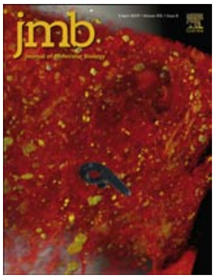


Visualizing Protein Folding and Unfolding☆

Jennifer Ferina and Valerie Daggett

Department of Bioengineering, *Box 355013, University of Washington, Seattle, WA 98195-5013, USA*

Correspondence to Valerie Daggett: *daggett@uw.edu*

https://doi.org/10.1016/j.jmb.2019.02.026
Edited by Sean Ignatius O'Donoghue

Abstract

Protein folding/unfolding is a complicated process that defies high-resolution characterization by experimental methods. As an alternative, atomistic molecular dynamics simulations are now routinely employed to elucidate and magnify the accompanying conformational changes and the role of solvent in the folding process. However, the level of detail necessary to map the process at high spatial–temporal resolution provides an overwhelming amount of data. As more and better tools are developed for analysis of these large data sets and validation of the simulations, one is still left with the problem of visualizing the results in ways that provide insight into the folding/unfolding process. While viewing and interrogating static crystal structures has become commonplace, more and different approaches are required for dynamic, interconverting, unfolding, and refolding proteins. Here we review a variety of approaches, ranging from straightforward to complex and unintuitive for multiscale analysis and visualization of protein folding and unfolding.

☆ This article is part of the special issue on "Biovisualization" edited by Donald E Ingber, Janet H Iwasa, Jodie Jenkinson, Gaël McGill, Seán O'Donoghue, Charles Reilly, Alain Viel.

Introduction

A denatured protein is typically 5–10 kcal/mol less stable than its native state; therefore, there is no single force contributing solely to destabilization, and there are numerous possible conformations between and within the denatured and native states [1]. This hints at the complexity of the problem and of the many different aspects of a protein that must be examined as potential contributors to protein unfolding/folding [2,3]. Experimental approaches have several limitations with respect to characterization of protein folding. Most importantly, they cannot provide a comprehensive high-resolution description of the temporal process, allowing us to visualize conformational changes at the nanosecond to microsecond scale and beyond with a high level of detail. Computational approaches allow us to simulate solvated proteins as a function of temperature or denaturant to obtain unfolding and folding trajectories that provide unprecedented amounts of data, which can be overwhelming and necessitates a variety of analysis and visualization approaches. Over the course of a simulation, the protein can be examined in several different ways in order to gain insight into the molecular mechanism behind these conformational changes. However, computational tools still have several limitations, especially regarding computing power and data storage.

Supercomputers can perform a finite number of calculations in a reasonable amount of time, and storage of big data from simulations also limits the number of simulations that can be performed, as well as introducing challenges for data analysis. Since simulations are ideally atomistic, performing molecular dynamics (MD) simulations of proteins with thousands of atoms has a high computational cost. Simulating the protein with surrounding solvent is critical to an accurate representation of the system, but it adds significantly to the simulation time and storage requirements [4–7]. MD simulations generate vast amounts of data, so determining the most relevant output parameters to examine is dependent on the system and its suspected mechanisms. Here we present different approaches and tools for visualizing protein unfolding/folding that allow the protein to be examined in a “residue-istic” manner and tools for MD analysis that can compare at scales ranging from atomic to collective motion involving multiple residues and protein regions.

Un/folding as a Movie Depiction

Direct visualization of the MD-derived protein coordinates as a function of time, as a movie, provides a wealth of information. These MD trajectories can be played frame by frame in the forward and reverse directions. In this way, MD simulations of protein denaturation, provided they are adequately validated [8–10], can be used to investigate both protein unfolding and folding. Thus, an unfolding MD trajectory can be played backward to visualize the structure in several different ways, allowing the user to see how individual interactions may contribute to the folding of the native structure. While structural visualization software, such as the widely used UCSF Chimera [11] and VMD [12] packages, provides a high level of visual insight into the mechanisms behind protein unfolding and folding, it is particularly helpful for driving more quantitative and directed approaches to elucidate features of the process. Images of structures at different time points throughout a trajectory can also enrich plots and other less intuitive analysis methods by displaying the current state of the protein and connecting it to other properties.

Among the common methods for analyzing MD simulations, there are several standard metrics, including atomic root-mean-square deviation (RMSD), atomic root-mean-square fluctuations (RMSF), solvent-accessible surface area (SASA), dihedral angles and associated conformational assignments, atomic and residue contacts, and correlated motion [13]. Here, we present a range of analyses of the folding/unfolding of several model protein systems, highlighting the importance of visualization as a bridge between both quantitative and abstract analyses to provide insight into protein un/folding. We begin by focusing on a well-characterized small protein that has been studied in a variety of different theoretical and experimental laboratories. We then present studies of a number of other proteins that highlight different aspects of folding, present interesting analyses, and/or demonstrate general features of folding across different protein folds.

Visualizing Aspects of Folding and Unfolding in MD Simulations in a 3-Helix Bundle Model Protein

The engrailed homeodomain protein (EnHD) (Fig. 1a, see 0-ns structure) is a particularly good, simple model system for comparing different analysis and visualization methods to obtain insight into the complicated and interconnected facets of protein folding. Various snapshots from MD simulation at 100 °C (373 K) and 225 °C (498 K) are presented in Fig. 1a, b, respectively. In this simple 3-helix bundle fold, it is easy to see the unpacking of the helices and loss of helical structure at high temperature in this cartoon representation, and heightened unfolding at higher temperature is particularly evident. Unfolding of EnHD is typically tracked experimentally through tryptophan (Trp) fluorescence, and the single Trp is colored magenta (Fig. 1); the repacking of the core, while containing helical structure, leaves the Trp exposed to solvent, thereby producing a ‘denatured’ or ‘unfolded’ signal.

This high-level cartoon view of unfolding is simple and easily grasped, but it is incomplete and imprecise. It takes a variety of approaches to characterize the process and quantify the critical interactions *en route*. One way to visualize the unfolding pathway is with an RMSD analysis. In its simplest form, a coordinate RMSD is calculated with respect to a reference structure, the crystal structure or other starting structure, over the trajectory illustrating how far the structure moves over time [14]. In this case, the native state 298 K simulation serves as a reference, and the Cα RMSD over time is shown for several simulations in Fig. 2a. A critical distance across the core of the protein is shown in Fig. 2b, closely following the RMSD plots. As the temperature increases, the RMSD and distance across the core increase, which is accompanied by the loss of contact pairs (Fig. 2c).

Plots of the all-*versus*-all Cα RMSD matrices over the trajectories illustrate times in the simulations when similar conformations were adopted as well as transitions during the process of unfolding (Fig. 2d) [14]. Boxes indicate conformational clusters visually on the matrix plots. The plots at 373 K in Fig. 2 can be compared with the snapshots in Fig. 1. The Cα RMSD increases and then decreases at 373 K (Fig. 2a), corresponding to the unpacking of the helices and subsequent collapse to a misfolded intermediate state with the loss of native contact pairs in the core.

A blow-up of a portion of the 323 K/2 simulation along with the corresponding 3D projection of the all-*versus*-all Cα RMSD matrix constructed through multidimensional scaling is shown in the projection in Fig. 3a, b [14]. This particular simulation is interesting because it is at the melting temperature (T_m) of EnHD, such that the free energy between the unfolded and folded states is 0 kcal/mol, and the protein unfolds and refolds over time. The points in the 3D plot represent structures over time, and the distance between the structures represents the Cα RMSD between them: points that cluster together in the projection are necessarily similar

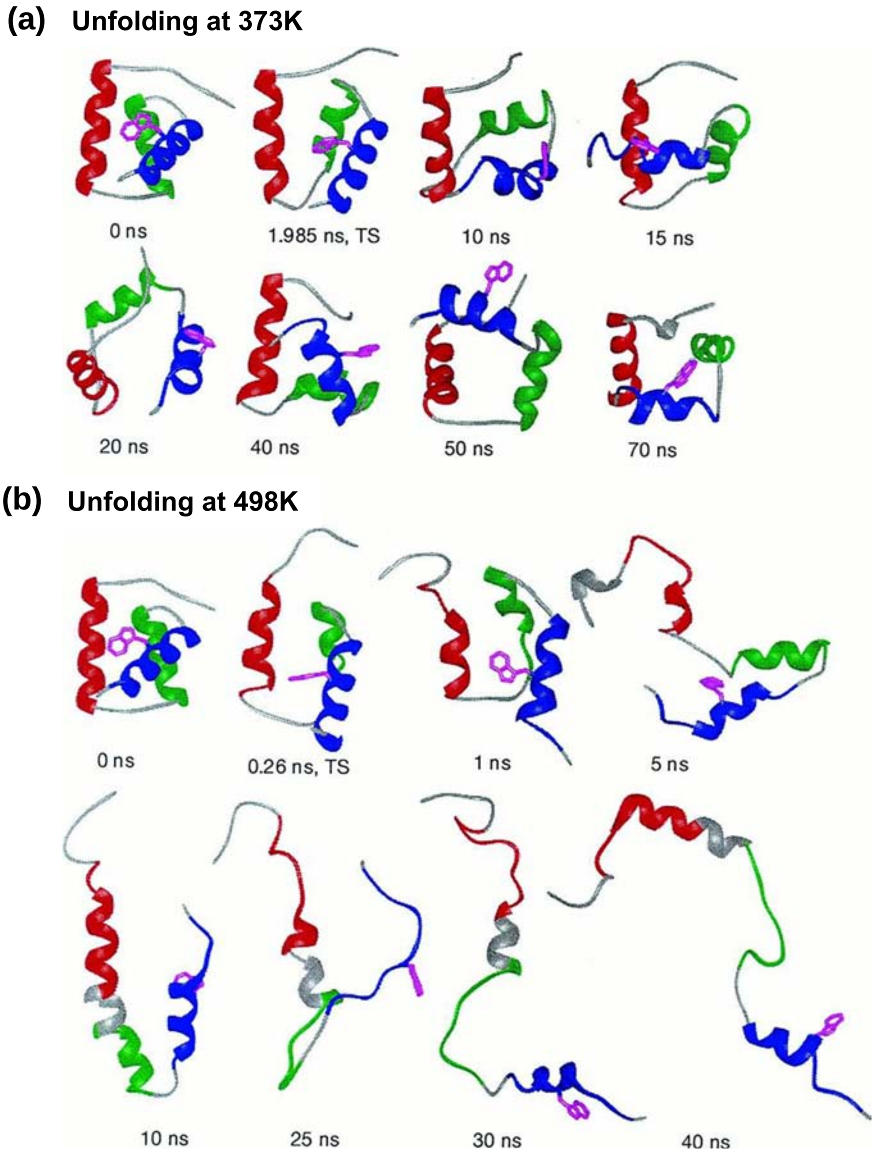


Fig. 1. Snapshots from EnHD MD unfolding trajectories at 373 and 498 K [15]. (a) Snapshots of structures from the unfolding trajectory over time at 373 K. The crystal structure is shown at the 0 ns time point, with Trp48, which becomes exposed to solvent at the transition state, shown in magenta. The native residues of helix I are shown in red (residues 10–22), helix II in green (residues 28–38), and helix III in blue (residues 42–55). The TS is shown immediately after the native structure. (b). Snapshots of structures from the unfolding trajectory over time at 498 K. The transition state structures have lower Cα RMSD than the starting structure and are similar in both cases. The cartoon structures provide the approximate shape of the protein over time, showing loss of secondary structure and allowing residues that become solvent exposed to be identified for further analysis. However, cartoon structures may make approximations to provide a smooth, continuous ribbon, and so are not necessarily representative of the exact atomic coordinates. Still, cartoon structures are a strong starting point for qualitatively identifying conformations and residues for further investigation. Figure and figure legend taken from Ref. [15] and used with permission.

(Fig. 3b) [14]. From this projection (beginning at 42.5 ns in red), the protein passed through the unfolding transition state (TS) (tan structure on right), and helix 3 pulled further away from the core (green structure) (Fig. 3b, c). Then the protein collapsed back down, passing through the same TS when refolding (blue structure) (Fig. 3c).

Increasing the temperature leads to more widespread unfolding of EnHD, but by a similar pathway through similar transition states. These transition states at different temperatures were first predicted by MD [15], and they were later confirmed thorough kinetic experiments verifying the timescale and extent of structure through experimental Φ-value

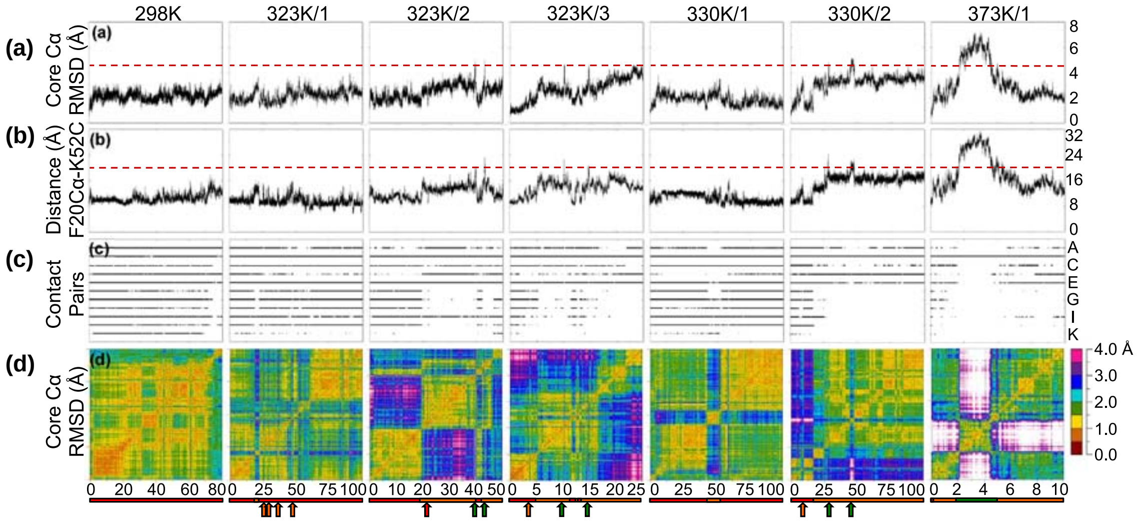


Fig. 2. EnHD core Cα RMSD, critical distance, and contact analysis for each simulation show unfolding and refolding behavior at high temperatures [14]. (a) Cα RMSD of the core of the protein (residues 8–53) calculated over time for each simulation relative to the native (0 ns) structure. The dashed red line at ~4.5Å indicates movement to D and is crossed for an extended period in the 373 K/1 simulation. (b) Distance between the Cα of Phe 20 and the backbone C of Lys 52 over time, representing the movement of HIII relative to the HI–HII scaffold. The N′ state is characterized by a distance of ~15 Å, while movement to D is characterized by distances greater than 20 Å (dashed red line). (c) Contacts between residue pairs, with labeled alternate pairs on 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, (K) Arg 53–Leu 26. In the N but not the N′ state, contacts between 49–24, 49–26, 52–20, and 53–24 are characteristically present, with additional contacts lost during D. (d) An all-*versus*-all core Cα RMSD matrix over time. 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 takes in each simulation: N (red), N′ (orange), and D (green). Arrows represent transiently occupied states. This more quantitative analysis helps identify specific states that the protein undergoes during unfolding and refolding; in particular, loss of several contact pairs is identified, the RMSD is shown to increase sharply over time, and the RMSD matrix shows higher deviation upon conversion to the D state. These three analysis types all support the identification of several structural states during unfolding simulations relative to the native simulations, which cannot be confirmed by the appearance of the structure alone. Figure and portions of figure legend taken from Ref. [14]. Adapted with permission from McCully *et al*. [14]. Copyright 2008 American Chemical Society.

analysis [16,17]. Furthermore, the intermediate state ensemble was identified and a representative structure is shown as the 10-ns snapshot at 498 K [15] (Fig. 1b). The structure of the intermediate state was determined by NMR 5 years later, and it was found to be essentially indistinguishable from the MD structure [18].

Unfolding sheds light on the folding pathway: microscopic reversibility and comparing unfolding *versus* folding simulations

There are many biological reasons to study unfolding instead of folding, and from a computational sampling point of view, unfolding is preferable. The assumption behind these unfolding studies is that studying unfolding sheds light on folding. This has been shown to be the case for EnHD through the simulations addressed above at the protein's T_m that allow for microscopic reversibility [14], as well as refolding simulations under native conditions beginning from structures taken from thermal unfolding trajectories, as in Fig. 1 [15,19].

Based on the known refolding rates, the number of MD simulations necessary to observe refolding was estimated. Consistent with the calculated probability, 1 of 43 independent refolding simulations in fact refolded. In addition to direct structural comparisons, several properties were monitored and salt bridge interactions frequently slowed folding, trapping it at the intermediate state [19]. Experimentally EnHD collapsed to the intermediate state in approximately a microsecond, and the correct docking of the helices took another 15 μs [16,20]. In addition, the refolding pathway mirrored the unfolding pathway [19], including passing through the same TSs in the forward and reverse directions, as was also observed in the simulations showing direct microscopic reversibility at the T_m, presented above [14]. Figure 4 shows the states from the native (N), nearly native

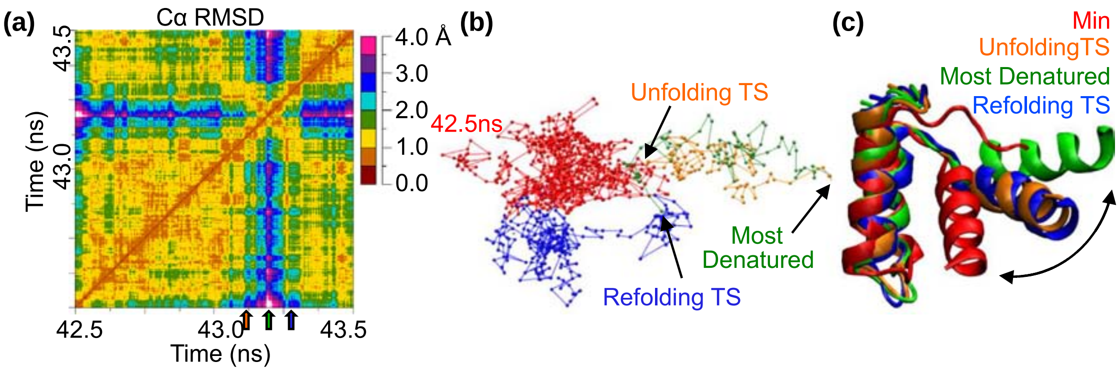


Fig. 3. Comparison of unfolding and refolding transition state ensembles and pathway for EnHD via RMSD matrix, corresponding projection and structures demonstrate that the protein adopts the same conformations during unfolding and refolding [14]. (a) An all-*versus*-all Cα RMSD matrix showing the unfolding TS (43.112 ns, marked by the orange arrow), most denatured state (43.195 ns, green arrow), and the refolding TS (43.269 ns, blue arrow). The visible "X" in the matrix 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) The 3D multidimensional scaling 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: 42.5 ns to the unfolding TS (red), unfolding TS to the most denatured conformation (orange), most denatured to the refolding TS (green), and 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 that the conformations the protein took were very similar and in reverse order. (c) Structures of the unfolding TS (orange), refolding TS (blue), most denatured conformation (green), and the starting minimized structure (red). The structures were fit based on the Cα atoms of HI–HII. The two TS structures are nearly identical, while they are distinct from both N and D. This fine-detail examination of the time period where the protein unfolds and refolds not only demonstrates the folding process, but that the process is reversible. In particular, the matrix projection shows that the protein takes the same path to go forward and backward. The superimposed cartoon structure further supports this, showing how close the refolding TS structure is to the unfolding TS. Figure and figure legend taken from Ref. [14]. Adapted with permission from McCully *et al.* [14]. Copyright 2008 American Chemical Society.

(N′), TS, and denatured state (D), and the pathway EnHD took between them during the refolding simulation. The structure of each state is shown along with a map depicting times at which the protein transitioned between states, illustrating the close correspondence between the folding and unfolding in continuous trajectories when directly sampling interconversion.

Multidimensional analysis for elucidating protein behavior: property space analysis and PCA

The preceding discussion introduces examples of several analysis methods for MD that provide insight into the folding/unfolding process of EnHD, but one is still frequently left with the problem of how to weight each of these properties of the protein and their relative effects on the un/folding pathway [21,22]. For unique folding pathways, folding events may be better explained by one data set or analysis method than another. Techniques for showing the aggregate effects of the output data from each property include property space analysis over a trajectory, where many properties are used together to determine overall effects, along with principal component analysis [22,23]. Property space values can be projected along their principal components in order to determine correlation of values of property space, highlighting the properties that work in tandem to contribute to the un/folding process [22]. This approach also has the advantage of objective weighting of properties. An example of such a projection is presented in Fig. 5a for the unfolding simulation shown in Fig. 1b. The different conformational states are clearly distinguishable in this abstract representation, and the individual properties of the clustered states can be readily retrieved through the trajectory's metadata.

An important aspect of the folding process that is largely ignored in MD simulations is the effect of intermolecular interactions with other proteins in the solution [24]. MD simulations typically only include one protein molecule. To explore the effect of neighboring molecules on the folding/unfolding process, 32 copies of EnHD were simulated at 298 K, as well as 498 K to unfold the protein in a "test-tube." Comparison of neighboring molecules in the test-tube yielded little effect on the EnHD native state, as shown by the N-state property space distribution obtained by plotting the first two principal components for the single molecule and test tube simulations (Fig. 5a and b, respectively, with representative structures overlaid). In general, each conformational state contains similarities both in profiles and in the fraction of structures; however, the states were conformationally broader in the test-tube simulations. Indeed, several contacts were identified in test-tube simulations that were

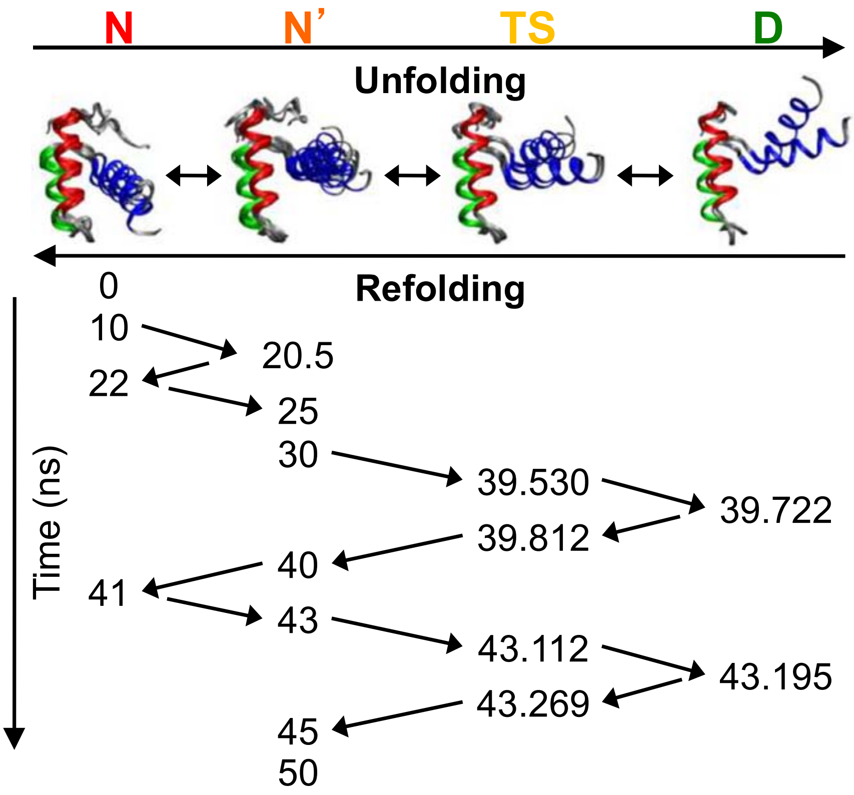


Fig. 4. Unfolding and refolding conformations of over time EnHd show microscopic reversibility between the states [14]. A states-*versus*-time plot shows switching between the states during a refolding MD simulation. The time each state is populated is shown in ns, and arrows display the direction of transition between states. HI (residues 8–20) is shown in red, HII (residues 26–36) in green, and HIII (residues 42–55) in blue. The states are distinct, although there is some variation within each state: in the N and N′ structures, HIII forms a 15° angle with the HI–HII scaffold, while in the TS, it reaches 30° and becomes wider in D. Within 41 ns, the structure has unfolded and refolded. The time-graph visualization of unfolding and refolding along with the corresponding structures shows that the path the protein takes is reversible and goes through the same conformational states each time, and that it may happen multiple times during a simulation. Figure and figure legend taken from Ref. [14]. Adapted with permission from McCully *et al.* [14]. Copyright 2008 American Chemical Society.

consistent with single-molecule simulations, including the refolding contacts. The test-tube unfolding simulations showed that the contacts were typically lost in the order similar to the reverse process. Nevertheless, the difference in the environment and the subtle increase in contacts with neighboring molecules leads to stabilization of the intermediate in the test-tube simulations, as can be seen by the increase in the population of the intermediate at higher concentration. This finding builds upon the intramolecular interactions in refolding simulations that stabilize the intermediate by supplementing them with intermediate-stabilizing intermolecular interactions with neighboring molecules, thereby increasing the population of the intermediate state.

Transforming contacts and fluctuations into spatial images via fast information matching

Another visual tool for MD simulation data is spatial heat maps, which were applied by Kovacs and Wriggers [25] to the 225 °C thermal unfolding simulation of EnHD presented in Fig. 1b. This approach maps properties such as dihedral pivot angle for Cα RMSF, or contact residue mutual information (MI) onto the crystal structure of a protein using fast information matching (FIM), shown as activity measures in Fig. 6a. RMSF is similar to RMSD; however, it uses the average structure over a designated time period as the reference, providing information about fluctuations and mobility of the chain [13]. The contacts in this case were defined in two ways: by two cutoff distances and by Generalized Masked Delaunay (GMD) tetrahedralization of a coarse-grained model, which is a global measure similar to RMSF [25]. The contact residue MI is plotted in Fig. 6b–e and projected on the structures in Fig. 6f–k. The heat map approach has the advantage of displaying contacts that are lost during the unfolding process. This is an intuitive way to determine which residues to examine during the un/folding process to determine those that are essential to the pathway.

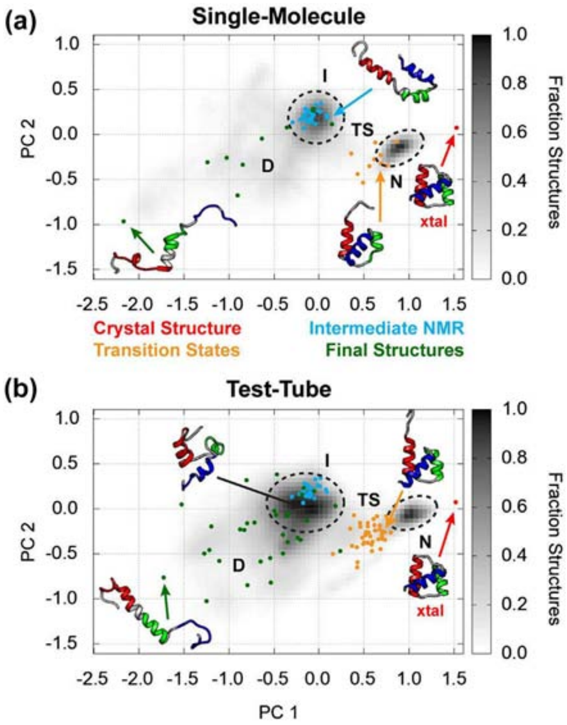


Fig. 5. Property space distributions for both single-molecule and test-tube simulations indicate that EnHD occupies similar conformational states in each [24]. The property-space distributions for the first two principal components of a 39-dimensional property space are shown for (a) single-molecule and (b) test-tube simulations. Simulations at 298 and 498 K were normalized independently to the bin with the maximum number of structures, rendered from white (0) to black (1). Structures and clusters show the native state N, transition state TS, intermediate I, and denatured state D. The crystal structure is shown in red, the TS in orange, the L16A ensemble in cyan, and final structures in green. It is clear that the test-tube simulations are mostly consistent with the single-molecule simulations, with very similar regions of property space populated. This suggests that the folding behavior in an *in vitro* environment with many proteins would behave almost the same as in a single-molecule simulation, showing that observations from a single-molecule simulation can be applied to expect similar experimental results. Superimposing the structures on the principal component graphs provides a clear idea of the direction of the protein through property space and shows the similar conformational states for each cluster, with some differences. The white to black map shows the conformational broadness of the single-molecule *versus* the test-tube simulations, with the intermediate state in particular being more broad for the test-tube simulations. Figure and figure legend taken from Ref. [24] and used with permission.

Indeed, one of the residues flagged using this process, Lys52, has a significant impact if mutated to Ala, which approximately doubles the folding rate. Using this approach, the time series is first characterized and broken into fast and slow degrees of freedom, where the statistical dependence between variables is measured and ranked in the order of importance for folding dynamics. The heat maps illustrate the locations of contact loss among the three helices, which are consistent with the Cα RMSF throughout the simulation.

Detecting transitions between protein states using SASA

The SASA is routinely used to examine unfolding, which can aid in detecting when a protein transitions between states and residues critical to the process, particularly to highlight when core hydrophobic residues are exposed to solvent [13]. Conversely, SASA can also readily identify when hydrophobic collapse and refolding occur [26,27]. Koulgi *et al.* [28] performed replica-exchange MD simulations, which provide sampling over a wide variety of temperatures and configurations, to examine the folding pathway of EnHD. However, note that replica exchange does not provide continuous trajectories, so pathways per se cannot be determined and instead are inferred. The structure with the lowest Cα RMSD from the simulations was examined with respect to the native structure (Fig. 7a). In Fig. 7b, the SASA difference between the two structures is plotted by residue for 10 significant hydrophobic amino acids, which are also portrayed as stick models in the visualization. The smallest differences in SASA and side-chain orientation between the native and lowest structures are experienced by the core residues, which are intact in the intermediate, as discussed above.

EnHD has been a good model system for establishing many of the methods to probe protein folding by both experiment and simulation. However, this leads to the question of whether other proteins with different structures display similar behavior, particularly those with more complicated architectures and changes in the sequence. Below we highlight findings from other systems that address this issue.

Characterizing Intermediate States and Their Interconversion

The FF domain is another fast-folding helical protein that has been well studied. A combined MD and experimental study showed that the protein forms identifiable folding intermediates [29]. The Cα pairwise RMSD matrices over all structures during five low-pH WT simulations were used to identify intermediates (Fig. 8a). Three structurally unique intermediates, I_1, I_2, and I_3, were found at low pH with interconversion through I_1 (Fig. 8b). Clusters on the diagonal represent temporal clusters of similar structures, while clusters off the diagonal represent similar conformations formed at different times during the simulation or in different simulations (Fig. 8c, d). The pathway was less complicated at neutral pH, with a combination $I_{1/2}$

intermediate (Fig. 8a). Thus, while other intermediates are populated, only one is necessary: I_1. Transitions between the intermediates are summarized in Fig. 8b and d with structure thickness scaling with the standard deviation of the Cα positions in each intermediate ensemble and structures colored to show location of helices. The visualization of the folding pathways in this manner provides a simple representation of the observed routes. Overall several different intermediates were observed; however, they were all very similar, which suggests that they may be conformational substates of the intermediate state (Fig. 8).

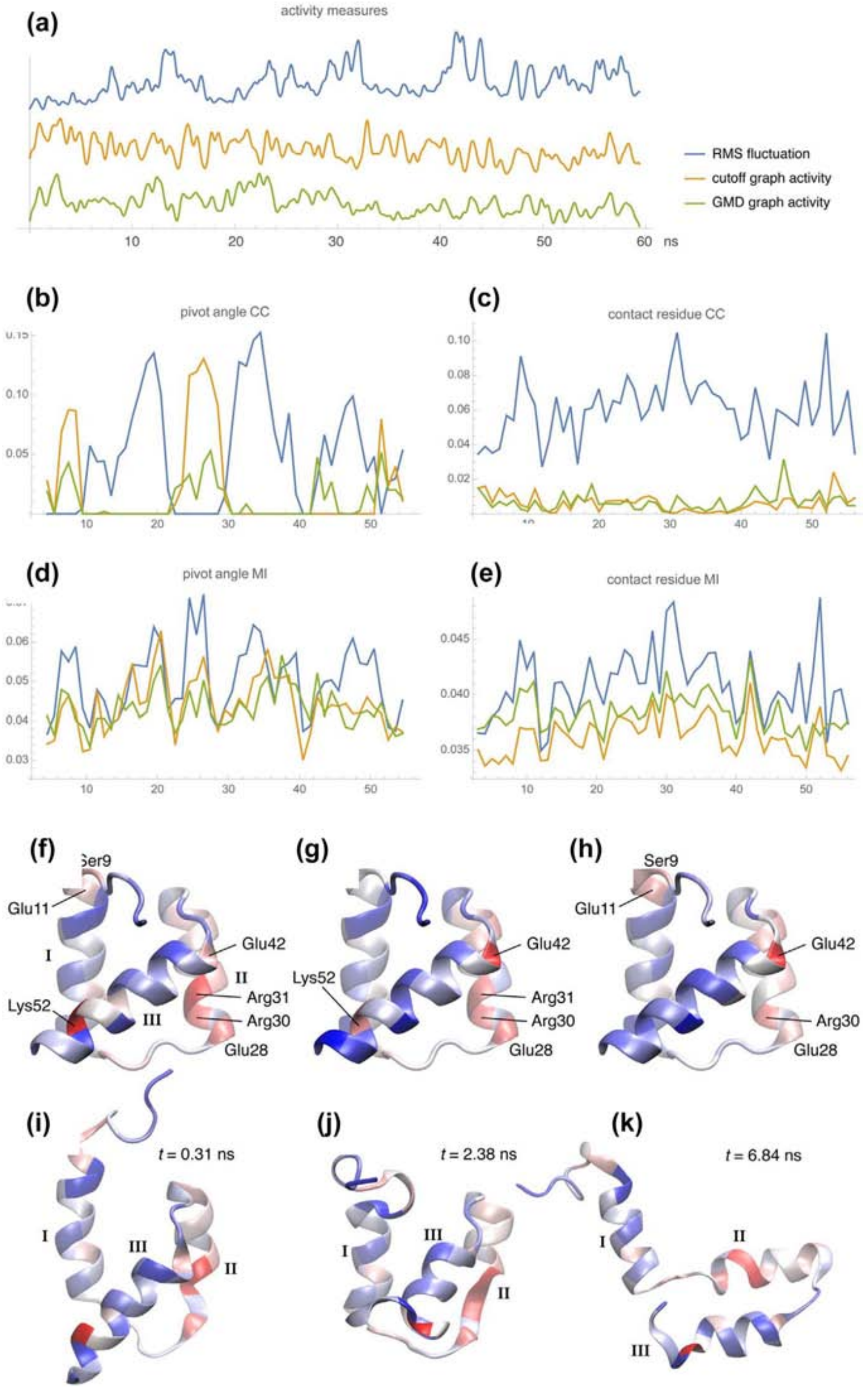


Fig. 6 (*legend on next page*)

The Folding Pathway and Mutations that Change the Pathway

Barnase is another model protein for folding studies with a more complicated structure than those presented above. The multidimensional scaling of the all-by-all Cα RMSD, as depicted in Fig. 3b for EnHD, is shown for thermal denaturation of barnase (Fig. 9a) [30]. Two distinct intermediate states are populated. When visualizing unfolding, superimposing representative structures from the clusters on a simple plot near their respective time points can clarify the structural differences and illustrate the mechanism for how a conformational change occurs during un/folding. This is demonstrated in Fig. 9b via a plot of the Cα RMSD over time with critical structures indicated.

Although we discussed above how property space analysis can reduce the number of dimensions while taking many properties into account simultaneously, we have not yet explored the 3D projection of the property space vectors onto the principal components in a path view. Unfolding trajectories of wild-type and mutant barnase are shown projected along the first three principal components in Fig. 9c [22,31]. All the trajectories follow similar paths in space except for Y17G. Consistent with the projections, I88V unfolds and folds by the wild-type pathway, while Y17G does not in experimental studies [31]. By plotting two of the physical properties contributing to the principal components in Fig. 9c, the plot in Fig. 9d shows that while the β-sheet unfolds before the helix in I88V and wild type, the unfolding of the helix and sheet is more coupled in Y17G with the helix leading [32].

Localized Unfolding Due to Mutation

The barnase mutations were engineered to interrogate the folding pathway, but the p53 protein represents a nice model system for localized unfolding in response to changes from naturally occurring mutations, in this case associated with cancer. One mutant of p53 that has been particularly well studied is R282W (Fig. 10a) [33]. In simulations of WT p53 and the R282W mutant at 310 K, a diverse range of MD output properties was compared (Fig. 10b–f).

The RMSF analysis shows which parts of the protein are moving and the degree of heterogeneity across three independent simulations, shown in Fig. 10b. This analysis based on atomistic motion is another way to extract major modes and filter out fast vibrations and fluctuations, which when displayed as a function of residue number, highlights the mobile portions of the protein. For example, the L1 loop region of p53 is destabilized and more mobile in the R282W mutant than in the WT simulation (yellow vertical stripe near preceding residue 125 in Fig. 10b).

A correlated motion analysis was also performed, with the all-*versus*-all correlated motion matrices displayed in Fig. 10c. There do not appear to be consistent, significant differences between the WT and mutant simulations. However, another useful method for identifying nonrandom motion during a trajectory for further targeted analysis is the continuous wavelet transform, a signal processing technique [13,34]. If a single atom has similar motion during the simulation to a specific wavelet function, that type of motion can be detected. A *p*-value can then be calculated, for one or many simulations, to quantify the atom's propensity for certain types of motion. The wavelet analysis in

Fig. 6. Activity measures, contact residue values, and structure heat maps throughout the EnHD unfolding trajectory identify important locations and residues for folding [25]. (a) Three activity measures as functions of the simulation time in ns. RMSF (blue line) were computed with the TimeScapes agility.py program using a sliding window of length δ = 500 ps, and increased throughout the simulation as the protein unfolded because of the loss of packing interactions. The parameters used by terrain.py were cut1 = 6 Å, cut2 = 7 Å, and δ = 500 ps for the cutoff graph activity (orange line) and cut1 = 2, cut2 = 3, and δ = 500 ps for the GMD graph activity (green line). The cutoff graph activity and the GMD activity were defined by the combined rate of contact-forming and -breaking events, with different definitions of contacts depending on cutoff distances and tetrahedralization of the coarse-grained model, respectively. The unfolding activity of the protein caused these measures to decrease with time, in contrast to the RMSF. (b) Pivot angle cross-correlation (CC) and MI profiles as functions of residue number. Pivot dihedral angles were attributed to the half-points between the residue indices of the center atoms. The three curves in the plots (b–e) correspond to the activity measures in panel a. Molecular graphics figures of heat maps and 3D conformations were created with the program VMD [13] using a linear red-white-blue color scale (from high to low values). EnHD heat maps computed from activities in FIM. Roman numerals label the three helices HI, HII, and HIII. (c) Contact residue MI heat map for RMSF. (d) Contact residue MI heat map for cutoff graph activity. (e) Contact residue MI heat map for GMD graph activity. (f–k) EnHD heat maps computed from activities in FIM. The three helices are labeled HI, HII, and HIII. (f–h) The corresponding maps from panel a mapped onto the starting structure. (f) Maps the RMSF, showing a high value for K52. (g) Maps the cutoff graph activity, pinpointing E42. (h) Maps the GMD graph activity, with E42 standing out as a high value as well. (i–k) Snapshots of unfolding events along the trajectory with the contact residue MI heat map from panel f superimposed. Showing the RMSF heat map superimposed on the structure provides spatial locations that are significant to unfolding and folding events. K52 (shown in panels f and g) was elucidated as an important contact from this analysis and had been previously identified as a residue that slowed the folding rate of the protein. Figure and figure legend taken from Ref. [25]. Adapted with permission from Kovacs and Wriggers [25]. Copyright 2016 American Chemical Society.

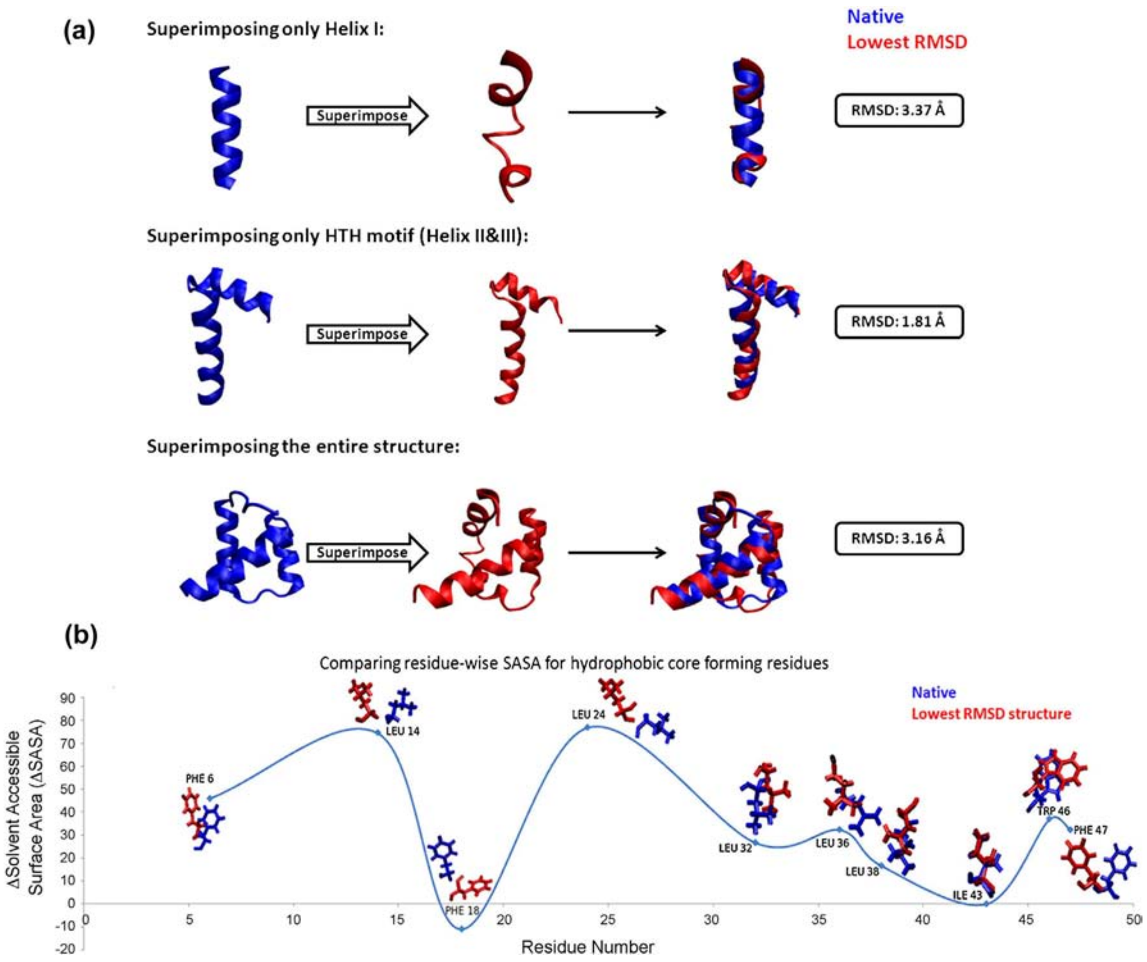


Fig. 7. Structural and SASA comparisons of the lowest RMSD structure of EnHD in a folding trajectory to the native structure [28]. (a) Partial and total structure superimposition of the lowest RMSD structure (red) and the native structure (blue). Overall, HII and HIII together have lower differences than only HI between the two structures. The structure from the total superimposition (bottom) was used for analysis of the hydrophobic core. (b) SASA difference between the lowest RMSD structure for the total superimposition and native structure for the 10 hydrophobic residues in the core (6, 14, 18, 24, 32, 36, 38, 43, 46, and 47), which are typically buried in the native structure. The native structure is portrayed in blue and the lowest RMSD structure in red. A smaller difference between the native structure and the lowest RMSD structure indicates similar hydrophobic packing, which occurred for L32, L36, L38, I43, W46, and F47. The stick structures of residues are most similar for these residues, with much greater differences shown for F6, L14, F18, and L24, the first four residues. The first four residues of the core were not part of the helix-turn-helix (HTH) motif present in the last six, indicating that the HTH structure has separate and stronger packing interactions from the rest of the protein. Figure and figure legend taken from Ref. [28] and used with permission.

Fig. 10d shows that there are significant bands of motion in WT and in the mutant in simulation 2, but there are less significant wavelet matches for the mutant overall. This is better visualized when mapped onto the structure, showing the loop regions of the protein in particular that have significant differences between the mutant and WT (Fig. 10e). This localizes the differences and effects of mutation and provides clues of where to compare contact differences to obtain higher-resolution mechanistic information. Figure 10f shows the DNA-binding surface of the protein and how normally exposed side chains become buried by the end of the simulation in the mutant, illustrating how a mutation can affect folding and protein function [35]. These results illustrate how a combination of detailed plots and lower resolution displays helps to communicate the complicated conformational changes experienced upon mutation and the necessity of including depictions of protein snapshots for an intuitive understanding of the process.

Protein Folding Networks

Similar to graph visualizations, Rao and Caflisch [36] mapped the conformational space of a peptide to

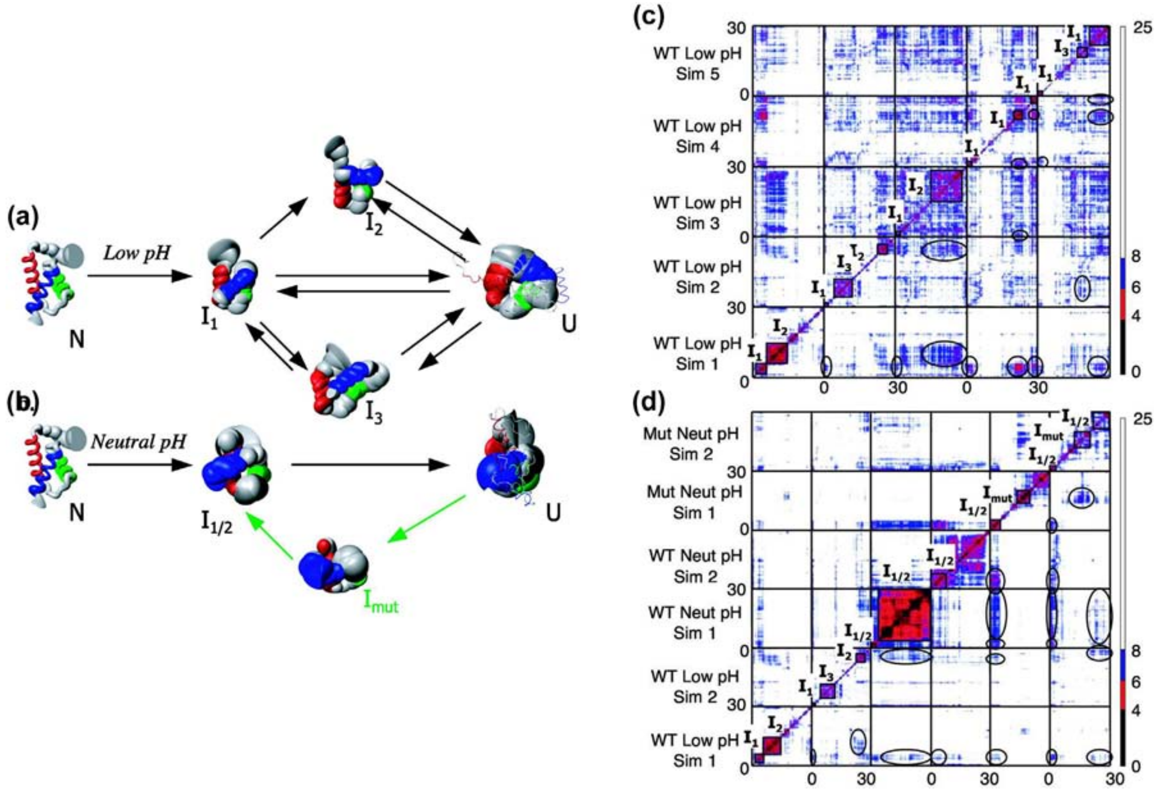


Fig. 8. Pairwise RMSD and transition states for the FF domain [29]. The native structures and reaction schemes observed from low (a) and neutral (b) pH simulations. Arrows indicate transitions between states observed during simulations; however, it is possible that some states may have not been observed and that this is not a complete representation of conformations. The green arrows and text for the neutral pH simulation indicate a transition to I_{mut} that is observed only in the simulations of the L25A/D46N mutant protein. The mean structure from each state is shown, with ribbon thickness dependent on standard deviation for each Cα's position in the ensemble. Each structure is colored by native helix position: the N-terminal helix is red, the middle helix green, and the C-terminal helix is blue. The native ensemble is taken from a 30-ns simulation at neutral pH, while intermediate ensembles are pooled from the clusters indicated in (c–d). The denatured ensembles are taken from regions of their respective simulations that do not fall into intermediate clusters. The denatured ensemble structures that are most similar and most dissimilar to the mean are overlaid over the mean denatured structure to depict the range of structures in the denatured state. (c–d) Pairwise Cα RMSD plots for multiple simulations. Each structure is compared with all other structures in all simulations at 50-ps resolution over the 30-ns simulations. The *x* and *y* coordinates indicate the time points being compared, with color-coded RMSDs. (c) Pairwise comparisons of the five simulations of the protein at low pH. The magnitude of the Cα-RMSD is given in the color scale to the right of the plot. On-diagonal clusters, indicative of local clusters in a simulation, are boxed off. Off-diagonal clusters are circled, with the on-diagonal clusters belong to a common unfolding intermediate. On-diagonal clusters are grouped into three structurally distinct intermediates as indicated by the labels I_1, I_2, and I_3 with structures shown in panel a. (d) Pairwise comparison between two simulations at low pH, two at neutral pH, and two of the L25A/D46N double mutant at neutral pH. An intermediate that is structurally similar to that seen at low pH ($I_{1/2}$) is seen in all neutral pH simulations, and a new intermediate (I_{mut}) is seen in the mutant simulations. The properties of the intermediates are very similar, with I_2, I_3, and I_{mut} all being variations on I_1. From observing the simulations, all went through I_1, but not all went through the other intermediate states, indicating that I_1 is a necessary intermediate state for the folding pathway. Figure and figure legend taken from Ref [29] and used with permission.

a network. Beta3s is a 20-residue peptide containing antiparallel β-strands. Structures from a 330 K simulation were grouped into nodes of a network according to secondary structure, connected by links (Fig. 11). A random heteropolymer of the same length was used as a control.

The network depicts free energy minima and the relationships between them without relying on an arbitrarily chosen reaction coordinate, and the network can be used to identify TS ensemble conformations. Secondary structure was calculated for each conformation. Depending on the free energy of the conformation and the number of snapshots with the same secondary structure, weights were assigned to each node, allowing sampling of only the conformations with significant weight. The nodes

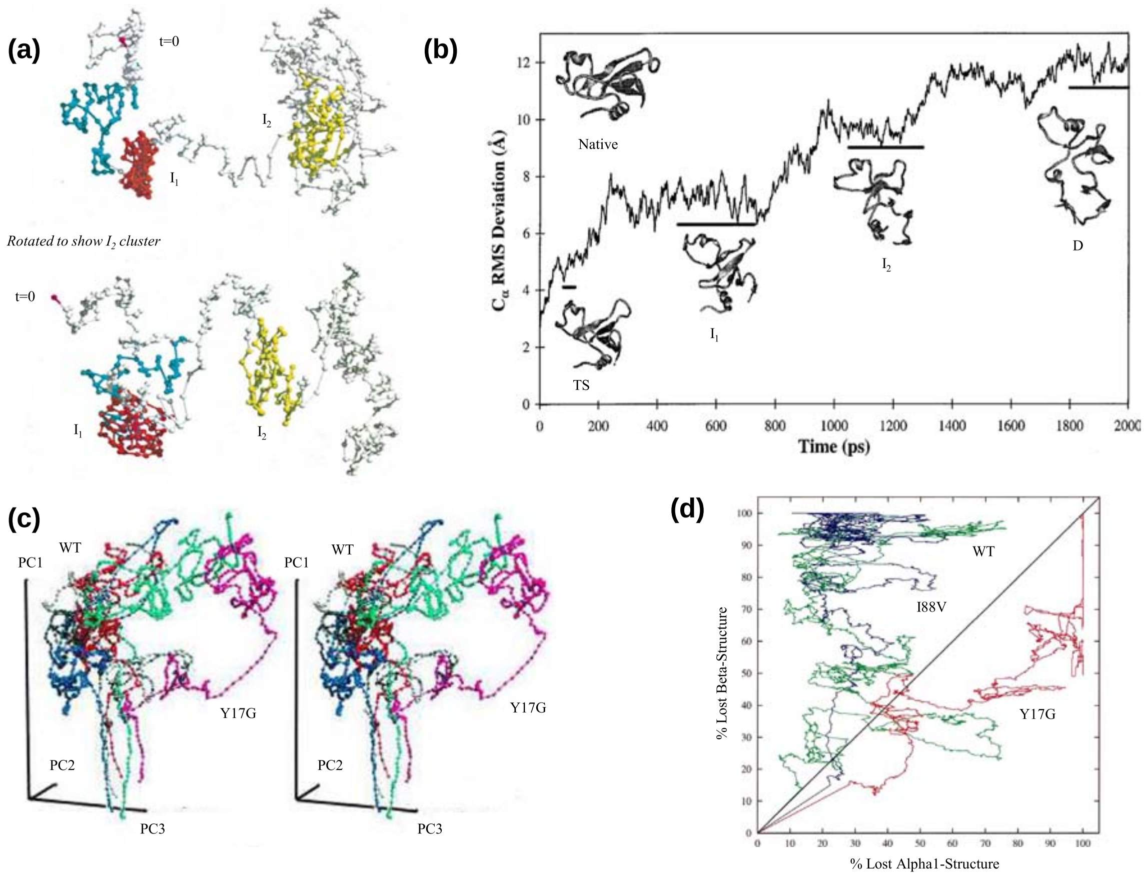


Fig. 9. Conformational clustering, RMSD, PCA, and DSSP of barnase demonstrate the secondary structure loss and pathway during unfolding [30,22,32]. (a) Conformational clustering analysis of protein structures sampled during the unfolding simulation [30]. The points are connected sequentially in time (3.2ps/point), and a farther distance between points indicates a higher RMSD between structures. The 250- to 468-ps time period is shown in cyan, 470 to 730 ps in red, and 1050 to 1300 ps is in yellow. The clustering allows for identification of two intermediate states(b) The Cα RMSD from the average NMR structure as a function of time [30]. Average protein structures for selected time periods during the MD simulation are shown. The native structure is provided for comparison. As the RMSD increases and plateaus at different states throughout the simulation, the protein enters the transition state TS, intermediates I_1 and I_2, and finally the unfolded state U. RMSD in context with structures presented in this way can provide clear information about the steps of the unfolding pathway. (c) Stereoview of five barnase unfolding trajectories (0–2 ns) projected along their first three principal components [22]. The three WT trajectories are white, cyan, and blue, while the I88V trajectory is red and the Y17G trajectory is magenta. The trajectories start following the first principal component as the structures expand, increasing SASA as tertiary contact loss occurs. When the protein passed through intermediate states, lower property space distance was observed between trajectories. Later in the projection, the Y17G mutant trajectory moves away from the other unfolding trajectories and travels along a different path, showing that the mutation has a greater effect on the unfolding pathway than I88V and supporting previous observations that Y17G folds via a different pathway. At the later time periods, the principal component projection fluctuates as the structures reach the denatured state and the ensemble becomes heterogeneous. (d) Loss of β-structure and helical content of α1, the main α-helix in barnase [32]. The points are connected sequentially in time and show that at the beginning of the simulation, between 20% and 40% of helix 1 structure was lost, followed by a severe loss of β-sheet structure. While both the WT and I88V mainly lost β-sheet structure for the remainder of the simulation until WT lost some α-helix 1 structure at the very end, the Y17G mutant lost virtually all of its α-helix 1 structure and then immediately lost the rest of its β-sheet structure. These analyses show that the Y17G mutant had complex effects such that it changed the unfolding pathway significantly even relative to the other mutant, I88V. Figures and figure legends taken from Refs. [30,22,32] and used with permission.

were linked together for snapshots close in time (within 20 ps) or separated by conformation(s) with less than 20 snapshots each. This network is robust for a wide range of threshold values, the simulation length, and the node definition. For each TS ensemble, a set of structures with the same probability of folding, 10 short trajectories were performed varying random seeds on a single node to determine if folding occurred. If the fraction of native contacts was greater than 22/26 at the start of the simulations, the trajectory led to folding. Overall, the network in Fig. 11 nicely illustrates that while there are several pathways for beta3s to reach a folded state, there are two main folding pathways.

Mapping and Clustering to Characterize Free Energy States

Ubiquitin is another protein that has been well studied; it contains a 5-stranded β-sheet, α-helix, and a 3/10 helix. Its folding behavior has been explored by a variety of groups but perhaps most comprehensively by the Shaw Lab [37]. As with the EnHD simulations described above, the simulations of ubiquitin were performed near its T_m, at 390 K, to observe both folding and unfolding events. Six of the simulations began from native structure, while two started from the unfolded state. Two of the simulations showed reversible folding and unfolding, as tracked by several analysis metrics.

Sudden shifts and plateaus in Cα RMSD can help identify unfolding and folding events. Figure 12a shows the Cα RMSD plotted with time from all residues, excluding the N and C termini, from the native structure in two reversible folding simulations. Over less than 1 ms in both simulations, the protein unfolds and refolds, as demonstrated by the large changes in Cα RMSD. The autocorrelation function between the RMSD and the secondary structure content of the protein is displayed in Fig. 12b. The Cα RMSD relaxation (red line) can be well modeled by a biexponential fit (black line). The timescale of the relaxation is consistent with the observed folding and unfolding rates, and supports ubiquitin's two-state folding behavior.

However, as shown for many other proteins, there are additional metastable states beyond the typical

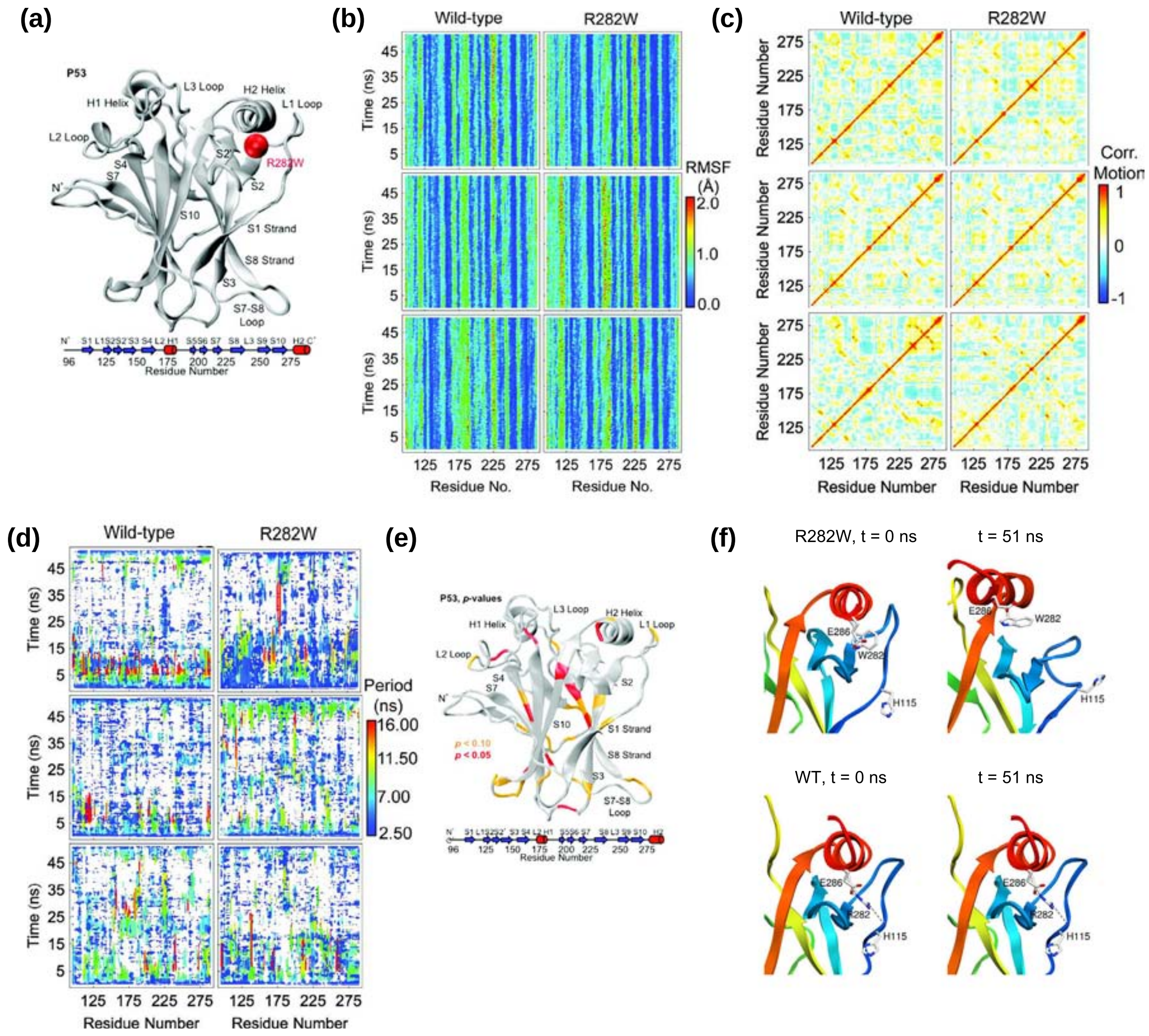


Fig. 10 (*legend on next page*)

two-state behavior, including three misfolded states (MF1–3) that comprise only a few percent of the structural ensemble. Figure 12c portrays the clustering analysis, showing two major unfolded states (U and U1), characterized by the formation of hairpin 1 (shown in red) in the U1 state. There are two folded structures shown in the clustering analysis as well, F and F1, where F is the native state and F1 is nearly native with the C-terminal region unstructured. Each state is displayed using 20 random structures. Such clustering analysis is advantageous because it helps identify patterns in possible overlooked states and substates.

The free energy surface was then projected along a one-dimensional reaction coordinate, indicating the energy barrier contributing to the likelihood a protein would be found at each state. The clustering analysis from Fig. 12c was particularly helpful in finding states at energy minima, shown in the free energy plot (black line) for each state in Fig. 13, with corresponding structures shown below; however, the intermediate state (I) shown in the free energy plot was not identified via clustering analysis because the relaxation rate was too fast and it was not heavily populated. Mapping the free-energy surface provided additional insight, helping to identify an additional F′ folded state where the C-terminal helix region was melted along with the intermediate not found via clustering analysis. Several other properties were projected along the reaction coordinate as well, including the number of helical residues (blue), sheet residues (orange), Cα RMSD (green), contact order (purple), and fraction of native contacts (red). Overall, there was a positive correlation between the reaction coordinate and these properties. The transition state occurred relatively late into the folding pathways. These observations suggest a homogeneous folding pathway, where the first β-hairpin forms along with part of the helix, next the C-terminal loop is consolidated, followed by the 3/10 helix and fifth β-strand. Several of these conformational states were not evident in individual analyses, demonstrating the necessity of using a variety of different approaches.

Proteins with High-Sequence Identity, but Different Folds: Visualizing the Denatured State and Determinants of Fold Switching

Although the sequence of a protein contains the information needed for folding, it has been very difficult to predict structure from sequence. In order to investigate the effects of sequence on structure, two proteins were derived to have a high sequence

Fig. 10. p53 and R282W mutant structure, unfolding simulations, and analysis [13,35]. (a) Minimized crystal structure and secondary structure map of the p53 DNA binding domain with the mutated residue R282W indicated by a red sphere. Helix H2 binds to the major groove of DNA; loop L3 and helix H1 also participate in binding and normally hold a zinc ion. In the wild-type simulations, the R282 residue forms contacts with the backbone of loop L1 holding H2 and the base of L1 close together, allowing the tip of L1 to flex considerably along a single axis. The mutant W282 residue, however, does not form these contacts, allowing H2 and L1 to separate. Residue H115 of L1 forms contacts with the residues of S2 in this situation, preventing L1 from becoming too disorganized but allowing it to swing much more randomly into solvent. (b) RMSF plots over time of each Cα atom by residue for each of the WT and R282W mutant p53 simulations. Notably, residue 120 has slightly higher fluctuations in the mutant. Slightly higher fluctuations can also be seen near residue 225 (L3 loop) in all three of the R282W simulations. Overall, the RMSF stays consistent throughout the simulations for each residue, so for p53, it is difficult to say much about the mutant *versus* WT for this analysis metric; there are not many consistent differences between the simulations shown, necessitating other analyses to discern details of the folding pathway. (c) Correlated motion plots of all simulations of WT and R282W p53. Although differences between all simulations can be observed, consistent differences between WT and R282W simulations are also difficult to find; none of the maps are similar enough for comparison. Anti-correlated motion is slightly more prevalent in the WT simulations than the mutant simulations, however. (d) Wavelet analysis of all simulations of the WT and R282W mutant of p53. The most significant wavelet match at each time is shown for each Cα atom with white indicating no significant match. Low-frequency motion is observed between 5 and 10 ns of WT simulation 1 and is scattered throughout but especially present at residue 120 in the S1 loop of WT simulation 2. There is also less significant motion in the mutant in general, but some motion near residue 180 of H1. (e) Significant differences in ordered motion between the WT and R282W p53 simulations, from the wavelet analysis, are mapped onto the p53 structure. The greatest differences in motion tend to occur in the loops and the strands near the polymorphic site (S1 and S1). (d) Graph distances plotted by residue for both WT and R282W variants of p53 over time. Graph distance represents the amount of change in the types of neighbors a particular graph node or residue has compared to the protein's crystal structure. It is calculated as a Euclidean distance in the space of contact probabilities and thus has arbitrary units. The region near residue 225 (loop L3) shows the greatest variation between WT and R282W simulations, with R282W simulations showing significantly greater deviation. This may be explained by changes in the structure of the H1 helix nearby. (f) Structure of p53 and mutant at the start of the simulation and at 51 ns [35]. Residues H115, R282, and E286 are shown as stick models. While in the WT simulation, H115 stays in a similar conformation, by 51 ns in the mutant simulation, H115 has flipped and the loop has lost structure. This DNA binding surface has been affected by the mutant structure and buried normally exposed side chains, showing the drastic effects of the mutant on folding by the end of the simulation. Figures and figure legends taken from Refs. [13,35]. Adapted with permission from Benson and Daggett [13] and Calhoun and Daggett [35]. Copyright 2012 and 2011, American Chemical Society.

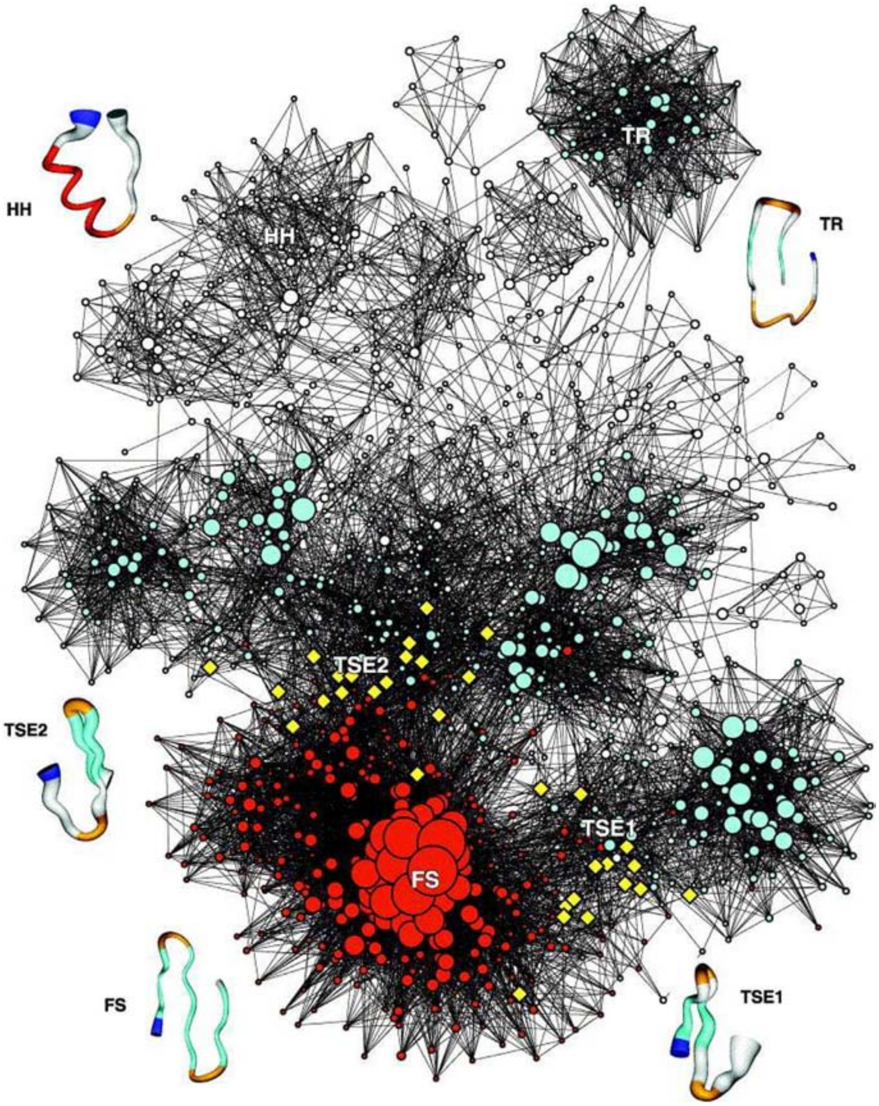


Fig. 11. Beta3s conformational space network showing conformational states during folding [36]. The size and color coding of the nodes reflect the statistical weight *w* and average neighbor connectivity k_{nn}, respectively. White, cyan, and red nodes have $k_{nn} < 30$, $30 \le k_{nn} \le 70$, and $k_{nn} > 70$, respectively. Representative conformations are shown by a pipe colored according to secondary structure: white stands for coil, red for α-helix, orange for bend, cyan for strand, and the N terminus is in blue. The variable radius of the pipe reflects structural variability within snapshots in a conformation. The yellow diamonds are folding TS conformations (TSE1, TSE2) characterized by a connectivity/weight ratio, a clustering coefficient $C < 0.3$, and $60 < k_{nn} < 80$. FS is the folded state, while the denatured state encompasses two conformations: HH and TR. HH is partially helical, and TR is the curl-like trap of conformations with low entropy and enthalpy. The network shows that the denatured state is very conformationally broad and identifies two unfolding pathways to reach the denatured state. Figure and figure legend taken from Ref. [36] and used with permission.

identity of 59%, but different secondary structure. These proteins were developed by Alexander and co-workers [38] using genetic selection from the streptococcal protein G (G_B), whose structure is like that of ubiquitin, and the staphylococcal protein A (G_A), another 3-helix bundle (Fig. 14a). To obtain insight into how such similar sequences give rise to completely different protein folds, high-temperature MD simulations were performed to study the unfolding processes of these two proteins (denoted G_A59 and G_B59, the protein A and protein G variants, respectively, with 59% sequence identity) in comparison to simulations of their native states [39]. The G_B protein unfolded quickly in the high-temperature simulations, and after reaching the denatured state, it formed nonnative α-helical structure, while the G_A

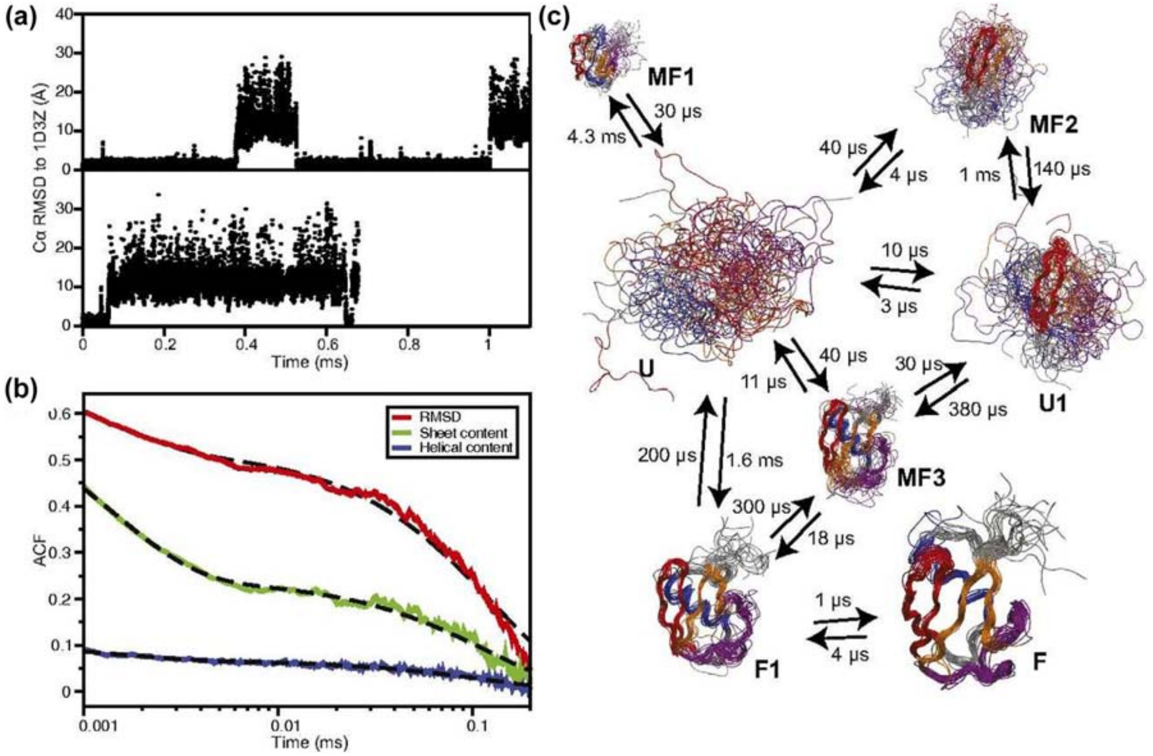


Fig. 12. Ubiquitin unfolding and refolding simulations [37]. (a) The Cα RMSD (for residues 2–71) *versus* time, using the native structure as a reference, for the two simulations with reversible folding. Two states are clearly shown from the simulations, where the RMSD starts low and jumps to over 10 Å before reverting back to the low RMSD of the native structure. (b) Autocorrelation function of the Cα RMSD (red), the number of helical residues (blue), and the number of residues in β-sheets (green). The dashed black lines are biexponential fits to the autocorrelation functions with characteristic times of ~1.5 and ~120 µs. This function was used for clustering and identification of conformational states. (c) Kinetic model of the folding free-energy surface. For each state, 20 random structures are displayed and scaled according to state population. Hairpin 1 is shown in red, hairpin 2 in orange, the α-helix in blue, and the C-terminal loop containing the 3/10 helix and the fifth β-sheet in purple. A higher number of states results in negligible improvement on the quality of the fit. The transition times between states are also reported. U combines two unstructured unfolded-state clusters that were generated by the fitting procedure and are here merged in a single state for simplification. Here, the RMSD provides evidence for unfolding and refolding, while the autocorrelation function aids with clustering, the effects of which are displayed in a kinetic model of the protein's conformational states. Figure and figure legend taken from Ref. [37] and used with permission.

protein was highly disordered upon denaturation, with very little helical structure after the first 10 ns (Fig. 14b) [39], shown for a further evolved pair with 88% sequence identity [40]. The contact maps in Fig. 14b illustrate the difference in the TS and denatured states of the two folds and the increase in nonnative contacts as the proteins unfold. For example, note the retention of α1 in G_B and loss of that same helix in G_A. Furthermore, G_B also adopted nonnative helical structure while the G_B denatured state lacked residual helix [39]. Based on these simulations, it was proposed that interactions in the denatured state direct folding and determine which fold is adopted. In particular the central helix in G_B segregates the residues on either side to allow the β-hairpins to loosely form through side-chain interactions, setting up the native topology.

The MD simulations for the 59% and 88% sequence identity pairs were quite similar, confirming the role of residual structure in determining which fold was adopted, narrowing down the search. Independent experiments confirmed the MD predictions [39]. Overall, by visualizing the denatured state, the effects of small sequence differences elucidate how events early in the folding pathway determine the ultimate native structure adopted, which would not have been possible without the MD simulations.

What is even more remarkable is that Alexander *et al.* [41] were able to further evolve the sequences so that a single residue change (Leu or Tyr at residue 45) determines whether the G_A or G_B fold is adopted (Fig. 15a). In an orthogonal approach, the interactions in the denatured states of G_A88 and G_B88 were analyzed in depth, and a Thr1–Glu19 hydrogen bond

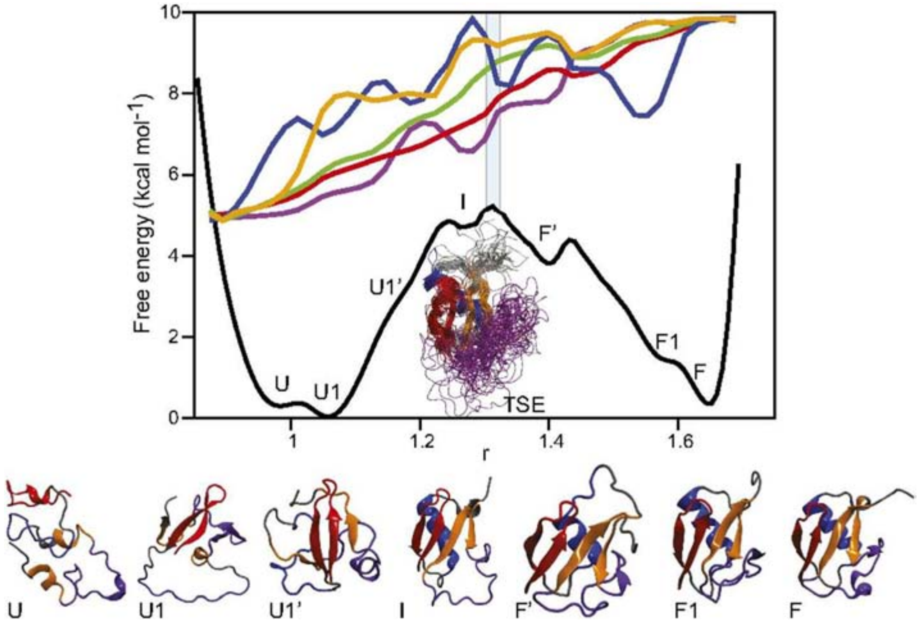


Fig. 13. Folding free-energy surface of ubiquitin with structures representing each identified state [37]. The folding free-energy surface of ubiquitin is projected along an optimal one-dimensional reaction coordinate. Representative structures for each well on the folding free-energy surface are visualized as a cartoon structure according to the color scheme of Fig. 12c. The U well represents the completely unstructured protein, while the F well represents the native structure, which is on average a 0 .8-Å deviation from the Cα RMSD from the x-ray structure. The same letter code is used from Fig. 12's clustering analysis to identify distinct structural states. The free-energy surface analysis, however, demonstrates the limitations of the clustering analysis in Fig. 12, which did not identify the I, F′, or U1′ states. I was not heavily populated enough to be shown as a distinct cluster, and F′ and U1′ interconverted too quickly between states to be recognized as unique clusters. Using the free-energy surface as an additional analysis metric confirms the existence of states found by the clustering analysis while providing additional states that were overlooked by clustering. The average values of several structural properties (blue, number of helical residues; orange, number of sheet residues; green, Cα RMSD; purple, contact order; red, fraction of native contacts) are also plotted as a function of the reaction coordinate. Structural properties were normalized to the same U and F states for accurate comparison. Figure and figure legend taken from Ref. [37] and used with permission.

was found in nearly all of the G_B88 denatured state structures at neutral pH [39,42]. At low pH, protonated Glu19 interacted with solvent, and the bond did not form. As a result, Thr1 formed other long-range interactions with Asp47 and Glu48. In the G_A88 denatured state, this interaction already occurs, stabilizing the α-helical structure. The analyses indicate that long-range interactions play a strong role in determining the nature of the residual structure in the denatured state, and Thr1–Glu19 interactions appeared to be critical [42,43]. Consequently, based on the MD, it was predicted that a single minor mutation, changing a Glu➔Gln, would break the interaction, thereby favoring the G_A fold. The mutant was made, and CD confirms that this single minor change leads to the formation of the helical G_A fold (Fig. 15b) [43]. This study demonstrates how coupled analysis and visualization of the unfolded states can lead to the design of a subtle change that triggers a fold switch, something that could not have been done by inspection of the native states.

Interrogating Proteins using Interactive Visual Analytics and beyond

The level of detail necessary to characterize protein folding/unfolding mechanisms and the associated data sets produced can make analysis both challenging and tedious. To deal with the analysis of big data, DIVE (Data Intensive Visualization Engine) was created; it is a software framework that provides visual analytics interrogate and characterize protein dynamics and other large data sets [44,45]. DIVE was initially created to examine the Dynameomics data set [46], which contains MD simulations for represents of essentially all protein folds. The main window of the DIVE user interface is shown in Fig. 16a, showing the structure of the protein superoxide dismutase 1 (SOD), a β-sandwich fold protein similar to p53, along with its contact map illustrating a particular contact. ContactWalker, a program included in DIVE, integrates visual graphs in order to show networks of contact changes between WT and mutant structures [47,48].

Residues are represented by graph nodes, while their contacts are represented by graph edges, which in this case are proportional to contact occupancy change. Target residues for examination can be manually identified if the region of interest is known, or the user can specify a tolerance for ContactWalker to filter high

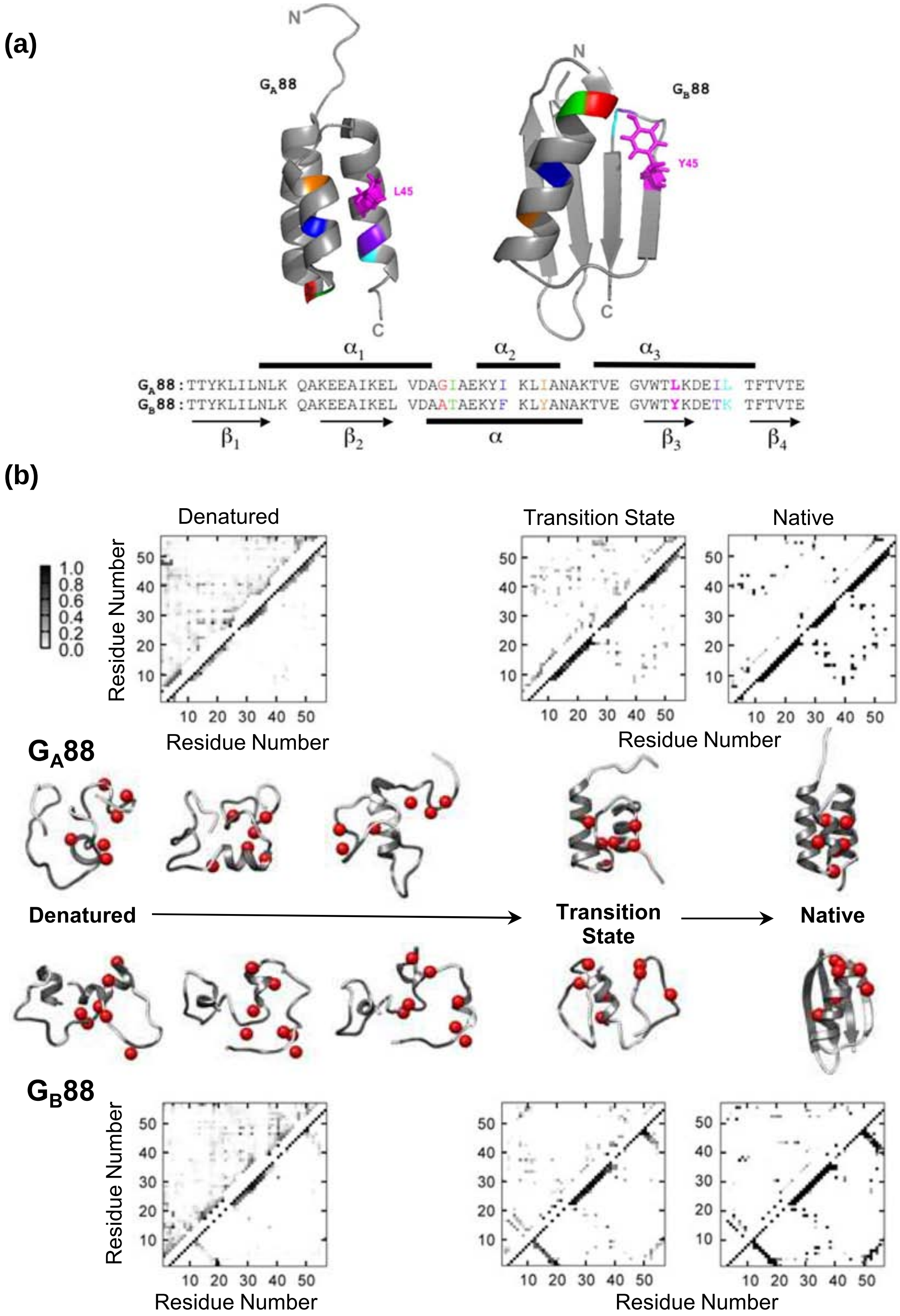


Fig. 14 (*legend on next page*)

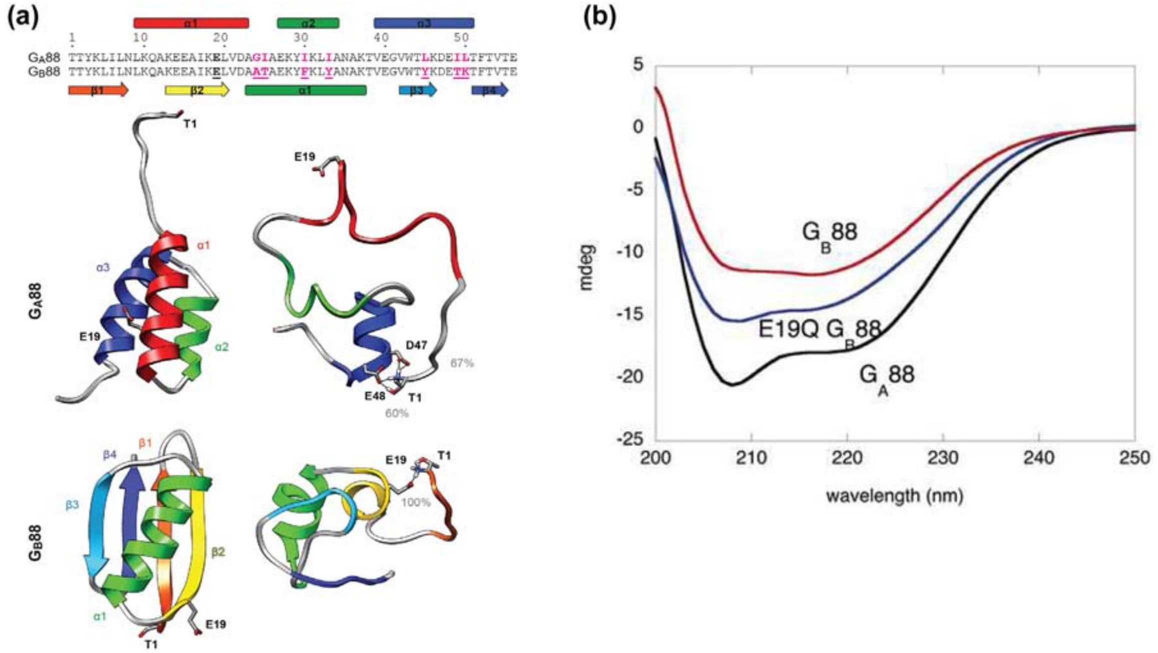


Fig. 15. G_A88 and G_B88 structures and unfolding trajectories, and G_B88 and G_B88 E19Q spectroscopic and thermodynamic properties [43]. (a) G_A88 and G_B88 structures, with sequence alignment and secondary structure shown above. The seven residues that differ between the two proteins are highlighted in magenta. The crystal structure of GA88 is shown on the left, with its denatured state shown on the right at 36 ns for a 498K simulation, showing the hydrogen bond network between T1–D47–E48. For GB88, the crystal structure is shown on the left, and the denatured state is shown on the right at 31 ns. The highlighted hydrogen bond between E19 and T1 promotes formation of the β1/β2 and β3/β4 hairpins. (b) Spectroscopic properties of WT G_B88 and G_B88 E19Q mutant. Comparison between the far-UV CD spectrum of G_A88 (black), G_B88 (red), and G_B88 E19Q (blue). It is evident from the shift of the spectrum toward G_A88 that the mutation in G_B88 causes an increase in α-helical content. Figure and figure legend taken from Ref. [43] and used with permission.

occupancy changes. The resulting occupancy graph uses edge color to indicate degree of mutant stabilization or destabilization, which is displayed in Fig. 16b for SOD. Contacts are rendered in several different visualization modes, including a network view at the top, a plot showing atom distances over time, and the color-coded contacts mapped onto the ribbon structure, allowing the frequency of contacts, the time points of contacts, and the spatial locations of contacts to all be visualized.

The data structuring of DIVE is based on ontologies made up of data nodes, which contain relationships, or references, to other nodes, as well as data edges, which allow for more complex and flexible relationships between nodes than in traditional networks [45]. Data nodes can also be dynamically altered at runtime, as can their inheritance aspects, allowing facile data clustering [45]. The graph theory representation of DIVE displays specific interaction correlations, and correlated motion between structures can be mapped easily onto proteins to determine effects of certain regions. Several ways of viewing motion and structural stability during a simulation are shown in Fig. 16c, where the top of the figure shows the secondary structure type for a variety of different mutants and whether the structure is stable or not. For one particular mutant, A4V, the Cα RMSD plot is shown *versus* the WT protein, where RMSD toward the end of the simulation is higher for the mutant. The amount of space between the two residues on the loop

Fig. 14. G_A88 and G_B88 structures, trajectories, and contact analysis [42]. (a) Ribbon diagrams of G_A88 and G_B88, with sequence below. Residues that are different between the two sequences are highlighted, with the corresponding residue on the structure highlighted in the same color. L45 on G_A88 and Y45 on G_B88 have corresponding stick structures shown. (b) Contact maps for G_A88 and G_B88 denatured, transition, and native states and representative structures with the differing residues shown as red balls. The denatured state of the contact map shows that there is little organized structure, with the exception of one helix, for G_A88. G_B88 also shows a small amount of helix structure in the denatured state, forming some tertiary structure in the transition state and β-sheet in the native state. Figure and figure legend adapted from Gianni *et al*. [42] and used with permission.

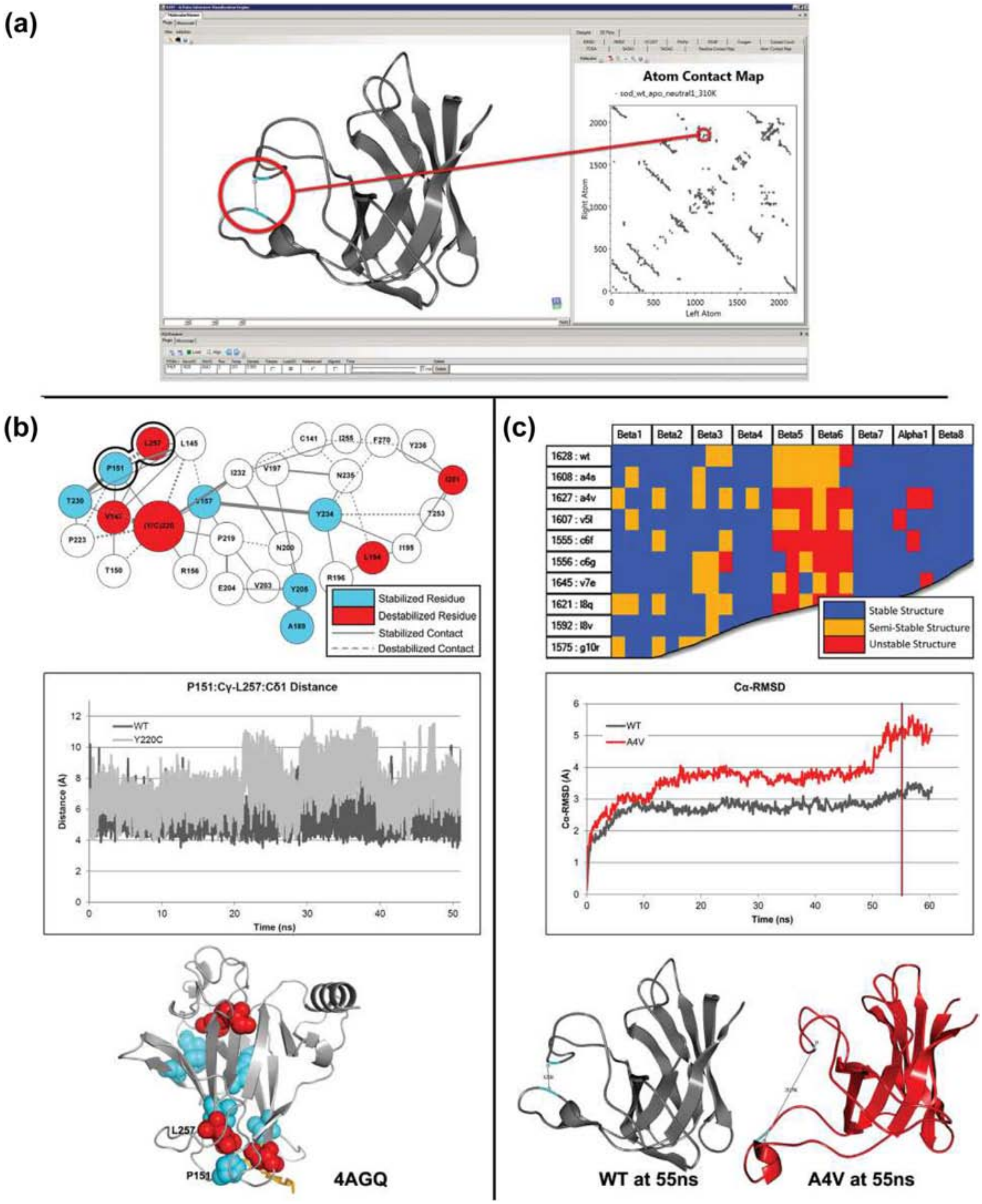
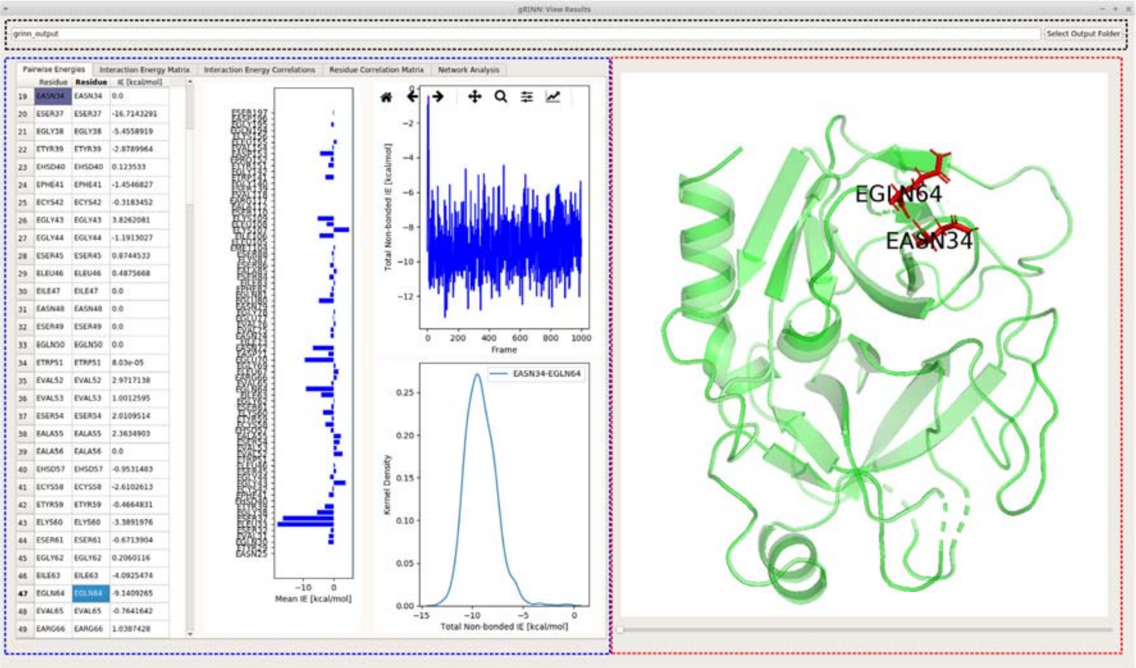


Fig. 16. Interactive visualizations in DIVE [44]. (a) Protein dashboard application within DIVE showing a viewer and interactive contact map for the superoxide dismutase (SOD) protein. The contact highlighted in the structure representation is pinpointed on the interactive contact map. (b) Top: ContactWalker summary of contact differences between WT p53 and mutant Y220C simulations. The highlighted residues have contacts with 50% occupancy change. Distances between P151 and L257 are outlined in black in the map. Middle: the distance is also shown between P151 and L257 over time in the WT and mutant Y220C structures; the Y220C shows a greater distance between the two residues. Bottom: p53 structure with ligand (stick figure at bottom) in proximity to disrupted colored residues. (c) Top: aggregated secondary structural data from mutant simulations of SOD. Middle: plot of the Cα RMSD of the wild-type and A4V mutant simulations. Bottom: associated MD structures at 55 ns. Figure and figure legend taken from Ref. [44] and used with permission.



	Residue	Residue	IE [kcal/mol]
19	EASN34	EASN34	0.0
20	ESER37	ESER37	-16.7143291
21	EGLY38	EGLY38	-5.4558919
22	ETYR39	ETYR39	-2.8789964
23	EHSD40	EHSD40	0.123533
24	EPHE41	EPHE41	-1.4546827
25	ECYS42	ECYS42	-0.3183452
26	EGLY43	EGLY43	3.8262081
27	EGLY44	EGLY44	-1.1913027
28	ESER45	ESER45	0.8744533
29	ELEU46	ELEU46	0.4875668
30	EILE47	EILE47	0.0
31	EASN48	EASN48	0.0
32	ESER49	ESER49	0.0
33	EGLN50	EGLN50	0.0
34	ETRP51	ETRP51	8.03e-05
35	EVAL52	EVAL52	2.9717138
36	EVAL53	EVAL53	1.0012595
37	ESER54	ESER54	2.0109514
38	EALA55	EALA55	2.3634903
39	EALA56	EALA56	0.0
40	EHSD57	EHSD57	-0.9531483
41	ECYS58	ECYS58	-2.6102613
42	ETYR59	ETYR59	-0.4664831
43	ELYS60	ELYS60	-3.3891976
44	ESER61	ESER61	-0.6713904
45	EGLY62	EGLY62	0.2060116
46	EILE63	EILE63	-4.0925474
47	EGLN64	EGLN64	-9.1409265
48	EVAL65	EVAL65	-0.7641642
49	EARG66	EARG66	1.0387428

Fig. 17. gRINN window showing UI and residue interaction data [49]. gRINN window showing the output folder outlined in black. The UI, outlined in blue, shows residue interaction data in tabular and graph form about the structure of bovine cationic trypsin, outlined in red. A residue is selected on the table, and its corresponding interaction is shown in the structure. Figure and figure legend taken from Ref. [49] and https://grinn.readthedocs.io/en/latest/tutorial.html, used with permission.

increased dramatically in the mutant relative to WT, corresponding to the increase in RMSD and the destabilized β5-β6 structure relative to WT, which is pinpointed in Fig. 16c. DIVE has the ability to render a structure by attribute, while examining its dynamics in simulation and simultaneously plotting data, allowing multiple protein properties to be displayed during the unfolding process.

gRINN (get Residue Interaction eNergies and Networks) is a program that has a similar visualization interface to DIVE that focuses on residue–residue interactions [49]. Like DIVE, it contains an embedded molecular viewer (an instance of PyMol) so that the structure can be viewed simultaneously with graphs and other analytics. In Fig. 17, the pairwise energies for a 50-ns simulation of trypsin are displayed in a table on the left, while a bar graph of each pairwise energy is shown to the right of the table. Graphs of the total interaction energy throughout the trajectory and the kernel density over the interaction energy are also provided in the UI. The corresponding interaction pairs can then be viewed on the molecular structure, and residue correlation matrices can be calculated. Examining residue correlations and interactions aids in finding areas of the structure for more focused analysis with other programs; this approach should be useful for analysis of protein unfolding simulations.

RIP-MD is another program that focuses on residue interaction networks [50]. Simulations of the human CX26, MD2, and gap-junction channel proteins were used for validation. MD2 is an immune response protein that undergoes a conformational change in MD, with a hydrophobic cavity that closes when the ligand is removed. A 20-ns simulation of MD2 was used, during which the conformational change occurred, and divided into three stages to characterize the behavior of the hydrophobic cavity. The open, closing, and closed interaction networks are shown in Fig. 18a–c, respectively. Cytoscape [51] was used to load and visualize the network file. Initially, there are more hydrogen bonds (blue edges) and vdW interactions (pink edges); then, the number of correlated pairs decreases as the cavity closes. Although this is unintuitive behavior for folding, RIP-MD showed that conformational changes can lead to lower correlations between residues.

A well-established program that has been widely used for computational analysis of structures is Visual Molecular Dynamics (VMD), developed with the primary focus of viewing and interacting with MD trajectories [12]. VMD's versatility at the time of initial release made it available to a wide user base, with both text command and mouse controls. One interesting feature is collaborative viewing of structures in stereo by correcting for differences between each person's perspective. Various analyses, including center of mass motion, RMSD, autocorrelation functions, and Ramachandran data, are available in VMD, with the option for the user to create and add additional analysis metrics. VMD uniquely connects to and visualizes a running simulation, allowing the user to make perturbations to atoms or modify parameters.

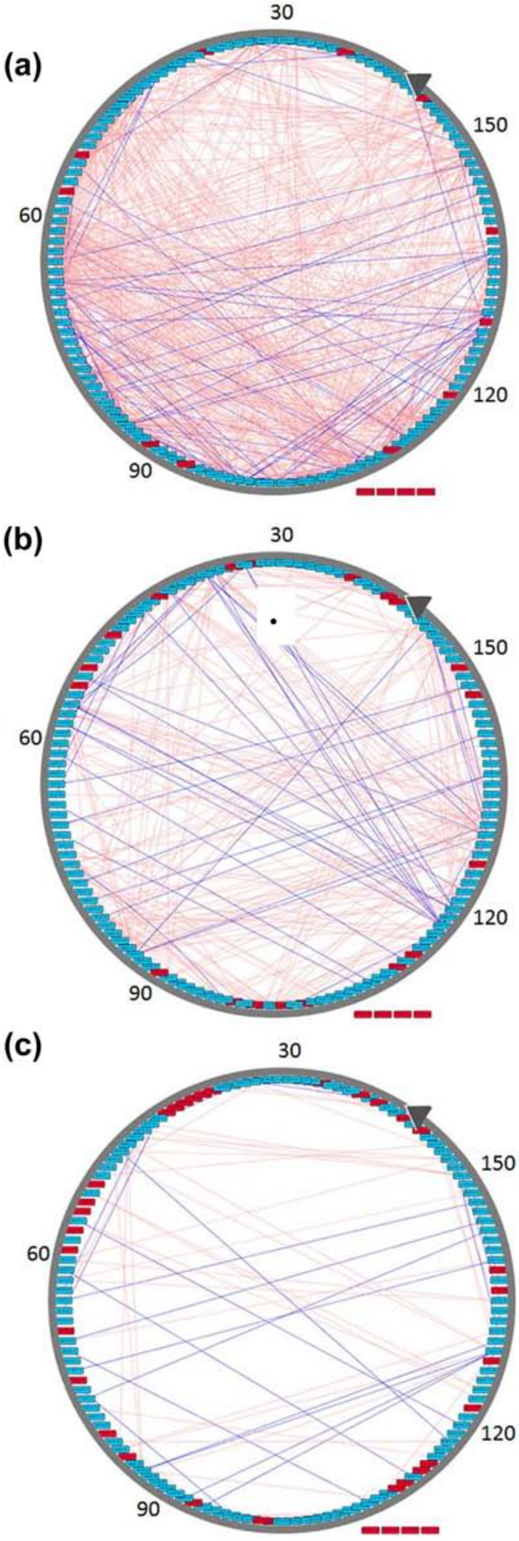


Fig. 18. RIP-MD residue interaction network examining MD2 conformational changes [50]. Changes in the residue interaction network during conformational change of open (a), closing (b), and closed (c) hydrophobic cavity in the MD2 protein. Pink edges connect vdW interactions, and blue connect hydrogen bonds between amino acids. An absolute Pearson's correlation coefficient of less than 0.5 for interactions is indicated by the red nodes, while nodes outside the circle do not form interactions in any portion of the simulation. Figure and figure legend adapted from Ref. [50] and used with permission.

Recently, the visualization of structures and trajectories in VMD moved into the realm of virtual reality (VR). Ratamero *et al*. [52] used the Unity3D Game Engine to accomplish this, providing more complex and larger displays (Fig. 19). However, the VR application is not yet mature enough to add labels or change views of the structure as a standalone program, and instead it relies on VMD to make those changes prior to VR viewing. The Mancera research group is also using the Unity Game Engine to develop Molecular Dynamics Visualization (MDV), a program that works on personal computers as well as the Hub for Immersive Visualization and Research (HIVE) cylindrical projection display (Fig. 20) [53]. The unique display method provides multi-user visualization of protein dynamics with fewer issues with perspective because of the curved screen. Other recent programs seek to provide easier access to, and sharing of, MD data and trajectories. For example, HTMoL allows users to visualize MD trajectories from a web browser, and authors can provide supplementary MD data and embed them in a webpage [54].

Conclusions

While there are many valuable analysis methods for visualizing folding and unfolding simulations, it is clear that more than one depiction is required to characterize the process. Although quantitative data are critical to analyzing the folding pathway, MD simulations provide an overwhelming amount of data such that weighting and accounting for all the data is a tedious process and communicating the findings in an intuitive and approachable manner is challenging. Combining data output from many different analyses, such as in property-space analyses, can provide a visual depiction of how the data aggregate to describe an un/folding pathway. Nonetheless, analyses of single properties also have merit. Superimposing structures at time points where there is a large switch or deviation in a property can be used to demonstrate the effect of that property on the pathway, thereby elucidating the relative importance of different physical properties

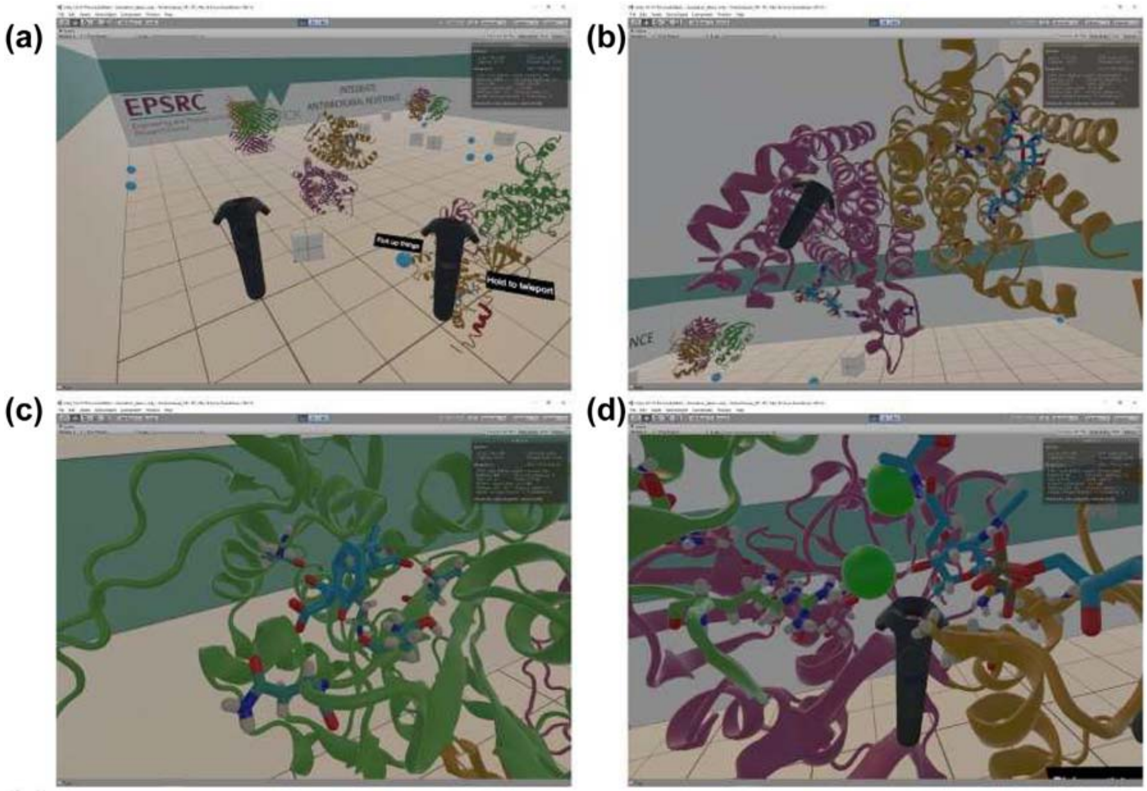


Fig. 19. The Unity Game Engine VR interface showing different protein structures [52]. (a) The VR room showing four different protein structures and the game controllers. (b–d) Specific details of residues and interactions of the structures that can be viewed at any angle. Figure and figure legend taken from Ref. [52] and used with permission.

for a specific system. Also, there is not a one-size-fits-all approach to the problem, and analyses need to be tailored to the system to provide a multiscale description of the process, and interactive data-agnostic visual analysis approaches and frameworks can be very helpful in this regard. Independent of system and analysis methods, visualizations are critical to obtaining a deeper understanding of

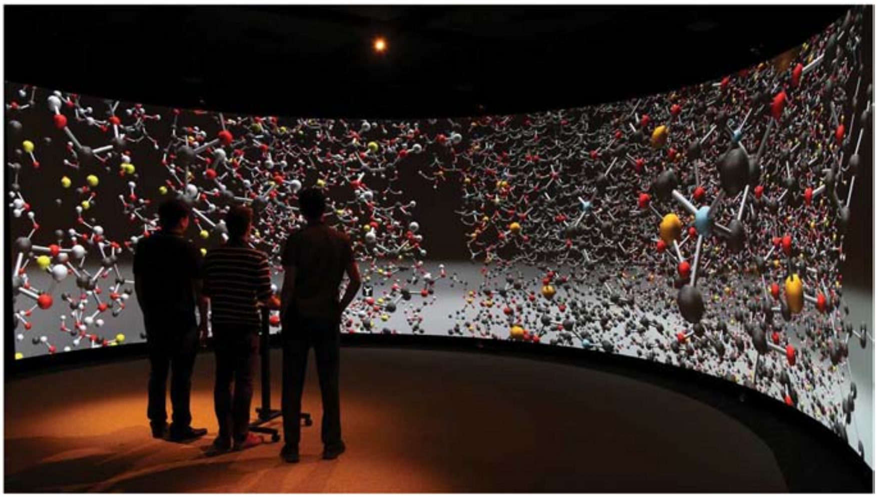


Fig. 20. HIVE display of structure in MDV [53]. The HIVE curved cylinder display showing a structure in MDV, examined by collaborators. Figure and figure legend taken from Ref. [53] and used with permission.

protein behavior and synthesizing the generated data and resulting analyses into a digestible and insightful format. Furthermore, the last year has brought exciting advances in collaborative visualizations, virtual reality implementations, and new web tools to share, stream, analyze, and visualize MD trajectories.

Received 15 September 2018;
Received in revised form 14 February 2019;
Accepted 18 February 2019
Available online 03 March 2019

Keywords:
molecular dynamics;
protein folding;
protein unfolding;
biovisualization

Abbreviations used:
MD, molecular dynamics; SASA, solvent-accessible surface area; RMSD, root-mean-square deviation; RMSF, root-mean-square flexibility; TS, transition state; D, denatured state; N, native state; EnHD, engrailed homeodomain; FIM, fast information matching; MI, mutual information; T_m, melting temperature.

References

[1] J.S. Yang, W.W. Chen, J. Skolnick, E.I. Shakhnovich, All-atom ab initio folding of a diverse set of proteins, Structure 15 (2007) 53–63.

[2] K.A. Dill, S.B. Ozkan, M.S. Shell, T.R. Weikl, The protein folding problem, Annu. Rev. Biophys. 37 (2008) 289–316.

[3] A.R. Fersht, V. Daggett, The present view of the mechanism of protein folding, Nat. Rev. Mol. Cell Biol. 4 (2003) 497–502.

[4] M. Levitt, R. Sharon, Accurate simulation of protein dynamics in solution, Proc. Natl. Acad. Sci. U. S. A. 85 (1988) 7557–7561.

[5] D.A.C. Beck, D.O.V. Alonso, V. Daggett, A microscopic view of peptide and protein solvation, Biophys. Chem. 100 (2003) 221–237.

[6] V. Daggett, Protein folding-simulation, Chem. Rev. 106 (2006) 1898–1916.

[7] R.D. Schaeffer, A. Fersht, V. Daggett, Combining experiment and simulation in protein folding: closing the gap for small model systems, Curr. Opin. Struct. Biol. 18 (2008) 4–9.

[8] M. Levitt, M. Hirshberg, R. Sharon, V. Daggett, Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution, Comput. Phys. Commun. 91 (1995) 215–231.

[9] M. Levitt, M. Hirshberg, R. Sharon, K. Laidig, V. Daggett, Calibration and testing of a water model for simulation of the molecular dynamics of proteins and nucleic acids in solution, J. Phys. Chem. 101 (1997) 5051–5061.

[10] M.C. Childers, V. Daggett, Validating molecular dynamics simulations against experimental observables in light of underlying conformational ensembles, J. Phys. Chem. B 122 (2018) 6673–6689.

[11] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, et al., UCSF chimera—a visualization system for exploratory research and analysis, J. Comput. Chem. 25 (2004) 1605–1612.

[12] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, J. Mol. Graph. 14 (1996) 33–38 (27-8).

[13] N.C. Benson, V. Daggett, A comparison of multiscale methods for the analysis of molecular dynamics simulations, J. Phys. Chem. B 116 (2012) 8722–8731.

[14] M.E. McCully, D.A. Beck, V. Daggett, Microscopic reversibility of protein folding in molecular dynamics simulations of the engrailed homeodomain, Biochemistry 47 (2008) 7079–7089.

[15] U. Mayor, C.M. Johnson, V. Daggett, A.R. Fersht, Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 13518–13522.

[16] U. Mayor, N.R. Guydosh, C.M. Johnson, J.G. Grossmann, S. Sato, G.S. Jas, et al., The complete folding pathway of a protein from nanoseconds to microseconds, Nature 421 (2003) 863–867.

[17] S. Gianni, N.R. Guydosh, F. Khan, T.D. Caldas, U. Mayor, G.W. White, et al., Unifying features in protein-folding mechanisms, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 13286–13291.

[18] T.L. Religa, J.S. Markson, U. Mayor, S.M. Freund, A.R. Fersht, Solution structure of a protein denatured state and folding intermediate, Nature 437 (2005) 1053–1056.

[19] M.E. McCully, D.A. Beck, A.R. Fersht, V. Daggett, Refolding the engrailed homeodomain: structural basis for the accumulation of a folding intermediate, Biophys. J. 99 (2010) 1628–1636.

[20] U. Mayor, J.G. Grossmann, N.W. Foster, S.M. Freund, A.R. Fersht, The denatured state of engrailed homeodomain under denaturing and native conditions, J. Mol. Biol. 333 (2003) 977–991.

[21] G.G. Maisuradze, A. Liwo, H.A. Scheraga, Principal component analysis for protein folding dynamics, J. Mol. Biol. 385 (2009) 312–329.

[22] S.L. Kazmirski, A. Li, V. Daggett, Analysis methods for comparison of multiple molecular dynamics trajectories: applications to protein unfolding pathways and denatured ensembles, J. Mol. Biol. 290 (1999) 283–304.

[23] K. Pearson, On lines and planes of closest fit to systems of points in space, vol. 2, 1901 559–572.

[24] M.E. McCully, D.A. Beck, V. Daggett, Multimolecule test-tube simulations of protein unfolding and aggregation, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 17851–17856.

[25] J.A. Kovacs, W. Wriggers, Spatial heat maps from fast information matching of fast and slow degrees of freedom: application to molecular dynamics simulations, J. Phys. Chem. B 120 (2016) 8473–8484.

[26] D.O. Alonso, V. Daggett, Molecular dynamics simulations of protein unfolding and limited refolding: characterization of partially unfolded states of ubiquitin in 60% methanol and in water, J. Mol. Biol. 247 (1995) 501–520.

[27] D.O. Alonso, V. Daggett, Molecular dynamics simulations of hydrophobic collapse of ubiquitin, Protein Sci. 7 (1998) 860–874.

[28] S. Koulgi, U. Sonavane, R. Joshi, Insights into the folding pathway of the engrailed homeodomain protein using replica exchange molecular dynamics simulations, J. Mol. Graph. Model. 29 (2010) 481–491.

[29] P. Jemth, S. Gianni, R. Day, B. Li, C.M. Johnson, V. Daggett, et al., Demonstration of a low-energy on-pathway intermediate in a fast-folding protein by kinetics, protein engineering, and simulation, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 6450–6455.

[30] A. Li, V. Daggett, Molecular dynamics simulation of the unfolding of barnase: characterization of the major intermediate, J. Mol. Biol. 275 (1998) 677–694.

[31] J.M. Matthews, A.R. Fersht, 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 (1995) 6805–6814.

[32] V. Daggett, A. Li, A.R. Fersht, 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 (1998) 12740–12754.

[33] A.C. Joerger, H.C. Ang, A.R. Fersht, Structural basis for understanding oncogenic p53 mutations and designing rescue drugs, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 15056–15061.

[34] N.C. Benson, V. Daggett, Wavelet analysis of protein motion, Int. J. Wavelets Multiresolution Inf. Process. 10 (2012), 1250040.

[35] S. Calhoun, V. Daggett, Structural effects of the L145Q, V157F, and R282W cancer-associated mutations in the p53 DNA-binding core domain, Biochemistry 50 (2011) 5345–5353.

[36] F. Rao, A. Caflisch, The protein folding network, J. Mol. Biol. 342 (2004) 299–306.

[37] S. Piana, K. Lindorff-Larsen, D.E. Shaw, Atomic-level description of ubiquitin folding, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 5915–5920.

[38] P.A. Alexander, D.A. Rozak, J. Orban, P.N. Bryan, Directed evolution of highly homologous proteins with different folds by phage display: implications for the protein folding code, Biochemistry 44 (2005) 14045–14054.

[39] K.A. Scott, V. Daggett, Folding mechanisms of proteins with high sequence identity but different folds, Biochemistry 46 (2007) 1545–1556.

[40] Y. He, D.C. Yeh, P. Alexander, P.N. Bryan, J. Orban, Solution NMR structures of IgG binding domains with artificially evolved high levels of sequence identity but different folds, Biochemistry 44 (2005) 14055–14061.

[41] P.A. Alexander, Y. He, Y. Chen, J. Orban, P.N. Bryan, A minimal sequence code for switching protein structure and function, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 21149–21154.

[42] A. Morrone, M.E. McCully, P.N. Bryan, M. Brunori, V. Daggett, S. Gianni, et al., The denatured state dictates the topology of two proteins with almost identical sequence but different native structure and function, J. Biol. Chem. 286 (2011) 3863–3872.

[43] S. Gianni, M.E. McCully, F. Malagrinò, D. Bonetti, A. De Simone, M. Brunori, et al., A carboxylate to amide substitution that switches protein folds, Angew. Chem. Int. Ed. Engl. 57 (2018) 12795–12798.

[44] D. Bromley, S.J. Rysavy, R. Su, R.D. Toofanny, T. Schmidlin, V. Daggett, DIVE: a data intensive visualization engine, Bioinformatics 30 (2014) 593–595.

[45] S.J. Rysavy, D. Bromley, V. Daggett, DIVE: a graph-based visual-analytics framework for big data, IEEE Comput. Graph. Appl. 34 (2014) 26–37.

[46] M.W. van der Kamp, R.D. Schaeffer, A.L. Jonsson, A.D. Scouras, A.M. Simms, R.D. Toofanny, et al., Dynameomics: a comprehensive database of protein dynamics, Structure 18 (2010) 423–435.

[47] D. Bromley, P.C. Anderson, V. Daggett, Structural consequences of mutations to the α-tocopherol transfer protein associated with the neurodegenerative disease ataxia with vitamin E deficiency, Biochemistry 52 (2013) 4264–4273.

[48] R. Koradi, M. Billeter, K. Wüthrich, MOLMOL: a program for display and analysis of macromolecular structures, J. Mol. Graph. 14 (1996) 29–32.

[49] O. Serçinoglu, P. Ozbek, gRINN: a tool for calculation of residue interaction energies and protein energy network analysis of molecular dynamics simulations, Nucleic Acids Res. 46 (2018) W554–W562.

[50] S. Contreras-Riquelme, J.A. Garate, T. Perez-Acle, A.J.M. Martin, RIP-MD: a tool to study residue interaction networks in protein molecular dynamics, PeerJ. 6 (2018) e5998.

[51] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.

[52] E.M. Ratamero, D. Bellini, C.G. Dowson, R.A. Römer, Touching proteins with virtual bare hands: visualizing protein-drug complexes and their dynamics in self-made virtual reality using gaming hardware, J. Comput. Aided Mol. Des. 32 (2018) 703–709.

[53] M. Wiebrands, C.J. Malajczuk, A.J. Woods, A.L. Rohl, R.L. Mancera, Molecular dynamics visualization (MDV): stereoscopic 3D display of biomolecular structure and interactions using the unity game engine, J. Integr. Bioinform. 15 (2018).

[54] M. Carrillo-Tripp, L. Alvarez-Rivera, O.I. Lara-Ramírez, F.J. Becerra-Toledo, A. Vega-Ramírez, E. Quijas-Valades, et al., HTMoL: full-stack solution for remote access, visualization, and analysis of molecular dynamics trajectory data, J. Comput. Aided Mol. Des. 32 (2018) 869–876.