Published in final edited form as:
J Am Chem Soc. 2010 February 10; 132(5): 1526–1528. doi:10.1021/ja9090353.

Molecular simulation of *ab initio* protein folding for a millisecond folder NTL9(1–39)

Vincent A. Voelz[†], **Gregory R. Bowman**[‡], **Kyle Beauchamp**[‡], and **Vijay S. Pande**[†,‡,§]

[†]Department of Chemistry, Stanford University, Stanford, CA 94305

[§]Department of Structural Biology, Stanford University, Stanford, CA 94305

[‡]Biophysics Program, Stanford University, Stanford, CA 94305

A complete understanding of how proteins fold, i.e. self-assemble to their biologically relevant "native state," remains an unattained goal[1]. Computer simulation, validated by experiment, is a natural means to elucidate this. There is over a million-fold range in folding rates, suggesting a possible diversity in mechanisms between slow and fast folding proteins[2]. Very fast (microsecond timescale) folding proteins[3,4] appear to fold via a large number of heterogeneous, parallel paths[5–7], potentially key for folding on such fast timescales. Does the folding of much slower proteins change this picture?

To date, the slowest-folding proteins folded *ab initio* by all-atom molecular dynamics simulations with fidelity to experimental kinetics have had folding times in the range of nanoseconds to microseconds. These include the designed mini-protein Trp-cage (~4.1 μs)[8], the villin headpiece domain (~10 μs)[9], a fast-folding variant of villin (<1 μs)[7], and Fip35 WW domain (~13 μs)[10]. In this communication, we report simulations of several folding trajectories, each from fully unfolded states, of the 39-residue protein NTL9(1–39), which experimentally has a folding time of ~1.5 milliseconds[11].

MD simulation

Trajectories were simulated via the Folding@Home distributed computing platform[12] at 300K, 330K, 370K and 450K from native, extended, and random-coil configurations using an accelerated version of GROMACS written for GPU processors[13], for an aggregate time of 1.52 ms. GPUs play a key role here, allowing for dramatically longer trajectories than previously possible. The AMBER ff96 forcefield[14] with the GBSA solvation model[15] was used, a combination previously shown to give good results folding Fip35 WW domain[10], and shown to exhibit a good balance of native-like secondary structure for a set of small helical and beta sheet peptides studied by replica exchange[17].

Prediction of native structure and folding rates

We find that the native state (taken from the N-terminal domain of the crystal structure of ribosomal protein L9[18]) is stable in this forcefield at 300K, exhibiting decreasing stability with increasing temperature (Figure 1a). RMSD-C_α distributions after 10 μs show well-defined native and collapsed unfolded basins near 3Å and 5Å, respectively. Of the ~3000 trajectories started from unfolded (extended and coil) states at 370K (Figure 1b), two reach an RMSD-

pande@stanford.edu.

Supporting Information Available. Detailed description of simulation methods, results, analysis, and Supporting Figures S1–S7 is available free of charge via the Internet at http://pubs.acs.org.

$C_\alpha < 3.5$Å and eight reach an RMSD-$C_\alpha < 4$Å. No productive folding trajectories were observed at lower temperatures, consistent with the enhanced forward folding rate expected by Arrhenius kinetics. Higher temperature trajectories (450K) exceed the melting temperature of NTL9 in the forcefield.

The observed number of folding events *n* is consistent with expectations from a simple model of parallel uncoupled folding simulations[19] in which folding is modeled as a two-state Poisson process: $\langle n \rangle = \int M(t) k \exp(-M(t)kt) dt$, where $M(t)$ is the number of simulations that reach time *t* (Figure 1b) and *k* is the experimental folding rate (~640/sec)[11]. This theory predicts (on average) ~1.8 folding trajectories for the amount of sampling performed, in agreement with the two folding trajectories found in practice. Posterior distributions of folding rates given the amount of simulation time and number of folding trajectories were computed using a Bayesian approach[16], which yield expectation values within an order of magnitude of the experimental folding rate.

In addition to native-like conformations, we see near-native configurations, which show heterogeneity in hydrophobic packing, most notably in alternative side chain arrangements in the beta-sheet structure (Figure 2). Most common of these is a non-native hydrophobic core involving residues I4, I18 and I37 (which normally contact the C-terminal helix in the full-length protein) with F5 solvent-exposed.

Insight into folding mechanisms

In order to describe the kinetics and mechanistic aspects of folding, we employ a new paradigm for sampling the global free energy landscape of folding, using Markov State Models (MSMs). MSM approaches, by automatically identifying a set of kinetically metastable states (such as foldons[20]) and efficiently sampling transitions between these states, can model long-timescale kinetics from much shorter trajectories[21–24].

Our strategy for simulating slow-folding proteins is first to generate an initial series of kinetically connected states from both the folding and unfolding directions, and then to use adaptive resampling techniques[25] to produce statistically converged estimates of metastable basins and the transition rates between them. In the remainder of this communication, we report progress toward the first goal, by constructing an MSM from the entire set of 370K trajectory data[26,27], which we will use to seed future rounds of transition sampling. While additional rounds of adaptive sampling could likely aid in increasing the quantitative power of this model, there are several notable observations which can be made with the current data set.

Key to accurately identifying metastable states is the clustering of trajectory conformations into *microstates* fine-grained enough to be used for lumping into groups of maximally metastable *macrostates*[26]. 100,000 microstate clusters were calculated using an approximate *k*-centers algorithm[28], each with an average radius of 4.5Å RMSD-backbone. Lag times ranging from 1 to 32 ns were used to build a series of MSMs. The implied time scales predicted by these models (obtained by diagonalizing the rate matrix) show a clear spectral gap separating the slowest relaxation time scale from the rest, indicative of single-exponential kinetics (see Figure S1). The implied time scale of the model levels off beyond a lag time of ~10 ns to an implied time scale of ~1 ms, close to the experimental folding time.

An important strength of MSMs is their ability to gain insight at coarser scales by “lumping” the kinetic transitions into a simpler model with fewer states. To gain a mesoscopic view of the folding free energy landscape, we lumped our 100,000- microstate MSM into a 2000-macrostate model. In this view, we find that the metastable states are diffuse collections of conformations over which multiple possible folding pathways can occur, indicating a vast heterogeneity of folding substates that need to be understood in greater detail. At the same

time, we can identify highly populated "native" (state *n*) and "unfolded" (state *a*) macrostates that dominate the observed relaxation rates (Figure 3 and Figure S2).

The ten pathways with the highest folding flux from macrostate *a* to *n* were calculated by a greedy backtracking algorithm (see SI) from the macrostate transition matrix using transition path theory[29,30] (TPT). The diversity of pathways demonstrates the power of the MSM approach: although we observe only a few folding trajectories directly, a network of many possible pathways can be inferred from the overlapping sampling of local transitions.

While NTL9(1–39) folds quickly for a two-state folder, it is similar in size to many ultrafast (sub-millisecond) folders that appear to exhibit so-called "downhill" folding. Hence, we would like to understand the structural features that limit the overall folding rate. As in a macroscopic two-state model, the highest-flux pathways in our mesoscopic model are $a \rightarrow m \rightarrow n$ and $a \rightarrow l \rightarrow n$ direct routes from disordered to structured macrostates, reminiscent of nucleation-condensation. These pathways by themselves, however, account for only ~10% of the total flux, and the structural diversity seen in all pathways is reminiscent of more hierarchical folding models such as diffusion-collision. Thus, we sought to more fully study the 15 macrostates transited by the top ten folding pathways.

To examine structural changes along the folding reaction, we considered three main native structural elements: the central helix (α), the pairing of strands 1 and 2 (β_{12}), and the pairing of strands 1 and 3 (β_{13}). To quantify the extent of native-like structuring for each of these elements we calculated Q_α, $Q_{\beta 12}$ and $Q_{\beta 13}$, respectively (see SI for details). The *Q*-value is a number between 0 and 1 that quantifies the extent of native-like contacts. We then examined, for each macrostate, the *Q*-values in relation to the p_{fold} value (committor), a kinetic reaction coordinate. The p_{fold} value is computed from the macrostate transition matrix[24,29,30].

This analysis yields several key insights into the folding mechanism of NTL9(1–39) on the mesoscale. We find the "unfolded" state *a* is compact, and contains a baseline level of residual native-like structure, with Q_α near 0.5, and $Q_{\beta 12}$ and $Q_{\beta 13}$ near 0.2. In general, across the fifteen macrostates studied, *Q*-values increase as p_{fold} values increase, although the relative balance of Q_α, $Q_{\beta 12}$ and $Q_{\beta 13}$ varies, indicating pathway heterogeneity: i.e. native-like structures can form in different orders (Figures S4–S6). An exception to this, however, is observed for β_{12} strand pairing. Only for macrostates with $p_{\text{fold}} > 0.5$ (states *g*-*n*) does appreciable β_{12} strand pairing occur (Figure 4). This suggests that the formation of a *local* strand pair (β_{12}), rather than a *nonlocal* strand pair (β_{13}), is rate-limiting. This effect is not predicted by strictly topological models of folding in which loop closure entropy loss dominates[31], but instead may result from sequence-specific details. Unlike the β_{13} strand pair, which has a small interaction surface stabilized by hydrophobic contacts, the β_{12} hairpin contains seven of the protein's eight lysine residues, and three of its five glycine residues in a flexible loop region, features which may imbue β_{12} with larger barriers to folding. This proposed role of β_{12} is also consistent with the large changes in kinetics and stability seen experimentally for mutations in the β_{12} hairpin[11].

It is natural to compare our results with previous unfolding simulations of NTL9(1–39) K12M by Snow et al.[32]. In that work, a detailed characterization of the transition state ensemble required the definition of strand-pairing reaction coordinates corresponding to β_{12} and β_{13} formation. In our MSM analysis, no such pre-definition is required. Snow et al. also note the difficulty in resolving kinetic intermediates not captured by the chosen order parameters. Indeed, our structural analysis can resolve subtle kinetic intermediates within the native basin, corresponding to alternative rearrangements of the β_{12} hairpin loop (Figure S7).

Conclusion

The above results suggest that existing forcefield models using implicit solvent are indeed accurate enough to fold proteins *ab initio* at long time scales (milliseconds), opening the door to simulating more structurally complex proteins. Moreover, our work demonstrates that there need not be a single pathway or single, dominant mechanism for the folding of a given protein: since the theories proposed for how proteins fold are based on broadly relevant physical principles, it is natural to imagine that multiple mechanisms could be *simultaneously* present, but that the sequence of the protein, coupled with the chemical environment would control the balance to which each mechanistic pathway is seen.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the NSF for support through FIBR grant EF-0623664, the NIH through R01-GM062868 and Simbios U54-GM072970, and NSF award CNS-0619926 for computer resources.

References

(1). Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. Current Opinion in Structural Biology 2007;17:342–346. [PubMed: 17572080]

(2). Plaxco KW, Simons KT, Baker D. Journal of Molecular Biology 1998;277:985–994. [PubMed: 9545386]

(3). Yang WY, Gruebele M. Nature 2003;423:193–197. [PubMed: 12736690]

(4). Kubelka J, Chiu TK, Davies DR, Eaton WA, Hofrichter J. Journal of Molecular Biology 2006;359:546–553. [PubMed: 16643946]

(5). Kubelka J, Hofrichter J, Eaton WA. Current Opinion in Structural Biology 2004;14:76–88. [PubMed: 15102453]

(6). Udgaonkar JB. Annual Review of Biophysics 2008;37:489–510.

(7). Ensign DL, Kasson PM, Pande VS. Journal of Molecular Biology 2007;374:806–816. [PubMed: 17950314]

(8). Pitera JW, Swope W. PNAS 2003;100:7587–7592. [PubMed: 12808142]

(9). Zagrovic B, Snow CD, Shirts MR, Pande VS. Journal of Molecular Biology 2002;323:927–937. [PubMed: 12417204]

(10). Ensign DL, Pande VS. Biophysical Journal 2009;96:L53–L55. [PubMed: 19383445]

(11). Horng J-C, Moroz V, Raleigh DP. Journal of Molecular Biology 2003;326:1261–1270. [PubMed: 12589767]

(12). Shirts M, Pande V. Science 2000;290:1903–1904. [PubMed: 17742054]

(13). Friedrichs MS, Eastman P, Vaidyanathan V, Houston M, Legrand S, Beberg AL, Ensign DL, Bruns CM, Pande VS. Journal of Computational Chemistry 2009;30:864–872. [PubMed: 19191337]

(14). Wang J, Cieplak P, Kollman PA. Journal of Computational Chemistry 2000;21:1049–1074.

(15). Onufriev A, Bashford D, Case D. Proteins 2004;55:383–394. [PubMed: 15048829]

(16). Ensign DL, Pande VS. Journal of Physical Chemistry B 2009;113:12410–12423.

(17). Shell MS, Ritterson R, A. K. Journal of Physical Chemistry B 2008;112:6878–6886.

(18). Hoffman DW, Davies C, Gerchman SE, Kycia JH, Porter SJ, White SW, Ramakrishnan V. The EMBO Journal 1994;13:205–212. [PubMed: 8306963]

(19). Shirts MR, Pande VS. Physical Review Letters 2001;86:4983–4987. [PubMed: 11384401]

(20). Panchenko AR, Luthey-Schulten Z, Wolynes PG. PNAS 1996;93:2008–2013. [PubMed: 8700876]

(21). Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC. Journal of Chemical Physics 2007;126:155101. [PubMed: 17461665]

(22). Noé F, Fischer S. Current Opinion in Structural Biology 2008;18:154–162. [PubMed: 18378442]
(23). Chodera JD, Swope WC, Pitera JW, Dill KA. Multiscale Modeling and Simulation 2006;5:1214–1226.
(24). Singhal N, Snow CD, Pande VS. Journal of Chemical Physics 2004;121:415–425. [PubMed: 15260562]
(25). Huang X, Bowman GR, Bacallado S, Pande VS. PNAS 2009;106:19765–19769. [PubMed: 19805023]
(26). Bowman GR, Huang X, Pande VS. Methods 2009;49:197–201. [PubMed: 19410002]
(27). Bowman GR, Beauchamp KA, Boxer G, Pande VS. Journal of Chemical Physics 2009;131:124101. [PubMed: 19791846]
(28). Dasgupta S, Long PM. J. Comput. Syst. Sci 2005;70:555–569.
(29). Metzner P, Schütte C, Vanden-Eijnden E. Multiscale Modeling and Simulation 2009;7:1192–1219.
(30). Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR. PNAS 2009;106:19011–19016. [PubMed: 19887634]
(31). Weikl TR. Archives of Biochemistry and Biophysics 2008;469:67–75. [PubMed: 17662688]
(32). Snow CD, Rhee YM, Pande VS. Biophysical Journal 2006;91:14–24. [PubMed: 16617068]

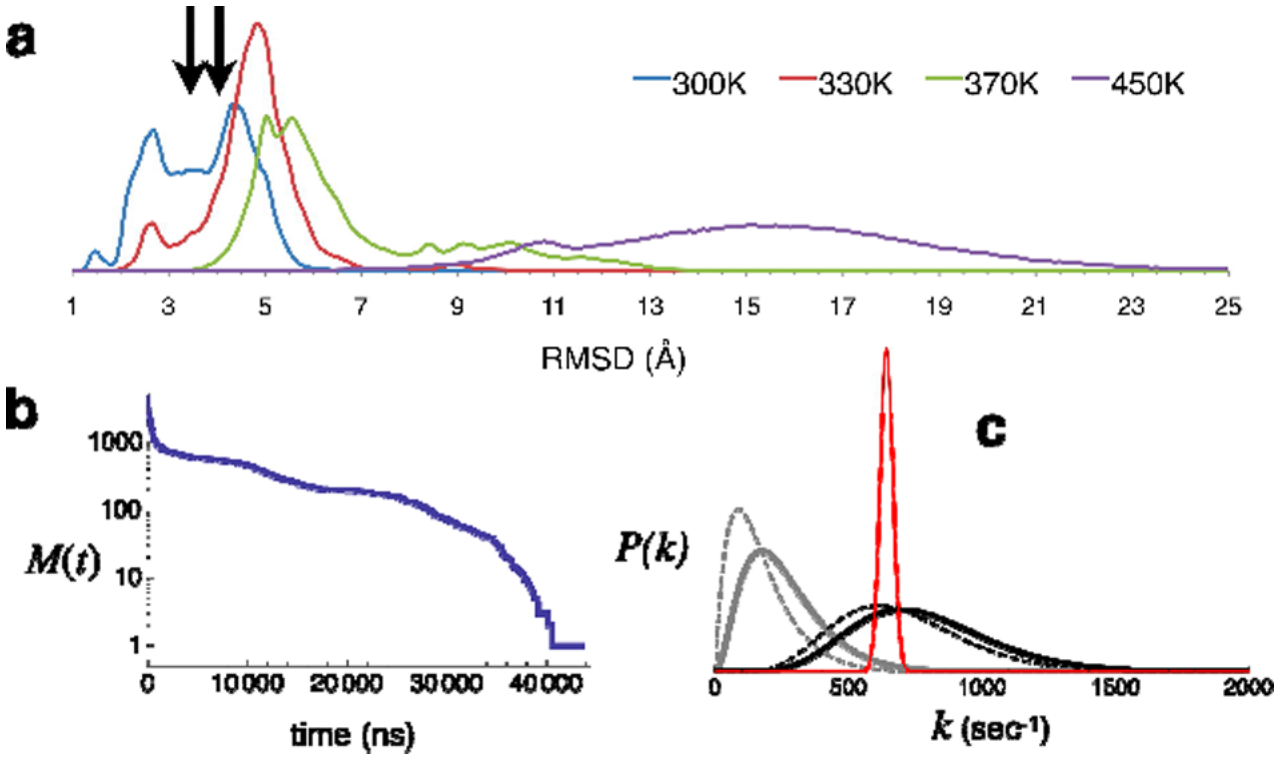

Figure 1.

(a) Distributions of RMSD-C_α for native-state simulations of NTL9(1–39) after 10 μs. The arrows indicate thresholds defined for the native basin at 3.5Å and 4Å. (b) The number of parallel simulations $M(t)$ started from unfolded states at 370K that reach time t. (c) Posterior predictions of the folding rate given the amount of simulation time and observed folding events for 3.5Å (dashed) and 4Å (solid) thresholds, using uniform (black) and Jeffrey's (gray) priors, using methods from[16]. In red is a Gaussian distribution representing the experimental rate mean and standard deviation.

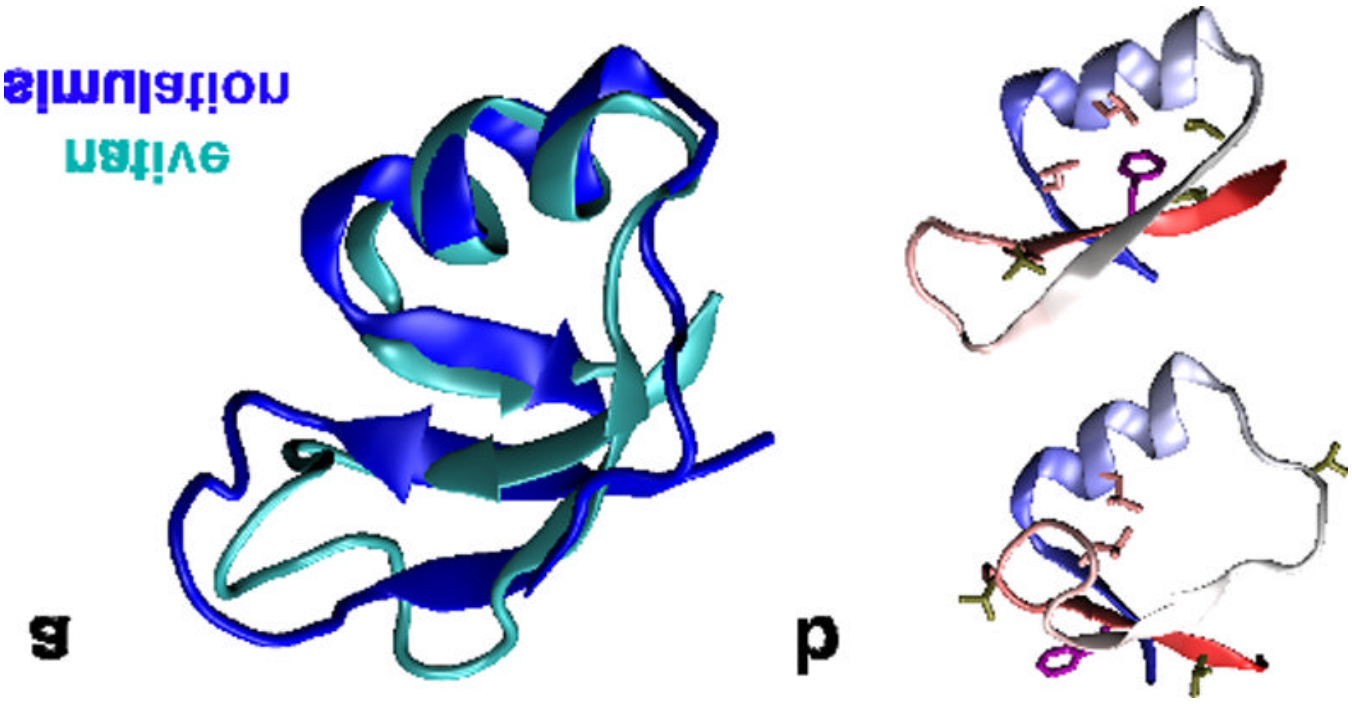

Figure 2.
(a) A snapshot from a folding trajectory (dark blue) achieves an RMSD-C_α of 3.1Å compared to the native state (cyan). (b) Non-native (top) and native-like (bottom) hydrophobic core arrangements observed in low-RMSD conformations of folding trajectories. Highlighted are sidechains of residues F5 (magenta), V3,V9,V21 (tan), and L30,L35 (pink).

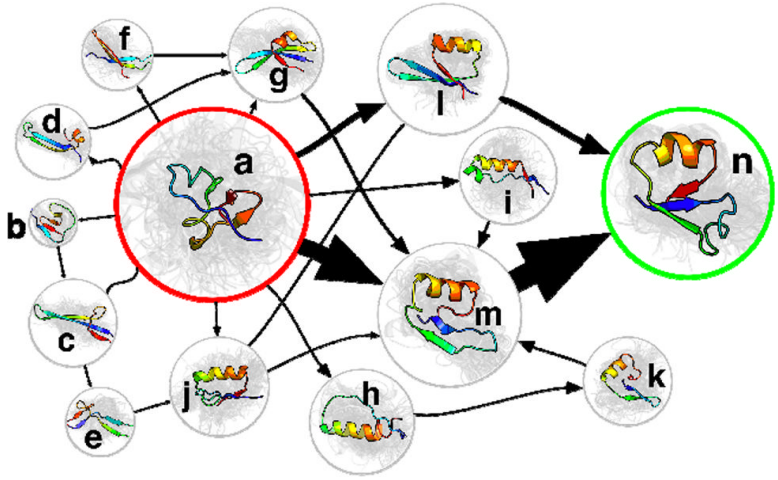

Figure 3.

A 2000-state Markov State Model (MSM) was built using a lag time of 12 ns. Shown is the superposition of the top 10 folding fluxes, calculated by a greedy backtracking algorithm (see Supporting Information). These pathways account for only about 25% of the total flux, and transit only 15 of the 2000 macrostates (shown labeled *a*-*n*, for convenient discussion). The visual size of each state is proportional to its free energy, and arrow size is proportional to the inter-state flux.

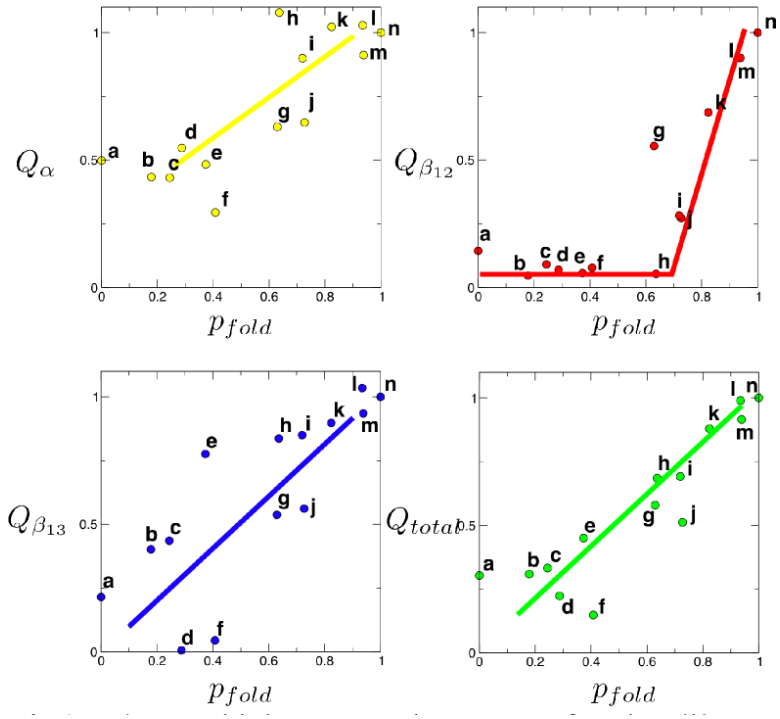

Figure 4.

Q-values, which capture the extent of native-like structures, plotted versus p_{fold} (committor) values. The lines are to guide to eye.