Published in final edited form as:
J Am Chem Soc. 2013 March 27; 135(12): 4729–4734. doi:10.1021/ja311077u.

Native states of fast-folding proteins are kinetic traps

Alex Dickson[†] and **Charles L. Brooks III**[†,‡,*]

[†]Department of Chemistry, The University of Michigan, Ann Arbor, MI

[‡]Biophysics Program, The University of Michigan, Ann Arbor, MI

Abstract

It has been suggested that the native state of a protein acts as a kinetic hub that can facilitate transitions between nonnative states. Using recently developed tools to quantify mediation probabilities ("hub scores"), we quantify hub-like behavior in atomic resolution trajectories for the first time. We use a data set of trajectory ensembles for 12 fast-folding proteins previously published by D. E. Shaw Research ("How Fast-Folding Proteins Fold", Lindorff-Larsen et al, *Science*, 334, 2011) with an aggregate simulation time of over 8:2 ms. We visualize the free-energy landscape of each molecule using configuration space networks, and show that dynamic quantities can be qualitatively understood from visual inspection of the networks. Modularity optimization is used to provide a parameter-free means of tessellating the network into a group of communities. Using hub scores, we find that the percentage of trajectories that are mediated by the native state is 31% when averaged over all molecules, and reaches a maximum of 52% for the Homeodomain and Chignolin. Furthermore, for these mediated transitions, we use Markov models to determine whether the native state acts as a facilitator for the transition, or as a trap (i.e. an off-pathway detour). Although instances of facilitation are found in 4 of the 12 molecules, we conclude that the native state acts primarily as a trap, which is consistent with the idea of a funnel-like landscape.

Introduction

In the late 1990s, our understanding of protein folding increased dramatically as the classical idea of folding pathways was replaced with the more complicated ideas of energy landscapes and funnels.[1–4] This transition in thought is represented iconically by schematic diagrams of rough, funnel-shaped landscapes, with high free-energy unfolded states at the top, and a low free-energy native state at the bottom. For 15 years, these have served as the *de facto* visualization of the free energy landscape of a protein. However, the true free energy landscape of a protein is many dimensional, and unsuited for visualization as a funnel.

Just as the transition from folding pathways to funnels was motivated by advances in experimental techniques, there is another revolution in understanding occurring, this time motivated by computation. Microsecond to millisecond simulations enabled by parallel computing,[5–7] distributed computing,[8–11] and GPU technology[12,13] are becoming increasingly common, and provide us with the ability to see multiple folding and unfolding

*To whom correspondence should be addressed brookscl@umich.edu.

Supporting Information Available

We examine the correlation of native state hub scores with protein size and folding time. We also show the representative structures of native-mediated communities, discuss the effects of increasing the number of microstates and the number of communities, and show the effect of removing nodes from the native state of the PRG network.

This material is available free of charge via the Internet at http://pubs.acs.org/.

events for a variety of fast-folding proteins. Enhanced sampling methods can also be incorporated with these technologies to describe events on even longer timescales.[14–19] Using network analysis, with specific conformations represented by nodes, and the transitions between conformations represented by edges, we can visualize the entire accessible free energy landscape of a protein in a single network graph.[14,20–24] However, questions still remain as to the best way to generate these graphs, and how we should interpret their main features.

In particular, questions have recently arisen as to the role of the native state, specifically whether it can be well-described as a kinetic hub.[9,22,25] This can be broken down into two key questions. First, does the native state mediate nonnative-to-nonnative transitions? Secondly, are native-state–mediated transitions faster or slower than transitions that are not mediated by the native state? To answer the first question, we introduced a metric ("hub scores") that quantifies transition path mediation in different regions of protein configuration space networks[26] by assigning each region *i* a score, between 0 and 1, which is the probability of region *i* being visited on a transition path between any other two regions *j* and *k*. Using this metric, we found that for a Gō model of protein A, the native state was not a strong mediator: the hub score of the native state was 0:12, indicating that 88% of the nonnative-to-nonnative paths do not involve the native state. To address the second key question, we use "native-state knockouts" (Markov models with the native state removed), to compare mean first passage times for nonnative-to-nonnative transitions with and without the native state. A longer MFPT in the knockout would indicate the native state acts as a facilitator, which provides a short pathway between nonnative states where none existed before. A shorter MFPT would indicate that the native state acts as a trap that prevents the system from reaching the final state.

We apply this analysis to an extensive data set: the set of folding trajectories created by D. E. Shaw Research as reported in Lindorff-Larsen et al.[6] This data set comprises simulations of 12 different fast-folding proteins ranging from 10 to 80 amino acids in length, with an aggregate sampling time of over 8.2 ms. The sampling was made possible using specialized hardware (the Anton supercomputer[27]), and software (the Desmond MD package) built for molecular dynamics. The analysis performed in the original manuscript mainly regarded the order of formation of different contacts in the folding process.

We examine here the segmentation of the configuration space into regions, the properties of transitions between these regions, and the visualization of the free energy landscapes. First, the simulation data is clustered into a set of microstates for each molecule, which act as the nodes of a network, and transitions in the underlying trajectories dictate the connections between the nodes. The structure of the configuration space for each molecule is visualized using network analysis, and we show that dynamic properties can be visualized by examination of network graphs. The microstates are further grouped into a smaller number of communities using a modularity optimization algorithm. We then use hub scores to determine how often the native state is used as an intermediate on transition paths connecting two nonnative states. In addition, we use native-state knockouts to determine whether the native state is acting as a facilitator, or a trap. Together, these two analyses provide a quantitative way to measure hub-like activity of the native state.

Results

Configuration space networks can reveal dynamic quantities

Network graphs are used to visualize the folding landscapes for each of the 12 molecules (Figures 1 and 2). The detailed procedure for creating the networks is given in Methods. In short, each node represents a microstate, which is a cluster of configurations that are close in

configuration space, and the size of each node is proportional to the statistical weight of that microstate. Links are shown between microstates if the transition probability from one state to another is greater than or equal to 0:001. Microstates are further grouped into communities, which are shown by color. We denote the community of microstates with configurations close to the native structure (usually colored light-blue, but in the case of BBL and BBA, grey as well) as the "native state", and refer to the remainder of the nodes as the "unfolded ensemble".

Upon inspection of the networks, it is clear that the unfolded ensembles for each molecule are not fragmented: there are many pathways connecting different parts of the unfolded ensemble that do not involve the native state. The unfolded ensembles of Villin, PRB, Chignolin, A3D, WW and NTL9 appear particularly homogeneous: most of each unfolded ensemble has aggregated into a densely packed circle that has little noticeable substructure. The unfolded ensembles of PRG, Lambda and BBL, in contrast, appear heterogeneous, with individual communities broken off from a dense central core. It has been previously shown by Beauchamp et al.,[10] upon analyzing the relaxation spectra of the 12 trajectory sets studied here, that two-state models can accurately describe the folding of the first set of molecules (Villin, PRB, Chignolin, A3D,WWand NTL9), whereas the latter set (PRG, Lambda and BBL) displayed multistate folding behavior. It is significant that for these two sets of molecules, a dynamic property (whether or not the dynamics are "two-state") can be rationalized from the appearance of the network graph. This is because the force minimization algorithm used to produce the network graphs allows for the visual identification of community structure. Of the three remaining molecules that have mild heterogeneity in the unfolded ensemble (UVF, BBA and Trp-Cage), UVF and BBA displayed multistate behavior, and Trp-Cage displayed two-state behavior.

For some of these molecules we can see large separation of the native state from the rest of the map, indicating weak links between the native and unfolded ensembles. This can be quantified with the measure $\phi = I_{nn}/2I_{un}$, where I_{nn} is the sum of the weights of directed native-native links and I_{un} is the sum of the weights of directed unfolded-native links. The graphs with the highest ϕ values are shown in Figure 2, and have the largest separation of the native state from the nonnative states. We also find that ϕ values are correlated with unfolding times. The four highest ϕ values predict which molecules have the four longest unfolding times, in order. A plot of ϕ versus τ_u, the mean unfolding time (as computed in Lindorff-Larsen et al[6]), is given in Figure 3. If Chignolin, which has the smallest network, is removed as an outlier, the top six ϕ values correspond to the longest six unfolding times, in order. This correlation of ϕ with τ_u allows one to predict which molecule will have a longer unfolding time from simply visually inspecting the two network graphs, although we caution that this correlation is better for larger ϕ values.

Hub scores are not larger for native states

We calculate the hub scores for each molecule, and they are given in Figure 4: the solid circles mark the hub score of the native state, while the open circles show the hub scores of other communities in the network. To determine which community corresponds to the native state, we find a cutoff (r_{nc}) such that only 5% of the snapshots in the trajectory are closer to the native structure than the cutoff. The values of r_{nc} are given for each molecule in Table 1. Using 5 representative structures of each microstate, the community with the most structures within the cutoff is defined as the native state. For 10 of the 12 molecules, all of the structures are in the community marked in Figure 4. For BBA, 89% of the structures within the cutoff are in the grey community, with the remainder in the light-blue community. For BBL, 62% of the structures within the cutoff are in the light-blue community, and 34% of the structures are in the grey community; these populations are considered close enough that both communities should be considered "native".

The hub scores show that a substantial number of nonnative-to-nonnative transitions do not go through the native state. The largest hub score obtained here for any native state is 0:52 for both UVF and Chignolin. This indicates that even for the molecule with the highest hub score, almost half of the nonnative-to-nonnative transitions are not mediated by the native state. We find that both the size of the protein, and the folding time correlate poorly with the hub scores of the native state (see Supporting Information). For 10 of the 12 molecules examined, the community with the highest hub score is nonnative. More significantly, the average hub score of the native states across all molecules (0:31) is slightly lower than the average hub score of the nonnative states (0:33). We can conclude from this that the native state should not, generally, be seen as exceptional in its role as a mediator.

We also find that the hub scores depend on the precise definition of the communities as determined by the modularity optimization algorithm. To demonstrate this, we started with the predicted communities for PRG, and created three modified community groups in which a progressively larger set of nodes are removed from the native community (light blue) and added to the first nonnative community (grey) (Figure S5). We then compute the hub scores for each set of communities. As the native community gets smaller, the hub score for the native state decreases from 0:41 to 0:07. This indicates that the native community predicted by the modularity optimization algorithm is not fast-equilibrating. As such, we emphasize that hub scores are only meaningful in the context of a given community assignment, and although one can determine rules for an “optimal” assignment to communities, the specific hub scores obtained are strongly dependent on the assignment rules used. More discussion on the choice of communities is given in Supporting Information.

Removal of the native state reveals trap-like behavior

Hub scores determine what percentage of transition paths use the native state as an intermediate. However, this does not fully determine the role of the native state. Does the native state act as a “facilitator” that connects previously disconnected regions of space? Or does the native state act instead as a trap: an unnecessary detour on the way to the final destination? To distinguish between these two possibilities, we can examine the mean first passage times of transition between different nonnative communities both in the normal network, and a network where the native state has been excised. If the native state acts as a facilitator between two states, the MFPT between those states will increase when the native state is removed. Conversely, the MFPT will decrease if the native state acts as an unnecessary detour.

The MFPT are determined using the method described in Dickson and Brooks.[26] In short, to determine the MFPT to a community B, a rate matrix is constructed ($\mathbb{R}_{B0}$) where all the states in B are turned into probability sinks (the columns, or “exit” terms are set to zero for each state). This matrix is then diagonalized, and the quantity $\langle i|e^{\mathbb{R}_{B0}t}|j\rangle = \langle i|\mathbb{A}e^{\mathbb{V}t}\mathbb{A}^{-1}|j\rangle$ is the population, starting in state j that is in state i at time t, where $\mathbb{R}_{B0} = \mathbb{A}\mathbb{A}^{-1}$. The MFPT from every state i to the community B is then determined by examining the population that has reached B at different points in time. The case where the native state is removed is created simply by removing from $\mathbb{R}_{B0}$ all rows and columns corresponding to states in the native state. The diagonal terms of the matrix are then recomputed in this case as the opposite of the sum of the elements in each column, such that the sum of each column is zero.

Let $\boldsymbol{\tau}$ be the normal MFPT between two states, and let $\boldsymbol{\tau}_k$ be the MFPT with the native state removed. The quantity $(\boldsymbol{\tau}_k - \boldsymbol{\tau})/\boldsymbol{\tau}$ measures fractional deviation of a MFPT upon removal of the native state. This quantity is computed for every pair of nonnative communities, and histograms for each molecule are shown in Figure 5. For 8 of the 12 molecules, $(\boldsymbol{\tau}_k - \boldsymbol{\tau})/\boldsymbol{\tau}$ (which we will denote hereafter as Δ) is strictly less than zero, indicating that the native

state acts as a trap for every pair of nonnative states: its removal, on average, speeds up transitions between these states.

For the remaining 4 molecules, the exceptions mainly involve a single community. In BBL, the native state acts as a facilitator (Δ = 0:68) for transitions to or from community 2 (grey). Figure 1 shows that the grey region in BBL lies adjacent to the native state, and as stated before, contains 34% of the native population. It is thus reasonable that the native state would act as a facilitator to community 2. In PRG, the native state acts as a strong facilitator (Δ = 1:3) to community 9 (which comprises the nodes circled in Figure 2). This can be predicted from the graph structure: there are three connections from community 9 to the native state, and only one connection to the remaining nonnative states. In Lambda, the native state facilitates transitions to community 7 (yellow) with Δ =0:3. Although the yellow community lies adjacent to the native state in Figure 1, this would be hard to tell from appearance alone, as there are many connections from the community 7 to both the native and the remaining nonnative communities. Similarly, the native state in UVF facilitates transitions to community 3 (purple) with Δ = 0:2. Structurally, all of the native-facilitated states have collapsed structures that are similar to the native state, but have stabilizing interactions which differ from those present in the native state (see Figure S2).

We emphasize that despite these few facilitating interactions, we find that the native state primarily does not act as a facilitator for transitions between nonnative communities. Out of 3262 nonnative-to-nonnative transitions, only 119 (or 3:6%) have $\Delta > 0$. Even the largest effect of native state removal is modest: a nonnative-to-nonnative transition in PRG takes 2:3 times longer without the native state. This not only reveals that the primary role of the native state is a trap, but also speaks directly to experiments that involve thermodynamic variables, such as pH, temperature, or pressure, which destabilize the native state. For experiments where folding is initiated by an abrupt change of a thermodynamic variable, these results suggest that the unfolded ensemble is ergodic. Consequently, if the thermodynamic variable is varied periodically to drive the proteins in and out of the folded state, the refolding kinetics will not depend on the specific path along which the protein denatures.

Discussion and Conclusions

Above we determined communities using an initial RMSD clustering using Ward's algorithm, followed by an aggregation of states using modularity optimization. Modularity optimization algorithms have been shown to have a "resolution limit", in that they have trouble detecting communities that are smaller than a certain size.[28] As a result, the communities predicted here for the native state could be too large. As such the hub scores for the native community can be seen as upper bounds, as reducing the size of the communities for the native state causes the hub scores to decrease, as demonstrated in Figure S5. Similarly, larger native communities will also overemphasize the effect of the native state as a facilitator. This can be directly observed in Figures S5 and 2, where decreasing the size of the native state removes the connections from native to community 9, and thus destroys its role as a facilitator. Therefore, we expect that any failure of the modularity optimization algorithm used here will not affect the main conclusions of this paper. To remain objective, we chose to define communities without incorporating any information of the native structure. It is possible that using an RMSD cutoff from a given native structure would prevent the definition of overly large communities for the native state. However, this would introduce many adjustable parameters to the analysis, as we found that it is unlikely that a single RMSD cutoff would work for all of the molecules studied here.

The simulation temperatures used in these experiments ranged from 290 K to 370 K, and the average population fraction in the unfolded ensemble is 55%. As we have previously shown, high simulation temperatures can decrease hub scores of the native state by stabilizing higher energy but higher entropy transition paths between unfolded states.[26] It remains to be seen whether lower simulation temperatures would affect the results presented here, and undoubtedly enhanced sampling methods would be required to obtain such results. However, we note that for the four molecules with the lowest population fraction in the unfolded ensemble (NTL9, Chignolin, WW and UVF, see Table 1), we do not see higher than average hub scores for the native state, nor a decrease in the connectivity in the unfolded ensemble.

The replacement of schematics of funnels with graphs of networks like those presented here would enhance our understanding of protein dynamics. Upon visual examination of a single graph, one can develop an intuitive understanding of the observed unfolding times, as well as the partitioning of the nonnative ensemble. The graphs produced above show that although there are a large number of possible pathways along which to fold, these pathways typically overlap and interconvert extensively. However, there is also significant variation from one molecule to the other. For BBL, UVF, Lambda, and PRG, we show that this analysis is able to detect native-state facilitation, in that we find nonnative states that are more quickly accessible through the native state than through nonnative states. Therefore, we expect network analysis to continue to be a useful tool to reveal the structure of configuration space networks for other biomolecules not studied here, particularly those that are not fast-folding.

Experimental

Trajectories

The simulation details for the trajectories are described at length in the original work, and we will not reproduce them here. We obtained the Cα coordinates for trajectories of the 12 molecules from D. E. Shaw Research. Details of the data set are given in Table 1. The trajectories contain snapshots of the Cα coordinates of the molecule sampled every 200 ps. There is between 104 and 2936 μs of sampling for each molecule, and each molecule has at least 10 folding and unfolding transitions. The simulations are all run using the CHARMM22* force field.[29]

Clustering into microstates

We use MSMBuilder2[30,31] to cluster the trajectories into microstates. We first subsample the data to obtain snapshots every 50 ns, then cluster the trajectories using root mean squared distance (RMSD) and Ward's algorithm.[32,33] The number of clusters chosen for most molecules is related to the number of conformations via $n_{\text{clusters}} = n_{\text{conformations}}/10$, which is slightly larger than the number of clusters used by Beauchamp et al.[10] For NTL9, a smaller number of clusters (3117, the same used by Beauchamp et al[10]) is found to be sufficient. The number of clusters used for each molecule is given in Table 1.

Determining communities

After clustering into microstates, the resulting networks have between 213 and 3117 nodes. For analysis, we partition each network into a much smaller number of communities. We use an algorithm that works by optimizing a quantity known as the modularity, which compares connections within and between different communities.[34] The modularity for directed networks[35] is given by

$$Q=\frac{1}{2m}\sum_{i,j}\left[A_{ij}-\frac{k_i^{\text{out}}k_i^{\text{in}}}{2m}\right]\delta(c_i.c_j). \quad (1)$$

A_{ij} is the number of links from i to j, k_i^{out} is the total number of links from i, k_i^{in} is the total number of links into j, and $2m$ is the total number of links in the network. c_i denotes the community to which node i is assigned, and $\delta(i, j)$ returns 1 if $i = j$ and 0 otherwise.

Lancichinetti and Fortunato[36] recently compared several community detection algorithms using a common benchmark. Here we use one of the most accurate and efficient algorithms tested therein: the modularity optimization algorithm of Blondel et al.,[37] which begins with each node in its own community and merges communities together that maximally increase the modularity. The algorithm has been applied to systems of over 100 million nodes,[37] and defines communities in the largest network used here (3117 nodes) almost instantaneously. Determining communities using modularity optimization algorithms has the advantage that one does not need to determine how many communities there are in the system beforehand. The number of communities determined for each molecule range from 8 to 31, and are given in Table 1.

Creating network graphs

We build the network graphs in Figures 1 and 2 using the program Gephi.[38] The size of the nodes are proportional to their statistical populations, however for each graph there is a minimum node size that is 30 times smaller than the size of the largest node. The orientation of the nodes is obtained using a force minimization algorithm built into Gephi (ForceAtlas), which introduces a repulsive force between all nodes, but attracts nodes that are linked together with a force that is proportional to the weights of the links. The weights of the links are determined as follows. First, weights of directed links with values between 1 and 1000 are determined as $w_{ij} = 1000p_{ij}$, where p_{ij} is the transition probability from i to j. Weights of undirected links are then determined as the average of the two directed links. The graph is first allowed to minimize without adjusting for node sizes (i.e. with overlapping nodes), and then a second minimization is subsequently performed with adjusting for node sizes to prevent overlap.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to D. E. Shaw Research for sharing their protein folding trajectories with us. We also acknowledge support from the Center for Multi-Scale modeling tools for structural biology (MMTSB), funded by the NIH (RR012255) and the Center for Theoretical Biological Physics (CTBP) funded by the NSF (PHY0216576).

References

1. Baldwin RL. J. Biomol. NMR. 1995; 5:103–109. [PubMed: 7703696]
2. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Proteins. 1995; 21:167–195. [PubMed: 7784423]
3. Dill KA, Chan HS. Nat. Struct. Biol. 1997; 4:10–19. [PubMed: 8989315]
4. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Annu. Rev. Phys. Chem. 1997; 48:545–600. [PubMed: 9348663]
5. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan YB, Wriggers W. Science. 2010; 330:341–346. [PubMed: 20947758]

6. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. Science. 2011; 334:517–520. [PubMed: 22034434]
7. Zhang B, Miller TFI. J. Am. Chem. Soci. 2012; 134:13700–13707.
8. Shirts M, Pande VS. Science. 2000; 290:1903–1904. [PubMed: 17742054]
9. Bowman GR, Voelz VA, Pande VS. J. Am. Chem. Soci. 2011; 113:664–667.
10. Beauchamp KA, McGibbon R, Lin YS, Pande VS. Proc. Natl. Acad. Sci. USA. 2012; 109:17807–17813. [PubMed: 22778442]
11. Voelz VA, Jäger M, Yao S, Zhu L, Waldauer SA, Bowman GR, Friedrichs M, Bakalin O, Lapidus LJ, Shimon W, Pande VS. J. Am. Chem. Soci. 2012; 134:12565–12577.
12. Eastman P, Pande VS. J. Comp. Chem. 2009; 31:1268–1272. [PubMed: 19847780]
13. Götz AW, Williamson MJ, Xu D, Poole D, Le Grand S, Walker RC. J. Chem. Theory Comput. 2012; 8:1542–1555. [PubMed: 22582031]
14. Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR. Proc. Natl. Acad. Sci. USA. 2009; 106:19011–19016. [PubMed: 19887634]
15. Adelman JL, Dale AL, Zwier MC, Bhatt D, Chong LT, Zuckerman DM, Grabe M. Biophys. J. 2011; 101:2399–2407. [PubMed: 22098738]
16. Dickson A, Maienschein-Cline M, Tovo-Dwyer A, Hammond JR, Dinner AR. J. Chem. Theory Comput. 2011; 7:2710–2720.
17. Liu Y, Strümpfer J, Freddolino PL, Gruebele M. J. Phys. Chem. Lett. 2012; 3:1117–1123. [PubMed: 22737279]
18. Pierce LCT, Salomon-Ferrer R, de Oliveira CA, McCammon JA, Walker RC. J. Chem. Theory Comput. 2012; 8:2997–3002. [PubMed: 22984356]
19. Vashisth H, Maragliano L, Abrams CF. Biophys. J. 2012; 102:1979–1987. [PubMed: 22768955]
20. Duan Y, Kollman PA. Science. 1998; 282:740–744. [PubMed: 9784131]
21. Rao F, Caflisch A. J. Mol. Biol. 2004; 342:299–306. [PubMed: 15313625]
22. Muff S, Caflisch A. Proteins. 2008; 70:1185–1195. [PubMed: 17847092]
23. Beauchamp KA, Ensign DL, Das R, Pande VS. Proc. Natl. Acad. Sci. USA. 2011; 108:12734–12739. [PubMed: 21768345]
24. Giambasu GM, Lee T, Scott WG, York DM. J. Mol. Biol. 2012; 423:106–122. [PubMed: 22771572]
25. Bowman GR, Pande VS. Proc. Natl. Acad. Sci. USA. 2009; 107:10890–10895. [PubMed: 20534497]
26. Dickson A, Brooks CL III. J. Chem. Theory Comput. 2012; 8:3044–3052.
27. Shaw DE, et al. Commun. ACM. 2008; 51:91–97.
28. Fortunato S, Barthélemy M. Proc. Natl. Acad. Sci. USA. 2007; 104:36–41. [PubMed: 17190818]
29. Piana S, Lindorff-Larsen K, Shaw DE. Biophys. J. 2011; 100:L47–L49. [PubMed: 21539772]
30. Beauchamp KA, Bowman GR, Lane TJ, Maibaum L, Haque IS, Pande VS. J. Chem. Theory Comput. 2011; 7:3412–3419. [PubMed: 22125474]
31. Bowman GR, Huang X, Pande VS. Methods. 2009; 49:197–201. [PubMed: 19410002]
32. Ward J Jr. J. Am. Stat. Assoc. 1963; 58:236–244.
33. Müllner D. arXiv:1109.2378v1 [stat.ML]. 2011
34. Newman MEJ. Phys. Rev. E. 2004; 70:056131.
35. Arenas A, Duch J, Fernández A, Gómez S. New J. Phys. 2007:9.
36. Lancichinetti A, Fortunato S. Phys. Rev. E. 2009; 80:056117.
37. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E. J. Stat. Mech. 2008:P10008.
38. Bastian, M.; Heymann, S.; Jacomy, M Gephi. An Open Source Software for Exploring and Manipulating Networks. In: Hamilton, M., editor. International AAAI Conference on Weblogs and Social Media; May 17–20, 2009; California, USA. California: AAAI Press; 2009.

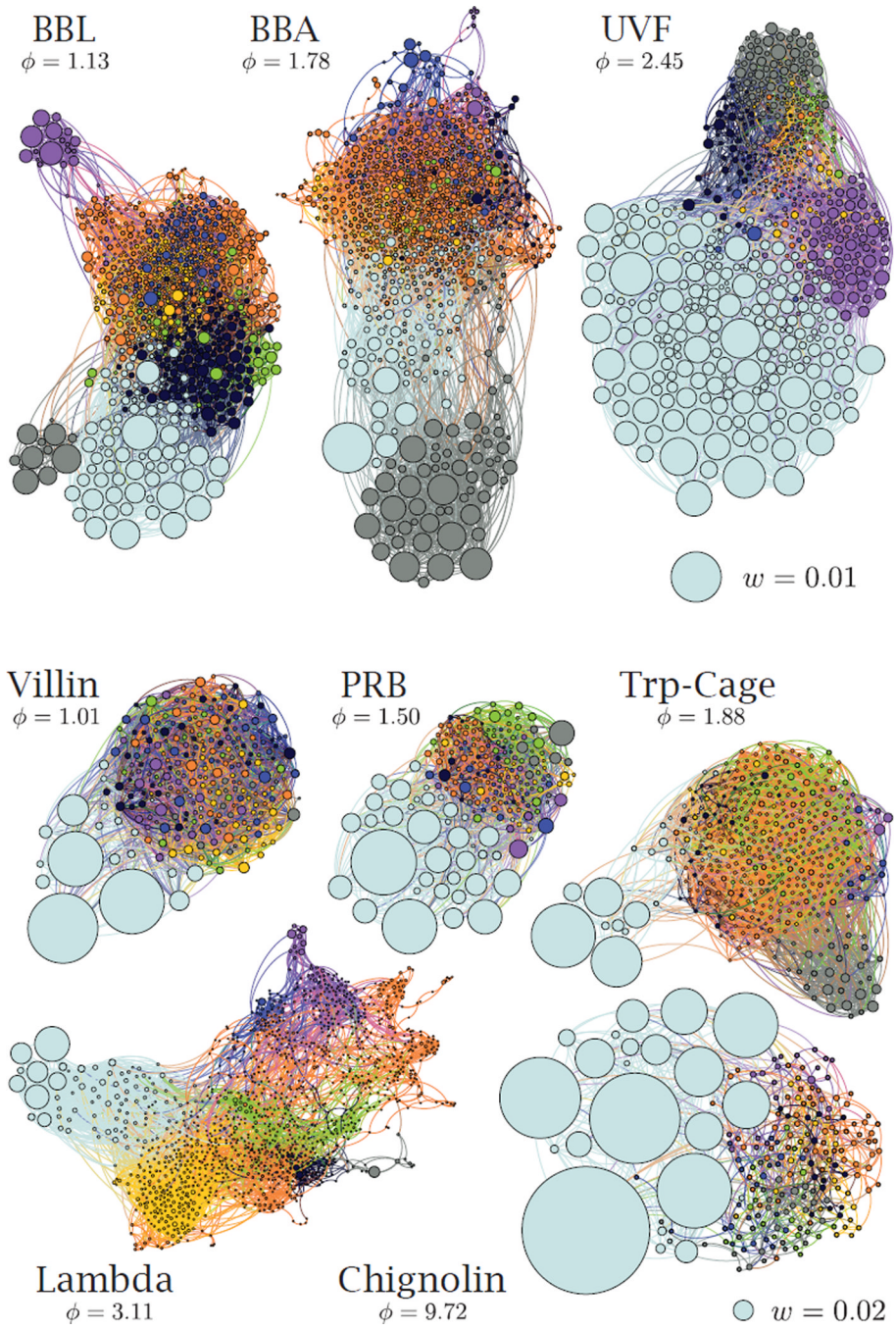

Figure 1.

Graphs of the eight molecules with the lowest ϕ values. The graphs are sized appropriately so that node sizes (statistical weights) can be compared across graphs. A reference node with a specified weight is given for each. In each figure the molecules are arranged in order of increasing ϕ. The nodes are assigned a color according to their community, and for each molecule the communities are sorted according to the weight of their highest-weighted node, with the highest weight given the lowest index. The colors are as follows: 1 - light blue, 2 - grey, 3 - purple, 4 - blue, 5 - dark blue, 6 - green, 7 - yellow, 8 - orange. Any community with an index greater than 8 is also assigned the color orange.

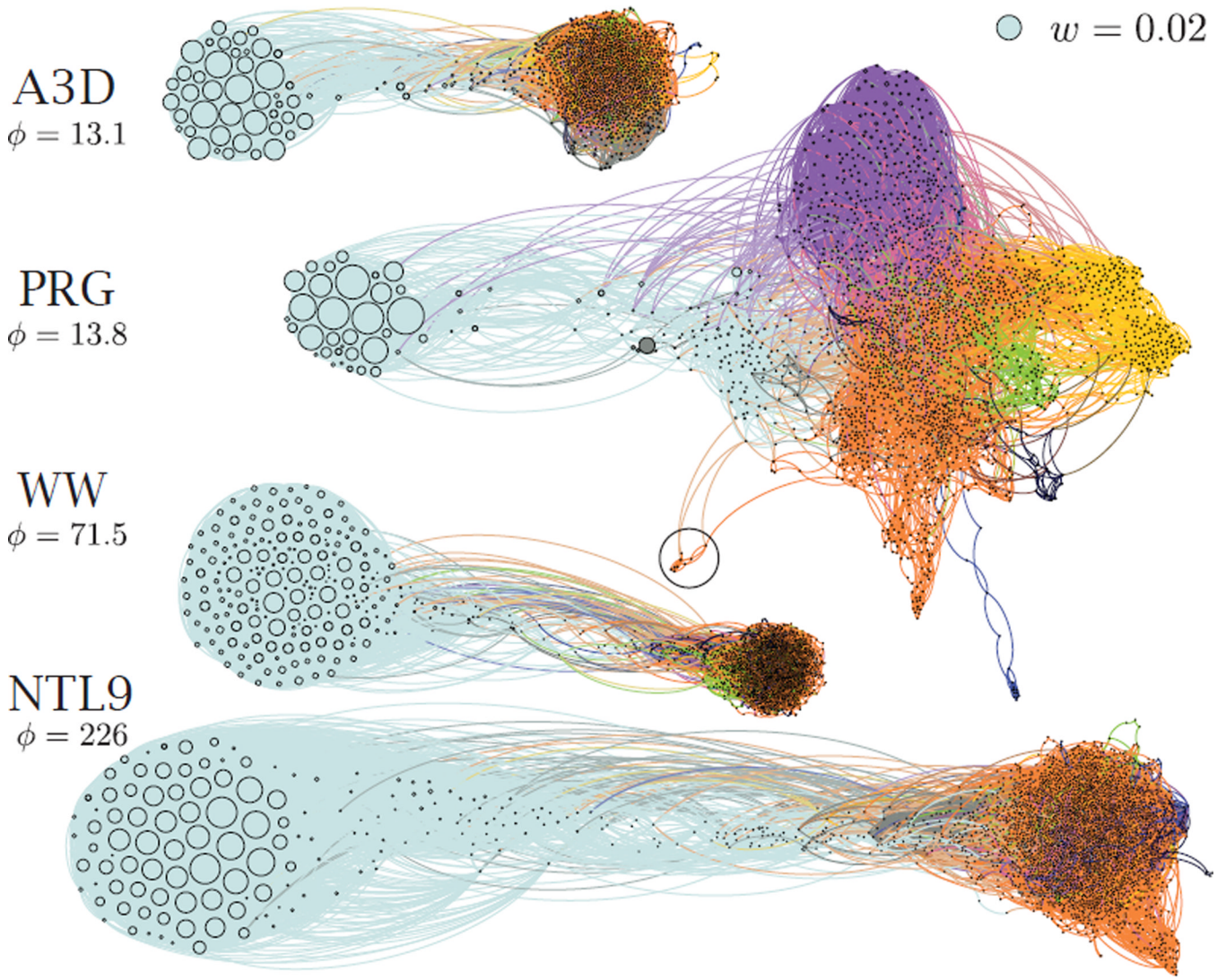

Figure 2.
Graphs of the four molecules with the highest ϕ values. The graphs are sized appropriately so that node sizes (statistical weights) can be compared across graphs. The molecules are arranged in order of increasing ϕ. The circled nodes in PRG are the members of community 9; transitions to and from this region are largely mediated by the native state. The colors of the nodes are assigned as in Figure 1.

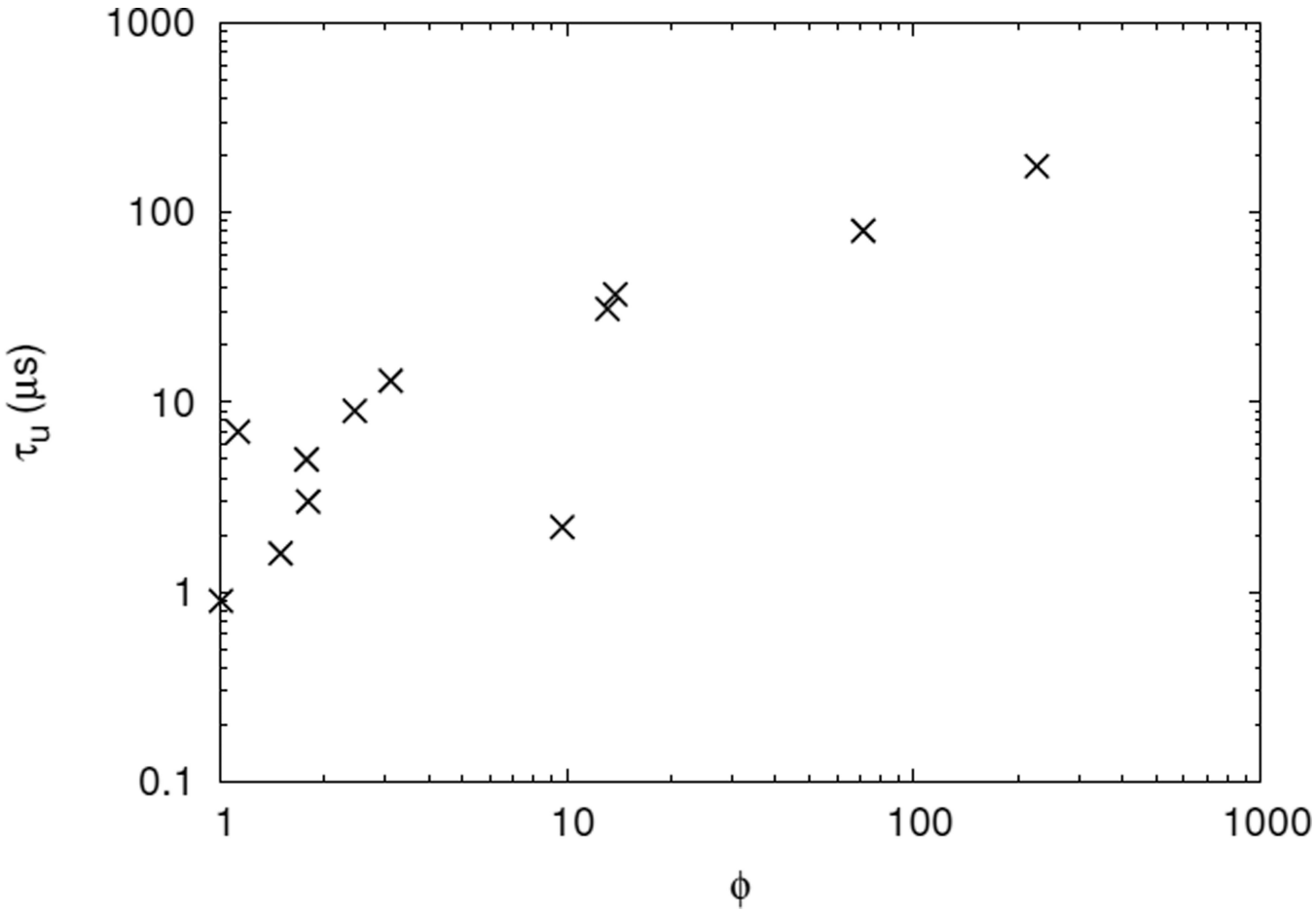

Figure 3.
The measure ϕ, which compares native-native links versus native-nonnative links, is plotted against the mean unfolding time (τ_u) as computed in Lindorff-Larsen et al,[6] and given in Table 1.

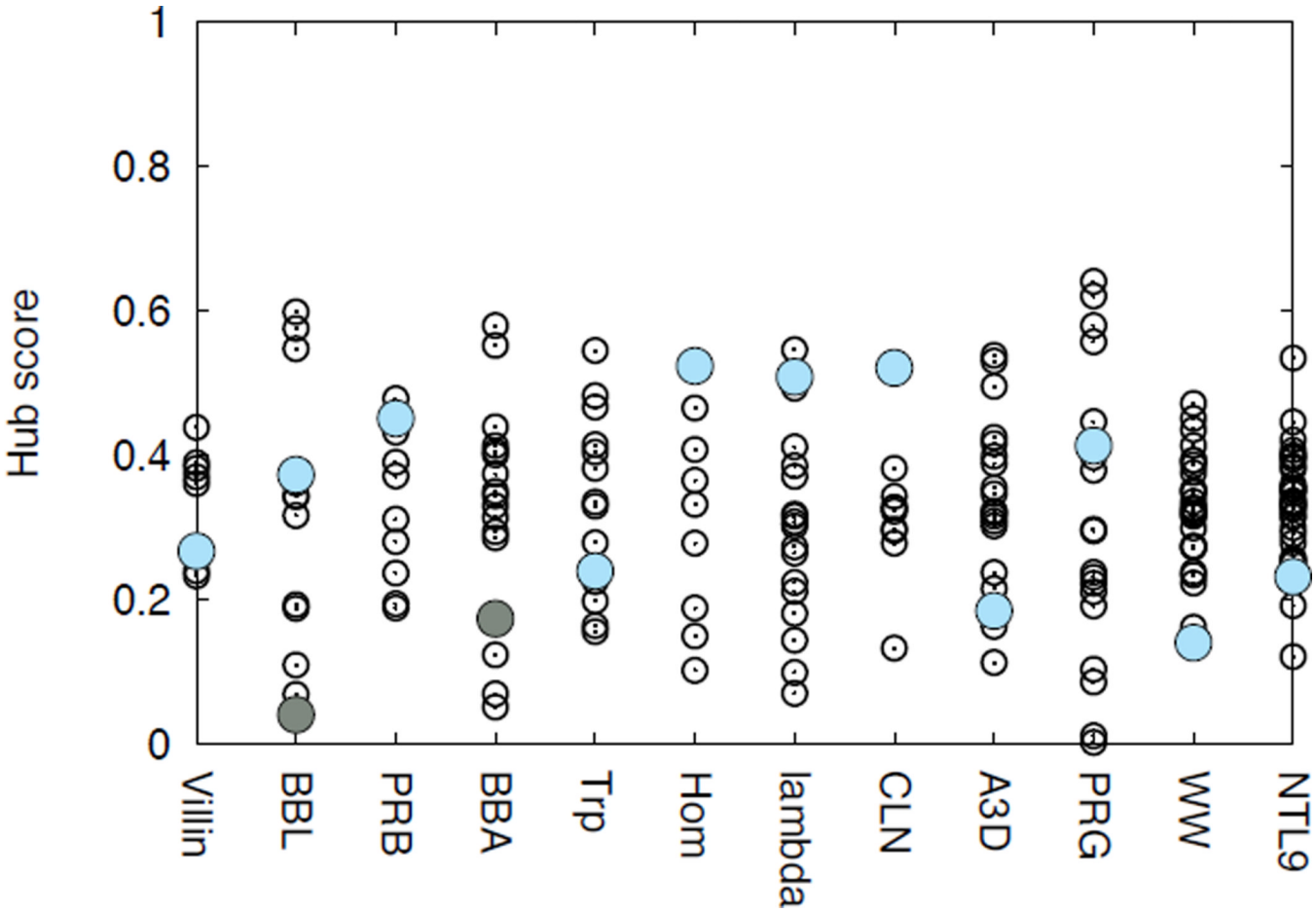

Figure 4.

Hub scores of all 12 molecules. The larger, solid circles show the hub scores of native states, and are colored (either light blue or grey) according to which community is native in Figures 1 and 2. The native states are determined as described in the text.

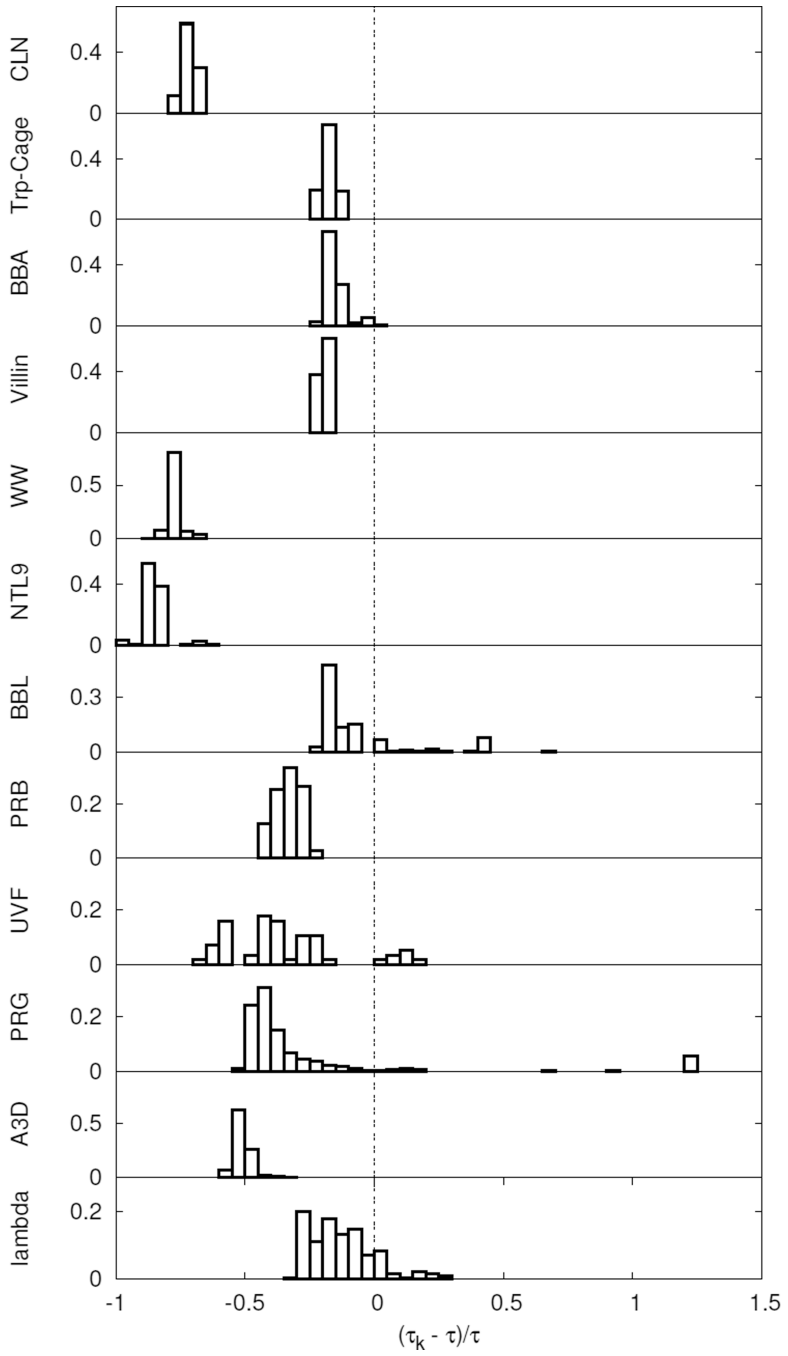

Figure 5.

Histograms of the quantity $(\tau_k - \tau)/\tau$, also referred to as Δ in the text, which measures the impact of removing the native state on the mean first passage time between two nonnative communities. The quantity is measured for transitions between all possible pairs of nonnative states. For a particular transition between two nonnative states, $(\tau_k - \tau)/\tau < 0$ indicates that the native state acts as a trap: its removal results in a faster transition, on average. Conversely, $(\tau_k - \tau)/\tau > 0$ indicates that the native state acts as a facilitator: its removal results in slower transitions between the two states.

Table 1

Molecule-specific properties and parameters. N_{res} is the number of amino acid residues; N_{traj} is the number of trajectories used for analysis, with combined duration t_{tot}; T is the simulation temperature; P_u is the fraction of the population in the nonnative ensemble, computed as $\tau_f/(\tau_u + \tau_f)$, where τ_u and τ_f are the mean unfolding and folding times, taken from Lindorff-Larsen et al;[6] $n_{clusters}$ is the number of microstates constructed using hierarchical clustering; n_{com} is the number of communities determined by the modularity optimization algorithm. The native structure used for each molecule is given in Native PDB, where the numbers in parentheses, when present, denote the subset of residues used. Configurations that are closer than r_{nc} to the native structure are considered "native" and used to determine which community corresponds to the native state.

	N_{res}	N_{traj}	t_{tot} (μs)	T (K)	P_u	τ_u(μs)	$n_{clusters}$	n_{com}	r_{nc}(Å)	Native PDB
Chignolin (CLN)	10	1	106	340	0.21	2.2	213	11	0:55	1UAO
Trp-Cage	20	1	208	290	0.82	3	417	14	1:27	2JOF
BBA	28	2	325	325	0.78	5	650	17	2:35	1FME
Villin	35	1	125	360	0.76	0.9	251	8	1:13	2F4K
WW	35	2	1137	360	0.21	80	2274	26	1:21	2F21 (4-39)
NTL9	39	4	2936	355	0.14	175	3117	31	0:53	2HBA (1-39)
BBL	47	2	429	298	0.81	7	860	15	4:69	2WXC
Protein B (PRB)	47	1	104	340	0.71	1.6	208	10	3:31	1PRB (7-53)
Homeodomain (UVF)	52	2	327	360	0.26	9	654	9	3:37	2P6J
Protein G (PRG)	56	4	1155	350	0.60	37	2312	19	1:15	1MI0 (10-65)
α3D (A3D)	73	2	707	370	0.47	31	1414	20	2:77	2A3D
λ-repressor	80	4	643	350	0.79	13	1294	19	2:43	1LMB (6-85)