BioMol Concepts, Vol. 1 (2010), pp. 335–344 • Copyright © by Walter de Gruyter • Berlin • New York. DOI 10.1515/BMC.2010.032

Review

Dynameomics: protein dynamics and unfolding across fold space

Amanda L. Jonsson[1], R. Dustin Schaeffer[2], Marc W. van der Kamp[1] and Valerie Daggett[1,2,*]

[1] Department of Bioengineering, University of Washington, Box 355013, Seattle, WA 98195-5013, USA
[2] Biomolecular Structure and Design Program, University of Washington, Box 355013, Seattle, WA 98195-5013, USA

* Corresponding author
e-mail: daggett@u.washington.edu

Abstract

All currently known structures of proteins together define 'protein fold space'. To increase the general understanding of protein dynamics and protein folding, we selected a set of 807 proteins and protein domains that represent 95% of the currently known autonomous folded domains present in globular proteins. Native state and unfolding simulations of these representatives are now complete and accessible via a novel database containing over 11 000 simulations. Because protein folding is a microscopically reversible process, these simulations effectively sample protein folding across all of protein fold space. Here, we give an overview of how the representative proteins were selected and how the simulations were performed and validated. We then provide examples of different types of analyses that can be performed across our large set of simulations, made possible by the database approach. We further show how the unfolding simulations can be used to compare unfolding of structural elements in isolation and in different structural contexts, using as an example a short, triple stranded β-sheet that forms the WW domain and is present in several larger unrelated proteins.

Keywords: dynameomics; molecular dynamics; protein folding; transition state; WW domain.

Introduction

The Protein Data Bank (PDB) (1) currently contains around 65 000 protein structures and is likely to continue to expand. The coordinates deposited in the PDB are static, average structures of proteins. Although these structures provide an abundance of information, they only represent a small part of the full story. In reality, proteins are dynamic entities that sample an ensemble of conformers in their folded (native) states. This dynamic behavior of proteins is crucial for understanding their function, their interactions and perhaps even their evolution (2–4). Obtaining a detailed picture of the dynamics of a protein can therefore lead to new insights. An example is provided by the relationship between cytochrome b_5 dynamics and formation of complexes with other cytochrome proteins: molecular dynamics (MD) simulation of the native state dynamics of this protein revealed cyclical formation of a cleft giving access to the buried heme group (5). We hypothesized that the cleft serves as a site for binding of its protein partners, providing a more protected site for electron transfer. Formation of the cleft and binding of cytochrome *c* to the cleft were confirmed later by experiment (6, 7). Thus, thermal motion can be crucial for molecular recognition and MD simulations can reveal important excursions from the static, average structures. For many proteins it is not yet understood how their movements affect their function, as well as how dynamics is related to the three-dimensional fold.

Apart from being important for the function of proteins, dynamics is also involved in the process of adopting the native three-dimensional fold and will determine available pathways of unfolding and misfolding. Folding and unfolding play an important part in the life cycle of a protein (e.g., adopting their functional form, translation through the cell and degradation), as well as the life cycle of an entire organism (e.g., aging and disease). Increased understanding of the protein folding/unfolding process can provide important rules to help predict structure from sequence, a major challenge to translate the ever-growing data from genome sequencing efforts into biologically relevant information. Understanding of protein unfolding will not only help to understand the many cellular processes that involve partial unfolding (8), it can also provide crucial insights into the growing number of amyloid diseases (9, 10) and the molecular basis of the consequences of amino acid mutations due to single-nucleotide polymorphisms linked to disease (11). The dynamics of a protein in its natively folded form provides an important benchmark for analyzing its unfolding.

Obtaining a detailed picture of the unfolding/folding pathway of a protein requires structural information on all the different conformational states involved (native, transition, intermediate and denatured), as well as the mechanism of interconversion between these states. Experimentally, this information is difficult to obtain, owing to the transient, dynamic and heterogeneous nature of partially folded states. Fundamentally, obtaining structural information on the

unfolding transition state (the state that defines the main energy barrier for forming or unfolding the native protein fold) is particularly difficult. Computer modeling, in particular physics-based simulations, can help to fill in information that is difficult or impossible to obtain through experiment (12). Specifically, all-atom MD simulations in explicit solvent are able to provide the detailed temporal and spatial resolution needed to understand protein folding/unfolding (13). By combining such simulations with experiment, the pathways of individual proteins can be mapped accurately and in great detail (14–16).

To characterize protein dynamics and unfolding more generally, we established a large-scale, high-throughput simulation project, named Dynameomics (www.dynameomics.org) (17–19). The main aim of this project is to simulate the native state and unfolding of a set of proteins and protein domains that represent all the known independent folding units that occur in globular proteins. In addition to the set of individual proteins that together represent protein fold space, we are simulating multiple members of certain fold-families, simulations of the GGXGG set of pentapeptides (20) and simulations of proteins with single-nucleotide polymorphisms that can provide insight into the causes of genetic disease (19). In total, we have over 11 000 simulations of over 2000 different protein and peptide systems totaling more than 390 μs of simulation time and over 4×10^8 structures, 10^4 times more than in the PDB (note that precise numbers are not given as we are constantly running and loading simulations into the database).

Given the vast amount of data produced in MD simulations, one of the biggest challenges is data management and organization. Consequently, we developed a novel hybrid relational/multidimensional database (12–14) that scales well with increasing simulations and is optimized for efficient queries across large datasets (200 terabytes and growing). By collecting physically realistic simulations of protein dynamics and unfolding in a structured database, we can perform analyses across all simulations and specific subsets thereof. Although the Dynameomics project is, to our knowledge, the only such project of this scale, fold coverage and database complexity, others are also now collecting and organizing biomolecular simulation data (21).

In this overview of the Dynameomics project, we describe our simulation data from a biophysical perspective, omitting the technical details involved in storing and accessing large simulation datasets. We first outline the origin of the consensus domain dictionary from which we assign fold representatives, or targets for simulation, and the protocols that were used to perform, characterize and validate the MD simulations. We illustrate different types of analyses that can be used to characterize large collections of native state simulations. Then, as a further example of what is made possible by our comprehensive database of simulations, we focus on the mechanism of unfolding of the β-hairpin structural motif. β-Hairpin unfolding in the context of the minimal WW domain (essentially a double hairpin) is compared with the unfolding of double hairpin motifs in unrelated protein folds, where this motif occurs in different structural contexts.

Selection of protein fold representatives

The goal of the Dynameomics project is to capture protein dynamics and unfolding across all known independently folding (or autonomous) protein domains. It is, however, not feasible to perform simulations of all protein structures deposited in the PDB. Apart from the fact that it is not feasible, it is also not necessary, as both the PDB and nature contain many structures that have very similar folds, and many PDB entries contain more than one independent folding unit. The aim therefore is to select a set of proteins and protein domains that is representative of all folds that occur in the structures in the PDB. To this end, we make use of three well-known protein fold classification systems, each with a different philosophy and methodology to distinguish fold similarity: SCOP (22), CATH (23) and Dali (24). We have devised a procedure to (i) take all domains in the PDB that are classified by at least two of the three systems; (ii) look up and match their classifications; (iii) filter the domains by sequence (<95% sequence identity); and (iv) identify what we call 'metafolds' as domains that share at least two of the three classifications (18, 25) (Figure 1). We call the result of this procedure a Consensus Domain Dictionary (CDD) and the latest version (2009) consists of 1695 metafolds across a total of 80 062 domains or 13 345 non-redundant domains (25). All metafolds, including all domains and their classifications, are available at www.dynameomics.org.

The next step is to select metafold representatives for simulation (Figure 1). First, metafolds that are not truly independently folded domains were filtered out, including: (i) domains that extend through domain-swapped dimers; (ii) domains that are part of a complex and show a large buried interface; (iii) domains with secondary structure elements that continue into other domains (i.e., a β-strand forming a β-sheet with a strand from another domain); and (iv) domains that lack regular secondary structure elements and/or were unstructured peptides. A surprisingly large number of metafolds falls into this category (672 metafolds), calling into question their use in bioinformatics studies addressing globular protein properties [see Figure 1 here and Figure 5 in reference (25) for examples]. Then, the quality of the experimentally determined structure was taken into account: structures with gaps extending more than seven residues and structures determined by X-ray crystallography with a resolution >3 Å were deemed insufficiently accurate for simulation. Including a recently retracted structure (1 BEF), 107 metafolds (representing 2% of the non-redundant domains) were rejected in this step. Because we aim to simulate folds occurring in globular proteins in water, we also omitted 27 metafolds that only occur in transmembrane regions.

Of the remaining metafolds (889), we did not simulate those with obligate cofactors other than Zn^{2+}, Ca^{2+} and heme. The cofactors in these domains, totaling 57 metafolds, representing <2% of the non-redundant domains, were often major structural elements (Figure 1). We also excluded 11 metafolds with domains containing over 450 residues. For the remaining 821 metafolds, we selected representative

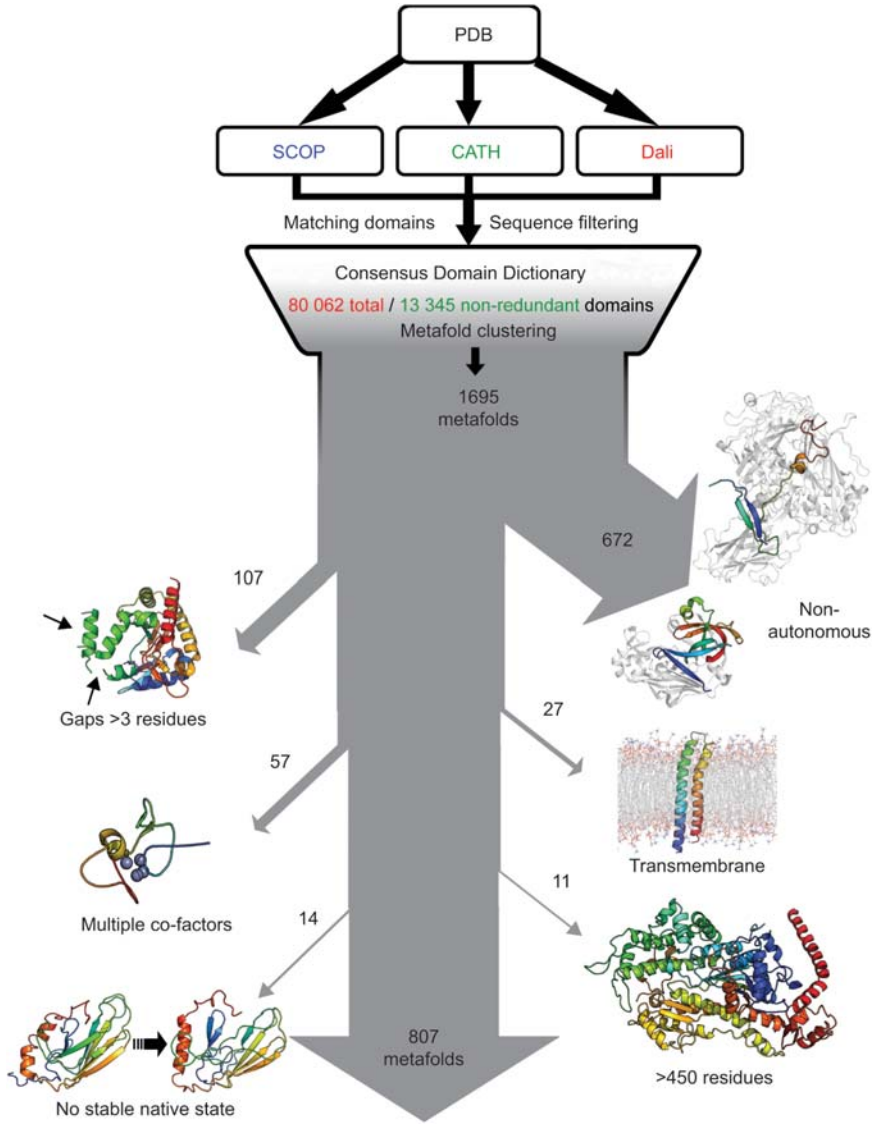

Figure 1 Selection of autonomous fold representatives for simulation.
First, a 'Consensus Domain Dictionary' was defined by matching domains between different fold dictionaries (SCOP, CATH and Dali) and filtering for sequence identity. Thereafter, clustering identified 1695 'metafolds', which represent all currently known protein structures (25). Several of these folds are not self-contained [672], only exists in membrane proteins [27] or contain non-protein cofactors that contribute significantly to the protein fold or did not have well-tested parameters for simulation [57]. Others were not suitable for simulation with our standard protocols due to large parts of unknown structure [107], large size [11] or unstable simulation (i.e., poor starting structure) [14]. (Each reason for rejecting a metafold for simulation is illustrated by an example.) This resulted in a set of 807 metafolds for which fold representatives were simulated, analyzed and collected in the Dynameomics database. Native state simulations of the Top 100 most populated folds are publicly available at our website (www.dynameomics.org).

structures for simulation. After simulating the 821 targets, the native states of 14 metafolds proved to be unstable and no other suitable representatives could be found (see discussion below).

Our full set of simulated protein domains, representing 807 metafolds and 10 848 non-redundant domains (95% of the autonomous folded domains), was ranked by the non-redundant domain population of their metafolds, i.e., fold-representative 1 (twitchin, an immunoglobulin-like β-sandwich) 'represents' 1279 non-redundant domains in the PDB, whereas metafolds 502–807 contain only one structure. Consequently, there is a sharp drop-off in coverage of domains (Figure 2), with the first six representatives covering 30% of the known, autonomous, globular protein domains and the Top 100 representatives (for which native state simulations are publicly available on the Dynameomics website) cover 72%. In the selection of representatives for metafolds with multiple domains, we preferred domains with biomedical relevance or with experimental folding studies available. Our representatives cover a wide range in size (29–417 residues),

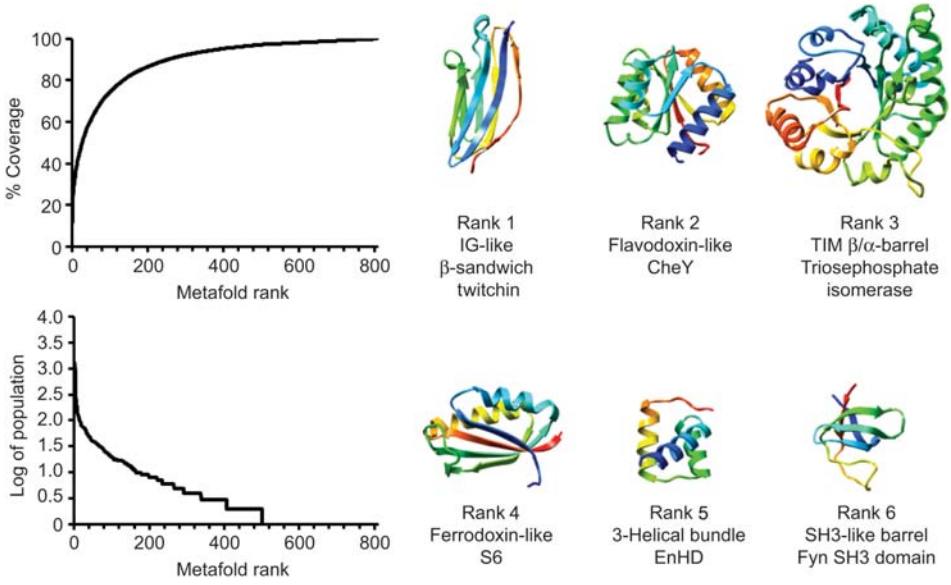

Figure 2 Dynameomics fold space.
(Top left) Cumulative percentage of all 13 345 non-redundant domains in the metafolds. (Bottom left) Log of the non-redundant population of each metafold. (Right) Metafold representatives of the top six metafolds. Structures are shown in ribbons colored in rainbow from red to blue. Metafold rank and name, along with the protein name are listed below each structure.

function (enzymes, enzyme inhibitors, transcription factors, structural proteins) and are derived from 218 source organisms [see ref. (19) and www.dynameomics.org for more details].

Protocols for simulation and analysis

To sample the native state dynamics of all representative proteins, we performed MD simulations at room temperature (298 K) (17). To obtain information on the folding/unfolding of domains, we use high-temperature MD simulations starting from the experimentally determined structure (26). Although simulations of protein folding (from extended or denatured states) are now possible for certain small, fast-folding proteins, it remains very challenging, both methodologically and computationally, and approximations that can affect the folding pathway are often necessary (27). Folding simulations are therefore not suitable for the high-throughput project described here. Fortunately, previous simulation studies have shown that protein folding is a microscopically reversible process, i.e., the unfolding pathway is essentially the same as the folding pathway both at a single temperature and comparing high temperature and quenched refolding (28, 29). In addition, previous studies have shown that the unfolding process is largely insensitive to changes in temperature and that raising the temperature merely accelerates the process and the same conformational states are visited in simulations at different temperatures (16, 28, 30). Consequently, the use of high temperature is a reasonable choice for our high-throughput simulation effort. Furthermore, it allows us to run multiple simulations for each representative, which can be used to capture the average properties of states along the unfolding pathway (31).

To sample the unfolding pathway and the denatured state, we have performed at least two simulations of all 807 fold representatives for 51 ns at 498 K. To obtain additional sampling of the early unfolding events, including the transition state, we performed at least three additional short simulations (≥ 2 ns). Because simulation at high temperature will rapidly cause significant conformational changes, it is important to sample the native state dynamics as a reference. For all our 807 representatives, we therefore performed at least one simulation of 51 ns or more at a temperature of 298 K. Preparation and simulation (at least six simulations for each representative) was performed according to a fixed protocol, as described previously (17). In brief, we obtain the starting structure from the PDB, add missing atoms, side chains and/or residues if necessary, perform a short energy minimization and solvate the structure in water (using the experimental density for 298 K or 498 K). All atoms are explicitly represented using fully flexible parameters for the protein as defined in our force field (32), with the flexible three-center water model (33). Our in-house modeling package, *in lucem* molecular mechanics (*il*mm), was used for all calculations (34).

Once simulations were complete, each trajectory was characterized through an extensive set of analyses. Broadly, these analyses serve to identify gross structural changes, monitor changes in secondary structure, determine the number of contacts between protein atoms, measure the solvent accessible surface area (SASA) of the protein, etc. For a comprehensive list of analyses performed see (17).

Validation of native state simulations

The simulations at 298 K are designed to sample the native state dynamics of the protein folds. As the selected fold representatives should be stable independent folded domains, we first assessed the stability of the proteins. This stability assessment was performed by calculating the structural deviation from the starting structure and fluctuation about the mean structure. Because conformational changes and flexibility in large loops or tails in the structure will influence these measurements significantly, we took these measurements only over the 'core' of the protein. Of our original 821 representatives, 19 were considered unstable by these metrics. All 19 started from older NMR structures (19). For five of these 19 metafolds alternative crystal structures were available and they were stable by MD.

We now compare the resulting 807 native state simulations to experimental data for further validation. Pairwise distance restraints obtained from Nuclear Overhauser Effect (NOE) crosspeaks are particularly informative in this respect. Previously, we reported on NOE comparison from a set of 27 proteins in our simulation set (17), yielding an overall NOE restraint satisfaction of 92% (based on 28 504 NOE restraints). There are now NOE lists available for 117 of our targets from the BioMagResBank (35). To remove contradictory NOEs, we used the parsed and filtered constraints as present in the Filtered REstraint Database (36). A total of 148 580 NOEs were obtained, averaging 13.6 NOEs per residue.

Comparison of these NOE data with our simulations revealed an average NOE restraint satisfaction per fold-representative of 91%. As expected, satisfaction of short-range constraints (i.e., $i \rightarrow i \leq 2$) was higher than that of long-range constraints ($i \rightarrow i \geq 5$). In general, the agreement between MD and experiment is good. For example, for the engrailed homeodomain (representative for the three-helical bundle fold, rank 5) and ubiquitin (representative of the β-grasp fold, rank 8), we used the available crystal structures as a starting point for simulation (1 ENH and 1 UBQ, respectively), and NMR data are also available. In both cases, the crystal structure satisfied fewer NOEs than the average satisfaction of the simulation, indicating how MD simulation moved the crystal structure closer to the solution ensemble probed experimentally.

Analysis of native state ensembles

The Dynameomics database provides an organizing framework, a repository and a variety of access interfaces for the data stored in it: the CDD, the coordinates obtained through MD simulation and the related analysis data or metadata (19). The database was designed to be a uniform, scalable and reliable data warehouse, making it possible to perform queries across data from thousands of simulations. It essentially is a hybrid database model, partly a relational database and partly a multidimensional on-line analytical processing database, which can be queried using the structured query language (37, 38). Together with the comprehensive set of simulations and analysis data collected in the database, this setup enables comparisons across all protein folds and subsets thereof. For example, the 807 different metafolds for which representatives are present in the database can be subdivided into broad classes based on their secondary structure or their size (Figure 3). Next, we can compare important summary statistics across the whole set and the different subsets (Table 1). This comparison indicates that there is no significant bias in our set of simulations; similar values are obtained throughout.

We have also used additional, non-standard analyses to investigate properties of the dynamics across the different fold representatives. One example is the analysis of flexibility, based on a method outlined by Teodoro et al. (39). This analysis allows one to obtain a general view of an entire simulation by showing the primary modes of every atom in the simulated protein. Using this technique, we can scan the native state simulations for regions in proteins that show flexibility that is uncharacteristic of their secondary structure. This revealed several unusually rigid loops with distinct properties that can constitute a new class of non-traditional secondary structure (40). By examining additional simulations of several metafolds, we determined that the backbone motions of proteins within a metafold are related and correlated with sequence similarity.

Apart from information on the dynamics of different protein folds, the acquired simulation data can also provide a comprehensive resource for conformational and dynamic properties of proteins in general. For example, we have captured the backbone conformations of all three to nine residue fragments from our simulations and clustered results provide fragment libraries that cover significant structural diversity, which are available for download on our website. We also collected a library of conformational preferences of amino acid side chains within proteins (as part of our Structural Library of Intrinsic Residue Propensities, also available through our website) and analyzed their dynamics (41).

An example of a 'fold independent' analysis of our protein dynamics dataset is the distribution of backbone torsion angles (i.e., Ramachandran distributions) per residue. In Figure 4, we show the distribution for Ile residues as an example; Ile residues are evenly distributed between α-helices, β-sheets and other structural elements in our 807 starting structures. We can compare the distribution from our Dynameomics simulations with those derived from static structures of proteins [from the ASTRAL-40 dataset (42)] and those from a sterically unrestrained Ile distribution, based on exhaustive simulation of Gly-Gly-Ile-Gly-Gly (GGIGG) pentapeptide (20). Relative to the static data, the Dynameomics data are more dispersed. In addition, there is increased sampling of the α-left Ramachandran region. The peaks in the two distributions, i.e., the most dominant structural preferences, are very similar. By contrast, the distribution for the Ile backbone in the sterically unrestrained pentapeptide shows marked shifts in the maxima in the α-helical and β-sheet regions, in addition to increased dispersion. This discrepancy indicates that in a hydrated, unhindered envi-

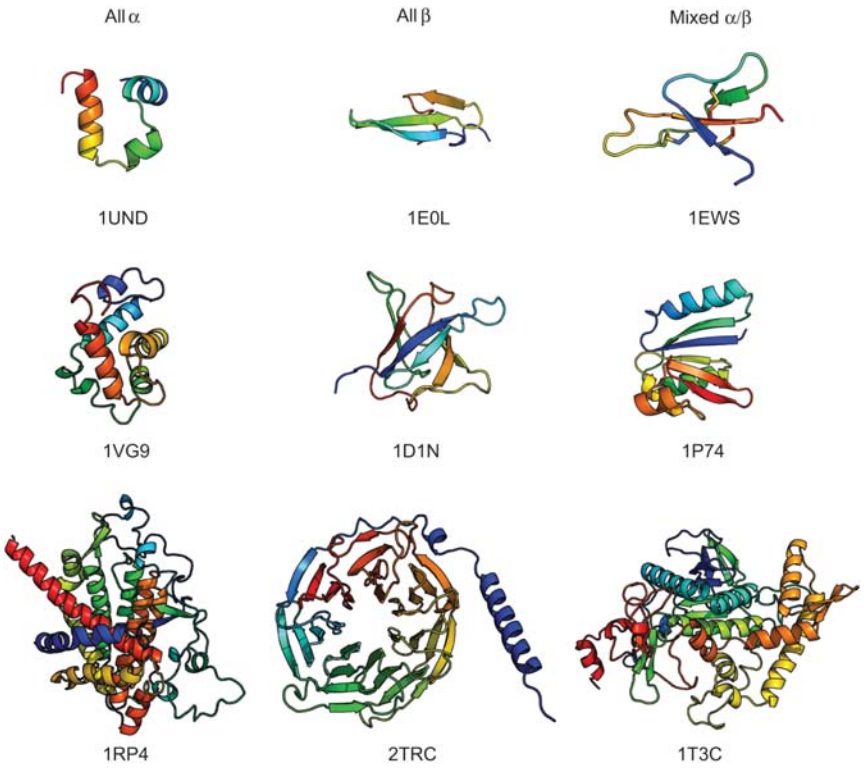

Figure 3 Examples of fold representatives of different structural class and size.
From top to bottom, examples (including PDB code) are given of small (<50), medium (≥50, <150) and large (≥150) representative protein domains. From left to right, different structural classes are depicted, as indicated.

ronment the conformations sampled by Ile are significantly different from those when the residue is found in the context of a folded protein. Similar results were found for the other 19 naturally occurring amino acids (20).

Using simulation to characterize protein unfolding: the WW domain

WW domains consist of a three-stranded, antiparallel β-sheet (a double hairpin) and have been used as a model system for β-hairpin folding by both simulation and experiment. Experimental studies suggest that WW domains have high amounts of structure in the turn of the first hairpin in the protein folding transition state (TS) (43–45). These results have been interpreted to mean that this turn is structured in the TS and acts as a structured nucleus for hairpin folding. We used MD simulations to study both the native state dynamics and the unfolding pathway of the WW domain to gain a better understanding of the role the first turn plays in hairpin formation (30, 46).

Multiple simulations of three WW domain proteins (FBP28, Pin1 and hYAP) were performed at 285 K and 298 K to study the native state dynamics (46). Structure in the first turn of each WW domain fluctuated in all of the native state simulations, sampling different regions of (φ, ψ)

Table 1 Average properties of native state simulations grouped by structural class and size.

Subset	No. of fold representatives	Cα RMSD (Å)	Total SASA per residue (Å^2)	Radius of gyration (Å)	Fraction of α-helical residues[a]	Fraction of β-sheet residues[a]
All	807	2.9±1.1	71.4±24.9	14.4±2.9	0.44±0.27	0.23±0.13
All α	216	2.8±1.2	74.1±22.3	14.1±3.1	0.71±0.25	0.05±0.02
All β	135	3.1±1.2	68.9±20.5	13.8±2.7	0.14±0.13	0.33±0.14
Mixed α/β	408	2.9±1.0	70.0±25.4	14.8±2.7	0.36±0.17	0.21±0.10
Other	48	2.8±1.0	76.1±30.7	13.8±3.3	0.35±0.25	0.21±0.15
Small (<50)	29	2.7±1.0	96.1±35.4	9.7±0.9	0.49±0.32	0.26±0.19
Medium (≥50, <150)	522	2.9±1.2	74.2±23.6	13.2±1.7	0.45±0.28	0.24±0.14
Large (≥150)	256	3.0±1.0	61.8±20.6	17.7±2.2	0.42±0.25	0.20±0.10

[a]Residues that are participating in (at least) three-residue motifs of secondary structure.

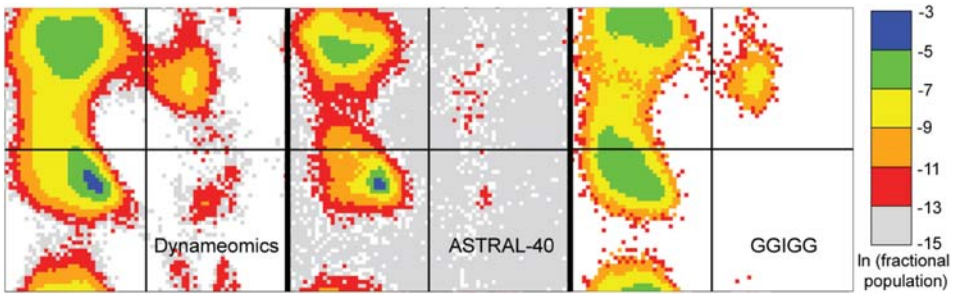

Figure 4 Ramachandran distributions for Ile in different structural contexts.
Represented are Dynameomics native state simulations, static protein structures and simulation of a sterically unrestrained amino acid (in a GGIGG pentapeptide). (Φ, ψ) space (from -180° to +180°) is divided into 72 bins with a width of 5°. Each bin is colored by fractional population on a logarithmic scale. From left to right, distributions are shown for (i) all Ile residues in all time samples in the Dynameomics native state simulation set, (ii) all Ile residues present in the ASTRAL-40 dataset (v1.74) (42) and (iii) all time samples from simulations of the (acetylated and amidated) pentapeptide Gly-Gly-Ile-Gly-Gly (20).

space, which was confirmed through NMR relaxation experiments and comparison with crystallographic B-factors. Consequently, the first turn of WW domains is flexible in the native state.

We also performed unfolding simulations of FPB28 WW domain under different conditions. Initially, simulations were run at low pH (protonation of Asp and Glu residues) at 333 K, 348 K and multiple simulations at 373 K (30, 44). As part of the Dynameomics project, unfolding simulations of FBP28 WW domain (rank 194) were also run at neutral pH and 498 K. The unfolding pathway was similar across all temperatures and conditions. The third strand was the first to unfold in the major unfolding pathway at all temperatures (Figure 5A).

The next step was to identify important states along the unfolding pathway, such as the TS ensemble, which was done using a conformational clustering method (15, 47). Using this method, we identified TS ensembles from all unfolding simulations of the WW domain. Structure indices, or S-values, were then calculated to quantify the amount of both secondary and tertiary structure present at each residue in the TS ensemble (48). The S-values can subsequently be

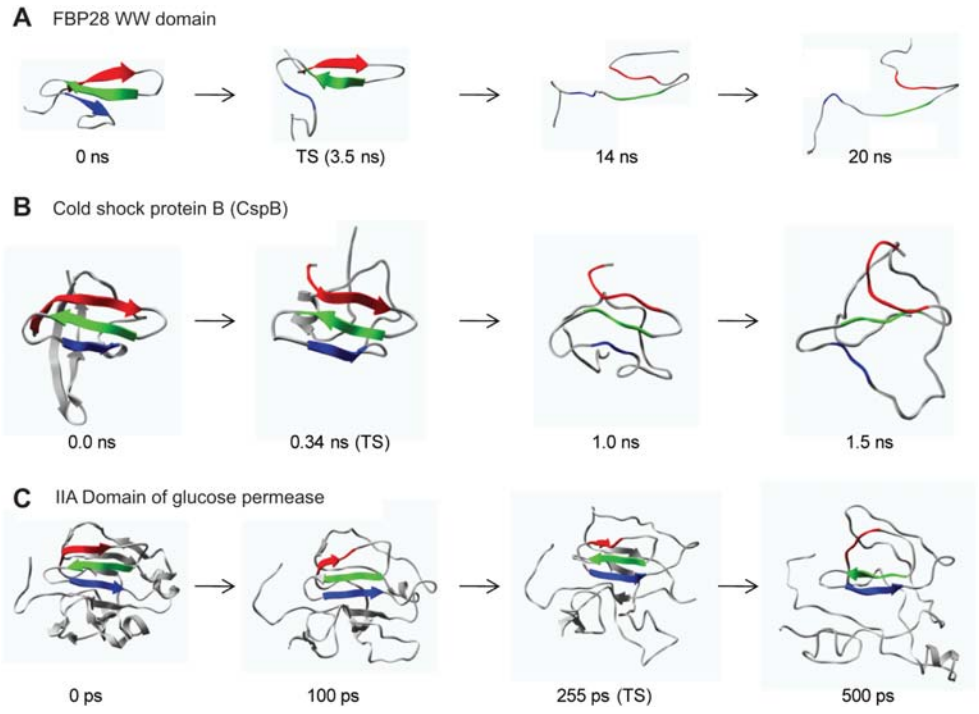

Figure 5 Unfolding of the double hairpin motif in different protein contexts.
(A) Snapshots from a 348 K unfolding simulation of FBP28 WW domain with β1 in red, β2 in green and β3 in blue. (B) Snapshots from a 498 K simulation of CspB with β1 in red, β2 in green and β3 in blue. (C) Snapshots from a 498 K simulation of IIA domain with β4 in red, β5 in green and β6 in blue.

compared with the experimental Φ-values to validate the unfolding simulations.

In all cases, the backbone was very flexible for the first turn of FBP28. Because this turn is flexible even in the native state, it is not surprising that the backbone angles fluctuate during unfolding. However, this artificially lowered the S-values in the turn, even in the presence of side chain interactions. Therefore, the tertiary component of the S-values better represents the amount of structure in the turn in the TS ensemble (30). The resulting S- and Φ-values are in good agreement for FBP28 WW domain, independent of temperature and pH (R of 0.7 between S- and Φ-values).

Examining the unfolding simulations of the WW domain in reverse to study folding revealed that the residues in the first turn formed a kink, allowing side chain contacts to form between residues in the first two strands. These contacts brought the backbones into contact and allowed the formation of hydrogen bonds, in no particular order, to form the first hairpin. The native state and unfolding simulations of the WW domain revealed that the first turn acts as a nucleus for folding via formation of tertiary interactions that pull the chain around so that it doubles back, but the precise formation of the structure of the first turn does not drive folding (30, 46).

Understanding β-hairpin unfolding within a larger protein structure

But how applicable are the results of a model system like the WW domain to the unfolding of β-hairpins in proteins? The WW domain is a convenient model system with both experimental and simulation data readily available. But the domain lacks a conventional hydrophobic core and the opportunity for contacts between the hairpins and other portions of the domain. We mined the Dynameomics database looking for double hairpin motifs structurally similar to the WW domain but part of larger proteins. Both the cold shock proteins in the OB-fold metafold (rank 9) and the IIA domain of glucose permease in the barrel-sandwich hybrid metafold (rank 352) have such double hairpin motifs.

The cold shock protein B (CspB), representing rank 9 in the CDD, has a double hairpin motif at the N-terminus (strands β1, β2 and β3) with contacts to the core as part of an OB-fold. CspB has many aromatic contacts within the double hairpin motif itself. The first event along the unfolding pathway is the separation of the double hairpin motif from the hydrophobic core and the rest of the protein structure. After this separation, the double hairpin motif proceeds to unfold in a manner similar to the WW domain, with contacts lost first between strands β2 and β3, whereas structure remained in the first hairpin (Figure 5B). Also similar to the WW domain, there was no order to loss of hydrogen within a hairpin. Instead, hydrogen bonds were often broken starting in the middle of a hairpin, with no evidence for a zipper-like unfolding mechanism.

TS ensembles from the unfolding simulations of CspB were identified using the conformational clustering method described above and compared them to the 12 available Φ-values (49, 50) with a $\Delta\Delta G_{D-N} > 0.7$ kcal/mol. The average S-values from four of the five unfolding simulations (one simulation, run #2, was problematic, unfolding very slowly which complicated the clustering) of CspB have a correlation coefficient of 0.7 to the Φ-values.

The Dynameomics database also contains simulations of cold shock protein A (CspA), another member of the OB-fold (metafold rank 9). CspA shares 60% sequence identity with CspB, including an aromatic cluster within the double hairpin motif. The unfolding simulations of CspA follow the same major unfolding pathway as CspB, with the double hairpin motif first breaking away from the rest of the protein followed by loss of structure in the second hairpin.

By contrast, the IIA domain of glucose permease is a larger protein with 13 β-strands and two helices (Figure 5C). The double hairpin motif (strands β4, β5 and β6) is in the middle of a larger β-sheet and the motif lacks the aromatic cluster seen in the cold shock proteins. Unlike the CspB unfolding pathway, the double hairpin motif in the IIA domain maintains contacts to the hydrophobic core and the other strands in the β-sheet as it unfolds. Unfolding within the double hairpin motif varies among the multiple unfolding trajectories of this protein. In some cases the second hairpin (strands β5 and β6) maintains contacts longer than the first hairpin in the motif, whereas other simulations lost all structure in the double hairpin motif at the same time. Instead, β-structure was consistently first lost further down in the β-sheet in strands β10, β2 and β1 (Figure 5C). Similar to the hairpin unfolding in the other proteins, there was no consistent order to the loss of hydrogen bonds.

Our existing Dynameomics database of simulations allowed us to look at the unfolding of a double hairpin motif in multiple contexts. We can examine the unfolding of the motif alone, in the model system WW domain, with minimal contacts to a proteins core in the cold shock protein and as part of a larger β-sheet in the IIA domain. In all cases, there was no evidence for a zipper-like unfolding mechanism, where hydrogen bonds would be lost starting from the hairpin ends, continuing along the strands until the last hydrogen bonds to break would be in the β-turn. Instead, the loss of backbone hydrogen bonds within the hairpin varied between simulations for all of the proteins studied.

This type of analysis made possible by our existing Dynameomics database of simulations increases our confidence in the results from the WW domain simulations in how the hairpins themselves unfold. But the unfolding behavior of the entire double hairpin motif is dependent on the surrounding protein context. The cold shock proteins lost contacts between the double hairpin motif and the rest of the protein very early in the unfolding pathway, allowing the motif to unfold in a manner similar to the WW domain. In contrast, the double hairpin motif in the IIA domain maintains contacts to the rest of the protein as it unfolds and therefore has a more varied unfolding pathway. The contacts to the rest of the protein for the IIA domain were dominant over the intrinsic unfolding behavior of the double hairpin motif, as reflected in the WW domain.

Conclusions and outlook

By defining and simulating a set of proteins representative of (almost) all known autonomous protein folds, we have created a bioinformatics resource for protein dynamics and protein unfolding. Organization of the data in a flexible and queryable database allows access to this resource, enabling comparisons of protein dynamics and folding across protein folds. Our analysis of the large dataset of simulations has already offered insights into native state protein dynamics, properties of protein folding transition states and the effect of environment on the unfolding of structural elements.

Altogether, we believe that our high-throughput simulation effort and storage of these data in an easily accessible structured repository, which can be linked to other sources of biological and experimental data, can be a valuable resource for researchers in biology, biochemistry and biophysics. Not only does our database provide high-resolution information on the dynamics and unfolding of individual proteins, it will allow the exploration of broader scientific questions by analyzing dynamics and unfolding across fold space in a way that was previously impossible or extremely cumbersome (51). We expect that such exploration will increase the general knowledge of protein dynamics and contribute to solving 'the protein-folding problem'. Currently, native state dynamics simulations of the Top 100 most frequently occurring protein folds are publicly accessible (see www.dynameomics.org).

Acknowledgements

Dynameomics simulations were performed using computer time through the DOE Office of Biological Research as provided by the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. We are grateful for support from Microsoft for development of our database and related mining methods. We are also grateful for financial support provided by the National Institutes of Health (GM50789).

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000; 28: 235–42.
2. Karplus M, Kuriyan J. Molecular dynamics and protein function. Proc Natl Acad Sci USA 2005; 102: 6679–85.
3. Smock RG, Gierasch LM. Sending signals dynamically. Science 2009; 324: 198–203.
4. Tokuriki N, Tawfik DS. Protein dynamism and evolvability. Science 2009; 324: 203–7.
5. Storch EM, Daggett V. Molecular-dynamics simulation of cytochrome b_5: implications for protein-protein recognition. Biochemistry 1995; 34: 9682–93.
6. Hom K, Ma QF, Wolfe G, Zhang H, Storch EM, Daggett V, Basus VJ, Waskell L. NMR studies of the association of cytochrome b_5 with cytochrome c. Biochemistry 2000; 39: 14025–39.
7. Storch MM, Daggett V, Atkins WM. Engineering out motion: introduction of a de novo disulfide bond and a salt bridge designed to close a dynamic cleft on the surface of cytochrome b_5. Biochemistry 1999; 38: 5054–64.
8. Prakash S, Matouschek A. Protein unfolding in the cell. Trends Biochem Sci 2004; 29: 593–600.
9. Chiti F, Dobson CM. Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem 2006; 75: 333–66.
10. Daggett V. α-sheet: the toxic conformer in amyloid diseases? Acc Chem Res 2006; 39: 594–602.
11. Rutherford K, Daggett V. Polymorphisms and disease: hotspots of inactivation in methyltransferases. Trends Biochem Sci 2010; 35: 531–8.
12. Fersht AR, Daggett V. Protein folding and unfolding at atomic resolution. Cell 2002; 108: 573–82.
13. Daggett V. Protein folding-simulation. Chem Rev 2006; 106: 1898–916.
14. Ladurner AG, Itzhaki LS, Daggett V, Fersht AR. Synergy between simulation and experiment in describing the energy landscape of protein folding. Proc Natl Acad Sci USA 1998; 95: 8473–8.
15. Li AJ, Daggett V. Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations. J Mol Biol 1996; 257: 412–29.
16. Mayor U, Guydosh NR, Johnson CM, Grossmann JG, Sato S, Jas GS, Freund SM, Alonso DO, Daggett V, Fersht AR. The complete folding pathway of a protein from nanoseconds to microseconds. Nature 2003; 421: 863–7.
17. Beck DA, Jonsson AL, Schaeffer RD, Scott KA, Day R, Toofanny RD, Alonso DO, Daggett V. Dynameomics: mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations. Protein Eng Des Sel 2008; 21: 353–68.
18. Day R, Beck DA, Armen RS, Daggett V. A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. Protein Sci 2003; 12: 2150–60.
19. Van der Kamp MW, Schaeffer RD, Jonsson AL, Scouras AD, Simms AM, Toofanny RD, Benson NC, Anderson PC, Merkley ED, Rysavy S, Bromley D, Beck DA, Daggett V. Dynameomics: a comprehensive database of protein dynamics. Structure 2010; 18: 423–35.
20. Beck DA, Alonso DO, Inoyama D, Daggett V. The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins. Proc Natl Acad Sci USA 2008; 105: 12259–64.
21. Rueda M, Ferrer-Costa C, Meyer T, Perez A, Camps J, Hospital A, Gelpi JL, Orozco M. A consensus view of protein dynamics. Proc Natl Acad Sci USA 2007; 104: 796–801.
22. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995; 247: 536–40.
23. Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, Orengo CA. The CATH classification revisited – architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res 2009; 37: D310–4.
24. Dietmann S, Holm L. Identification of homology in protein structure classification. Nat Struct Biol 2001; 8: 953–7.
25. Schaeffer RD, Jonsson AL, Simms AM, Daggett V. Generation of a consensus protein domain dictionary. Bioinformatics 2010; in press: DOI: 10.1093/bioinformatics/btq625.
26. Beck DA, Daggett V. Methods for molecular dynamics simulations of protein folding/unfolding in solution. Methods 2004; 34: 112–20.
27. Schaeffer RD, Fersht A, Daggett V. Combining experiment and

simulation in protein folding: closing the gap for small model systems. Curr Opin Struct Biol 2008; 18: 4–9.
28. Day R, Daggett V. Direct observation of microscopic reversibility in single-molecule protein folding. J Mol Biol 2007; 366: 677–86.
29. McCully ME, Beck DA, Daggett V. Microscopic reversibility of protein folding in molecular dynamics simulations of the engrailed homeodomain. Biochemistry 2008; 47: 7079–89.
30. Petrovich M, Jonsson AL, Ferguson N, Daggett V, Fersht AR. φ-Analysis at the experimental limits: mechanism of β-hairpin formation. J Mol Biol 2006; 360: 865–81.
31. Day R, Daggett V. Ensemble versus single-molecule protein unfolding. Proc Natl Acad Sci USA 2005; 102: 13445–50.
32. Levitt M, Hirshberg M, Sharon R, Daggett V. Potential-energy function and parameters for simulations of the molecular-dynamics of proteins and nucleic-acids in solution. Comput Phys Commun 1995; 91: 215–31.
33. Levitt M, Hirshberg M, Sharon R, Laidig KE, Daggett V. Calibration and testing of a water model for simulation of the molecular dynamics of proteins and nucleic acids in solution. J Phys Chem B 1997; 101: 5051–61.
34. Beck DA, McCully ME, Alonso DO, Daggett V. In lucem molecular mechanics. Seattle, WA: University of Washington, 2000–2010.
35. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao H, Markley JL. BioMagResBank. Nucleic Acids Res 2007; 36: D402–8.
36. Doreleijers JF, Nederveen AJ, Vranken W, Lin JD, Bonvin A, Kaptein R, Markley JL, Ulrich EL. BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. J Biomol NMR 2005; 32: 1–12.
37. Kehl C, Simms AM, Toofanny RD, Daggett V. Dynameomics: a multi-dimensional analysis-optimized database for dynamic protein data. Protein Eng Des Sel 2008; 21: 379–86.
38. Simms AM, Toofanny RD, Kehl C, Benson NC, Daggett V. Dynameomics: design of a computational lab workflow and scientific data repository for protein simulations. Protein Eng Des Sel 2008; 21: 369–77.
39. Teodoro ML, Phillips GN, Kavraki LE. Understanding protein flexibility through dimensionality reduction. J Comput Biol 2003; 10: 617–34.
40. Benson NC, Daggett V. Dynameomics: large-scale assessment of native protein flexibility. Protein Sci 2008; 17: 2038–50.
41. Scouras AD, Daggett V. Rotamer library. Protein Sci 2010; in press: DOI.
42. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. Nucleic Acids Res 2004; 32: D189–92.
43. Deechongkit S, Nguyen H, Powers ET, Dawson PE, Gruebele M, Kelly JW. Context-dependent contributions of backbone hydrogen bonding to β-sheet folding energetics. Nature 2004; 430: 101–5.
44. Ferguson N, Pires JR, Toepert F, Johnson CM, Pan YP, Volkmer-Engert R, Schneider-Mergener J, Daggett V, Oschkinat H, Fersht A. Using flexible loop mimetics to extend Phi-value analysis to secondary structure interactions. Proc Natl Acad Sci USA 2001; 98: 13008–13.
45. Jager M, Nguyen H, Crane JC, Kelly JW, Gruebele M. The folding mechanism of a beta-sheet: the WW domain. J Mol Biol 2001; 311: 373–93.
46. Sharpe T, Jonsson AL, Rutherford TJ, Daggett V, Fersht AR. The role of the turn in β-hairpin formation during WW domain folding. Protein Sci 2007; 16: 2233–9.
47. Li AJ, Daggett V. Characterization of the transition-state of protein unfolding by use of molecular-dynamics: chymotrypsin inhibitor-2. Proc Natl Acad Sci USA 1994; 91: 10430–4.
48. Daggett V, Li AJ, Itzhaki LS, Otzen DE, Fersht AR. Structure of the transition state for folding of a protein derived from experiment and simulation. J Mol Biol 1996; 257: 430–40.
49. Garcia-Mira MM, Boehringer D, Schmid FX. The folding transition state of the cold shock protein is strongly polarized. J Mol Biol 2004; 339: 555–69.
50. Garcia-Mira MM, Schmid FX. Key role of coulombic interactions for the folding transition state of the cold shock protein. J Mol Biol 2006; 364: 458–68.
51. Schaeffer RD, Daggett V. Protein folds and protein folding. Prot Eng Des Selec 2011; 24: 11–9.