## **Supporting Information**

## **Felitsky** *et al.* **10.1073/pnas.0710641105**

## **SI Text**

For two residues *i* and *j* in a given macrostate, the effective contact order (ECO) is defined as the number of monomeric links along the shortest path between *i* and *j* traced either covalently or topologically (1). In our model, ECO is used as an effective sequence separation for evaluating distance distributions in the presence of one or more topological restraints. Calculations are performed by replacing the contour length  $l_c$ and root mean square end-to-end distance for the unperturbed chain  $\langle r_0^2 \rangle^{1/2}$  determined by using the sequence separation  $|i-j|$ with one determined by using the ECO (e.g.,  $l_c = b_0 ECO$ ) in the set of logistically corrected distance distribution functions (Eq. **4** in *Results*). These ECO-based distance distribution functions are then used in two separate contexts. The first of these is in calculating  $R_{2P}$  values for spin label–amide proton pairs in individual macrostates via the ECO-based analog of Eq. **2** (*Results*) The second is the calculation of conditional free energies for estimation of the energetic penalty for loop closure via Eq. **5** (*Results*).

**Contributions to ECO from Nonlocal Clusters.** In our model, topological constraints arise from the presence of nonlocal clusters, which bring different regions of the chain into close proximity and simultaneously promote local chain compaction. Calculating the ECO exactly for two residues within or on opposite ends of such a cluster would require more detailed knowledge of the physical structure of such clusters than is presently available. In developing a suitable approximation, we observe that there are several instances in the experimental data set where introduction of a spin label into one local cluster causes complete line broadening  $(Q < 0.01)$  of amide cross-peaks in another local cluster (e.g., the effects of R139C\* on G region cross-peaks). This can only be accounted for via our ECO model if the maximum ECO for any pair of residues within a nonlocal cluster is 4 or less. An ECO of 5 corresponds to an  $R_{2P}$  of 61 s<sup>-1</sup> (calculated via *R*esults Eqs. **2**–**4**) and a minimum *Q* of 0.04 (if the contact were fully populated). Cross-peaks with a *Q* of 0.04 are still detectable in the paramagnetic spectrum and thus incompatible with the aforementioned experimental observation of complete line broadening across some nonlocal clusters. ECO values smaller than five predict more extensive PRE and thereby can account for the experimental results. In our analysis, we set the maximum ECO contribution when tracing the chain into or across a nonlocal cluster to four, the maximal value consistent with the experimental results within the context of our model. As such, this choice provides a roughly minimal estimate of the intracluster PRE and thus an upper bound estimate of the contact probabilities and coalesced macrostate populations.

**Example ECO Calculation in the Presence of a Single Nonlocal Cluster.** As a concrete example of the procedure for calculating ECO, we determine the ECO between residue  $i = 1$  and residue  $j = 6, 10, 20$ , 26, or 30 for macrostate AB-C-G-H, which contains a single nonlocal cluster (AB; residues 4–16 and 27–35). The ECO for the  $j = 6$  case is simply  $|6 - 1| = 5$ . Because residue 10 is more than four residues into the nonlocal cluster AB, the ECO for the  $j = 10$  case is  $|4 - 1| + 4 = 7$ ; this value is obtained by counting from *i* to the edge of cluster AB (i.e.,  $|4 - 1|$ ) and then adding 4 as the maximum contribution from the nonlocal cluster. The  $j = 20$  case is similar, except that we now must include contributions from linkers on both sides of the cluster:  $|4 - 1| + 4 + |20 - 16| = 11$ . The  $j = 26$  case demonstrates the importance of tracing all possible paths. In this

Felitsky *et al.* [www.pnas.org/cgi/content/short/0710641105](http://www.pnas.org/cgi/content/short/0710641105) **1 of 11**

case, there are two paths between residue 26 and cluster AB; while residue 26 and A have a sequence separation of  $|26 - 16| = 10$ , this residue is immediately adjacent to B ( $|27 - 26| = 1$ ); this latter value is used in calculating the ECO, which is  $|4 - 1| + 4 + |27 - 26| =$ 8. In the final example, where  $j = 30$  is in the nonlocal cluster AB cluster, we again set the maximum count within the nonlocal cluster to four and the ECO is identical to the  $j = 10$  case,  $|4 - 1| + 4 =$ 7. Note that the ECO values for the different cases are not related in any simple way to their sequence separations  $|i - j|$ .

**Example <sup>R</sup>2P Calculation in the Presence of Multiple Nonlocal Clusters.** We next consider the extent to which a spin-label at position 57 enhances relaxation of an amide proton at position 107 in the macrostate defined by nonlocal clusters AGH and BC. The local cluster boundaries needed for this calculation are given in the legend to Fig. 1. First, we consider the direct pathway along the chain. The number of links to reach the G region is  $|98 - 57|$  = 41. Because residue 107 is more than four residues into the interior of AGH (107  $-$  98  $>$  4), the nonlocal cluster contributes an additional four links, giving a total of 45 for this pathway. However, a shorter pathway exists which sums contributions between position 57 and BC (8), through BC (4), between BC and AGH (11), and into the interior of AGH (4). The ECO which determines the distribution function to be used for this interaction is therefore  $8 + 4 + 11 + 4 = 31$ . The corresponding  $R_{2P}$ is  $\approx 0.09$ ; if the entire ensemble were to consist of this particular macrostate, *Q* for the interaction would be 0.98.

**Calculation of Loop Closure Free Energies GL in Macrostates with Topological Restraints.** The nonlocal clusters in a particular macrostate act as a set of topological restraints that reduce the configurational entropy of the protein chain. We calculate the entropy loss for each macrostate relative to the completely unfolded reference macrostate through the following series of steps: (*i*) enumeration of the pathways by which the particular set of nonlocal clusters may form; (*ii*) calculation of ECO values for the individual steps in each pathway; (*iii*) calculation of the capture radius and conditional free energy change  $\Delta G'(q|i, j, k \dots)$  for each coalescence event *q* in each pathway given that events  $i, j, k...$  have occurred; and  $(iv)$  selection of the pathway with the largest sum of  $\Delta G'$  (to ensure that the configurational entropy monotonically decreases as additional topological constraints are introduced into the chain).

In the model, a pathway is defined as a minimal series of pairwise coalescence events by which a particular macrostate can be reached from the unfolded (completely dissociated) macrostate. All pathways leading to a particular macrostate have an identical number of steps, corresponding to the *n* pairwise coalescence events required for formation. All possible pathways leading to two example macrostates are listed in [Table S1](http://www.pnas.org/cgi/data/0710641105/DCSupplemental/Supplemental_PDF#nameddest=ST1) and [Table S2.](http://www.pnas.org/cgi/data/0710641105/DCSupplemental/Supplemental_PDF#nameddest=ST2)

Calculation of  $\Delta G_L$  requires the definition and calculation of ECO values for clusters rather than individual residues. For a local cluster, the ECO endpoint is taken to be that of the central residue in the cluster. In the case where a local cluster has an even number of residues, the ECO endpoint is calculated for each of the two centermost residues; the average of these is used (with the resulting ECO being rounded down to the nearest integer). For example, the ECO between the C local cluster (residues 40–49) and G local cluster (residues 98–117) is  $|110.5 - 44.5| = 66$ . For a nonlocal cluster (involving two or more local clusters), intermediate calculations are carried out for each of ECO endpoints of the local clusters which compose the nonlocal cluster; the endpoint which gives the smallest ECO for the coalescence event involving the nonlocal cluster is then used (the four-residue maximum discussed above still applies). ECO values for the coalescence events along all possible pathways leading to the macrostates ABC-G-H and ACG-BH are given in [Table S1](http://www.pnas.org/cgi/data/0710641105/DCSupplemental/Supplemental_PDF#nameddest=ST1) and [Table S2.](http://www.pnas.org/cgi/data/0710641105/DCSupplemental/Supplemental_PDF#nameddest=ST2)

The capture radius  $r_c$  is defined as the average distance to which the central residues of two clusters must approach before coalescence occurs. We model  $r_c$  for cluster formation using the diffusion-collision approximation, which proceeds as follows.

NAS PNAS

For each of the two clusters undergoing coalescence, an effective volume is obtained by summing over the van der Waals volumes of the residues in the two clusters and multiplying by  $3\sqrt{2}/\pi$  to account for irregular packing (2). The radii of the two clusters are calculated from the resulting volumes assuming spherical geometry and added together to obtain the capture radius *r*c.

Conditional free energies  $\Delta G'$  for each coalescence event in all pathways leading to the macrostates ABC-G-H and ACH-BG, calculated through Eq. **5** (*Results*), are tabulated in [Table S1](http://www.pnas.org/cgi/data/0710641105/DCSupplemental/Supplemental_PDF#nameddest=ST1) and [Table S2,](http://www.pnas.org/cgi/data/0710641105/DCSupplemental/Supplemental_PDF#nameddest=ST2) respectively. The largest path summation  $\Delta G'$ (path) is then identified as  $\Delta G_L$  for each macrostate.

<sup>2.</sup> Bashford D, Cohen FE, Karplus M, Kuntz ID, Weaver DL (1988) Diffusion-collision model

<sup>1.</sup> Fiebig KM, Dill KA (1993) Protein core assembly processes. *J Chem Phys* 98:3475–3487.

for the folding kinetics of myoglobin. *Proteins* 4:211–227.



**Fig. S1.** Paramagnetic enhancement to nuclear spin relaxation for unfolded apoMb at pH 2.3 in the presence of 8 M urea for all apoMb variants investigated. As in Fig. 2*A*, the histograms show the experimental intensity ratios; data for overlapping resonances are omitted from the PRE profiles and analysis. The red fitted curves correspond to the corrected model described in the text. The analysis includes previously published data for the E18, K77, and K133 mutants [Lietzow MA, Jamin M, Dyson HJ, Wright PE (2002) Mapping long-range contacts in a highly unfolded protein. *J Mol Biol* 322:655–662]. In the presence of 8 M urea (at pH 2.3), the extent of PRE in the apoMb ensemble markedly decreases throughout the chain and interactions between the two chain termini are no longer detectable [Lietzow MA, Jamin M, Dyson HJ, Wright PE (2002) *J Mol Biol* 322:655–662]. Residual interactions show no obvious indication of specific tertiary contacts (cf. R139C\* vs. K140C\* and L11C\* vs. A15C\*), consistent with our restriction of solvation to interactions involving both termini. Nonspecificity of interactions is also supported in that all spin-labels reporting on the same pairwise contact (e.g., between the A and B regions) exert similar magnitude reciprocal effects. For example, PRE effects from L11C\* on B region residues are similar to those from R31C\* on A region residues.

TAS.



**Fig. S2.** ApoMb sequence variation of the Kyte–Doolittle hydropathy, plotted by using a seven-residue moving average.

**SVNAS** 

 $\mathbf{A}$ 



**Fig. S3.** Paramagnetic enhancement to nuclear spin relaxation for unfolded apoMb at pH 2.3 in the absence of urea for A57C\* and K77C\*. Each panel is labeled with the residue for which the spin-labeled cysteine was substituted; arrows indicate the sites of substitution. As in Fig. 2*A*, the histograms show the experimental intensity ratios. The fitted curves correspond to the initial (green) and corrected (red) models described in the text. The locations of the local clusters (regions of high AABUF) are indicated by blue bars at the top of the figure. Data for overlapping resonances are omitted from the PRE profiles and analysis.

IAS.

 $\overline{\mathbf{C}}$ 

**Table S1.** *G***L calculations for formation of macrostate ABC-G-H, which requires two coalescence events**

Path	Step 1	Step 2	ECO for step 1	ECO for step 2 $\Delta G'(1)$ , kcal·mol <sup>-1</sup>	$\Delta G'(2 1)$ , kcal·mol <sup>-1</sup>	$\Delta G'$ (path), kcal·mol <sup>-1</sup>
		$A + B \rightarrow AB$ $AB + C \rightarrow ABC$		4.00	2.01	6.01
		$B + C \rightarrow BC$ $A + BC \rightarrow ABC$		2.88	3.32	6.20
		$A + C \rightarrow AC$ $AC + B \rightarrow ABC$	34	4.69	2.06	6.75

**Table S2.** *G***L calculations for formation of macrostate ACH-BG, which requires three coalescence events**

Path	Step 1	Step 2	Step 3	ECO step 1	ECO step 2	ECO step 3	$\Delta G'(1)$ , $kcal$ -mol $^{-1}$	$\Delta G'(2 1)$ , $kcal$ -mol <sup>-1</sup>	$\Delta G'(3 1,2)$ , $kcal·mol-1$	$\Delta G'$ (path), $kcal·mol-1$
	$A + C \rightarrow AC$	$B + G \rightarrow BG$	$AC + H \rightarrow ACH$	34	71	33	4.69	5.42	4.33	14.44
	$A + C \rightarrow AC$	$AC + H \rightarrow ACH$	$B + G \rightarrow BG$	34	92	39	4.69	5.67	4.65	15.01
	$A + H \rightarrow AH$	$B + G \rightarrow BG$	$AH + C \rightarrow ACH$	127	45	28	6.41	4.85	3.97	15.23
4	$A + H \rightarrow AH$	$AH + C \rightarrow ACH$	$B + G \rightarrow BG$	127	32	39	6.41	4.19	4.65	15.25
5.	$C + H \rightarrow CH$	$B + G \rightarrow BG$	$A + CH \rightarrow ACH$	93	39	30	6.19	4.65	4.05	14.89
6	$C + H \rightarrow CH$	$A + CH \rightarrow ACH$	$B + G \rightarrow BG$	93	34	39	6.19	4.25	4.65	15.09
	$B + G \rightarrow BG$	$A + C \rightarrow AC$	$AC + H \rightarrow ACH$	76	30	33	5.50	4.50	4.33	14.34
8	$B + G \rightarrow BG$	$A + H \rightarrow AH$	$AH + C \rightarrow ACH$	76	41	28	5.50	5.15	3.97	14.62
9	$B + G \rightarrow BG$	$C + H \rightarrow CH$	$A + CH \rightarrow ACH$	76	34	30	5.50	5.03	4.05	14.58











## **Table S4. continued**

PNAS PNAS



Blank cells indicate residues for which peak overlap prevented accurate measurements. BBD is used to designate residues that are either broadened beyond detection in the paramagnetic state or otherwise not quantifiable due to low signal intensities.