Supplementary Information for *Modulation of a protein free-energy landscape by circular permutation*

Gaël Radou, Marta Enciso, Sergei Krivov and Emanuele Paci
Astbury Centre for Structural Molecular Biology, University of Leeds, UK

**Generation of folding trajectories**

The reference structure for WT*T4L is the 1.25 Å resolution crystal structure 3DKE in the PDB. The reference structure for CP13*T4L has been obtained by introducing a peptide bond between the original termini and by removing the peptide bond between residue 12 and 13 (Figure 1); the termini being naturally within a peptide bond length the bond does not require any modification of the structure; thus in the definition of the structure based potential only the bonded terms involving the termini are modified. This choice is consistent with the experimental observation that the X-ray crystal structure of the wild type and the circular permutant are identical within experimental resolution [?].

Both structure have been subjected to energy minimization (2000 steepest descent) using a united atom model with implicit solvent [3].

Folding simulations of WT*T4L and CP13*T4L have been performed using a structure-based model that has one centre of interaction per amino acid in the $C_\alpha$ position [1] implemented in CHARMM [2]. Interactions are attractive if they are present in the reference structure and repulsive otherwise. The magnitude and range of the interactions depend on the chemical properties of the residues, the separation between side-chains and the presence of hydrogen-bonds in the reference structure.

Langevin simulations have been performed using friction coefficient of 1 ps$^{-1}$ and a timestep of 15 fs. Coordinates were saved for the analysis every 500 steps (7.5 ps). Long simulations (more than 1 µs) in a range of temperatures between 200 K and 400 K show that each domain of T4 lysozyme (either the wild type or the circular permutant) has different stability (Figure S1), which makes it impossible to simulate folding and refolding of both domains at equilibrium at one given temperature. Instead, we have performed non-equilibrium folding simulations: for each variant we have computed 100 folding events at 250 K, starting from completely unfolded structures obtained from equilibrium simulations at 400K. As the root mean square deviation of the protein from the native state falls down lower than 1.5 Å, the protein was considered refolded and the trajectory stopped 75 ns after.

![Figure S1: Heat capacity of both WT*T4L (black) and CP13*T4L (red) as a function of the temperature $T$, calculated with WHAM [4]. Peaks correspond to the unfolding of the N- and C-domain. The melting temperature of the N-domain is lower than the melting temperature of the C-domain for both species.](image-url)
Construction of the optimal reaction coordinate

Identification of the single coordinate that accurately describes complex folding process is challenging. In many cases, the standard progress variables (e.g. number of native contacts, radius of gyration, root mean square distance from the native structure) are not good reaction coordinates, because they do not preserve the barriers on the free-energy surface (FES) and thus may mask the inherent complexity of the latter [5]. A number of methods to construct good reaction coordinates have been suggested [6-12]. Here we employ a method which constructs the coordinate based on system dynamics by optimizing its cut-based free-energy profile (cFEP) [7, 13]. Given a set of trajectories $\tilde{Y}_j(i\Delta t)$ recorded with time interval $\Delta t$ and a functional form defining the reaction coordinate $P$, the time series of $P$ is computed $X_j(i\Delta t) = P(\tilde{Y}_j(i\Delta t))$. The cFEP is $F_C(x)/k_B T = -\ln Z_C(x)$, where the partition function $Z_C(x)$ equals half the number of transitions performed by $X_j(i\Delta t)$ through point $x$. The $F_C$ together with the conventional histogram-based free-energy profile $F_H$ determine the coordinate-dependent diffusion coefficient $D(x)$, and completely specify diffusive dynamics [7]. The optimal coordinate is constructed by numerically optimizing the parameters of the functional form $P$ to make the mean first passage time (MFPT) to the native state largest. The MFPT is computed by numerically integrating the Kramers equation [7]

$$\langle t_{AB} \rangle = \int_A^B dx \frac{e^{\beta F(x)}}{D(x)} \int_{-\infty}^x dy e^{-\beta F(y)} = \frac{\Delta t}{\pi} \int_A^B dx \frac{Z_H(x)}{Z_C^2(x)} \int_{-\infty}^x dy Z_H(y). \qquad (1)$$

It is reasonable to assume that any bad projection that results in overlapping of different parts of the configuration space will result in faster kinetics, i.e., in a smaller MFPT. Clearly the longest MFPT is obtained on the original free-energy surface or from a projection where no such overlapping occurs. Hence, the maximum value of the integral in Eq. 1 can serve as a definition of the best one-dimensional projection. The functional form of the reaction coordinate was obtained by generalising the conventional number of native contacts: $X_i = P(\tilde{Y}_i) = \sum_{i<j} a_{ij} \Theta(\Delta_{ij} - r_{ij})$, where $a_{ij}$ is either 1 or $-1$, $r_{ij}$ is the distance between atoms $i$ and $j$ and $\Delta_{ij}$ is the distance threshold, when contact between the atoms is considered to be formed; $\Theta$ is the Heaviside step function, whose value is zero for a negative argument and one for a positive argument. Numerical optimization was performed by iteratively picking a random pair of atoms $ij$, then scanning the whole parameter space for the pair ($a_{ij} = \pm 1$ and $\Delta_{ij} = 0, 0.5, 1, \ldots, 30$) and finally selecting the one that gives the highest MFPT.

Equilibrium free energy profile

The equilibrium free-energy landscape is computed from non-equilibrium trajectories by reweighting the transition network, which describes the dynamics along the reaction coordinate [14]. A Markov network is constructed by binning the reaction coordinate. The transition probabilities are estimated as $p_{ij} = n_{ij}/\sum_i n_{ij}$, where $n_{ij}$ is the number of transition from bin $j$ to bin $i$. Assuming that the network describes the kinetics, the equilibrium population are found as $p_{ij}^{eq} = \sum_j p_{ij} p_{ij}^{eq}$. The equilibrium number of transitions $n_{ij}^{eq} = p_{ij} p_{ij}^{eq}$ is used to compute the equilibrium free-energy profile. Such computed landscape provides an accurate description of the folding dynamics as confirmed by agreement of the reaction rates computed from the profile and directly from the simulation. The success of the reweighting procedure also confirms that the reaction coordinate faithfully reproduces the folding dynamics. The mean first passage time between the basins estimated by the Kramers equation 1 from the free-energy profile and coordinate dependent diffusion coefficient are relatively in good agreement with that estimated directly from the trajectories (with a factor of 3).
Dynamical correlation between the N-domains and C-domains

To explore the cooperative nature of our results we performed a quasi harmonics principal component analysis [15]. We calculated the covariance matrix for the atomic fluctuations along the folding trajectory for each domain separately, obtaining $\lambda_i$ eigenvalues for the N-domain and C-domain, respectively. After projecting these trajectories onto the modes with the smallest eigenvalues we calculated the cross-correlation between the timeseries for the two domains. The higher the cross-correlation, the more coupled the relevant dynamics of the two domains; we propose this as a measure of the coupling between the two subdomains.

Additional analysis of the intermediate and transition states

The system dynamics in each basin and transition state have been quantitatively characterised by the calculation of RMSF (root mean square fluctuations) for each particle $i$, $\text{RMSF}_i = \sqrt{1/T \sum_{j=1}^{T} (x_i(t_j) - \bar{x}_i)^2}$ where $T$ is the total number of frames of the trajectory; B-factors were calculated as $B_i = \frac{8}{3} \pi \text{RMSF}_i^2$. For each of the considered ensembles, we have computed the average structure and coloured it according to B-factors (Figures 2 and 3). In Figure S2, the RMSF values for the intermediates and the transition states of the rate limiting step are shown ($TS_{D-I}$ and $TS_{I-N}$ for wild type and the circular permutant, respectively). For both the wild type and the circular permutant, the intermediate presents an ordered, folded C-domain (low RMSF) and a disordered N-domain (high RMSF). At the transition state of rate limiting step, the RMSF values for the N-domain are significantly lower, indicating the achieving of the ordered native structure.

Clues on the nature of the transition state can also be gathered from the fraction of the native contacts present on average in the corresponding ensemble (Figure S3). These can be considered as “structural $\phi$-values”, and related to the experimentally measured $\phi$-values [16]. For the wild type, $\phi$-values is close to 0 ($\phi \approx 0.2$) confirm that the N-domain is almost completely unfolded. The C-domain (residues 1-12 and 60-164), on the contrary, is mainly folded, though not completely, with a $\phi$-values closer to 1 ($\phi \approx 0.8$). The structural $\phi$-values of the circular permuant are higher than for the wild-type, reflecting the proximity of the transition state to the native state. In particular, it is true for the C-domain, which is almost folded.

Interestingly, the transition state of CP13*T4L is again compact (see the blue region in the N-domain of $TS_{I-N}$ in Figure 2 and the high $\phi$-values for residues around 20 in Figure S3). This implies that, in spite of differences in the free-energy landscape, there is a common feature in the two variants: the closing of the protein determines the rate limiting step in T4 lysozyme and its circular permutant.

Low friction does not affect the folding mechanisms

In this study, low friction ($\gamma = 1 \text{ ps}^{-1}$), has been used. This choice improved the sampling efficiency, but in principle may affect the folding mechanisms of proteins by changing the proportion between protein-protein and protein-solvent.
Figure S3: Structural phi values for the rate limiting transition state of WT*T4L (blue squares) and CP13*T4L (red circles) TS$_{D-I}$ and TS$_{I-N}$, respectively. Note that this fraction is systematically higher for the circular permutant, showing its proximity to the native state.

frictions [17, 18]. To check that the findings reported here do not depend on the choice of friction coefficient we repeated simulations with higher friction coefficients ($\gamma = 3$, 5 and 10 ps$^{-1}$) and analysed them using the optimal reaction coordinate found at $\gamma = 1$ ps$^{-1}$ (Figure S4).

Figure S4 shows that FEP is mostly unchanged at higher frictions, i.e., some minor differences are observed at the transition states. To show that discrepancies between the profiles (about 0.5 k$_B$T) are due to statistical fluctuations, we compared the FEP of CP13*T4L at $\gamma = 1$ ps$^{-1}$ using only the first or the second half of the trajectory (Figure S5). Differences between the FEPs is of the same order of magnitude as the discrepancies observed on Figure S4. A careful reader may notice difference in populations of native states. It is due to a decrease of proportion of time spent by the protein in the native basin and in the rest of the profile. The native basin was sampled with 10,000 frames for every friction coefficient, while the sampling time of the non-native fraction of the conformation space was increasing with increasing friction.

From the projection onto the optimal reaction coordinate, we calculated the MFPT of the events D$\rightarrow$N ($\tau_{D\rightarrow N}$), D$\rightarrow$I ($\tau_{D\rightarrow I}$) and I$\rightarrow$D ($\tau_{I\rightarrow N}$) as:

$$\tau_{D\rightarrow N} = \frac{1}{n} \sum_{j=1}^{n} t_{j}^{N}$$

$$\tau_{D\rightarrow I} = \frac{1}{n} \sum_{j=1}^{n} t_{j}^{I}$$

$$\tau_{I\rightarrow N} = \tau_{D\rightarrow N} - \tau_{D\rightarrow I}$$

where $t_{j}^{X}$ is time when the trajectory $j$ reaches the bottom of the basin $X$ and $n$ is the total number of trajectories. The Figure S6 shows that MFPT
Figure S5: FEP of WT*T4L at $\gamma = 1 \text{ ps}^{-1}$ along the optimal reaction coordinate built with either all the frames of the trajectory (black) or the first half of the trajectory (cyan) or the second half of the trajectory (purple). The fluctuation due to the statistic have the same order of magnitude as the fluctuations observed for different frictions on Figure S4.

increases linearly with the friction in the studied interval, confirming that change in friction does not affect the folding mechanisms. The fits do not converge to 0 at friction 0 due to the internal friction of the protein [17, 18], which is independent of $\gamma$. Assuming that linearity holds at higher friction of 50 ps$^{-1}$ (water), the MFPTs of WT*T4L and CP13*T4L were estimated as 9.8 $\mu$s and 24.9 $\mu$s, respectively. These actual MFPTs have been measured experimentally as 54 ms for WT*T4L and 45 ms for CP13*T4L [19]. The time-scale from the real proteins cannot be compared directly with the time-scale from the simulations because the model used is coarse-grained without explicit solvent (comparison of time-scales from coarse-grained model and experiment is discussed by Kouza et al. in [20]).

Figure S6: Friction dependence of the different mean first passage times (MFPT) of WT*T4L (red line) and CP13*T4L (blue lines). The MFPT have been calculated using the projections of the trajectories onto the optimal reaction coordinate found for $\gamma = 1 \text{ ps}^{-1}$. Error bars correspond to standard errors. This plot confirms that MFPT are linear in the friction in the range $\gamma \in [1,10]$. 
References

1. J. Karanicolas and C. L. Brooks, *J. Mol. Biol.*, 2003, *334*, 309–325.

2. B. R. Brooks, C. L. Brooks, A. D. Mackere, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovechinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schäfer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, *J. Comp. Chem.*, 2009, *30*, 1645–1614.

3. T. Lazaridis and M. Karplus, *Proteins*, 1999, *35*, 133–152.

4. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, *J. Comp. Chem.*, 1992, *13*(8), 1011–1021.

5. S. V. Krivov and M. Karplus, *Proc. Natl. Acad. Sci. USA*, 2004, *101*, 14766–14770.

6. S. V. Krivov and M. Karplus, *J. Phys. Chem. B*, 2006, *110*, 12689–12698.

7. S. V. Krivov and M. Karplus, *Proc. Natl. Acad. Sci. USA*, 2008, *105*, 13841–13846.

8. A. Ma and A. R. Dinner, *J. Phys. Chem. B*, 2005, *109*, 6769–6779.

9. R. B. Best and G. Hummer, *Proc. Natl. Acad. Sci. USA*, 2005, *102*, 6732–6737.

10. L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, *J. Chem. Phys.*, 2006, *125*, 24106.

11. S. V. Krivov, *J. Phys. Chem. B*, 2011, *115*, 11382–11388.

12. S. V. Krivov, *J. Phys. Chem. B*, 2011, *115*, 12315–12324.

13. S. V. Krivov, *PLoS Comput. Biol.*, 2010, *6*, e1000921.

14. S. V. Krivov, *Phys. Rev. E*, 2011, *84*, 011135.

15. A. Andricioaei and M. Karplus, *J. Chem. Phys.*, 2001, *115*, 6289–6292.

16. H. Kato, N. Vu, H. Feng, Z. Zhou, and Y. Bai, *J. Mol. Biol.*, 2007, *365*, 870–880.

17. R. B. Best and G. Hummer, *Proc. Natl. Acad. Sci. USA*, 2010, *107*, 1088–1093.

18. Stephen J. Hagen, *Curr. Protein Pept. Sci.*, 2010, *11*, 385–395.

19. M. Sagermann, W. A. Baase, B. H. M. Mooers, L. Gay, and B. W Matthews, *Biochemistry*, 2004, *43*, 1296–1301.

20. M. Kouza, M. S. Li, E. O’Brien, C.-K. Hu, and D. Thirumalai, *J. Phys. Chem.*, 2006, *110*, 671–676.