Hydrogen bonding heterogeneity correlates with protein folding transition state passage time as revealed by data sonification

Carla Scaletti,[1*] Premila P. Samuel Russell,[2] Kurt J. Hebel,[1] Meredith M. Rickard,[2] Mayank Boob,[2] Franz Danksagmüller,[9] Stephen A. Taylor,[7] Taras V. Pogorelov,[2,3,4,5,6] and Martin Gruebele[2,3,5,8*]

[1]Symbolic Sound Corporation, Champaign, IL 61820, United States; [2]Department of Chemistry, University of Illinois Urbana-Champaign, IL 61801, United States; [3]Center for Biophysics and Quantitative Biology, University of Illinois Urbana-Champaign, IL 61801, United States; [4]School of Chemical Science, University of Illinois Urbana-Champaign, IL 61801, United States; [5]Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, IL 61801, United States; [6]National Center for Supercomputer Applications, University of Illinois Urbana-Champaign, IL 61801, United States; [7]School of Music, University of Illinois Urbana-Champaign, IL 61801, United States; [8]Department of Physics, University of Illinois Urbana-Champaign, IL 61801, United States; [9]Musikhochschule Lübeck, 23552 Lübeck, Germany

Author Information

Corresponding authors:
Martin Gruebele *Email: mgruebel@illinois.edu*
Carla Scaletti *E-mail: carla@symbolicsound.com*

Disclosure: Carla Scaletti and Kurt Hebel are the developers of the Kyma software used for data analysis and sonification/visualization in this work.

Author ORCID IDs (when available):
Carla Scaletti (0009-0002-0471-9584)
Premila P. Samuel Russell (0000-0001-6690-0886)
Kurt J. Hebel (0009-0007-3035-7640)
Meredith Rickard (0000-0002-4571-9743)
Mayank Boob (0000-0002-0287-3257)
Stephen A. Taylor (0009-0009-4616-5648)
Taras V Pogorelov (0000-0001-5851-7721)
Martin Gruebele (0000-0001-9291-8123)

ABSTRACT

Protein-protein and protein-water hydrogen bonding interactions play essential roles in the way a protein passes through the transition state during folding or unfolding, but the large number of these interactions in molecular dynamics (MD) simulations makes them difficult to analyze. Here we introduce a state space representation and associated 'rarity' measure to identify and quantify transition state passage (transit) events. Applying this representation to a long MD simulation trajectory that captured multiple folding and unfolding events of the GTT WW domain, a small protein often used as a model for the folding process, we identified three transition categories: Highway (faster), Meander (slower), and Ambiguous (intermediate). We developed data sonification and visualization tools to analyze hydrogen bond dynamics before, during, and after these transition events. By means of these tools, we were able to identify characteristic hydrogen bonding patterns associated with 'Highway' versus 'Meander' versus 'Ambiguous' transitions and to design algorithms that can identify these same folding pathways and critical protein-water interactions directly from the data. Highly cooperative hydrogen bonding can either slow down or speed up transit. Furthermore, an analysis of protein-water hydrogen bond dynamics at the surface of WW domain shows an increase in hydrogen bond lifetime from folded to unfolded conformations with Ambiguous transitions as an outlier. In summary, hydrogen bond dynamics provide a direct window into the heterogeneity of transits, which can vary widely in duration (by a factor of 10) due to a complex energy landscape.

SIGNIFICANCE STATEMENT

Hydrogen bonds play a major role in how proteins fold, especially in terms of organizing secondary structure and the protein-water interface. Here we use sonification to listen to and analyze time series obtained from long, all-atom simulations in which a small protein folds and unfolds many times to examine, in atomistic detail, the role played by hydrogen bonds to water and within the protein during the 'transition state passage' — that fleeting moment when a protein switches between folded and unfolded states. Although non-native hydrogen bonds play a key role in slowing down some passages, they can also speed up unfolding and show cooperativity, in the sense that they can form and break in highly correlated groups.

KEYWORDS: solvation shell; cooperativity; tertiary structure; secondary structure

Introduction

Proteins represent an extraordinary state of matter, traversing a multidimensional energy landscape between disorder and order. Fast-folding globular proteins have proven particularly effective at revealing the essential feature of folding (1–3): entropy-enthalpy compensation that leads to a funneled yet partly frustrated energy landscape with unusually low free energy barriers when water is excluded from the hydrophobic core (4). Numerous fast folding studies have focused on protein mutants and the dynamics of the protein backbone when residues form secondary structure (e.g., turns and helices) and pack cooperatively into tertiary structure (5). Far less studied among fast folders are the mechanisms by which protein-water and water-water hydrogen bonding organize folding (6, 7).

Transition state passage (hereafter referred to as "transit") can be resolved on a microsecond time scale for fast-folding proteins (8–11); during transit, intra-protein hydrogen bonds contribute to the defining native contacts of the folded topology (12–15), but they can also form non-native contacts that slow down folding or result in the trapping of folding intermediates (16). The typical free energy of a single hydrogen bond (ca. 15-20 kJ/mole) is on the same order as the entire folding free energy of many proteins under physiological conditions. Thus, protein and water interact on a free energy scale comparable to the folding free energy, making them less easily disentangled than the solvent and solute in a typical small molecule chemical reaction. Furthermore, it is difficult to find experimental data with the sub-microsecond time resolution required for studying both protein-water dynamics and transitionary intra-protein contacts during folding, and it can be challenging to visualize and sort through the numerous water molecules present near the protein in a folding simulation.

To investigate hydrogen bonding patterns during transit, we analyze a molecular dynamics trajectory of the FiP25 GTT mutant (17, 1) of WW domain, a widely used model for fast folding experiments (18, 19). The WW domain (ca. 35 amino acids) contains a small hydrophobic core and three beta strands connected by loops 1 and 2 (Fig. 1). Fast-folding WW domains can fold via two parallel pathways: one where loop 1 forms first, and another where loop 2 forms first. Transit times of WW domains, measured by barrier-top relaxation of low-barrier proteins or by single molecule experiments in viscous solvents, lie between 300 ns and 3 μs (9, 10, 19, 20), as is typical of small model proteins. In MD simulations, we expect them to be about a factor of 3 faster due to the reduced shear viscosity of the TIP3P water model (21).

We applied a combination of data sonification (22–24), visualization, and data analytics to identify rapidly changing hydrogen bond patterns in 28 folding/unfolding events of GTT, and to classify the events into three categories according to their hydrogen bonding dynamics — Highway, Meander, and Ambiguous. To arrive at the classification, we first devised a 2-dimensional state space representation and associated 'rarity' measure to identify and label transit events; next we developed sonification/visualization tools for analyzing protein-protein

and protein-water hydrogen bond dynamics before, during, and after these events. We found that the three categories, with their distinct hydrogen bond signatures, are associated with different average transit times.

Results and Discussion

Protein transit in $\mu\pm\sigma$ space. Fig. 1A-D shows the structure of GTT and snapshots of three transit types identified by our analysis of 28 all-atom events from the 650.8 μs full trajectory of Ref. (1) (Methods). We pay particular attention to the area around loop 1, which connects β-strands 1 and 2 (residues 12-18), and loop 2, which connects β-strands 2 and 3 (residues 21-29) because in WW domain, folding nucleates at the loops (25). We avoided reaction coordinates that require prior knowledge of the folded state structure (e.g., Q, the fraction of native contacts, or *RMSD*, the root-mean-square deviation from the folded structure). Instead, the folding transitions were tracked in terms of hydration-related coordinates such as the solvent-accessible surface area (*SASA*), the radius of gyration (R_g), and the exposed hydrophobic area (*EHA*) (Methods). These coordinates produce folding/unfolding time series that look like noisy binary steps, suggesting that GTT is in the regime connecting downhill and fast two-state folding (Fig. S1).

For each of the three coordinates R_g, *SASA*, and *EHA,* smaller values correspond to the folded states. We found these coordinates to be significantly different from one another in terms of their running averages and standard deviations (Fig. S2). We generated a set of six $\mu_i(t)$ (running average) and $\sigma_i(t)$ (running standard deviation) time series, which could then be reduced to two principal components PC1(t) and PC2(t) that account for over 97.5% of the variability in the complete data (Fig. S3). PC1(t) weights all μ_i and σ_i values with the same sign, and the vertical axis PC2(t) weights all the μ_i values with the opposite sign to that of the σ_i values; these two coordinates can be roughly thought of as $\mu\pm\sigma$ combinations (Table S1). PC1(t) distinguishes the folded and unfolded ensembles; PC2 correlates with transit time (next subsection). Fig. 2A (top) shows all trajectories plotted in the $\mu\pm\sigma$ space.

Rarity to identify the transition region. The full PC1 trajectory spends roughly 70% of the time in the unambiguously folded ensemble of states and about 10% of the time in the unambiguously unfolded states. To determine the centers of transits in the $\mu\pm\sigma$ space, we sorted the PC1 (μ) coordinates of the full trajectory into a histogram from smallest (native-like) to largest (unfolded-like) PC1 value and then defined 'Rarity' as the moving-window standard deviation of that histogram (Fig. S4 and SI Methods). Thus, Rarity is at its maximum for values of PC1 where the protein spends the least time, i.e. during transits. Depending on the Rarity function decision threshold, the time intervals near each transit are partitioned into Folded, Transit, or Unfolded states. For example, a Rarity decision threshold of 33% results in the color-coded partitioning shown in Fig. 2A (top) with the normalized Rarity plotted below it.

For comparison, we applied the forward-backward transit identification method described in ref. (1) (see SI Methods) to the standard RMSD coordinate, obtaining an average transit time of 0.43 μs (Fig. S5), which agrees with the 0.5±0.1 μs mean duration published in (1). Setting the Rarity threshold to 5% of its maximum yields a

comparable average transit time of 0.33 μs (Fig. S5). To study hydrogen bonding in the middle of the transits, we henceforth use a more stringent Rarity threshold of 33% (Fig. S6), partitioning each folding/unfolding transition into the three segments shown in Fig. 2A (top): folded side (lavender), center of transit (orange, blue, or gray), and unfolded side (pale green).

Through sonification (next subsection), three classes of transitions were identified: smooth arcs with large PC2 (orange), convoluted paths with a low value of PC2 (blue), and paths that did not fall into either category (gray). We henceforth refer to these as Highway=H, Meander=M, and Ambiguous=A transitions. We confirmed that a *k*-means clustering based on features of the $\sigma\pm\mu$ trajectories produces a similar classification (Fig. 2B, Fig. S7). This tripartite clustering agrees with the findings reported in (1).

The transition regions we studied (SI Methods, Fig. S6) range in duration from ≈19 ns for the fastest Highway to ≈180 ns for the slowest Meander (Table S2). We found that the distribution of transit durations was not consistent with an exponential distribution (Fig. S8). Clustering on the transit time alone yields a classification that agrees with the *k*-means clustering of $\sigma\pm\mu$, with one exception: a transition at 552.322 μs was classified as Highway by the duration clustering, whereas it was classified as Ambiguous by the authors (using sonification) as well as by the automated $\sigma\pm\mu$ clustering in Fig. 2B.

Inspection of the mid-transition GTT structures (Rarity ≈1, Fig. 1BCD) revealed that H-transitions often proceed rapidly via early formation or late dissolution of loop 1 with native-like hydrogen bonding, M-transitions are more likely to involve strand misregistration or non-native helical structure associated with non-native hydrogen bonding. A few A-transitions also include non-native helical structure (Fig. 1C), while a subset of A-transitions is further distinguished by long range non-native contacts of the C-terminus (discussed in the last Result sub-section), which reduces the *SASA* of the three C-terminal 'PSG' residues.

Sonified protein-protein and protein-water hydrogen bond dynamics: methodology. Once we identified different transition classes of GTT, we investigated the hydrogen bond dynamics in the neighborhood surrounding each transition. In an animated molecular visualization of water molecules within a small cut-off distance around the protein, it proved difficult to visually track the formation of individual hydrogen bonds among the numerous sidechain, backbone, and water O/H atoms. However, using the sonification tool we developed (Methods, Fig. S9), we were easily able to identify time-dependent patterns of hydrogen bond formation (Movie S1). In the sonification, a sound event is triggered each time a hydrogen bond forms with a contact in or near loop 1 (residues 12-18) or loop 2 (residues 21-29), including the three mutant residues GTT (residues 26-28). Each combination of bond position and type is associated with a distinct frequency, making it possible to hear when bonds form on multiple residues simultaneously and to independently track the state of each bond type as GTT folds or unfolds (Movie S2, S3).

The hydrogen bond sonification listening results from the 'discovery phase' were compiled, summarized, and then recapitulated by quantitative computation directly from the data (Methods, Tables S2, S3, S4). In the

quantitative analysis, associations emerged between the H-, A-, or M-transition classification and patterns of hydrogen bond formation; for example, in Table S4 the slowest M-transitions are markedly different from the faster H-transitions both in (A) protein-protein and (B) protein-water bonds. Fig. S10 and Table S5, obtained by averaging the likelihood of each specific hydrogen bond over the course of each transit passage duration, confirm that the hydrogen bond profile of H-transitions tends to be most like that of the native state, and that the likelihood of non-native hydrogen bonds increases with increasing transit duration in A- and M-transitions.

Correspondence between $\mu\pm\sigma$ plot and hydrogen bond formation. Kinks and loops in the $\mu\pm\sigma$ trajectory of a transition correspond to time intervals of change in key hydrogen bonds. For example, Fig. 3 compares an H- to an M-transition. Although the averaged hydrogen bond likelihoods (10% threshold) of the H-transition at 300.800 μs are indistinguishable from those in the M-transition at 332.898 μs, the M-transition takes 5.5× longer to pass through the transition state. In Fig. 3A the $\mu\pm\sigma$ trajectory of the H-transition (black) is a smooth curve with large average PC2, while the M-transition (red) traces out a meandering path in the plot before finally folding. Fig. 3B shows a 'piano roll' of hydrogen bond formation involving loops by amino acid *vs.* time, and the associated sonification facilitates the recognition of temporal patterns in hydrogen bonding (Movie S4). The M- and H-trajectories reach alignment by point 'x' in Fig. 3A, when most of the expected native hydrogen bonds (Q29 N→Y20 O / Y20 N→Q29 O, Q29 NE2→E31 O and others) have formed. Prior to that, the M-transition lacks several native hydrogen bonds at point 'u' and instead several non-native or misregistered bonds form, involving residues R14, T27 and T28 (Fig. 3B). Between 'u' and 'v' more of the native-like structure breaks down and is replaced by non-native bonds, most notably, the non-local bond M12 N→E31 O. M12 does not usually participate in protein-to-protein hydrogen bonding in GTT WW domain (Table S7C). Following 'w', the native structure begins to reassemble and is characterized at 'y' by higher than usual hydrogen bonding among the side chains of S13, D15, and R17, residues which frame the 'RDG' loop 1 (Movie S5 and Movie S6).

Piano Rolls such as Fig. 3B include only a subset of the residues (those in or near loop 1 and loop 2) and are thus insufficient to explain all changes visible in the $\mu\pm\sigma$ trajectories. However, as demonstrated by this example, the time evolution of loop hydrogen bonding is an important feature to consider when identifying state passage transition heterogeneity. (See Fig. S11 for additional examples and Fig. S12 for a comparison of Piano Rolls to RMSD and secondary structure.)

Overview of protein-protein hydrogen bonds. Fig. 4 shows the hydrogen bonds formed >5% of the time in the folded state, during H-transitions (orange), A-transitions (gray), M-transitions (blue), and in the unfolded state (see Methods). As expected, the fastest H-transitions maximize native-like hydrogen bond formation even during the transit, whereas the A- and M-transitions form significant non-native hydrogen bonds, such as at residues M12 or V18. Thus, the likelihood of non-native hydrogen bonds increases, and the bonding patterns become more heterogeneous as the transition durations increase from the fastest Highway (≈19 ns) to the slowest Meander (≈180 ns) transitions (Table S2). In the unfolded state just before or just after a transition, certain native hydrogen bonds persist with high probability (>40%) (Fig. S10A). Of the residues S13, D15, R17, F21, all but one (F21) originate in

or around loop 1 — evidence that some loop 1 structure is pre-formed prior to folding or persists after an unfolding transition. These same residues also participate in the only bonds to exceed a 5% threshold during the long-term unfolded state (Fig. 4, green bars).

Importance of multi-hydrogen-bond donors and acceptors. We hypothesize that evolution favors multi-hydrogen bond donors and acceptors (26, 27) in regions that are critical to the early stages of folding. Among all hydrogen bonds, the overall prevalence of residues E9 and R11 (Table S7A, S7D), capable of forming reciprocal native bonds F21 N→E9 O / E9 N→F21 O (Table S5A) and R11 N→Y19 O / Y19 N→R11 O (Table S5B) — both of which frame the central 'RDG' loop 1, which is the most likely nucleation site for WW domain folding (28) — is consistent with this hypothesis. In addition, of the 3 residues in the short 'RDG' loop, D15 is the most active participant in protein-protein hydrogen bonds (Table S7A). Likewise, in the longer 'HITG' loop 2, H23 makes/breaks a key protein-protein hydrogen bond during folding/unfolding (H23 N→G7 O and, less often, H23 ND1→9 OE1, see Table S5A). In both cases, residues on the immediate C-terminal side of the loops (residues 13 and 21, 22) are strongly involved, indicating asymmetric zipping of both β-strands. Table S5B also shows hydrogen bonds formed outside the two loops, and Tables S7B-F show the likelihoods of hydrogen bonds aggregated by residue, sidechain type, and in several other ways.

Cooperative non-native hydrogen bonds associated with a decrease or increase in speed. In one type of non-native hydrogen bond, neither amino acid forms a native protein-protein hydrogen bond over the course of the transition, for example, M12 N→Y19 O / Y19 N→M12 O and R14 N→R17 O / R17 N→R14 O (leading to misregistration of strands 1 and 2), Q29 N→ F21 O and H23 N→T27 O (leading to misregistration of strands 2 and 3), or Q29 N→P6 O (leading to incorrect structure between strands 1 and 3). The latter is an example of a long-range non-native contact, which nonetheless ends up being corrected during transit (Table S5). Other non-local non-native contacts, e.g., those involving the C-terminal 'PSG' residues in A*-transitions discussed in the next subsection, are also associated with a slow-down in the speed of transit.

Even when at least one amino acid in the hydrogen-bonding pair ultimately forms a native protein-protein hydrogen bond, its non-native hydrogen bond can block native protein-protein hydrogen bonds to a varying degree. For example, R14 N→R17 O, R32 NH1→R17 O and S13 N→V18 O, S13 N→E9 O all block the native S13 N→ R17 O bond, one of the most prevalent (65%), in the folded WW domain (Table S8). In the native state, each of these non- or half-native bonds is replaced by the native reciprocating β-sheet hydrogen bond pattern such as F21N →E9 O / E9 N→F21O and Q29 N→Y20 O / Y20 N→Q29 O (Table S5A). Non-native hydrogen bonds, especially those involved in non-native helical structure (Fig. 1C), could increase internal friction and slow down torsional motions during folding (29, 39) as well.

Sonification followed by targeted visualization reveals several instances of cooperative formation of multiple non-native (as well as native) hydrogen bonds during the transition period (Fig. 5, and Fig. S14). For example, in the folding trajectory at 557.4 μs in Fig. 5, the non-native hydrogen bond Y19 N→M12 O forms 55% of the time, concurrently with several others (on R14, R17, N22, G26, T27 and Q29), preventing native hydrogen bonds (e.g.,

S13 N→R17 O and G16 N→S13 O) from forming during the transition and resulting in a 'Meander' transition with an exceptionally slow transit time of ≈160 ns. Non-native cooperative hydrogen bonds have been shown to slow down overall folding rates (30); in this work, they appear to modulate transition passage times.

It is not always the case that non-native bonds act to slow the transition. Somewhat surprisingly, in the very fastest *unfolding* H-trajectory (which, at 18.8 ns transit time is 20% faster than the next fastest H-transition) a non-native, non-local hydrogen bond forms between Q29 OE1 and W8 NE1 during the transition (Tables S5A and S7D). In this instance, a non-native non-local bond appears to accelerate the unfolding, perhaps by breaking a native bond to Q29.

Principal Component Analysis of hydrogen bonding patterns. An SVD principal components analysis of the 69 most common protein-protein hydrogen bonds formed during transit (Table S9) reveals that 5 principal components (PCa through PCe) account for 91% of the variations, hinting that there may be additional sub-categories within the three H, M, and A already identified. 70% of the variance is captured by a single component, PCa, which encompasses the full set of native bonds (row 1 of Table S9A). Not surprisingly, sorting all transitions by the contribution of PCa to that transition roughly organizes the transitions in order of increasing transit time as given in Table S2 — additional evidence that the absence of native-like hydrogen bond patterns during a transition is associated with increased transit time.

A negative weighting of PCc is a strong indicator of an Ambiguous transition (See Table S9C). Three of the A-type transitions (552.332, 413.378, and 598.638 μs) stand out as having particularly strong negative contributions from PCc (all are less than -3 in Table S9C). These transitions differ from both H- and M-transitions in two respects: the non-native S34 N→V18 O, R32N→Q29 O, and Y20 N→R32 O are highly likely, whereas the likelihoods of the native E9 N→F21 O, Y20 N→Q29 O, and R11 NH1/NH2→E9 OE1 are greatly reduced during transit, suggesting that the sub-group of three A-transitions may represent a fourth category: A* (See Table S5AB, Fig. S14AB). In Fig. 1C the structure of one such A*-transition shows the characteristic S34 N→V18 O bond as well as Q29 N→Y20 O. As discussed in the next section, the protein-water hydrogen bond half-life is longer for the group of A* transitions than it is for either the Highway or the Meander transitions, exceeding even that of the unfolded state: further evidence that these transitions may represent a distinct subcategory.

It has been previously shown that inter-strand hydrogen bonds foster structural organization but do not have significantly more favorable free energy when compared to protein-water hydrogen bonds (31). Although in this paper we focus on the loops that nucleate folding, the β-strands, as expected, show a lower probability of hydrogen bonding to water and a higher total number of overall hydrogen bonds. Finally, Q29 is unique among all loop residues in that it undergoes strong backbone-backbone hydrogen bonding with Y20 in the middle of β-strand 2 (Fig. S13 and Table S5A).

Water dynamics in the folded, unfolded, and transition classes. Dynamics of protein-water hydrogen bonds during solvation and desolvation of protein also play a critical role both in guiding protein folding and in stabilizing the native state of proteins in general (32), and our sonifications and visualizations confirm this to be true for the

WW domain (Table S4B, and Fig. S14CD). Of the three residues in the short 'RDG' loop 1 (Fig. 1), the most likely nucleation site for folding, both R14 and D15 remain bonded to water as the protein folds (see Table S7G and Fig. S14C), constraining the number of possible protein-protein hydrogen bonds. Lysine residues (K3 and K10), which occur only in the beta strands, participated in protein-water hydrogen bonds more than the other basic amino acids (Table S7G and H), even though arginine has a higher hydrogen-bonding capacity.

Protein-water bonds, like protein-protein non-native bonds, can also slow down a transition. We found that S13 N→Water, D15 N→Water, G16 N→Water, Water→R17 O, Water→F21 O, and H23 N→Water are associated with slower folding (Fig. S14C and Table S4B). Most of these hydrogen bonds involve backbone amino groups which was unanticipated, given that O···HO hydrogen bonds are generally stronger than NH···O hydrogen bonds due to the higher electronegativity of oxygen (33).

Hydrogen bonds between the protein and the first hydration shell break and re-form frequently during the folding transitions as observed across the Piano Rolls (Fig. S14CD). Water dynamics (diffusion, protein-water hydrogen bonds, water-water hydrogen bonds, water orientation, etc.) take place quickly, on the picosecond timescale, and the original data (sampled every 200 ps) does not provide the temporal granularity needed to resolve these changes (1).

To further characterize the hydrogen bond dynamics, we selected three transits of each type (H/M/A*), two folding and one unfolding, from among the 28 transits we studied. In addition, we also selected two time points from elsewhere in the trajectory: one when the protein was unfolded and another when it was folded. From these short simulation trajectories, we determined that the 'continuous' hydrogen bond half-lives for both protein-water (Fig. 6) and water-water (Fig. S15) are on the order of picoseconds, as expected from previous protein folding studies (34). (See Methods for further details). Overall, the continuous water-water hydrogen bonding half-lives depended more strongly on the distance of water molecules from the protein than on the transition class (Fig. S15); however, we did observe an association between transition class and protein-water hydrogen bonding half-lives for water molecules in the solvation shell at the protein surface (3 Å in Fig. S15).

Fig. 6 shows that the shortest protein-water hydrogen bond half-life values of water molecules directly solvating the protein (≤3 Å shell) are associated with the folded state. Fig. 6 also suggests that water plays a significant role in defining the A* transitions and slowing down the A* transit, a solvent effect that could contribute to the prefactor (friction) for folding and affect the folding kinetics of fast-folding proteins with low free energy barriers (35). The *SASA* of the three terminal residues 'PSG' is much smaller for the A* transitions than for any of the other states (Fig. S16), as can also be seen in Fig. 1C (where a S34→V18 hydrogen bond that may help block solvent accessibility of the terminus is visible) and in Table S7A. We also observed that the hydrogen bond lifetimes of the A-transitions, all drawn from subgroup A* whose hydrogen bond signatures differed most from H and M, were significantly longer (10-12 ps) than those of the Folded, Highway, Meander transitions, and even longer than those of the unfolded states of the protein. Longer half-life generally corresponds to a stronger hydrogen bond (36); for example, hydrogen bonds between protein carboxylate groups and water have been observed to be stronger relative

to those involving other protein atom groups (33, 37), and Tables S7E/S7F show that water bonds to the E9 sidechain are particularly likely in A* transitions.

Conclusion

WW transits between folded and unfolded states over a wide range (10-fold) of transit durations and with distinct signatures in the protein-protein and water-protein hydrogen bonding patterns. We identified at least three transition state passage (transit) classes in an MD simulation of GTT protein, a variant of WW: Highway (faster), Meander (slower), and Ambiguous (intermediate). In most cases, cooperative formation of non-native hydrogen bonds is associated with slowing down the passage time, although occasionally non-native hydrogen bonds can help speed up the transition (38). We computed average transit times in the range of 0.33 to 0.43 μs, which, when corrected for the ≈3 lower viscosity of TIP3P water, are on the order of 0.9 to 1.3 μs, in good agreement with previous estimates (9, 10, 19, 20). Here we focused on the Rarity>33% region to look at hydrogen bonding in the middle of the transition.

The temporal dynamics of hydrogen bond formation during transits proved difficult to characterize using visualization alone. To overcome some of these difficulties we devised the $\mu \pm \sigma$ plot (a projection of the high-dimensional protein conformation, via observables, to a 2-D "state space" representation), and a rarity function with an adjustable cut-off criterion to automatically identify transition time intervals, enabling us to isolate and study individual transit time intervals during the protein folding/unfolding reaction. We then developed multiple data sonification tools for the detection of hydrogen bond formation during a transit and a Piano Roll for visualizing continuous changes in the likelihood of each bond during the transition.

Our measurement of simulated water-water hydrogen bond half-lives reveals that they are more sensitive to distance from the polypeptide chain than they are to the state (folded, unfolded, or transitional). In the hydration shell closest to the protein, however, protein-water hydrogen bonds were characterized by a longer half-life (Fig. 6) and a more compact structure (Fig. 1C) during the A* transitions, suggesting that these transitions may be subject to higher internal friction (29, 39) — one possible factor contributing to slower folding.

Protein-protein and protein-water hydrogen bonds are important, not just for organizing structure and for maintaining free energy neutrality as they form and break; they are also correlated with how quickly a polypeptide chain can switch from the unfolded to the folded state or vice-versa. This process occurs via surprisingly organized and cooperative patterns that recur through multiple folding and unfolding events.

The methods reported here have potential applications in analyzing other systems, such as protein-protein and protein-ligand models as well as in analyzing long simulation trajectories. We are also enthusiastic about the expanding role of data sonification as a legitimate tool for scientific exploration and discovery.

Methods

The trajectory in ref. (1) was generated using the CHARMM22* force field and TIP3P water model. At the simulation temperature of 360 K, near the unfolding temperature, GTT had an average residence time <40 μs in the folded and <8 μs in the unfolded states. We used VMD (40) to obtain time series for three structural parameters: *EHA*, *SASA*, and R_g. Each time series has 3,254,126 rows and each row represents a time step of 200 ps. For R_g, the units are Ångstroms, and for *SASA* and *EHA*, the units are Å^2. Following principal component analysis of the three coordinates and their standard deviations in gaussian-smoothed 125, 250 or 500 ns moving windows, the PC1 $\mu+\sigma$ and PC2 $\mu-\sigma$ were plotted (Fig. S2, S3 and SI 'Dimensionality reduction' section). Time intervals for the 28 transition regions are shown in Fig. S6; the time ranges for sampling the folded and unfolded states are shown in the caption of Fig. S13. We used CPPTRAJ to calculate total residue-residue hydrogen bonds and residue-water hydrogen bonds, based on atom type, present for the different transition categories (41).

Sonification was implemented in the sound design language Kyma (42, 43). Hydrogen bond formation was represented by a polyphonic, pitch-dependent Geiger counter, allowing for multiple hydrogen bonds to be tracked simultaneously. The user interface allows for interactive selection of individual residues and types of hydrogen bonds (e.g., side chain-backbone) for auditioning, while the PC1 and PC2 components described above could be followed simultaneously in a phase plot or time-series plot (Fig. S9 and associated SI text). The initial classification by human listeners (consensus of all authors) was then algorithmically reproduced directly from the hydrogen bond data (Fig. S17, Tables S3-4 and SI 'Sonification of hydrogen bond dynamics' section).

Water hydrogen bonding. We simulated 200 ps of hydrogen bond dynamics using NAMD (44), sampled every 200 fs, at the midpoints of the 11 categories selected in Results (Table S10). Hydrogen bond half-lives were calculated (45), as done previously (34, 46), by fitting an averaged auto-correlation decay curve for hydrogen bond presence across the simulation frames to a sum of two exponentials (See SI 'Hydrogen Bond Lifetimes' section for further details).

Associated Content

Supporting Information

A PDF file of Supporting Information, containing: (1) links to the sonified movies of folding reactions, (2) further figures and tables with details of the computational results, (3) supporting text, (4) supporting code and data.

Acknowledgements

This work was supported by the James R. Eiszner Chair in Chemistry and NSF grant MCB 2205665 (M.G., P.S., M.M.R, M.B.) and Symbolic Sound Corporation (C.S. and K.H.). T.P. acknowledges support from the Department of Chemistry (UIUC) and NIH grant R01-GM141298.

References

1. K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, How Fast-Folding Proteins Fold. *Science* **334**, 517–520 (2011).

2. V. Muñoz, M. Cerminara, When fast is better: protein folding fundamentals and mechanisms from ultrafast approaches. *Biochem. J.* **473**, 2545–2559 (2016).

3. M. Gruebele, Downhill protein folding: evolution meets physics. *C. R. Biol.* **328**, 701–712 (2005).

4. J. D. Bryngelson, J. N. Onuchic, N. D. Socci, P. G. Wolynes, Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins Struct. Funct. Genet.* **21**, 167–195 (1995).

5. H. Gelman, M. Gruebele, Fast protein folding kinetics. *Q. Rev. Biophys.* **47**, 95–142 (2014).

6. Y. M. Rhee, E. J. Sorin, G. Jayachandran, E. Lindahl, V. S. Pande, Simulations of the role of water in the protein-folding mechanism. *Proc. Natl. Acad. Sci.* **101**, 6456–6461 (2004).

7. S. J. Kim, B. Born, M. Havenith, M. Gruebele, Real-Time Detection of Protein-Water Dynamics upon Protein Folding by Terahertz Absorption Spectroscopy. *Angew. Chem. Int. Ed.* **47**, 6486–6489 (2008).

8. W. Y. Yang, M. Gruebele, Folding at the speed limit. *Nature* **423**, 193–197 (2003).

9. F. Liu, M. Nakaema, M. Gruebele, The transition state transit time of WW domain folding is controlled by energy landscape roughness. *J. Chem. Phys.* **131**, 195101 (2009).

10. H. S. Chung, K. McHale, J. M. Louis, W. A. Eaton, Single-Molecule Fluorescence Experiments Determine Protein Folding Transition Path Times. *Science* **335**, 981–984 (2012).

11. P. Tripathi, A. Firouzbakht, M. Gruebele, M. Wanunu, Direct Observation of Single-Protein Transition State Passage by Nanopore Ionic Current Jumps. *J. Phys. Chem. Lett.* **13**, 5918–5924 (2022).

12. C. N. Pace, *et al.*, Contribution of hydrogen bonds to protein stability. *Protein Sci. Publ. Protein Soc.* **23**, 652–661 (2014).

13. A. E. Mirsky, L. Pauling, On the Structure of Native, Denatured, and Coagulated Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **22**, 439–447 (1936).

14. L. Pauling, R. B. Corey, H. R. Branson, The Structure of Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **37**, 205–211 (1951).

15. L. Pauling, R. B. Corey, Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds. *Proc. Natl. Acad. Sci. U. S. A.* **37**, 729–740 (1951).

16. M. B. Prigozhin, S.-H. Chao, S. Sukenik, T. V. Pogorelov, M. Gruebele, Mapping fast protein folding with multiple-site fluorescent probes. *Proc. Natl. Acad. Sci.* **112**, 7966–7971 (2015).

17. S. Piana, *et al.*, Computational design and experimental testing of the fastest-folding β-sheet protein. *J. Mol. Biol.* **405**, 43–48 (2011).

18. J. C. Crane, E. K. Koepf, J. W. Kelly, M. Gruebele, Mapping the transition state of the WW domain β-sheet. *J. Mol. Biol.* **298**, 283–292 (2000).

19. M. Szczepaniak, *et al.*, Ultrafast folding kinetics of WW domains reveal how the amino acid sequence determines the speed limit to protein folding. *Proc. Natl. Acad. Sci.* **116**, 8137–8142 (2019).

20. H. S. Chung, W. A. Eaton, Protein folding transition path times from single molecule FRET. *Curr. Opin. Struct. Biol.* **48**, 30–39 (2018).

21. M. A. González, J. L. F. Abascal, The shear viscosity of rigid water models. *J. Chem. Phys.* **132**, 096101 (2010).

22. C. Scaletti, A. B. Craig, Using sound to extract meaning from complex data in E. J. Farrell, Ed. (1991), pp. 207–219.

23. G. Kramer, *Auditory display: Sonification, audification and auditory interfaces* (CRC Press, 1994).

24. C. Scaletti, *et al.*, Sonification-Enhanced Lattice Model Animations for Teaching the Protein Folding Reaction. *J. Chem. Educ.* **99**, 1220–1230 (2022).

25. H. Nguyen, M. Jäger, A. Moretto, M. Gruebele, J. W. Kelly, Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation. *Proc. Natl. Acad. Sci.* **100**, 3948–3953 (2003).

26. W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

27. L. Shimoni, J. P. Glusker, Hydrogen bonding motifs of protein side chains: descriptions of binding of arginine and amide groups. *Protein Sci. Publ. Protein Soc.* **4**, 65–74 (1995).

28. M. Jäger, H. Nguyen, J. C. Crane, J. W. Kelly, M. Gruebele, The folding mechanism of a β-sheet: the WW domain. *J. Mol. Biol.* **311**, 373–393 (2001).

29. W. Zheng, D. De Sancho, T. Hoppe, R. B. Best, Dependence of Internal Friction on Folding Mechanism. *J. Am. Chem. Soc.* **137**, 3283–3290 (2015).

30. F. Bruno Da Silva, V. G. Contessoto, V. M. De Oliveira, J. Clarke, V. B. P. Leite, Non-Native Cooperative Interactions Modulate Protein Folding Rates. *J. Phys. Chem. B* **122**, 10817–10824 (2018).

31. C. Narayanan, C. L. Dias, Hydrophobic interactions and hydrogen bonds in β-sheet formation. *J. Chem. Phys.* **139**, 115103 (2013).

32. Y. Levy, J. N. Onuchic, Water and proteins: A love–hate relationship. *Proc. Natl. Acad. Sci.* **101**, 3325–3326 (2004).

33. G. A. Jeffrey, W. Saenger, *Hydrogen Bonding in Biological Structures* (Springer Berlin Heidelberg, 1991) https:/doi.org/10.1007/978-3-642-85135-3 (September 23, 2023).

34. M. M. Boob, S. Sukenik, M. Gruebele, T. V. Pogorelov, TMAO: Protecting proteins from feeling the heat. *Biophys. J.* **122**, 1414–1422 (2023).

35. B. A. Dalton, *et al.*, Fast protein folding is governed by memory-dependent friction. *Proc. Natl. Acad. Sci.* **120**, e2220068120 (2023).

36. K. M. Reid, H. Poudel, D. M. Leitner, Dynamics of Hydrogen Bonds between Water and Intrinsically Disordered and Structured Regions of Proteins. *J. Phys. Chem. B* **127**, 7839–7847 (2023).

37. M. Petukhov, G. Rychkov, L. Firsov, L. Serrano, H-bonding in protein hydration revisited. *Protein Sci.* **13**, 2120–2129 (2004).

38. P. R. Mouro, V. de Godoi Contessoto, J. Chahine, R. Junio de Oliveira, V. B. Pereira Leite, Quantifying Nonnative Interactions in the Protein-Folding Free-Energy Landscape. *Biophys. J.* **111**, 287–293 (2016).

39. B. N. Markiewicz, H. Jo, R. M. Culik, W. F. DeGrado, F. Gai, Assessment of Local Friction in Protein Folding Dynamics Using a Helix Cross-Linker. *J. Phys. Chem. B* **117**, 14688–14696 (2013).

40. W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics. *J. Mol. Graph. Model.* **14**, 33–38 (1996).

41. D. R. Roe, T. E. Cheatham, Parallelization of CPPTRAJ enables large scale analysis of molecular dynamics trajectory data. *J. Comput. Chem.* **39**, 2110–2117 (2018).

42. C. A. Scaletti, R. E. Johnson, An interactive environment for object-oriented music composition and sound synthesis in *Conference Proceedings on Object-Oriented Programming Systems, Languages and Applications*, (ACM, 1988), pp. 222–233.

43. C. Scaletti, Computer Music Languages, Kyma, and the Future. *Comput. Music J.* **26**, 69–82 (2002).

44. J. C. Phillips, *et al.*, Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **153**, 044130 (2020).

45. R. T. McGibbon, *et al.*, MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528–1532 (2015).

46. R. J. Gowers, P. Carbone, A multiscale approach to model hydrogen bonding: The case of polyamide. *J. Chem. Phys.* **142**, 224907 (2015).

Figure Captions

Fig. 1. WW domain structures during the transition between folded and unfolded states. Key hydrogen bonds are shown as black dotted lines. As discussed in the text, 'Highway' (H), 'Ambiguous' (A) and 'Meander' (M) transitions correspond to increasing state passage time and increased numbers of non-native bonds. (A) Representative structure of the folded state; the C terminus interacts with the N-terminal region. (B) Structure near the middle of a 'Highway' transition (which starts at 169.5338 μs) showing native bond Y19 N→R11 O. (C) Structure near the middle of an 'Ambiguous' transition subtype A* (start time = 552.622 μs) showing the non-native, non-local bond S34 N→V18 O. (D) Structure near the middle of a 'Meander' transition (start time = 557.6732 μs) showing a misregistered bond M12 N→Y19 O. (E) WW sequence showing amino acids single letter codes, where strands are in yellow, loop structures (loop 1 and loop 2) are cyan. We use the residue numbering of ref. (18).

Fig. 2. Folded, unfolded and transitional states of WW domain projected onto a reduced-dimensional $\mu\pm\sigma$ space. (A) All folding trajectories analyzed as a function of PC1 and PC2. Transitions shown in orange (H), dark blue (M) and gray (A) are identified by high rarity. Bottom: the rarity as a function of PC1; the yellow/green region (rarity > 33%) defines the transition region. (B) *k*-means clustering in 2 dimensions reveals three transition regions, separated by dashed blue line separators, which are generally in agreement with subjective assignment of Highway, Ambiguous and Meander. The 'Ambiguous' trajectory at t = 552.322 μs that is within the range of 'Highway' transit times is marked by an arrow and is a member of sub-category A* described in the main text.

Fig. 3. Comparison of a Highway (black) and Meander (red) folding transition in $\mu\pm\sigma$ space (A) to side-by-side Piano Rolls (B). The H-transition starts at ≈300.8 μs and the M-transition starts at ≈332.9 μs in the data from Ref. 1. See SI for details on transition identification and hydrogen-bond sonification. (A) The $\mu\pm\sigma$ trajectory of the H-transition (black) has a consistently high value of PC2 and is smooth, while the M-transition (red) partially unfolds (between points 'y' and 'z') before finally folding. By point 'x' (orange dot in H-transition, lime dot in M-transition), most of the expected native bonds have formed in both the Highway

and the Meander transitions. (B) Intensity on the piano rolls encodes the likelihood of a hydrogen bond from a donor (blue = NH) or acceptor (red = O) in the residue listed on the vertical axis (for detailed key, see Table S6B). The H-transition is on the left and is separated from the M-transition by a black vertical line; each transition is labeled by its start time in μs in the simulation. For reference, Fig. S12 compares the RMSD, secondary structure, and Piano Roll time series for the same H- and M- transitions.

Fig. 4. Likelihoods of selected hydrogen bonds as a function of transition category (H, A, or M). For each bond, its likelihood in the folded state (sampled from 3.6 to 80.6 μs) is shown in purple, its likelihood during the Highway transitions is shown in orange, Ambiguous in gray, Meander in blue, and its likelihood in the unfolded ensembles (sampled from each of the unfolded simulation segments) is shown in green. Native bonds show decreasing likelihood (purple) with increasing transition passage duration from the fastest, H (orange), to A (gray), to the slowest, M (blue). Non-native bonds show increasing likelihood with increasing transition passage duration. Included here are protein-protein bonds involving at least one residue in one of the loops with a 5% threshold applied. For a histogram of all protein-protein hydrogen bonds with a 10% threshold applied, see Fig. S13.

Fig. 5. Cooperativity among protein-protein hydrogen bonds represented as a 'Piano Roll' which shows the relative onset and duration of each type of hydrogen bond within each of the folding transitions (See Table S6B for a detailed Y-axis key; blue is NH donor, red is O acceptor). Transitions are displayed in order from shortest to longest transit duration. Labels below the x-axis refer to the approximate start time of each transition in μs. F=folding, H=Highway, A=Ambiguous, M=Meander. Unfolding transitions are shown in Fig. S14B.

Fig. 6. Hydrogen bond half-life with water at the protein surface. Continuous hydrogen bond lifetime at the protein surface during an example folded state (lavender), at the midpoint of three Highway transitions (orange), three Meander transitions (blue), three Ambiguous (subgroup A*) transitions (gray) and an example unfolded state (green). Within each bar is a label giving the timepoint that was resampled from the original WW domain mutant GTT 650.8 μs full trajectory of ref. (1) These snapshots were equilibrated for 1,000 ps and run for 200 ps to generate the required water dynamics for analysis. The text at the top of each transition state bar indicates the direction of the transition with U → F indicating a folding transition and F → U indicating an unfolding transition.

Figure 1

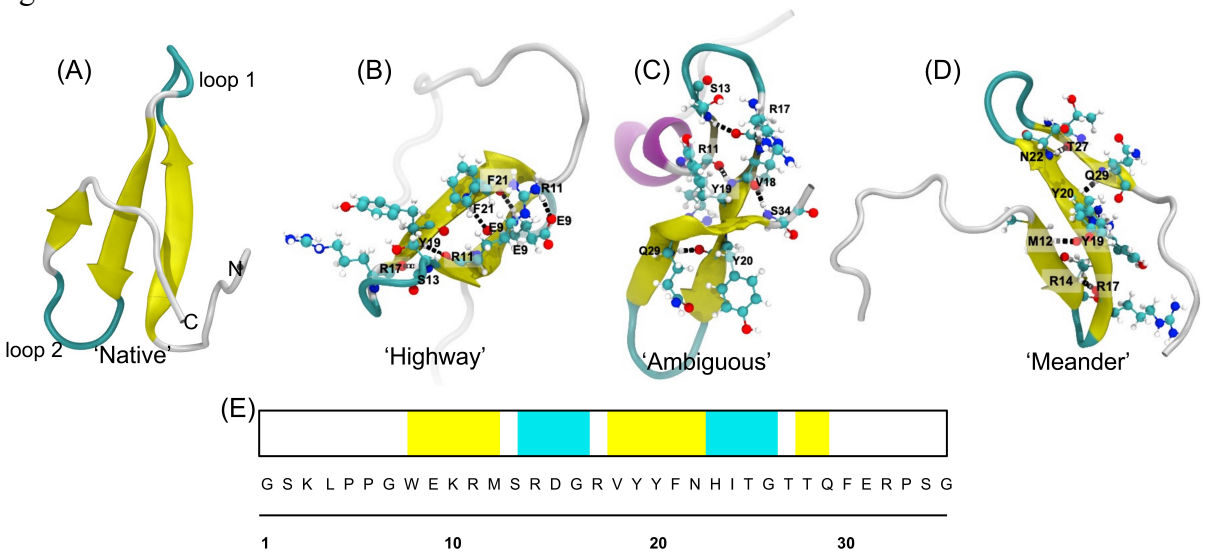


Figure 2

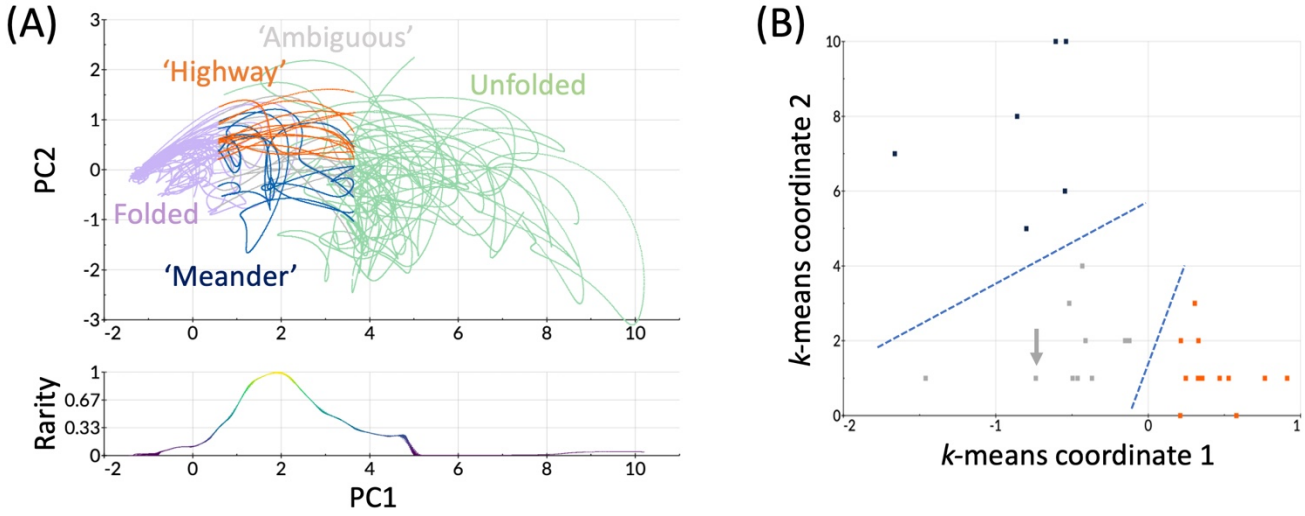


Figure 3

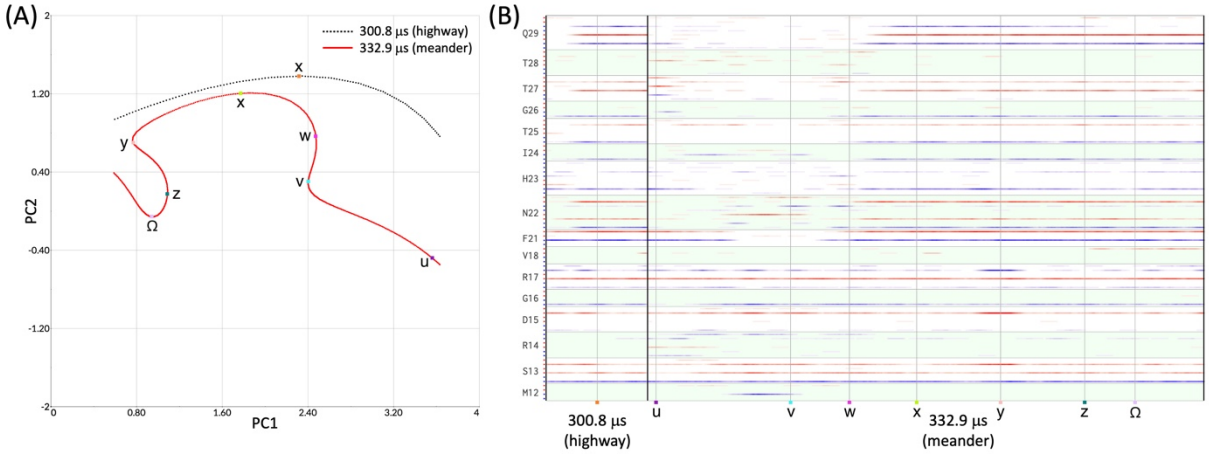


Figure 4

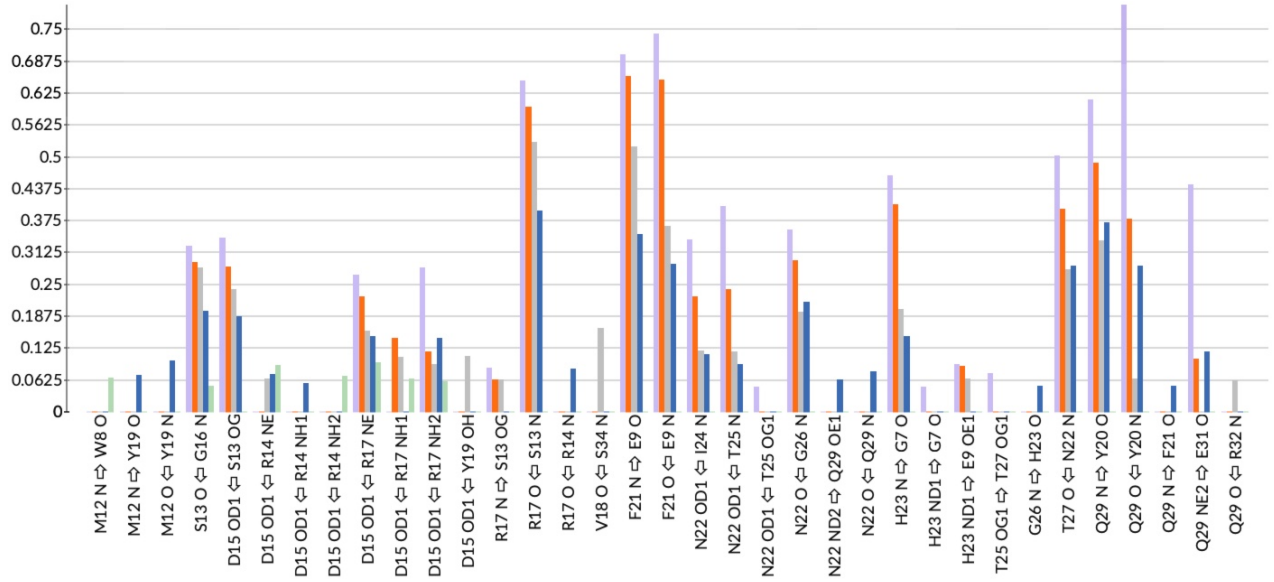


Figure 5

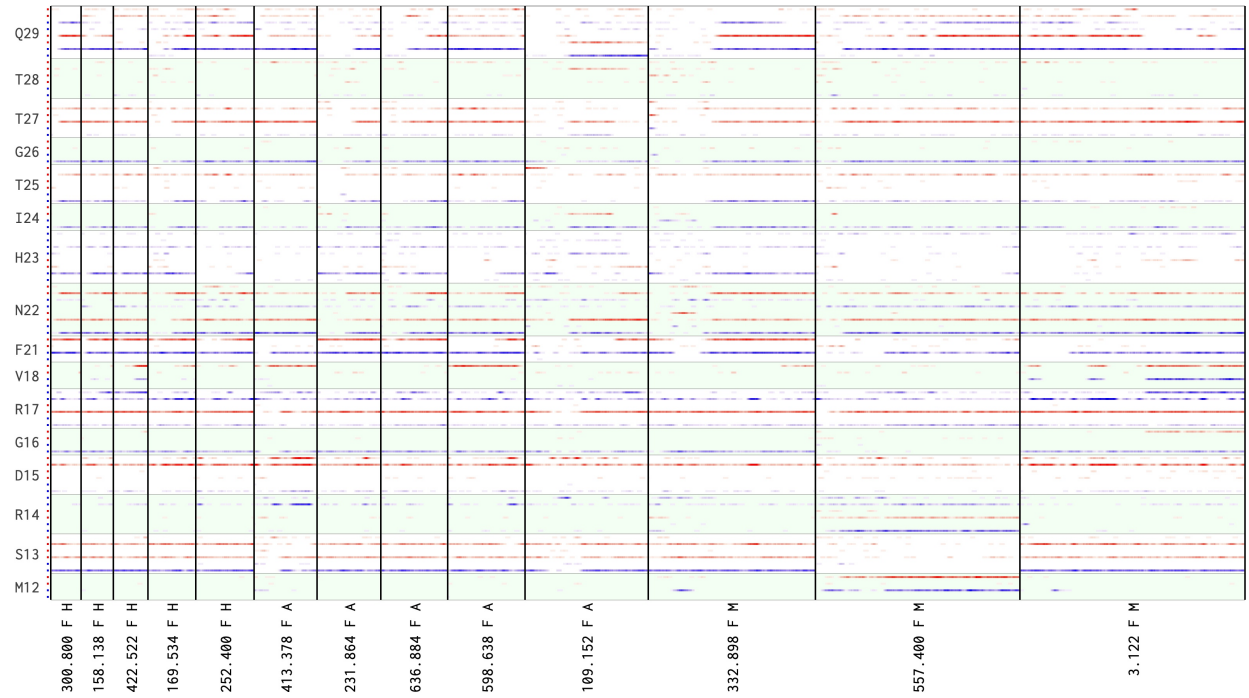
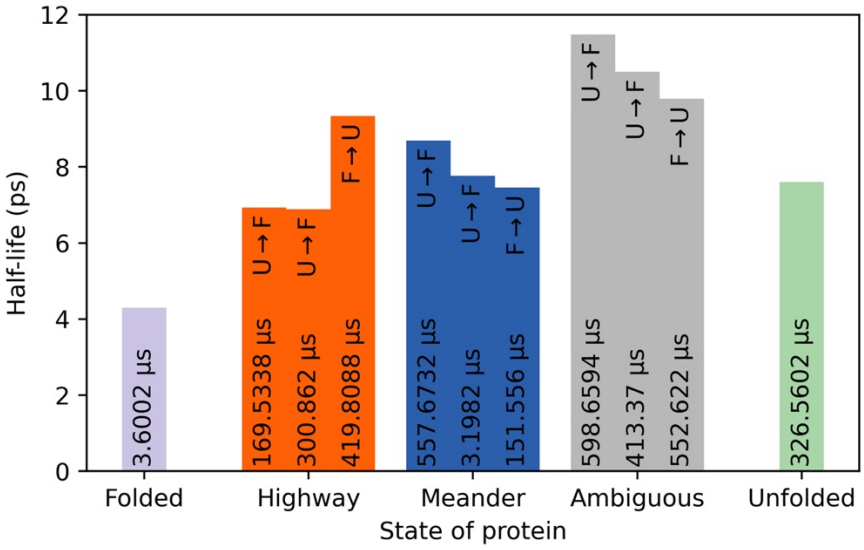


Figure 6


Half-life (ps)
0
2
4
6
8
10
12
3.6002 µs
169.5338 µs
U → F
300.862 µs
U → F
419.8088 µs
F → U
557.6732 µs
U → F
3.1982 µs
U → F
151.556 µs
F → U
598.6594 µs
U → F
413.37 µs
U → F
552.622 µs
F → U
326.5602 µs
Folded
Highway
Meander
Ambiguous
Unfolded
State of protein