, and (*C*) native contacts derived from native and thermal unfolding simulations. The normalization factors for *A–C* were 266,000 for 298 K, 396,000 for 310 K, 266,000 for 323 K, and 674,000 for 498 K. Two-dimensional property space histograms for native and unfolding reference simulations of EnHD: (*D*) Cα-RMSD versus *Rg*, (*E*) native contacts versus *Rg*, and (*F*) Cα-RMSD versus native contacts. TSE positions from our method are shown as yellow circles.

use of an expanded reaction coordinate where all properties are considered on a large test set of several hundred proteins (41,42).

The TSs identified by our clustering method and our P_{fold}-like quantity are separated in time by 5 ps. On such a timescale only minor changes in protein structure were observed: methyl group rotations, minor fluctuations in SC and MC dihedral angles, and exchange of waters in low residence time hydration sites. It is not surprising that the two sets of TSEs were almost indistinguishable from each other (Fig. 6 *B*). For

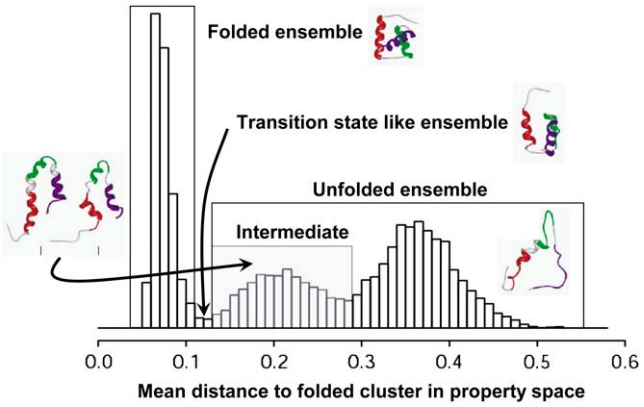

FIGURE 5 Histogram of mean distances to the native cluster from 9.8 × 106 samples of folded and thermally unfolded simulation data. The histogram was constructed by taking each structure in the reference set and computing the mean of the mean distance in property space (in accordance with Eqs. 1, 2, and 3) to every structure in the native cluster. The resultant distribution is trimodal with the left-most mode (i.e., those with small mean distances to the native cluster) corresponding to the folded ensemble. The second and third modes of decreased population represent an unfolding intermediate and the more denatured states represent the unfolded ensemble. The valley between the folded and first unfolded mode (at 0.12 units) signifies a relatively unpopulated set of structures that are structurally like the TSE.

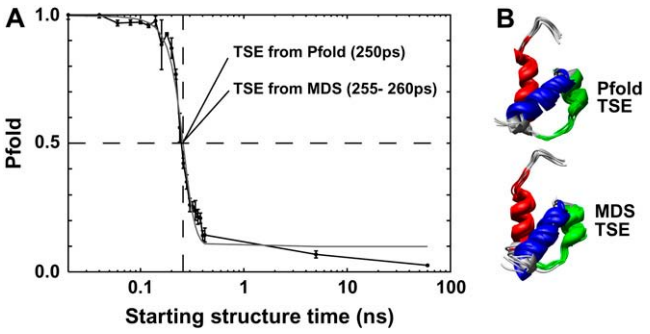

FIGURE 6 P_{fold} for starting structures from validated thermal unfolding trajectory and structures for TSE identified by two different methods. (*A*) The mean probability of folding for each structure is plotted in black with error bars indicating standard deviation of the three simulations, one each at 298, 310, and 323 K. In red, the resultant fit of a sigmoidal function to the data is displayed. The previously identified TS by MDS of the pairwise Cα RMSD matrix is centered in the range 255–260 ps. The TS identified by the sigmoid fit (i.e., the value in the domain where the fitted function is 0.5) is 250 ± 2 ps. The time axis is on a logarithmic scale, thus the nature of the fit at its lower bound. (*B*) Structures of the TSE identified in this study (*top*) and with our clustering method (*bottom*). Helix one is in red, helix two in green, and helix three in blue.

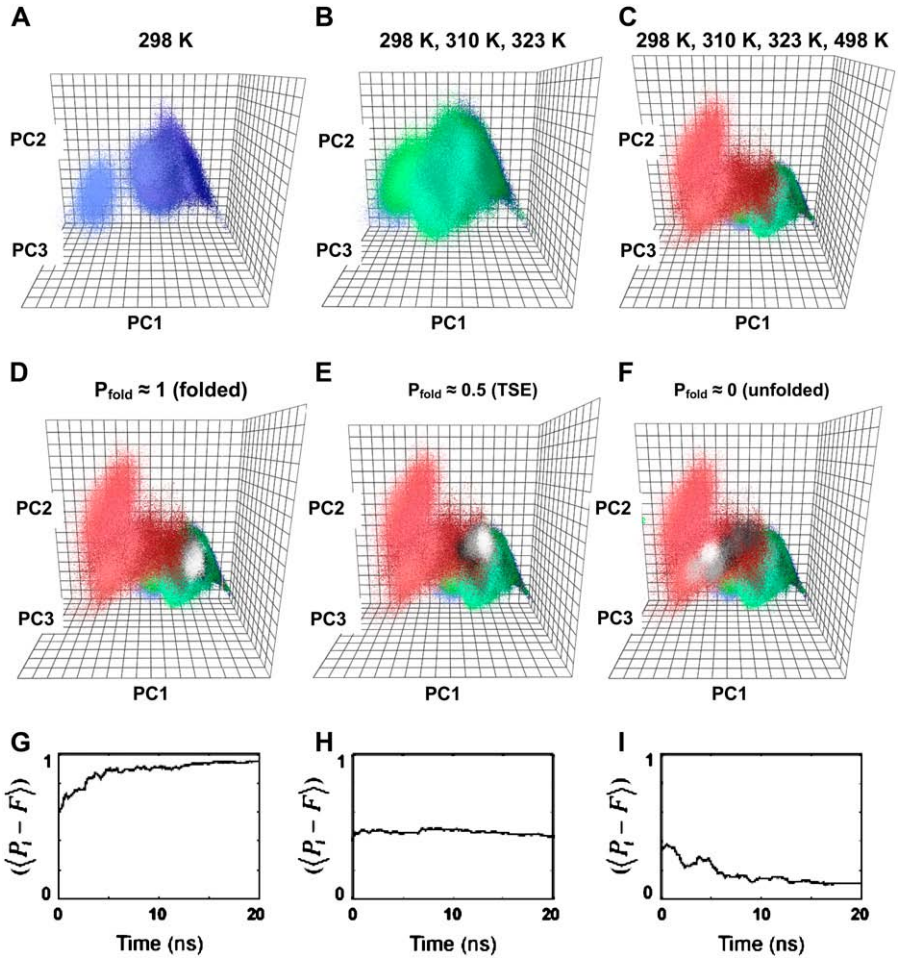

FIGURE 7 Property space and one-dimensional reaction coordinate ($\chi(\langle\langle P_i - F\rangle\rangle)$) projections of EnHD simulations. PCA projections of simulation data with principal components 1, 2, and 3 as PC1, PC2, and PC3, respectively. (*A*) 298 K reference simulation data, (*B*) 298 K (*blue*), 310 K (*green*), and (*C*) 323 K (*cyan*) reference simulation data. (*D–F*) Colored as in *A–C* with the addition of an example simulation from the P_{fold} set and 498 K thermal unfolding (*red*) simulation data. Simulation time is denoted by blending a gray scale from black (simulation start) to white (simulation end) onto the base color for each temperature. In *D–F*, a sample P_{fold} simulation has been depicted in gray scale: black (start) to white (end). (*D*) The 298 K P_{fold} simulation of the 160 ps simulation structure from the validated thermal unfolding trajectory is depicted. (*E*) The 323 K P_{fold} of the 260 ps structure. Notice that the simulation tends to the region defined as the interface of the native and unfolded clusters (i.e., about the TS). (*F*) The 310 K P_{fold} simulation of the 380 ps structure. Even under strongly folding conditions, this simulation rapidly diverges from the native state cluster and ends in the unfolding ensemble cluster. (*G–I*) One-dimensional reaction coordinate of P_{fold} simulations versus simulation time for the trajectories projected in *D–F*.

example, the P_{fold} TSE (248–253 ps) has an average Cα RMSD to the TSE identified by our method (255–260 ps) of 1.28 ± 0.22 Å. This value is statistically indistinguishable from the Cα RMSD within the originally identified TSE (1.12 ± 0.24 Å). Where there are structural differences between the original TS and the one identified in this study, they tend to be only in SC conformations at residues shown to be unstructured in the TS (i.e., in regions of low Φ_F).

The TSEs using the two different methods were also quite similar in their physical properties. For comparison to the values reported for the P_{fold} TSE in the Results section (119 hydrogen bonds to water, 42 intramolecular hydrogen bonds, and 548 intramolecular nonpolar contacts), the original TSE had 116 ± 6 hydrogen bonds to water, 44 ± 2 intramolecular hydrogen bonds, and 537 ± 20 intramolecular hydrophobic contracts. It is not possible to distinguish between these two sets.

We often compare our semiquantitative S-values (43) derived from the MD trajectories to experimentally measured Φ_F values. The Φ_F values, determined on a per residue basis via mutation, are interpreted as the extent of native tertiary structure in the TS. A value of 0 indicates nonnative whereas 1 implies native-like extent of structure in the TS at the given sequence position. Using the rules previously described for calculation of S-values for EnHD various residues (10–12), S-values were calculated for the ensemble of structures from 248 to 253 ps. The correlation coefficient for the S-values to the experimental Φ_F values is 0.77. This correlation is similar to that of the originally identified TSE, which is 0.79 (11).

The most important difference between the TSE identified using the two different methods was the time required to identify them. Using our clustering method requires 1), calculating the pairwise Cα RMSD between all structures in the unfolding trajectory; 2), MDS of the resulting symmetric, zero diagonal matrix; and 3), visual inspection of the resulting three-dimensional representation or the use of an automated clustering algorithm to replace visual inspection. In practice, we often limit these calculations to the first several thousand structures from the trajectory taken at 0.2 ps granularity. For all but the most thermostable proteins, experience dictates that the TS occurs on that timescale at 498 K.

The first step of our method, on a 2.1 GHz Advanced Micro Devices (AMD, Sunnyvale, CA) Athlon MP, executes in under 10 (wall clock) min using *il*mm's RMSD analysis module. The second step, essentially a problem in finding the

eigensystem for the input matrix, can be computed in no more than 10 min (on the same hardware) using the algorithms as implemented in the GNU Scientific Library (39). Finally, visual inspection of the MDS to three dimensions to identify the first cluster exit can require as long as 20 min. Thus, the total time to identify the TS from an average unfolding simulation is ~40 min.

In contrast, the time required for the P_{fold}-like approach is significantly longer. For each structure, drawn at sufficient granularity from the trajectory under examination, a simulation of appropriate length must be conducted. For slowly folding systems these simulations may need to be hundreds of ns long. In the case of EnHD on a 2.1 GHz AMD Athlon MP, a 20.2 ns simulation of the 240 ps structure required ~13 days of wall clock time. In this study, 72 such trajectories (and the 7 that were extended by 10 ns) were calculated for a total of 985 CPU days. Furthermore, this estimate neglects the time required for analyses, calculating the fits, performing extensive reference simulations, etc. Based on these timings, the MDS method of identifying the TSE is >36,000 times faster than our P_{fold}-like approach while producing the same results. Our estimate is very likely a significant understatement of the cost for a true P_{fold} computation where each putative TS would have 100–400 Bernoulli trials.

We note that despite their computational expense, P_{fold} calculations do provide sampling of the conformations and dynamics near the TS. With our method, no further simulations are required, and so this more in-depth level of sampling is not obtained. Other types of studies employing MD for protein folding/unfolding and identification of the TSE, such as long timescale simulations of a protein at or near its T_m, are also less expensive than P_{fold} while offering enhanced sampling of early unfolding/late folding events (54). Nevertheless, many consider P_{fold} the gold standard for assignment of the TS from MD simulation. In this study, we have shown that our method yields the same TS at a fraction of the computational cost.

University of California, San Francisco, chimera was used to prepare protein images.

This work was supported by the National Institutes of Health (GM 50789 to V.D.). D.B. was supported by a National Institutes of Health Molecular Biophysics Training Grant (National Research Service Award 5 T32 GM 08268).

REFERENCES

1. Ladurner, A. G., L. S. Itzhaki, V. Daggett, and A. R. Fersht. 1998. Synergy between simulation and experiment in describing the energy landscape of protein folding. *Proc. Natl. Acad. Sci. USA*. 95:8473–8478.
2. Li, A., and V. Daggett. 1994. Characterization of the transition state of protein unfolding by use of molecular dynamics: chymotrypsin inhibitor 2. *Proc. Natl. Acad. Sci. USA*. 91:10430–10434.
3. Daggett, V., A. Li, L. S. Itzhaki, D. E. Otzen, and A. R. Fersht. 1996. Structure of the transition state for folding of a protein derived from experiment and simulation. *J. Mol. Biol.* 257:430–440.
4. Kearsley, S. K. 1989. On the orthogonal transformation used for structural comparisons. *Acta Crystallogr. A*. 45:208–210.
5. Kazmirski, S. L., A. J. Li, and V. Daggett. 1999. Analysis methods for comparison of multiple molecular dynamics trajectories: applications to protein unfolding pathways and denatured ensembles. *J. Mol. Biol.* 290: 283–304.
6. DeMarco, M. D., D. O. V. Alonso, and V. Daggett. 2004. Diffusing and colliding: the atomic level folding/unfolding pathway of a small helical protein. *J. Mol. Biol.* 341:1109–1124.
7. Sander, J., M. Ester, H. P. Kriegel, and X. W. Xu. 1998. Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. *Data Min. Knowl. Disc.* 2:169–194.
8. Daggett, V., A. J. Li, and A. R. Fersht. 1998. Combined molecular dynamics and phi-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.
9. De Jong, D., R. Riley, D. O. V. Alonso, and V. Daggett. 2002. Probing the energy landscape of protein folding/unfolding transition states. *J. Mol. Biol.* 319:229–242.
10. Mayor, U., C. M. Johnson, V. Daggett, and A. R. Fersht. 2000. Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc. Natl. Acad. Sci. USA*. 97:13518–13522.
11. Gianni, S., N. R. Guydosh, F. Khan, T. D. Caldas, U. Mayor, G. W. White, M. L. DeMarco, V. Daggett, and A. R. Fersht. 2003. Unifying features in protein-folding mechanisms. *Proc. Natl. Acad. Sci. USA*. 100:13286–13291.
12. Mayor, U., N. R. Guydosh, C. M. Johnson, J. G. Grossmann, S. Sato, G. S. Jas, S. M. Freund, D. O. V. Alonso, V. Daggett, and A. R. Fersht. 2003. The complete folding pathway of a protein from nanoseconds to microseconds. *Nature*. 421:863–867.
13. Clarke, N. D., C. R. Kissinger, J. Desjarlais, G. L. Gilliland, and C. O. Pabo. 1994. Structural studies of the engrailed homeodomain. *Protein Sci.* 3:1779–1787.
14. Kissinger, C. R., B. S. Liu, E. Martinblanco, T. B. Kornberg, and C. O. Pabo. 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.
15. Islam, S. A., M. Karplus, and D. L. Weaver. 2002. Application of the diffusion-collision model to the folding of three-helix bundle proteins. *J. Mol. Biol.* 318:199–215.
16. Religa, T. L., J. S. Markson, U. Mayor, S. M. V. Freund, and A. R. Fersht. 2005. Solution structure of a protein denatured state and folding intermediate. *Nature*. 437:1053–1056.
17. Li, L., and E. I. Shakhnovich. 2001. Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations. *Proc. Natl. Acad. Sci. USA*. 98:13014–13018.
18. Gsponer, J., and A. Caflisch. 2002. Molecular dynamics simulations of protein folding from the transition state. *Proc. Natl. Acad. Sci. USA*. 99:6719–6724.
19. Rao, F., and A. Caflisch. 2004. The protein folding network. *J. Mol. Biol.* 342:299–306.
20. Lenz, P., B. Zagrovic, J. Shapiro, and V. S. Pande. 2004. Folding probabilities: a novel approach to folding transitions and the two-dimensional Ising-model. *J. Chem. Phys.* 120:6769–6778.
21. Du, R., V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich. 1998. On the transition coordinate for protein folding. *J. Chem. Phys.* 108:334–350.
22. Shimada, J., and E. I. Shakhnovich. 2002. The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *Proc. Natl. Acad. Sci. USA*. 99:11175–11180.
23. Beck, D. A. C., and V. Daggett. 2004. Methods for molecular dynamics simulations of protein folding/unfolding in solution. *Methods*. 34:112–120.
24. Boczko, E. M., and C. L. Brooks. 1995. First-principles calculation of the folding free-energy of a 3-helix bundle protein. *Science*. 269:393–396.
25. Sheinerman, F. B., and C. L. Brooks. 1998. Molecular picture of folding of a small alpha/beta protein. *Proc. Natl. Acad. Sci. USA*. 95:1562–1567.
26. Sheinerman, F. B., and C. L. Brooks. 1998. Calculations on folding of segment B1 of streptococcal protein G. *J. Mol. Biol.* 278:439–456.

27. Dokholyan, N. V., L. Li, F. Ding, and E. I. Shakhnovich. 2002. Topological determinants of protein folding. *Proc. Natl. Acad. Sci. USA.* 99:8637–8641.

28. Alonso, D. O. V., and V. Daggett. 2000. Staphylococcal protein A: unfolding pathways, unfolded states, and differences between the B and E domains. *Proc. Natl. Acad. Sci. USA.* 97:133–138.

29. Yee, D. P., and K. A. Dill. 1993. Families and the structural relatedness among globular proteins. *Protein Sci.* 2:884–899.

30. Klimov, D. K., and D. Thirumalai. 2004. Progressing from folding trajectories to transition state ensemble in proteins. *Chem. Phys.* 307:251–258.

31. Tenenbaum, J. B., V. de Silva, and J. C. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science.* 290:2319–2323.

32. Das, P., M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi. 2006. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. USA.* 103:9885–9890.

33. Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* 43:59–69.

34. Kell, G. S. 1967. Precise representation of volume properties of water at one atmosphere. *J. Chem. Eng. Data.* 12:66–69.

35. Haar, L., J. S. Gallagher, and G. S. Kell. 1984. NBS/NRC Steam Tables. Hemisphere, New York.

36. Levitt, M. 1990. ENCAD, Computer Program for Energy Calculation and Dynamics. Stanford University, Palo Alto, CA.

37. Pardi, A., M. Billeter, and K. Wuthrich. 1984. Calibration of the angular-dependence of the amide proton-C-α proton coupling-constants, 3jhn-α, in a globular protein—use of 3jhn-α for identification of helical secondary structure. *J. Mol. Biol.* 180:741–751.

38. Lee, B., and F. M. Richards. 1971. Interpretation of protein structures—estimation of static accessibility. *J. Mol. Biol.* 55:379–400.

39. Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi. 2005. GNU Scientific Library Reference Manual. Network Theory, Bristol, UK.

40. Snow, C. D., Y. M. Rhee, and V. S. Pande. 2006. Kinetic definition of protein folding transition state ensembles and reaction coordinates. *Biophys. J.* 91:14–24.

41. Day, R., D. A. C. Beck, R. S. Armen, and V. Daggett. 2003. A consensus view of fold space: combining SCOP, CATH, and the Dali domain dictionary. *Protein Sci.* 12:2150–2160.

42. Beck, D. A. C., A. L. Jonsson, D. Schaeffer, K. A. Scott, R. Day, R. D. Toofanny, D. O. V. Alonso, and V. Daggett. 2007. Dyanameomics: Mass annotation of protein dynamics and unfolding in water by high-throughput all-atom molecular dynamics simulations. *Genome Biol.* In press.

43. Day, R., and V. Daggett. 2005. Sensitivity of the folding/unfolding transition state ensemble of chymotrypsin inhibitor 2 to changes in temperature and solvent. *Protein Sci.* 14:1242–1252.

44. Gutmanas, A., and M. Billeter. 2004. Specific DNA recognition by the Antp homeodomain: MD simulations of specific and nonspecific complexes. *Proteins.* 57:772–782.

45. Sato, K., M. D. Simon, A. M. Levin, K. M. Shokat, and G. A. Weiss. 2004. Dissecting the engrailed homeodomain-DNA interaction by phage-displayed shotgun scanning. *Chem. Biol.* 11:1017–1023.

46. Simon, M. D., K. Sato, G. A. Weiss, and K. M. Shokat. 2004. A phage display selection of engrailed homeodomain mutants and the importance of residue Q50. *Nucleic Acids Res.* 32:3623–3631.

47. Ades, S. E., and R. T. Sauer. 1995. Specificity of minor-groove and major-groove interactions in a homeodomain-DNA complex. *Biochemistry.* 34:14601–14608.

48. Duan, J. X., and L. Nilsson. 2002. The role of residue 50 and hydration water molecules in homeodomain DNA recognition. *Eur. Biophys. J.* 31:306–316.

49. Fraenkel, E., M. A. Rould, K. A. Chambers, and C. O. Pabo. 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.

50. Grant, R. A., M. A. Rould, J. D. Klemm, and C. O. Pabo. 2000. Exploring the role of glutamine 50 in the homeodomain-DNA interface: crystal structure of engrailed (Gln50 → Ala) complex at 2.0 Å. *Biochemistry.* 39:8187–8192.

51. Ledneva, R. K., A. V. Alexeevskii, S. A. Vasil'ev, S. A. Spirin, and A. S. Karyagina. 2001. Structural aspects of interaction of homeodomains with DNA. *Mol. Biol.* 35:647–659.

52. Tucker-Kellogg, L., M. A. Rould, K. A. Chambers, S. E. Ades, R. T. Sauer, C. O. Pabo. 1997. Engrailed (Gln50→Lys) homeodomain-DNA complex at 1.9 Å resolution: structural basis for enhanced affinity and altered specificity. *Structure.* 5:1047–1054.

53. White, G. W., S. Gianni, J. G. Grossmann, P. Jemth, A. R. Fersht, and V. Daggett. 2005. Simulation and experiment conspire to reveal cryptic intermediates and a slide from the nucleation-condensation to framework mechanism of folding. *J. Mol. Biol.* 350:757–775.

54. Day, R., and V. Daggett. 2007. Direct observation of microscopic reversibility in protein folding. *J. Mol. Biol.* 366:677–686.