

Discrete Kinetic Models from Funneled Energy Landscape Simulations

Nicholas P. Schafer^{2,3,6}, Ryan M. B. Hoffman^{1,2,3}, Anat Burger^{2,3,4}, Patricio O. Craig^{1,2,3}, Elizabeth A. Komives¹, Peter G. Wolynes^{2,3,5*}

1 Chemistry/Biochemistry, University of California San Diego, La Jolla, California, United States of America, **2** Center for Theoretical Biological Physics, University of California San Diego, La Jolla, California, United States of America, **3** Rice University, Houston, Texas, United States of America, **4** Physics, University of California San Diego, La Jolla, California, United States of America, **5** Chemistry, Rice University, Houston, Texas, United States of America, **6** Physics and Astronomy, Rice University, Houston, Texas, United States of America

Abstract

A general method for facilitating the interpretation of computer simulations of protein folding with minimally frustrated energy landscapes is detailed and applied to a designed ankyrin repeat protein (4ANK). In the method, groups of residues are assigned to foldons and these foldons are used to map the conformational space of the protein onto a set of discrete macrobasins. The free energies of the individual macrobasins are then calculated, informing practical kinetic analysis. Two simple assumptions about the universality of the rate for downhill transitions between macrobasins and the natural local connectivity between macrobasins lead to a scheme for predicting overall folding and unfolding rates, generating chevron plots under varying thermodynamic conditions, and inferring dominant kinetic folding pathways. To illustrate the approach, free energies of macrobasins were calculated from biased simulations of a non-additive structure-based model using two structurally motivated foldon definitions at the full and half ankyrin repeat resolutions. The calculated chevrons have features consistent with those measured in stopped flow chemical denaturation experiments. The dominant inferred folding pathway has an “inside-out”, nucleation-propagation like character.

Citation: Schafer NP, Hoffman RMB, Burger A, Craig PO, Komives EA, et al. (2012) Discrete Kinetic Models from Funneled Energy Landscape Simulations. PLoS ONE 7(12): e50635. doi:10.1371/journal.pone.0050635

Editor: Vladimir N. Uversky, University of South Florida College of Medicine, United States of America

Received: September 18, 2012; **Accepted:** October 23, 2012; **Published:** December 12, 2012

Copyright: © 2012 Schafer et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The project described was supported by Grant P01 GM071862 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of National Institute of General Medical Sciences or the National Institutes of Health. Additional support was also provided by the D. R. Bullard-Welch Chair at Rice University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pwolynes@rice.edu

These authors contributed equally to this work.

Introduction

Energy landscape theory and the principle of minimal frustration, which provide both simple models and interpretative frameworks [1,2], have contributed greatly to our understanding of the protein folding process. Proteins have evolved to minimize the effects of roughness of their energy landscapes by ensuring a significant stability gap between the unfolded ensemble and the native state. This leads to landscapes that resemble the high-dimensional analog of a rugged funnel. Protein folding can therefore be understood as a diffusive process across a rugged, biased, and structurally correlated energy landscape with weak transient trapping. Translating the ruggedness and stability gap ideas into mathematical terms has allowed self-consistent optimization methods to learn predictive potentials from structural data [3,4]. Coarse-grained models based directly on known protein structures have been derived that are computationally tractable, yet able to provide insight into, and generally show qualitative and often even quantitative agreement with, experimental results [5]. All-atom simulations of fast folding proteins are just now becoming reliable [6] and give results largely consistent with the rugged funnel landscape picture [7]. However, model building is only part of the challenge facing theorists working on protein folding since,

even on a minimally frustrated landscape, many seemingly distinct detailed mechanisms of folding are possible.

In order to interpret raw simulation results in ways that deepen our understanding of folding, researchers can either take advantage of the connection between structure and energy implied by theory and experiment to exist for natural proteins (using the principle of minimal frustration) or try to remain agnostic as to whether such a connection exists. The former choice leads to free energy based methods that use global, structure based reaction coordinates to calculate free energy profiles [8]. This global description facilitates comparison across a wide range of systems and development of physical intuition about details of specific systems. Furthermore, these free energy based methods can be combined with semi-analytical perturbation methods [9] to extrapolate existing simulation data to new simulation conditions. The more agnostic schemes sometimes start by using approximate reaction coordinates suggested by landscape theory but often rely on clustering strategies to define macrobasins. Such agnostic schemes generally have only provided predictions of rates for each given set of simulation conditions independently, in contrast with experiments that usually scan a range of thermodynamic conditions. Such schemes thus entail a significant computational

load when comparing with experiment. Recently some suggestions have emerged of how such general methods can be extended to combine data from parallel tempering simulations to yield kinetic models at arbitrary temperatures [10].

In this paper, we describe a free energy based method that can be used to derive kinetic equations that are similar to those derived using clustering based approaches but that take into account what has been learned about natural protein folding. This method maintains the attractive features of both free energy based methods using smooth reaction coordinates and clustering algorithms to provide predictions about rates and insight into folding mechanisms under a continuous range of conditions. The resulting folding mechanisms are expressed in terms of the cumulative flux through the network of macrobasins [11].

Methods

1.1 Foldons and reaction coordinates

The most basic criterion for defining a hierarchy of states in a kinetic model is that a separation of time scales should exist. Dynamics within a defined macrobasin should ideally be fast compared to the interconversion between the macrobasins. Many clustering strategies attempt to directly apply this criterion to simulation data. However, for folding models based on minimally frustrated landscapes we can take advantage of the connection between structure and energy to help choose natural ways of coarse-graining a protein's conformational space without already knowing the results of the simulation. These methods are necessarily approximate, but may, in many cases, be sufficient as well as efficient. Even on a rough energy landscape, if there are correlations, geometrical distances between structures are a good guide as to the barriers between them [12].

For this study, we will define foldons as contiguous regions of primary structure that may fold independently. This corresponds to a *putative* foldon as defined by Panchenko and others, which requires the contiguous primary structural regions to be kinetically competent [13]. The word "foldon" is sometimes employed to describe the notion of cooperatively folding substructures with no constraint on primary structural contiguity [14]. Such a scheme can also be useful but the first guess that contiguous regions reconfigure most rapidly is often correct.

The study of ankyrin repeat proteins has already revealed that the choice of folding units can be non-trivial. We use the designed ankyrin repeat protein 4ANK [15] as illustration. We adopt structurally motivated schemes for defining foldons in this system, namely that each repeat, or each half repeat, is one foldon [16,17]. For other types of proteins, different schemes may be more appropriate, and general schemes for approximate foldon assignment exist [13].

To measure the foldedness of the individual foldons, we use the reaction coordinate Q given in Equation 1.

$$Q = \frac{1}{N_p} \sum_i \sum_j \exp \left[\frac{-(r_{ij} - r_{ij}^u)^2}{2\sigma_{ij}^2} \right] \quad (1)$$

In Equation 1, i and j are residue indices, N_p is the total number of pairs (i,j) , r_{ij} is the distance between the C_α atoms of residues i and j , r_{ij}^u is the same distance in the experimentally determined native structure, and σ_{ij} is a sequence separation dependent width. We define the degree of foldedness of a foldon as the instantaneous value Q as given in Equation 1 where the summation over i is taken over all residues within a foldon and j goes over all residues within the same foldon and those in native contact with residue i

as defined by an 8\AA $C_\alpha - C_\alpha$ distance cutoff. Q has a range between 0 and 1, with 0 being completely unfolded and 1 being completely folded.

1.2 Macrobases and free energy calculations

For the purposes of defining a set of discrete macrobasins, we set a foldedness threshold above which a foldon may be considered essentially folded. Below this threshold, the foldon is considered to be unfolded. For the results shown in Section 3, this threshold has been set to $Q=0.6$. Using this scheme, any arbitrary structure from a simulation of a protein with, for example, 4 foldons can be assigned to a macrobasin such as 0101, indicating that the second and fourth (but not the first and third) foldons exceed the foldedness threshold. A protein with N_f foldons therefore has 2^{N_f} macrobasins, though not all such macrobasins would necessarily be observed in each set of simulations. This scheme is very analogous to the Ising model schemes used extensively by Munoz and Eaton [18].

We performed molecular dynamics simulations in the canonical ensemble, employing a biasing potential to umbrella sample along a global reaction coordinate Q_{tot} , defined as the value of Q (Equation 1) obtained by summing over all unique (i,j) pairs. We then used the multistate Bennet acceptance ratio (MBAR) [19] to compute the relative free energies of all sampled macrobasins over a range of temperatures. MBAR is a method that can be used to combine data from multiple equilibrium simulations at different thermodynamic states to obtain unbiased free energy differences and expectation values.

1.3 Transition rates and kinetic equations

Before considering the transition rates between macrobasins, it is necessary to define the connectivity of the discrete macrobasin space. It is reasonable to assume that locality of dynamics would imply that each macrobasin is directly connected to other macrobasins for which only a single $0 \rightarrow 1$ or $1 \rightarrow 0$ reconfiguration event is required to change the starting state into the final state. That is to say the direct transition $1010 \rightarrow 1011$ is allowed, but $1010 \rightarrow 1111$ and $1010 \rightarrow 1001$ are not directly allowed because they both require two local reconfiguration events and would in all likelihood be made up of composites of the simpler local moves. This is an example of a locally connected landscape; the effects of local connectivity on folding have been discussed previously [20].

The transition rate for going from macrobasin i to macrobasin j , K_{ij} , is given in Equation 2 where $\Delta F_{ij} = F_j - F_i$ is the free energy difference between the macrobasins' free energies, k_B is the Boltzmann constant, T is the absolute temperature and $k_0 = 1\mu\text{s}^{-1}$ is the assumed universal downhill transition rate. A similar rate scheme was adopted by Zheng *et al.* [21] when studying Trp-Cage using stochastic simulations on a kinetic network. The value of k_0 is motivated by a consideration of the ultimate speed limit of folding and measurements of the kinetics of downhill folding domains, as has emerged from numerous studies starting with the Eaton group [22–24]. The diagonal values of the matrix are defined so as to conserve probability, $K_{ii} = -\sum_{i \neq j} K_{ij}$, where K_{ij} refers to the element in the i th column and the j th row of matrix \mathbf{K} .

$$K_{ij} = \begin{cases} k_0, & \Delta F_{ij} < 0 \\ k_0 e^{-\frac{\Delta F_{ij}}{k_B T}}, & \Delta F_{ij} > 0 \end{cases} \quad (2)$$

From these microscopic rates it is well known how to derive the overall kinetics by diagonalizing the rate matrix [25,26]. The set of

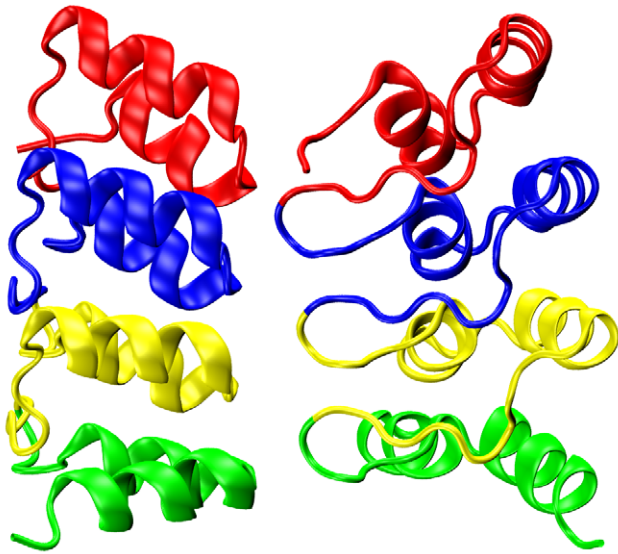


Figure 1. The protein 4ANK, comprised of 4 identical consensus ANK repeats. Each ANK is colored distinctly. The N-terminal repeat is colored red. The Visual Molecular Dynamics (VMD) software package [60] was used to visualize the structures in this work. doi:10.1371/journal.pone.0050635.g001

$\dim(\mathbf{K})$ eigenvalues, $\{r_\lambda\}$, and corresponding eigenvectors, $\{\mathbf{u}_\lambda\}$, are used in Equations 3–4. The instantaneous population of state i at time t is denoted $p_i(t)$. The time dependence of $\mathbf{p}(t)$, given in Equations 3 and 4, is then a function of the rate matrix, \mathbf{K} , and the initial concentrations $\mathbf{p}(0)$ via the coefficients $\mathbf{c} = \mathbf{U}^{-1}\mathbf{p}(0)$ where \mathbf{U} is the matrix of eigenvectors.

$$\frac{d\mathbf{p}(t)}{dt} = \mathbf{K}\mathbf{p}(t) \quad (3)$$

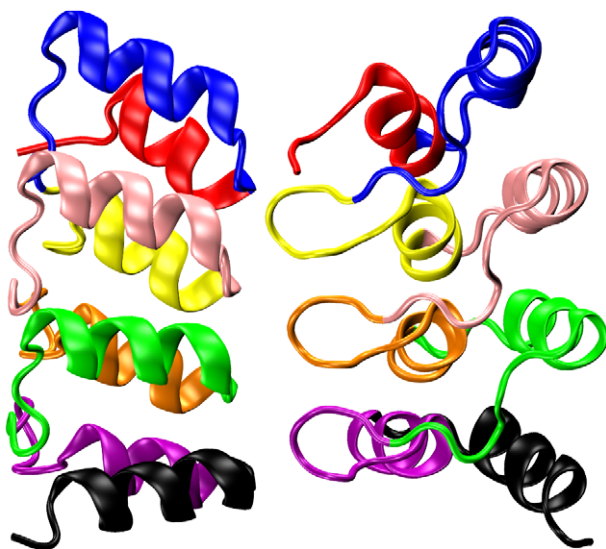


Figure 2. A structurally motivated foldon definition that splits each ANK element into two parts (8 foldons total). β -hairpins are contained within a single foldon (odd-numbered foldons). doi:10.1371/journal.pone.0050635.g002

$$\mathbf{p}(t) = \sum_{\lambda=1}^{\dim(\mathbf{K})} c_\lambda e^{r_\lambda t} \mathbf{u}_\lambda \quad (4)$$

For systems obeying detailed balance the eigenvalues $\{r_\lambda\}$ are all real and less than or equal to zero. Ordering them from largest to smallest, the resulting eigenvalue spectrum falls into two limiting scenarios [27]. If the largest non-zero eigenvalue (r_2) is well-separated from the next-largest eigenvalue, the system will initially rapidly relax in a multi-exponential fashion, then will be dominated by a single exponential. If several non-zero eigenvalues are all similar in magnitude, multi-exponential decays may be apparent.

The expression that we used to evaluate the cumulative flux between any two macrobasins i and j over a time interval $[0, t]$ is given in Equation 5.

$$J_{ij}(t) = \int_0^t [K_{ij}p_i(t') - K_{ji}p_j(t')] dt' \quad (5)$$

We evaluated Equation 5 from an initial concentration vector corresponding to a completely folded or unfolded state $t = \infty$, yielding equilibrium fluxes. We used the GraphViz software [28] to visualize the fluxes between each pair of directly connected macrobasins. Several examples of resulting flux diagrams are given in Section 3.

Model

2.1 Hamiltonian

The model used for the simulations reported in Section 3 has been previously described [29]. We only reiterate a few important aspects here. It is an explicit chain, coarse-grained, structure based, non-additive model. To avoid excessive computational burden, our model is coarse-grained to the level of three atoms per residue and does not explicitly represent solvent molecules. Attractive interactions are dictated by the experimentally determined native structure and are of a uniform strength (independent

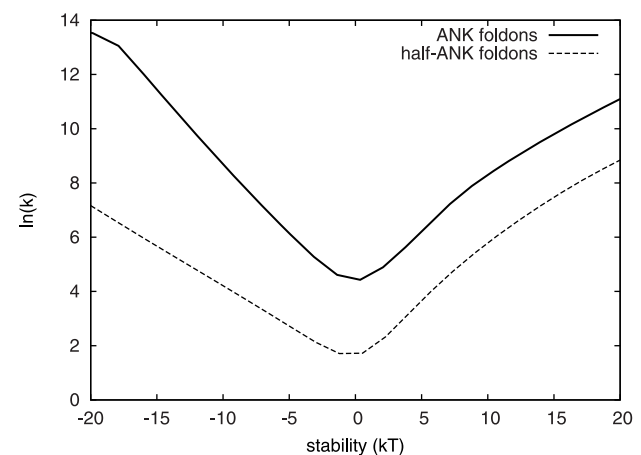


Figure 3. Thermal chevron plots obtained using two foldon definitions. For the ANK foldon definition (solid line) the maximum folding rate, at greatly stabilizing conditions, approaches the maximum downhill rate (10^6 s^{-1}). The minimum rate for the ANK definition is around 83 s^{-1} , for the half-ANK definition, about 5 s^{-1} . doi:10.1371/journal.pone.0050635.g003

of the amino acid identities). A consequence of the principle of minimal frustration [1] is that native contacts should be significantly more favorable than non-native contacts so that only those pairs of residues in contact in the experimentally determined native structure are assigned attractive interactions during the simulation. Although in reality non-native interactions are certainly present, their primary effect is to provide an additional source of friction, slowing the progression through the partially native manifold [30,31]. Structure based models have generally shown agreement with a variety of protein folding experiments although there are a few systems such as Im7 where specific non-native effects are quite apparent [32]. In our model, non-additive forces are approximated by introducing a non-additivity exponent p as shown in Equation 6, where H_{na} is the non-additive term in the Hamiltonian, E_i is a pairwise additive energy term and p is the non-additivity exponent. For the current study, a value of $p=2.0$ was used. Previous work indicates that adding a modest amount of non-additivity improves predictions of experimentally determined rate constants for both global and sub-global folding events of natural proteins [33,34].

$$H_{na} = -\frac{1}{2} \sum_i |E_i|^p \quad (6)$$

2.2 Example system

The ankyrin repeat (ANK) is a pervasive 33-residue motif found predominantly in eukaryotes [35]. It has been an excellent basis for constructing model systems for protein folding [16,36–38] and

engineering [39–44]. Through detailed comparison of ANK sequences, a consensus sequence – one that best represents the entire family – has been defined [15]. The secondary structure of a consensus ANK runs β -strand \rightarrow α -helix \rightarrow α -helix \rightarrow loop \rightarrow β -strand. The resulting tertiary structure contains a β -hairpin comprised of two rather short β strands coming from the N- and C-terminal ends of consecutive repeats. Previous work has shown that single ankyrin repeats in isolation do not adopt stable tertiary structures [15]. Our example system, 4ANK (RCSB PDB [45] ID: 1N0R [15]), is shown in Figure 1. The short β -strands are shown as coil in this particular representation. Not all published coordinates of ANK proteins are annotated as having β -strands elements. However, these extended loops typically populate the β -strand region of the Ramachandran plot. Variations in secondary structure detection algorithms (for example, consideration of hydrogen bonding geometry) may account for these apparent discrepancies.

Different groups have arrived at diverse descriptions of specific ANK protein folding mechanisms. Marchetti Bradley and Barrick, studying the Notch ankyrin domain (comprised of 7 ANK repeats), concluded that the central three ANKs of that protein formed the (early) transition state, based on ϕ value analysis. [46] Ferreiro and coworkers, who computationally evaluated the folding of ANK proteins ranging from 3 to 7 repeats, concluded that the folding nucleus consists not of an integer number of repeats but of one ANK plus the first helix of the following ANK repeat [16]. In order to remain agnostic regarding the nature of the nucleus without introducing unnecessary complexity, we have chosen to characterize the foldon macrobasins at both the ANK and the half-ANK resolution. To avoid subtleties associated with how sequence differences between the repeats can change the folding

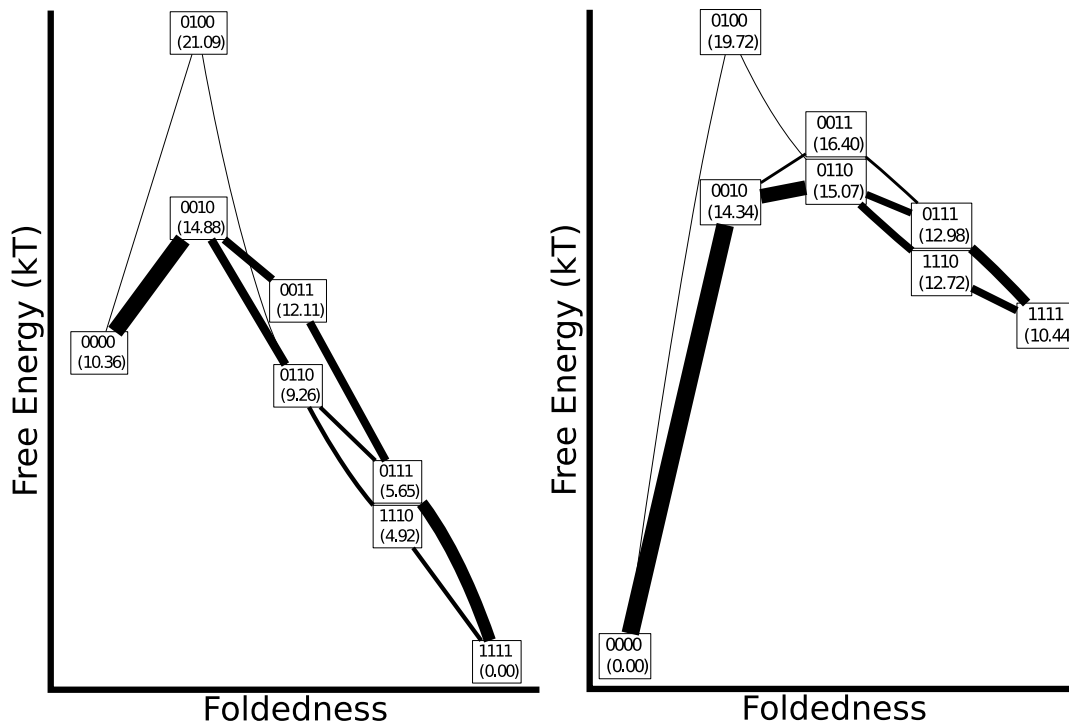


Figure 4. Flux diagrams for the full-ANK mechanism. The vertical coordinate approximates the free energy of each macrobasin and the precise relative free energies in units of kT are given in parentheses. The horizontal coordinate approximates the global reaction coordinate. A line is drawn between each pair of connected macrobasins, and the width of the line is proportional to the flux. A minimum line width is enforced for clarity. Folding conditions are shown on the left, unfolding on the right. doi:10.1371/journal.pone.0050635.g004

mechanism, we have chosen to study a consensus ANK protein (containing identical repeats) and simulate a model with uniform stabilizing contact energies.

4ANK is a designed ANK protein consisting of three identical, consensus repeats followed by a fourth consensus ANK lacking its final β -strand (which usually frays and promotes aggregation) [15]. A C-terminal tyrosine is the only non-consensus residue in the protein as constructed in the laboratory. Figures 1 and 2 show the experimentally determined structure of 4ANK and the two different foldon definitions we explored. One foldon definition assigns each ANK to its own foldon, while the second one divides the protein into 8 foldons of length 12, 19, 14, 19, 14, 19, 14, and 15 residues. The second definition was chosen so that the β -turn elements are contained within a single foldon. This allows us to monitor the formation of previously proposed [16] folding nuclei without deciding beforehand which ANKs would be involved.

Results

In Figure 3 we show the calculated characteristic rate coefficients for the protein 4ANK as a function of the relative stability of the completely folded and completely unfolded macrobasins. At lower temperatures (more negative stabilities) the characteristic rate reflects the rate of formation of the folded state – this parallels the experimental scenario where denatured protein is rapidly equilibrated in stabilizing conditions. At higher temperatures (more positive stabilities), the unfolding process dominates the relaxation kinetics. The rates become smallest when the folding and unfolding rate eventually meet near the folding temperature. For strictly two-state folders (with a transition state that does not vary with the stability) this sort of $\ln(k)$ vs. stability plot has a sharp V-shape and is therefore called a “chevron plot”. Deviations from a strict V-shape are expected for folding mechanisms with intermediates. Experimentally, chevron plots

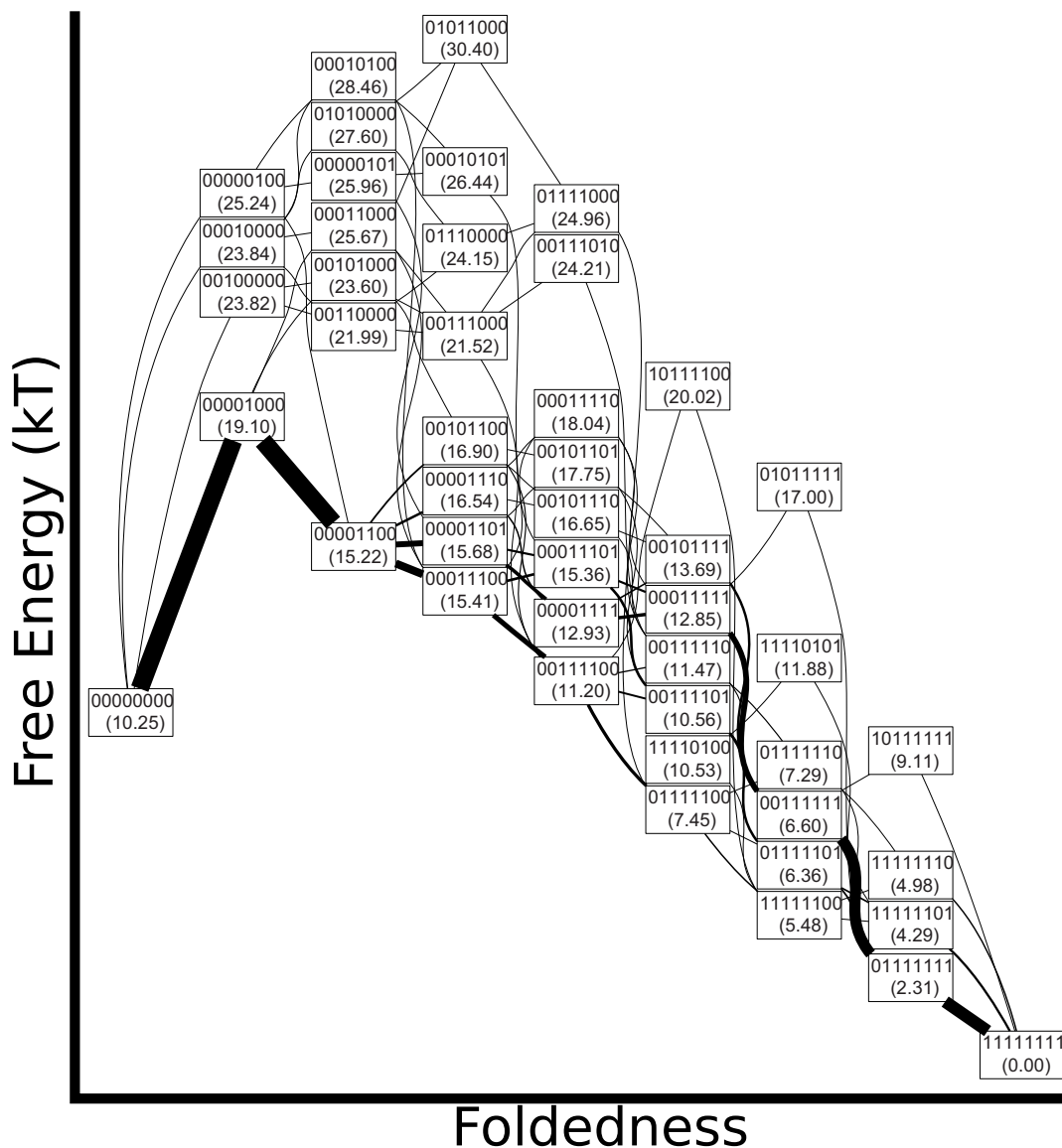


Figure 5. Folding flux diagram for the half-ANK mechanism. Symbols have the same meaning as in Figure 4.
doi:10.1371/journal.pone.0050635.g005

are typically obtained by using chemical denaturant to change the relative stability of the folded and unfolded states. In computer models that lack an explicit representation of chemical denaturants, it is necessary to find other ways to change the relative stability of the folded and unfolded states, and temperature is a common choice. Although not guaranteed to behave identically, calculated thermal chevron plots have been fruitfully compared to experimental chemical denaturant chevron plots to shed light on specific questions related to real biological systems [47,48].

Figure 3 shows chevron plots calculated using the ANK and half-ANK foldon definitions. Both foldon definitions give similar chevron plots although the rates obtained using the half-ANK definitions are lower. Using either foldon definition, the plots show curvature in the unfolding arm.

For each foldon definition, we calculated the cumulative folding and unfolding fluxes using Equation 5 (Figures 4, 5 and 6). The relative stability of the folded and unfolded macrobasins was chosen to be in the range of $10kT$ in all cases, about half way up

the folding or unfolding arm. The flux calculation was started with 100% of the population in either the folded or unfolded state, and Equation 5 was evaluated at $t = \infty$, yielding the equilibrium fluxes.

The mechanism inferred using ANK foldons (Figure 4) goes through a transition state with the third repeat folded. At high folded state stability, folding continues downhill in free energy through several competing pathways. The unfolding mechanism at high temperature is approximately the reverse of the folding process at low temperature, but it differs in that a single pathway dominates, proceeding through a broad transition state that contains both the 0010 and 0110 macrobasins. In contrast to folding conditions, relatively little flux flows through 0011.

Figure 5 shows the fluxes for folding according to the half-ANK foldon definition. With 46 macrobasins sampled, the half-ANK mechanisms are more elaborate. Flux goes through multiple pathways that are closely related to each other and similar to the previously discussed pathways for the ANK foldon definition. Most of the flux goes through the macrobasin 00001000, which

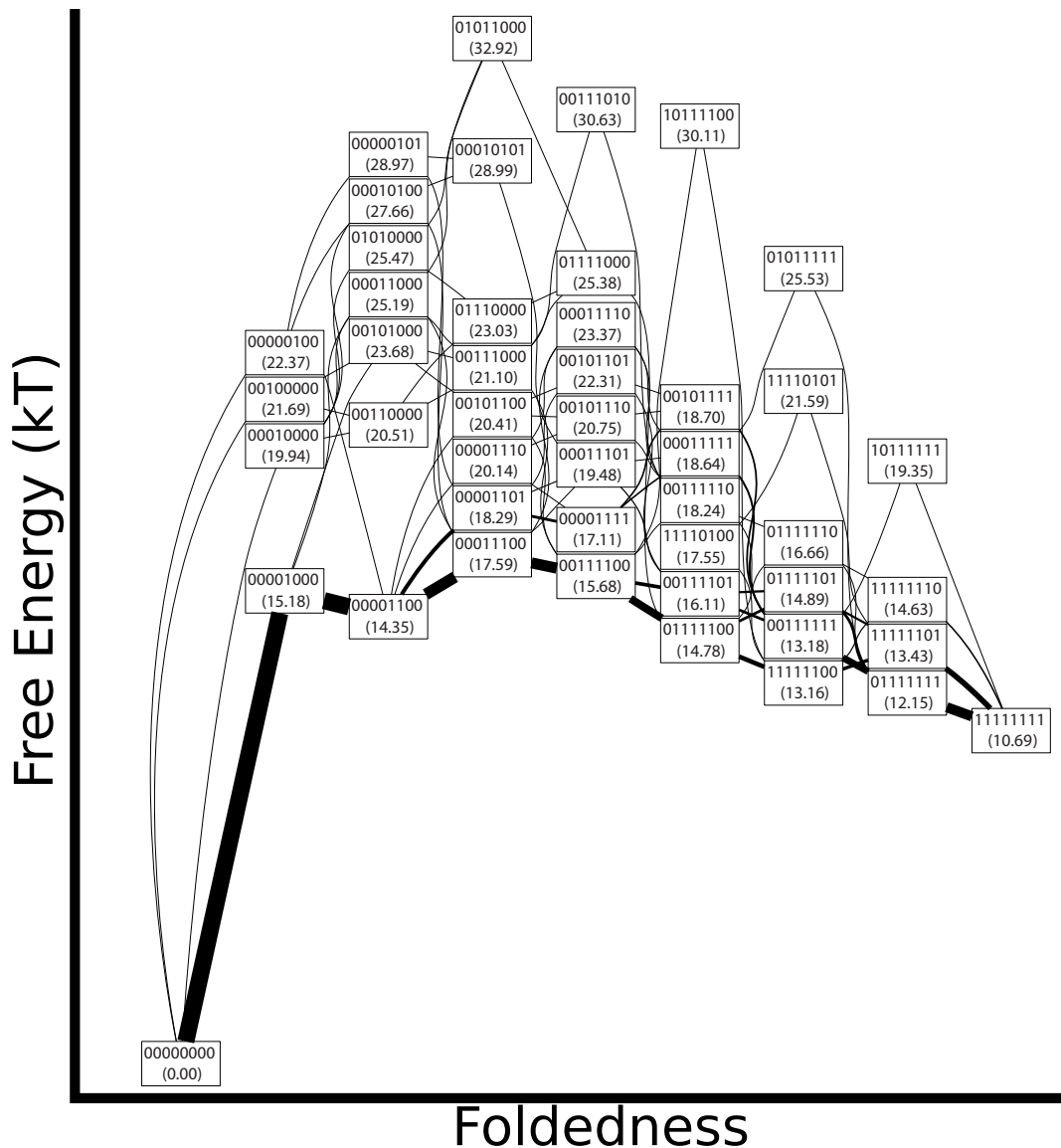


Figure 6. Unfolding flux diagram for the half-ANK mechanism. Symbols have the same meaning as in Figures 4 and 5.
doi:10.1371/journal.pone.0050635.g006

has the N-terminal helix of repeat 3 folded, and then through 00001100 to complete the folding of the 3rd repeat. While we predict a relatively high stability for the macrobasin 01111101, the flux analysis shows that this macrobasin is not kinetically significant. The mechanism does not follow trivially from the thermodynamics; the locality of transitions matters.

The unfolding fluxes under the half-ANK foldon assignment are shown in Figure 6. Unfolding is initiated at the termini. As with the ANK foldon case, the half-ANK mechanism goes through an intermediate with the center two repeats folded. For levels of global foldedness where an even number of half ANK units are folded, those macrobasins with all full ANK units either completely folded or unfolded (such as 00111100 and 00001111) are always found to be more stable than those with partially folded ANKs (such as 00011101 and 00101110). As a result, these states tend to have a larger fraction of the flux, although the exact amount of flux depends on the detailed connectivity of the model.

Discussion

Kinetic equation formalisms are useful as a way of coarse-graining protein folding landscapes and extracting measurable kinetics [10,11,49,50]. Here we develop an approach wherein umbrella sampling over a global folding reaction coordinate allows for accurate quantification of the free energies of the folding intermediates. A similar method was used by Ferreiro *et al.* to study TPR repeat proteins [17]. The current method extends that work by calculating folding kinetics and fluxes using simple assumptions about the kinetic connectivity of the network of intermediates and the universal rate for downhill transitions between macrobasins.

References

- Bryngelson JD, Wolynes PG (1987) Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci U S A* 84: 7524–7528.
- Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* 21: 167–195.
- Goldstein R, Luthey-Schulten Z, Wolynes PG (1992) Optimal protein-folding codes from spin-glass theory. *Proceedings of the National Academy of Sciences* 89: 4918.
- Davtyan A, Schafer N, Zheng W, Clementi C, Wolynes P, et al. (2012) AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *The Journal of Physical Chemistry B* 116: 8494–8503.
- Chan HS, Zhang Z, Wallin S, Liu Z (2011) Cooperativity, local-nonlocal coupling, and nonnative interactions: Principles of protein folding from coarse-grained models. *Annual Review of Physical Chemistry*, Vol 62 62: 301–326.
- Lindorff-Larsen K, Piana S, Dror R, Shaw D (2011) How fast-folding proteins fold. *Science* 334: 517–520.
- Best R (2012) Atomistic molecular simulations of protein folding. *Current Opinion in Structural Biology* 22(1): 52–61.
- Socci N, Onuchic J, Wolynes P, et al. (1998) Protein folding mechanisms and the multidimensional folding funnel. *Proteins Structure Function and Genetics* 32: 136–158.
- Eastwood M, Hardin C, Luthey-Schulten Z, Wolynes P (2002) Statistical mechanical refinement of protein structure prediction schemes: Cumulant expansion approach. *The Journal of chemical physics* 117: 4602.
- Prinz J, Chodera J, Pande V, Swope W, Smith J, et al. (2011) Optimal use of data in parallel tempering simulations for the construction of discrete-state markov models of biomolecular dynamics. *The Journal of chemical physics* 134: 244108.
- Berezhevskii A, Hummer G, Szabo A (2009) Reactive flux and folding pathways in network models of coarse-grained protein dynamics. *Journal of Chemical Physics* 130: 205102.
- Wang J, Plotkin S, Wolynes P (1997) Configurational diffusion on a locally connected correlated energy landscape; application to finite, random heteropolymers. *Journal de Physique I* 7: 395–421.
- Panchenko A, Luthey-Schulten Z, Cole R, Wolynes PG (1997) The foldon universe: a survey of structural similarity and self-recognition of independently folding units. *Journal of molecular biology* 272: 95–105.
- Lindberg MO, Oliveberg M (2007) Malleability of protein folding pathways: a simple reason for complex behaviour. *Current Opinion In Structural Biology* 17: 21–29.
- Mosavi LK, Minor DL, Peng ZY (2002) Consensus-derived structural determinants of the ankyrin repeat motif. *Proceedings of the National Academy of Sciences of the United States of America* 99: 16029–16034.
- Ferreiro DU, Cho SS, Komives EA, Wolynes PG (2005) The energy landscape of modular repeat proteins: topology determines folding mechanism in the ankyrin family. *J Mol Biol* 354: 679–692.
- Ferreiro DU, Walczak AM, Komives EA, Wolynes PG (2008) The energy landscapes of repeat-containing proteins: Topology, cooperativity, and the folding funnels of one-dimensional architectures. *PLOS Computational Biology* 4: e1000070.
- Muñoz V (2001) What can we learn about protein folding from ising-like models? *Current opinion in structural biology* 11: 212–216.
- Shirts M, Chodera J (2008) Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics* 129: 124105.
- Plotkin S, Wang J, Wolynes P (1997) Statistical mechanics of correlated energy landscape models for random heteropolymers and proteins. *Physica D: Nonlinear Phenomena* 107: 322–325.
- Zheng W, Gallicchio E, Deng N, Andreu M, Levy R (2011) Kinetic network study of the diversity and temperature dependence of trp-cage folding pathways: Combining transition path theory with stochastic simulations. *The Journal of Physical Chemistry B* 115(6): 1512–1523.
- Hagen S, Hofrichter J, Szabo A, Eaton W (1996) Diffusion-limited contact formation in unfolded cytochrome c: estimating the maximum rate of protein folding. *Proceedings of the National Academy of Sciences* 93: 11615.
- Kubelka J, Hofrichter J, Eaton W (2004) The protein folding ‘speed limit’. *Current opinion in structural biology* 14: 76–88.
- Kubelka J, Chiu T, Davies D, Eaton W, Hofrichter J (2006) Sub-microsecond protein folding. *Journal of molecular biology* 359: 546–553.
- Widom B (1965) Molecular transitions and chemical reaction rates: The stochastic model relates the rate of a chemical reaction to the underlying transition probabilities. *Science* 148: 1555–1560.
- Widom B (1971) Reaction kinetics in stochastic models. *The Journal of Chemical Physics* 55: 44–52.
- Widom B (1974) Reaction-kinetics in stochastic-models. II. *J Chem Phys* 61: 672–680.
- Ellson J, Gansner E, Koutsofios E, North S, Woodhull G (2004) Graphviz and dynagraphic and dynamic graph drawing tools. *Graph Drawing Software*: 127–148.
- Eastwood M, Wolynes PG (2001) Role of explicitly cooperative interactions in protein folding funnels: a simulation study. *The Journal of Chemical Physics* 114: 4702.

Acknowledgments

NPS thanks John Chodera for his technical support with the pymbar package. RMBH thanks Robert Konecny for essential computational support.

Author Contributions

Conceived and designed the experiments: NPS RMBH EAK PGW. Performed the experiments: NPS RMBH. Analyzed the data: NPS RMBH POC EAK PGW. Contributed reagents/materials/analysis tools: NPS RMBH AB POC. Wrote the paper: NPS RMBH EAK PGW.

30. Bryngelson J, Wolynes P (1989) Intermediates and barrier crossing in a random energy model (with applications to protein folding). *The Journal of Physical Chemistry* 93: 6902–6915.
31. Wang J, Saven J, Wolynes P (1996) Kinetics in a globally connected, correlated random energy model. *The Journal of chemical physics* 105: 11276.
32. Sutto L, Lätzer J, Hegler J, Ferreiro D, Wolynes PG (2007) Consequences of localized frustration for the folding mechanism of the im7 protein. *Proceedings of the National Academy of Sciences* 104: 19825.
33. Ejtehadi MR, Avall SP, Plotkin SS (2004) Three-body interactions improve the prediction of rate and mechanism in protein folding models. *Proceedings of the National Academy of Sciences of the United States of America* 101: 15088–15093.
34. Craig P, Lätzer J, Weinkam P, Hoffman RMB, Ferreiro DU, et al. (2011) Prediction of Native-State Hydrogen Exchange from Perfectly Funneled Energy landscapes. *J Am Chem Soc* 133: 17463–17472.
35. Mosavi LK, Cammett TJ, Desrosiers DC, Peng ZY (2004) The ankyrin repeat as molecular architecture for protein recognition. *Protein Science* 13: 1435–1448.
36. Devi VS, Binz HK, Stumpp MT, Pluckthun A, Bosshard HR, et al. (2004) Folding of a designed simple ankyrin repeat protein. *Protein Science* 13: 2864–2870.
37. Barrick D, Ferreiro DU, Komives EA (2008) Folding landscapes of ankyrin repeat proteins: experiments meet theory. *Current Opinion In Structural Biology* 18: 27–34.
38. Wetzel SK, Settanni G, Kenig M, Binz HK, Plückthun A (2008) Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *Journal of Molecular Biology* 376: 241–257.
39. Forrer P, Stumpp MT, Binz HK, Pluckthun A (2003) A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *Febs Letters* 539: 2–6.
40. Kohl A, Binz HK, Forrer P, Stumpp MT, Pluckthun A, et al. (2003) Designed to be stable: Crystal structure of a consensus ankyrin repeat protein. *Proceedings of the National Academy of Sciences of the United States of America* 100: 1700–1705.
41. Forrer P, Binz H, Stumpp M, Plückthun A (2004) Consensus design of repeat proteins. *Chembiochem* 5: 183–189.
42. Ferreiro DU, Cervantes CF, Truhlar SME, Cho SS, Wolynes PG, et al. (2007) Stabilizing IkappaB alpha by “consensus” design. *Journal of Molecular Biology* 365: 1201–1216.
43. Boersma YL, Plueckthun A (2011) Darpins and other repeat protein scaffolds: advances in engineering and applications. *Current Opinion in Biotechnology* 22: 849–857.
44. Tamaskovic R, Simon M, Stefan N, Schwill M, Plueckthun A (2012) Designed ankyrin repeat proteins (darpins): From research to therapy. *Methods In Enzymology: Protein Engineering For Therapeutics*, Vol 203, Pt B 503: 101–134.
45. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, et al. (2000) The protein data bank. *Nucleic acids research* 28: 235–242.
46. Marchetti Bradley C, Barrick D (2006) The notch ankyrin domain folds via a discrete, centralized pathway. *Structure* 14: 1303–1312.
47. Shen T, Hofmann C, Oliveberg M, Wolynes PG (2005) Scanning malleable transition state ensembles: comparing theory and experiment for folding protein u1a. *Biochemistry* 44: 6433–6439.
48. Zong C, Wilson C, Shen T, Wolynes PG, Wittung-Stafshede P (2006) θ -value analysis of apo-azurin folding: Comparison between experiment and theory. *Biochemistry* 45: 6458–6466.
49. Levy Y, Jortner J, Berry RS (2002) Eigenvalue spectrum of the master equation for hierarchical dynamics of complex systems. *Physical Chemistry Chemical Physics* 4: 5052–5058.
50. Buchete NV, Hummer G (2008) Coarse master equations for peptide folding dynamics. *Journal of Physical Chemistry B* 112: 6057–6069.
51. Matouschek A, Fersht A (1993) Application of physical organic chemistry to engineered mutants of proteins: Hammond postulate behavior in the transition state of protein folding. *Proceedings of the National Academy of Sciences* 90: 7814–7818.
52. Jonsson T, Waldburger C, Sauer R (1996) Nonlinear free energy relationships in arc repressor unfolding imply the existence of unstable, native-like folding intermediates. *Biochemistry* 35: 4795–4802.
53. Sánchez I, Kiefhaber T (2003) Evidence for sequential barriers and obligatory intermediates in apparent two-state protein folding. *Journal of molecular biology* 325: 367–376.
54. DeVries I, Ferreiro DU, Sanchez IE, Komives EA (2011) Folding kinetics of the cooperatively folded subdomain of the ikb alpha ankyrin repeat domain. *Journal of Molecular Biology* 408: 163–176.
55. Tang K, Fersht A, Itzhaki L (2003) Sequential unfolding of ankyrin repeats in tumor suppressor p16. *Structure* 11: 67–73.
56. Lowe A, Itzhaki L (2007) Rational redesign of the folding pathway of a modular protein. *Proceedings of the National Academy of Sciences* 104: 2679–2684.
57. Werbeck N, Rowling P, Chellamuthu V, Itzhaki L (2008) Shifting transition states in the unfolding of a large ankyrin repeat protein. *Proceedings of the National Academy of Sciences* 105: 9982–9987.
58. Kaya H, Chan HS (2003) Origins of chevron rollovers in non-two-state protein folding kinetics. *Physical Review Letters* 90: 258104.
59. Kaya H, Liu ZR, Chan HS (2005) Chevron behavior and isostable enthalpic barriers in protein folding: Successes and limitations of simple go-like modeling. *Biophysical Journal* 89: 520–535.
60. Humphrey W, Dalke A, Schulten K (1996) VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* 14: 33–38.