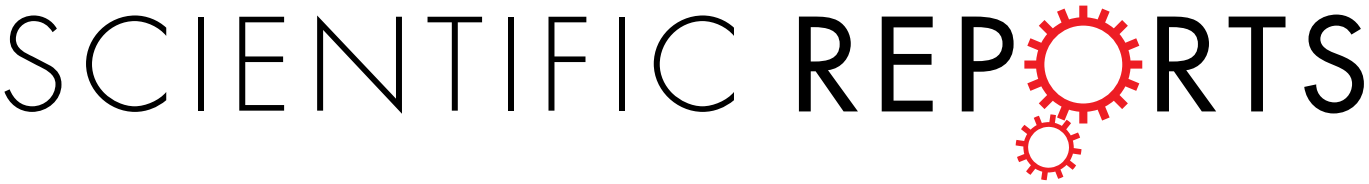

OPEN

PFDB: A standardized protein folding database with temperature correction

Received: 1 February 2018
Accepted: 22 November 2018
Published online: 07 February 2019

Balachandran Manavalan[1], Kunihiro Kuwajima[1,2,3] & Jooyoung Lee[1]

We constructed a standardized protein folding kinetics database (PFDB) in which the logarithmic rate constants of all listed proteins are calculated at the standard temperature (25 °C). A temperature correction based on the Eyring–Kramers equation was introduced for proteins whose folding kinetics were originally measured at temperatures other than 25 °C. We verified the temperature correction by comparing the logarithmic rate constants predicted and experimentally observed at 25 °C for 14 different proteins, and the results demonstrated improvement of the quality of the database. PFDB consists of 141 (89 two-state and 52 non-two-state) single-domain globular proteins, which has the largest number among the currently available databases of protein folding kinetics. PFDB is thus intended to be used as a standard for developing and testing future predictive and theoretical studies of protein folding. PFDB can be accessed from the following link: http://lee.kias.re.kr/~bala/PFDB.

Protein folding is one of the most difficult problems in biophysics and molecular biology. Due to the accumulation of over half a century's experimental data on reversible folding-unfolding mechanisms[1,2], at least 16 protein folding kinetics datasets have been reported[3–19]. However, there are many problems in these datasets, including variations in temperatures (from 5 °C to 75 °C) used in kinetic folding experiments, redundant data entries, and inadequate reported data. A more complete dataset of protein folding kinetics with corrections for the above problems is thus required, and once we have such a dataset, it will be very useful for developing and testing future predictive and theoretical studies of protein folding.

Here, we thus carefully examined the existing protein folding datasets, and introduced the necessary corrections. Among the available datasets, ACPro[19] and the dataset by Garbuzynskiy *et al.*[17] (hereinafter referred to as the Garbuzynskiy dataset) were the most recent ones, which contained the most updated and largest entries. Therefore, we utilized these two datasets in the current study to construct a new database called PFDB. Furthermore, we added new protein data into the PFDB from our own collection based on extensive literature search, which resulted in the entry size of 141 globular proteins in our dataset; whose size is the biggest among the currently available protein folding datasets.

In this study, we also developed a new temperature correction method for the proteins whose kinetic folding and unfolding experiments had been carried out at a temperature different from the standard temperature (25 °C). Our temperature correction method is based on the Eyring–Kramers equation[20], and the logarithmic rate constants of folding and unfolding, $\ln(k_f)$ and $\ln(k_u)$, respectively, at 25 °C is provided for all proteins in PFDB. Interestingly, the present study is the first to introduce the temperature corrections into the protein folding dataset, and we show that the introduction of the temperature correction has improved the quality of the database. PFDB is thus currently the most updated database of protein folding kinetics, and hence it can be used as a standard for developing future predictive and theoretical studies of protein folding.

Results and Discussions

Database construction and descriptions. We first combined the two most recent datasets of protein folding, the ACPro and Garbuzynskiy datasets, to construct the combined dataset (hereafter called "the AG dataset") in which redundant or inappropriate entries were filtered out. We excluded the proteins containing disulfide linkages or covalently bound prosthetic groups, because the presence of these linkages or groups can significantly affect the folding kinetics. Small polypeptides with less than 34 residues were also excluded. We

[1]School of Computational Sciences, Korea Institute for Advanced Study (KIAS), Seoul, Korea. [2]CPIS, the Graduate University for Advanced Studies (Sokendai), Hayama, Japan. [3]Department of Physics, School of Science, the University of Tokyo, Tokyo, Japan. Correspondence and requests for materials should be addressed to K.K. (email: kuwajima@ims.ac.jp) or J.L. (email: jlee@kias.re.kr)

PFDB: A standardized protein folding database with temperature correction

HOME N2S 2S DOWNLOAD DATASET CONTACT

							Present dataset									AG dataset				
No.	Protein short name	PDB	Class	Fold	L_{pdb}	L	pH	Temp (°C)	Folding type	$\ln(k_f)$	$\ln(k_f)$ (25°C)	$\ln(k_l)$	$\ln(k_u)$	$\ln(k_u)$ (25°C)	β_T	pH	Temp (°C)	Folding type	$\ln(k_f)$	Comments
1	Apomyoglobin (Whale) [1]	1A6N	α	Globin-like	151	153	6.2	5	N2S	1.1	4.5	NA	-3.8	5.5	0.72	–	–	–	–	
2	Pit1 [2]	1AU7 (103–160)	α	DNA/RNA-binding 3-helical bundle	58	63	5.5	25.0	N2S	9.7		12.6	5.5		0.74	–	–	–	–	
3	4-helix bundle protein FRB [3]	1AUE (Chain B: 2022–2115)	α	Four-helical up-and-down bundle	94	95	7.5	10	N2S	5.4	6.7	NA	-5.2	-1.0	0.83	– –	– –	–	– 6.0	Both the AG and our datasets adopted the same reference [3]. The ACPro and our datasets reported the same $\ln(k_f)$ value, but the Garbuzynskiy dataset reported a different value.
4	IM7 [4]	1AYI (1–86)	α	HIV-1 gp41 fragments	86	86	7.0	10	N2S	5.7	6.9	8.0	-0.84	3.0	0.90	–	25	2S	7.2	The $\ln(k_f)$ value reported in the AG dataset was based on the 2S model [5]. However, the N2S nature of this protein is well established [4], so that our reported value is based on the N2S model.
5	Apomyoglobin (Horse) [6]	1DWR (1–152)	α	Globin-like	152	153	6.0	26	N2S	2.9	2.9	5.3	NA		NA	NA	Na	NA	NA	The value of $\ln(k_f)$ was taken from the rate constant of the fast phase of the bi-phasic refolding kinetics reported.
6	Engrailed Homeodomain [7]	1ENH	α	DNA/RNA-binding 3-helical bundle	54	54	5.7	25.0	N2S	10.6		NA	7.6		0.83	–	–	–	–	
7	FF domain from human HYPA/FBP11 [8]	1UZC (3–71)	α	3-Helical bundle	69	71	5.7	25	N2S	8.0		9.9	3.4		0.91	– –	– 10	–	– 7.6	The same experimental group reported the $\ln(k_f)$ values at 25°C [8] and at 10°C [9]. The ACPro and our datasets reported the value at 25°C, but the Garbuzynskiy dataset reported the value at 10°C.

Figure 1. A snapshot of our dataset in the PFDB homepage. For each protein, our dataset lists (i) protein short name, (ii) PDB code, (iii) structural class (α, β, α/β, and α + β), (iv) folds in the SCOP classification, (v) the number of residues in the PDB structure (L_{PDB}), (vi) the actual number of residues of the protein used in the folding experiment (L), (vii) experimental conditions (pH and temperature), (viii) folding type (2S or N2S), (ix) $\ln(k_f)$ reported, (x) $\ln(k_f)$ after temperature correction, (xi) $\ln(k_l)$ (only for N2S proteins), (xii) $\ln(k_u)$ reported, (xiii) $\ln(k_u)$ after temperature correction, and (xiv) Tanford β (β_T). The AG dataset is also included in our database for comparison. A comment section is provided in the final column.

carefully examined each data in the AG dataset. For instance, if there is no updated protein folding kinetics data available for a protein, we included those proteins as such in PFDB, otherwise replaced with the updated data. Furthermore, we added the data of 33 new proteins into the PFDB from our own collection based on extensive literature search, resulting in the entry size of 141 globular proteins (89 two-state (2S) and 52 non-two-state (N2S) proteins) in our dataset (see Methods for details of the database construction).

Our dataset lists the following items: (i) the protein short name with a reference to the original experimental paper(s) on the folding kinetics, (ii) the PDB code, (iii) the structural class (α, β, α/β, and α + β), (iv) folds in the SCOP classification[21] (http://scop.mrc-lmb.cam.ac.uk/scop/), (v) the number of residues in the PDB structure (L_{PDB}), (vi) the actual number of residues of the protein used in the folding experiment (L), (vii) the experimental conditions (pH and temperature), (viii) the folding type (2S or N2S), (ix) the $\ln(k_f)$ value reported, (x) the $\ln(k_f)$ value after the temperature correction for the proteins whose folding experiments were carried out at a temperature other than 25 °C, (xi) the logarithmic rate constant of formation of a folding intermediate, $\ln(k_l)$, when the value is available in the literature (only for N2S proteins), (xii) the $\ln(k_u)$ value reported, (xiii) the $\ln(k_u)$ value after the temperature correction, and (xiv) the Tanford β (β_T) value, which is defined as $\beta_T = 1 - (m_u^{\ddagger}/m_{NU})$, where $m_u^{\ddagger}$ (kJ/mol/M) and m_{NU} (kJ/mol/M) are the denaturant concentration dependence of the activation free energy of unfolding and the denaturant concentration dependence of the unfolding free energy from the native (N) to the fully unfolded (U) state, respectively[22]. The $\ln(k_f)$, $\ln(k_l)$ and $\ln(k_u)$ values listed in PFDB are those in the absence of denaturant, usually obtained by linear extrapolation of the logarithmic rate constant along denaturant concentration.

In PFDB, the folding type is thus clearly specified. The proteins that exhibited a stable folding intermediate during the kinetic folding process were classified as N2S proteins, while the proteins, exhibiting the single-exponential kinetics of folding without stable intermediates, were classified as 2S proteins even if the existence of an unstable high-energy intermediate was expected from the unfolding-limb or the folding-limb curvature of the chevron plot[23]. To discriminate the 2S proteins with a high-energy intermediate from the other 2S proteins, the former proteins were denoted by 2S*. Each entry of the AG dataset is also included in PFDB for comparison. A comment section is provided in the final column of the dataset and interprets discrepancies between the present and the AG datasets if any/necessary. Figure 1 depicts a snapshot of our dataset shown in the PFDB homepage.

The protein composition in PFDB in terms of the folding type and the structural class is given in Table 1. It shows that both the 2S and N2S proteins cover all four structural classes of globular proteins. However, the 2S proteins contain only one α/β protein.

Temperature correction. Figure 2A shows a distribution of the temperature at which the $\ln(k_f)$ was determined experimentally for the proteins in our dataset. Among the 141 proteins in PFDB, 99 were measured at the standard temperature of T_0 (25 °C (=298.15 K)), but the other 42 (24 2S and 18 N2S proteins) were measured at different temperatures (T_x). The T_x value ranged from 5 °C to 75 °C. To maintain the consistency of folding temperature in PFDB, we developed a method for temperature correction. The predicted shape of the Eyring plot of a particular protein is determined by two parameters of the folding or unfolding reaction, the activation heat

Folding type	Structural class				
	α	β	α+β	α/β	Total
2S	24	39	25	1	89
N2S	10	13	16	13	52
Total	34	52	41	14	141

Table 1. The composition of the PFDB in terms of structural and folding class is shown.

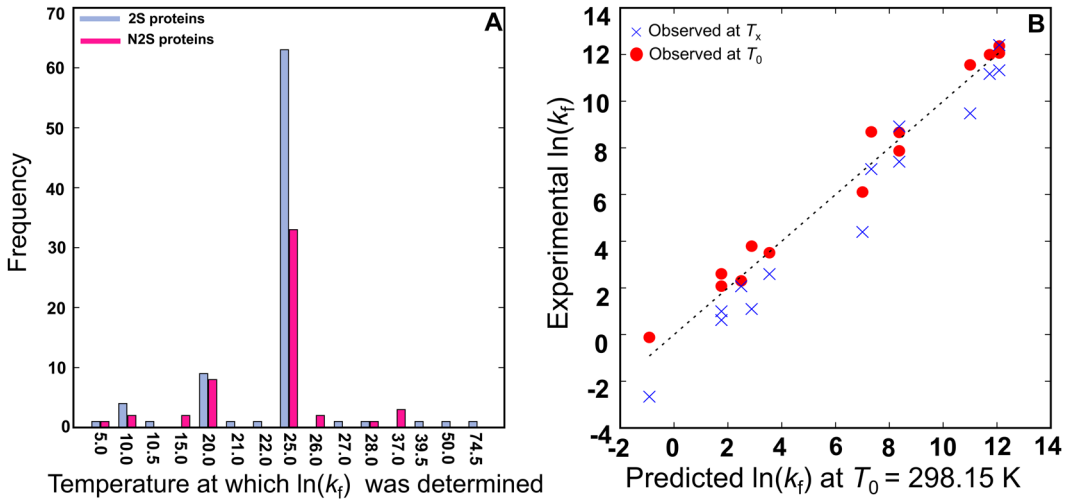

Figure 2. (**A**) The temperature at which $\ln(k_f)$ experimentally determined for 2S and N2S is shown. (**B**) Experimentally observed $\ln[k_f(T_0)]$ and predicted ones after temperature correction (red circles) are shown. Observed $\ln[k_f(T_x)]$ values are also shown for comparison (blue crosses).

capacity ($\Delta C_p^{\ddagger}$) and the temperature (T_H) where the activation enthalpy is zero (see Methods for more details). The predicted logarithmic rate constant at T_0 (298.15 K) is thus given by the following equation:

$$\ln[k(T_0)] = \ln[k(T_x)] + \left[1 + \frac{\Delta C_p^{\ddagger}}{R}\right]\ln\left(\frac{T_0}{T_x}\right) + \frac{\Delta C_p^{\ddagger}}{R}\left[\left(\frac{1}{T_0} - \frac{1}{T_x}\right) \cdot T_H\right] \quad (1)$$

where R is the gas constant, T_0 and T_x are given by the absolute temperature, and $\ln[k(T_x)]$ is the logarithmic rate constant measured at T_x; the detailed derivation of Eq. (1) is given in Methods. We assumed that $\Delta C_p^{\ddagger}$ is proportional to the heat capacity change (ΔC_p) of the equilibrium protein unfolding. The ΔC_p is approximately proportional to the protein chain length in the PDB structure (L_{PDB}) and empirically given by[24]:

$$\Delta C_p = 0.062 \cdot L_{PDB} - 0.53 \text{ [kJ/mol/K]} \quad (2)$$

Now, it follows that:

$$\Delta C_p^{\ddagger} = \beta \cdot \Delta Cp = \beta(0.062 \cdot L_{PDB} - 0.53) \text{ [kJ/mol/}K\text{]} \quad (3)$$

where β is a proportionality constant. Therefore, once we have reasonable estimates of T_H and β, we can evaluate $\ln[k(T_0)]$ from $\ln[k(T_x)]$ and T_x by Eqs (1) and (3). It is worth mentioning that Eq. 2 is an empirical one, and theoretically, the ΔC_p diminishes to zero when L_{PDB} tends to zero. A regression equation between ΔC_p and L_{PDB} with the zero intercept has thus also been reported in the original literature as given by $\Delta Cp = 0.058 \cdot L_{PDB}$[24]. Whether we used this equation or Eq. 2, the results of temperature correction were essentially identical for the proteins in our dataset, where $L_{PDB} \geq 34$.

Temperature correction for folding. We introduced the temperature corrections into the proteins whose k_f values were measured at a temperature other than the standard temperature (298.15 K). First, we found that the Eyring plot or the equivalent plot of folding was well described in 14 2S proteins and 3 N2S proteins; the k_f values were measured at every few degrees absolute from ~280 K to ~320 K for most of these proteins[25–41]. Both the T_H and β values for folding kinetics, T_{Hf} and β_f, respectively, were more or less common among the different 2S proteins (Table 2) and also among the different N2S proteins (Table 3), except for two 2S proteins (1K9Q[40] and 1PIN[41]), for which $-\Delta C_p^{\ddagger}$ for folding was larger than ΔC_p. Therefore, we employed the 12 2S proteins except for these two and the 3 N2S proteins, and from their Eyring plots, we calculated the T_{Hf} and $\Delta C_{pf}^{\ddagger}$. Examples of the Eyring plot for three proteins (1APS[34], 1D6O[35], and 1AVZ[37]) are shown in Figure S1. For folding kinetics, the Eyring

PDB	L_{PDB}	Temp. (K)	$\Delta H^{\ddagger}$ (kJ/mol)	$\Delta C_{pf}^{\ddagger}$ (kJ/mol/K)	T_{Hf} (K)	ΔC_p (kJ/mol/K)	β_f
1APS[34]	98	301.15	40.70	−2.57	316.99	5.55	−0.46
1D6O[35]	107	298.15	48.53	−2.80	315.46	6.10	−0.46
1E0G[28]	48	298.15	28.45	−1.76	314.34	2.45	−0.72
1HDN[30]	85	293.15	86.10	−3.22	319.89	4.74	−0.68
2VH7[29]	94	301.15	23.60	−2.48	310.67	5.30	−0.47
3CI2[39]	64	298.00	53.55	−2.05	324.12	3.44	−0.60
1EHB[62]	82	298.15	42.40	−3.60	309.93	4.55	−0.79
1CSP[38]	67	298.15	31.60	−2.70	309.85	3.62	−0.74
1AVZ[37]	57	293.00	43.09	−1.86	316.20	3.00	−0.62
1SHG[36]	57	298.00	37.00	−2.30	314.09	3.00	−0.77
1HCD[31]	118	293.15	57.74	−4.39	306.29	6.79	−0.65
2JMC[25]	77	298.15	45.00	−2.20	318.60	4.24	−0.52
Mean ± SE					314.70 ± 1.44		−0.62 ± 0.03

Table 2. List of proteins used to estimate T_{Hf} and β_f for two-state proteins.

PDB	L_{PDB}	Temp. (K)	$\Delta H^{\ddagger}$ (kJ/mol)	$\Delta C_{pf}^{\ddagger}$ (kJ/mol/K)	T_{Hf} (K)	ΔC_p (kJ/mol/K)	β_f
2CRO[26]	65	293.15	40.70	−3.05	310.50	3.50	−0.87
1PGB[32]	56	298.15	16.80	−1.90	306.99	2.94	−0.64
1L63[33]	162	285.15	92.05	−6.84	298.61	9.51	−0.72
Mean ± SE					305.369 ± 3.526		−0.746 ± 0.067

Table 3. List of proteins used to estimate T_{Hf} and β_f for non-two-state proteins.

plot is convexed, and hence, T_{Hf} corresponds to the temperature of the maximum point in the Eyring plot. The $\Delta C_{pf}^{\ddagger}$ is given by the curvature of the Eyring plot, and the β_f was thus evaluated by $\beta_f = \Delta C_{pf}^{\ddagger}/\Delta C_p$, where ΔC_p was obtained by Eq. (2); $\Delta C_{pf}^{\ddagger}$ and β_f are negative because the Eyring plot is convexed. The T_{Hf} and β_f values thus obtained were averaged for the 12 2S proteins and for the 3 N2S proteins (Tables 2 and 3). The T_{Hf} and β_f values thus obtained are 315 ± 1 (standard error estimate) K and −0.62 ± 0.03 for the 2S proteins, and 305 ± 4 K and −0.75 ± 0.07 for the N2S proteins.

For the proteins whose T_{Hf} and $\Delta C_{pf}^{\ddagger}$ were not available directly, we employed Eqs (1) and (3) to predict $\ln[k_f(T_0)]$ by assigning the T_{Hf} and β_f values to T_H and β in the equations. However, for the proteins whose T_{Hf} and $\Delta C_{pf}^{\ddagger}$ were available (1E0G[28], 1HDN[30], 2VH7[29], 1EHB[27], 1HCD[31], and 2CRO[26]), we directly calculated the $\ln[k_f(T_0)]$ values by Eq. (1). To distinguish $\ln[k_f(T_0)]$ predicted by using the averaged T_{Hf} and β_f and that directly calculated by Eq. (1) with the known T_{Hf} and $\Delta C_{pf}^{\ddagger}$, the latter values are indicated in boldface type in our dataset. It should be also noted that the above T_{Hf} and β_f estimates were based on the folding data of the proteins from mesophilic organisms, and hence some care may be required when applied to the thermophilic proteins.

Next, we compared predicted $\ln[k_f(T_0)]$ after the temperature correction with the experimentally observed $\ln[k_f(T_0)]$. For 9 2S and 5 N2S proteins (Table 4), which were not included in those used for estimating T_{Hf} and β_f, the experimental $\ln(k_f)$ was available at both T_0 and T_x. We thus applied the temperature correction to the $\ln[k_f(T_x)]$ values using the above T_{Hf} and β_f, and compared predicted $\ln[k_f(T_0)]$ with the experimentally observed $\ln[k_f(T_0)]$. From Fig. 2B, the predicted $\ln[k_f(T_0)]$ values show good agreement with the experimentally observed ones, showing the validity of our temperature correction. Although the number of data points used for this analysis is not very large (only 14 proteins), it may be enough to suggest that the temperature corrections have improved the quality of the database of protein folding.

Denaturant *m* values, the dependence of the free energy of unfolding on denaturant concentration, are well correlated with the ΔC_p of unfolding[42]. Therefore, we can reasonably assume that β_f is equivalent to $-\beta_T$ for 2S proteins. Therefore, for the 2S proteins for which the β_T is available, we also calculated the $\ln[k_f(T_0)]$ values by assigning the T_{Hf} and $-\beta_T$ values to T_H and β in Eqs (1) and (3). The $\ln[k_f(T_0)]$ values thus obtained are also listed in PFDB and indicated in italic type to distinguish them from those (in roman type) predicted on the basis of T_{Hf} and β_f. As seen from the PFDB dataset, these two types of predicted $\ln[k_f(T_0)]$ are reasonably coincident with each other.

Temperature correction for unfolding. We introduced the temperature corrections into the proteins whose k_u values were measured at a temperature other than the standard temperature (298.15 K), and the T_H and β values for unfolding kinetics, T_{Hu} and β_u, respectively, were required for temperature correction. For unfolding kinetics, the Eyring plot is usually concaved with a positive β_u. For 2S proteins, there is only a single transition state between U and N with a β_f of −0.62 ± 0.03, and we can reasonably assume that $\beta_u = 1 + \beta_f$. Therefore, we find that $\beta_u = 0.38 \pm 0.03$. For N2S proteins, this simple relationship may not hold, because of a contribution from an intermediate (I) state. For the N2S proteins, however, $(1 - \beta_T)$ is expected to be equivalent to β_u, because β_T represents the relative position of the transition state between U and N in terms of the denaturant *m* values. The β_T was

PDB	$\ln[k_f(T_x)]$	T_x (K)	$\ln[k_f(T_0)]$ observed	$\ln[k_f(T_0)]$ predicted
1FNF[63]	−2.66	278.15	−0.92	−0.12
1IMQ[13,43]	7.09	283.15	7.33	8.69
1K9Q[40,44]	8.92	311.15	8.37	8.67
1K9Q[40,44]	7.41	351.15	8.37	7.87
1RFA[45]	4.40	281.15	7.00	6.11
1SS1[46]	12.41	323.15	12.08	12.07
1SS1[46]	11.33	283.15	12.08	12.37
1U4Q[47,48]	9.48	283.15	11.00	11.56
2WXC[49,50]	11.17	283	11.73	12.00
1BNI[51]	**2.07**	**318.15**	**2.50**	**2.31**
1DWR*[64,65]	**1.10**	**281.15**	**2.88**	**3.79**
1NFI[66]	**1.00**	**288.15**	**1.76**	**2.08**
1NFI[66]	**0.62**	**283.15**	**1.76**	**2.60**
1EKG[52]	**2.60**	**288.15**	**3.54**	**3.51**

Table 4. List of Proteins used for predicting $\ln(k_f)$ at 25 °C. *T_0 for 1DWR was 299.15 K (26 °C). Normal font and bold, respectively, represent the 2S and N2S proteins.

PDB	$\ln[k_u(T_x)]$	T_x (K)	$\ln[k_u(T_0)]$ observed	$\ln[k_u(T_0)]$ predicted
1IMQ[13,43]	−4.42	283.15	−1.87	−1.79
1K9Q[40,44]	10.92	351.15	6.66	6.30
1K9Q[40,44]	7.38	311.15	6.66	6.33
1RFA[45]	−3.10	281.15	−1.17	−0.45
1SS1[46]	7.40	323.15	3.40	4.20
1SS1[46]	0.92	283.15	3.40	2.61
1U4Q[47,48]	−3.37	298.15	0.26	0.06
2WXC[49,50]	6.65	283	7.65	7.98
1BNI[51]	**−3.13**	**318.15**	**−10.55**	**−9.51**
1EKG[52]	**−11.02**	**288.15**	**−8.87**	**−7.42**
1ENH[53]	**10.78**	**325.3**	**7.00**	**6.79**

Table 5. List of proteins used for predicting $\ln(k_u)$ at 25 °C. Normal font and bold, respectively, represent the 2S and N2S proteins.

reported for 38 N2S proteins in PFDB, and their average was estimated at 0.79 ± 0.02, and hence $\beta_u = 0.21 \pm 0.02$ for N2S proteins; 1FTG was excluded in this calculation because the I state was mostly off-pathway in this protein.

The T_{Hu} corresponds to the temperature of the minimum point of the Eyring plot, but this is usually located at far below an observable temperature range of unfolding kinetics, leading to a large error in estimation of T_{Hu} due to a long extrapolation along temperature. Furthermore, the Eyring plot of unfolding is not available for many of the proteins used above for estimation of T_{Hf} and β_f. Therefore, we had to use a different way to estimate T_{Hu}. We thus chose 6 2S proteins (1IMQ[13,43],1K9Q[40,44], 1RFA[45], 1SS1[46], 1U4Q[47,48], and 2WXC[49,50]) and 3 N2S proteins (1BNI[51], 1EKG[52], and 1ENH[53]), for which the experimental $\ln(k_u)$ is available at both T_0 and T_x (Table 5). First, we assumed appropriate T_{Hu} values (e.g., 200 K and 150 K) for 2S and N2S proteins, and assigned these T_{Hu} values and the above β_u values to T_H and β in Eqs (1) and (3) to calculated tentative predictions of $\ln[k_u(T_0)]$ for 2S and N2S proteins. Then, the T_{Hu} values were gradually increased or decreased until the root-mean-square deviation between the experimentally observed $\ln[k_u(T_0)]$ and the predicted $\ln[k_u(T_0)]$ values was minimized. The optimized T_{Hu} values thus obtained were 224 K and 119 K for the 2S and N2S proteins, respectively. Figure 3 shows a comparison between the experimental $\ln[k_u(T_0)]$ values and those predicted by using the above T_{Hu} and β_u values, which indicates a reasonable coincidence between the experimental and predicted values.

For the proteins whose T_{Hu} and $\Delta C_{pu}^{\ddagger}$ were not available directly, we thus employed Eqs (1) and (3) to predict the $\ln[k_u(T_0)]$ by assigning the T_{Hu} and β_u values to T_H and β in the equations. However, for the proteins whose T_{Hu} and $\Delta C_{pu}^{\ddagger}$ were available (1EHB[27] and 1HCD[31]), we directly calculated the $\ln[k_u(T_0)]$ values by Eq. (1). To distinguish the $\ln[k_u(T_0)]$ predicted by using the optimized T_{Hu} and β_u and that directly calculated by Eq. (1) with the known T_{Hu} and $\Delta C_{pu}^{\ddagger}$, the latter values are indicated in boldface type in our dataset.

For the 2S proteins for which the β_T is available, we also calculated the $\ln[k_u(T_0)]$ values by assigning the T_{Hu} and $(1 - \beta_T)$ values to T_H and β in Eqs (1) and (3). The $\ln[k_u(T_0)]$ values thus obtained are also listed in PFDB and indicated in italic type to distinguish them from those (in roman type) predicted on the basis of T_{Hu} and β_u. As seen from the PFDB dataset, these two types of predicted $\ln[k_u(T_0)]$ are reasonably coincident with each other.

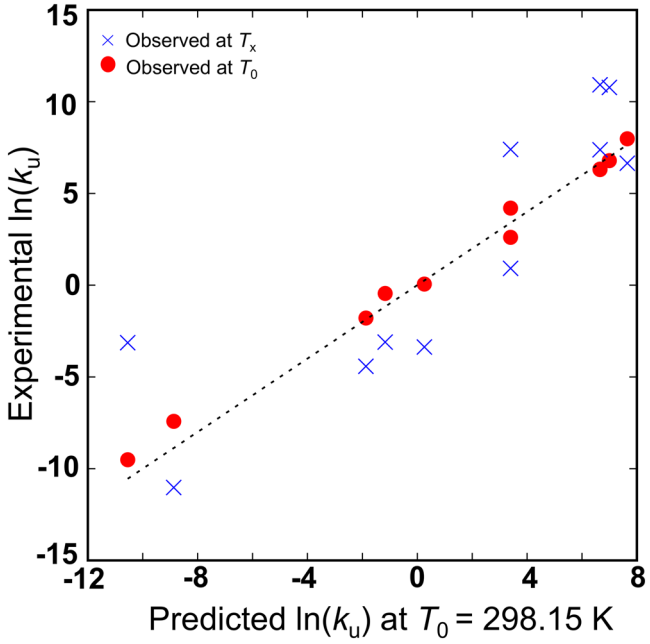

Figure 3. Experimentally observed $\ln[k_u(T_0)]$ and predicted ones after temperature correction (red circles) are shown. Observed $\ln[k_u(T_x)]$ values are also shown for comparison (blue crosses).

Availability of PFDB. As a user-friendly database, PFDB is freely available at http://lee.kias.re.kr/~bala/PFDB. The database main page contains the following options: HOME, N2S, 2S, DOWNLOAD DATASET, and CONTACT. Our dataset can be downloaded by clicking the "DOWNLOAD DATASET" button.

Conclusions

In this study, we have constructed PFDB, a systematically compiled standardized database of protein folding kinetics. It is currently the most updated one with the highest number of unique entries. The quality of the dataset has been improved significantly by our temperature correction method. Therefore, our dataset can be used as a standard for developing and testing future predictive and theoretical studies of protein folding kinetics.

Methods

Construction of the AG dataset. The most recent datasets of protein folding kinetics are ACPro[19] and the Garbuzynskiy dataset[17]. Prior to the filtering processes shown below, the ACPro dataset contained 126 proteins. Among these, we weeded out proteins with less than 34 residues (1PGB (41–56), 1L2Y and 3M48), proteins with disulfide bonds (2HQI, 1HEL, 1E65 and 1HMK), proteins with a covalently-bound prosthetic group (1YCC, 1YEA, 256B and 1HRC), proteins with irrelevant rate constants (*i.e.*, the rate constant for formation of an intermediate instead of the actual folding rate constant (k_f) for a few proteins (1AON, 1BD8 and 1JON)), and proteins whose k_f was reported in the presence of denaturant (1QOP chain B). In the case of ileal lipid binding protein, the actual folding experiment was performed on the rat protein, but its PDB coordinates were not available at the time of our database creation. Instead, the reported PDB ID of 1EAL is the pig protein that is of 71.1% sequence identity with the rat protein. Since the exact PDB coordinates were not available, we excluded this protein as well as another protein without experimental references (1PSF). Furthermore, 6 proteins had duplicate entries (1NTI–2FDQ, 1SRL–1FMK, 1BF4–1BNZ, 1POH–2HPR, 1O6X–1PBA and 1EAL–2EAL) which we corrected. These filtering processes resulted in the reduction of the size of the ACPro dataset from 126 to 102 proteins. We then applied the same filtering scheme to the Garbuzynskiy dataset (107 proteins) where we weeded out proteins with less than 34 residues (1L2Y, 1T8J, 1PGB (41–56), and the 3rd entry in the Garbuzynskiy dataset), proteins with irrelevant rate constants (1AON and 1BD8), the protein 1EAL (the reason is given above), and a protein with a covalently-bound prosthetic group (256B). This change reduced the size of the Garbuzynskiy dataset from 107 to 99 proteins. When we compared the updated Garbuzynskiy (99 proteins) and ACPro (102 proteins) datasets, 6 unique proteins (1IFC, 1CBI, 1IGS, 1OPA, 2MYO and 3H08) were identified in the Garbuzynskiy dataset. Therefore, we added these 6 proteins to the ACPro dataset, and collectively named it the AG dataset (108 proteins).

Data collection and construction of PFDB. We manually collected the data of protein folding and unfolding kinetics by extensive literature search. Then we compared our collected data with those of the AG dataset. We carefully examined the data of each entry of the AG dataset, and when newer updated data did not exist, the data of that entry were included as such in our dataset of PFDB, otherwise replaced by the updated data. Finally, we added the data of 33 new proteins into the PFDB from our own collection. Of these 33 proteins, 19 are 2S proteins (1DKT, 1FGA, 1IO2, 1KDX, 1NFI,1QAU, 1RG8, 2BKF, 2GA5, 2J5A, 2JMC, 2LLH, 2L6R, 2WQG,

3O48, 3O49, 3O4D, 3ZRT (N-terminal), and 3ZRT (C-terminal)) with the remaining 14 being N2S proteins (1DWR, 1EKG, 1FA3, 1HRH, 1OKS, 1THF, 1UCH, 2BJD, 2FS6, 2KDI, 2KLL, 2X7Z, 3BLM, and 5L8I).

For 4 proteins (1RA9, 1B9C, 1FA3, and 2PQE), the presence of multiple parallel pathways of folding has been reported[54–56], and the k_f value was obtained by averaging the rate constant values along the individual pathways:

$$k_f = \sum_{i=1}^{n} f_i k_i \tag{4}$$

where f_i and k_i are the fractional amplitude and the observed rate constant, respectively, for the i^{th} pathway of folding, and the $\ln(k_f)$ values thus obtained are listed in our dataset.

The $\ln(k_f)$, $\ln(k_I)$ and $\ln(k_u)$ values listed in PFDB are those in the absence of denaturant, usually obtained by linear extrapolation of the logarithmic rate constants along molar denaturant concentration. However, for 5 N2S proteins (1PHP (1–175)[57], 1PHP (186–394)[58], 1L63[59], 1HNG[60], and 1TTG[61]), the equilibria and kinetics of folding and unfolding were analyzed in terms of denaturant activity rather than the molar concentration. Whether we use the activity or the concentration in our calculation seriously affects the $\ln(k_u)$ estimation, because a long extrapolation from high concentrations of denaturant back to the native condition is required. To keep consistency of our dataset, we used the linear extrapolation along the molar concentration, as far as such data were available, to estimate the $\ln(k_u)$.

Derivation of Eq (1) for the temperature correction. In this study, we introduced a method for temperature correction, which gives the folding and unfolding rate constants at 25 °C ($k(T_0)$ where $T_0 = 298.15$ K) for a protein whose rate constant at any temperature (T_x) is known. The following section will describe the derivation of Eq. (1).

According to the Eyring–Kramers equation[20], we find that:

$$\ln\left(\frac{k}{T}\right) = C - \frac{1}{RT}\left[\Delta H^{\ddagger}(T_H) - T\Delta S^{\ddagger}(T_H) + \Delta C_p^{\ddagger} \cdot \left\{T - T_H - T \cdot \ln\left(\frac{T}{T_H}\right)\right\}\right] \tag{5}$$

where $\Delta H^{\ddagger}(T_H)$ and $\Delta S^{\ddagger}(T_H)$ are the activation enthalpy and the activation entropy, respectively, at a reference temperature T_H, and $\Delta C_p^{\ddagger}$ is the activation heat capacity; we assume that $\Delta C_p^{\ddagger}$ is a constant independent of temperature (T). When we set T_H to the temperature where $\Delta H^{\ddagger}$ is zero, i.e., the maximum or minimum point of the Eyring plot, Eq. (5) is rewritten as:

$$\ln\left(\frac{k}{T}\right) = C_2 - \frac{\Delta C_p^{\ddagger}}{RT} \cdot \left[T - T_H - T \cdot \ln\left(\frac{T}{T_H}\right)\right] \tag{6}$$

where C_2 is a temperature-independent constant ($C_2 = C + \Delta S^{\ddagger}(T_H)/R$). When $\Delta C_p^{\ddagger}$ and the $\Delta H^{\ddagger}(T_a)$ at a particular temperature (T_a) are known, T_H is simply given by $T_H = [T_a - \Delta H^{\ddagger}(T_a)/\Delta C_p^{\ddagger}]$. From Eq. (6), we can obtain the temperature dependence of $\ln(k/T)$, once we have T_H and $\Delta C_p^{\ddagger}$. The difference in $\ln(k/T)$ between T_0 (=298.15 K) and T_x is thus given by:

$$\ln\left[\frac{k(T_0)}{T_0}\right] - \ln\left[\frac{k(T_x)}{T_x}\right] = \frac{\Delta C_p^{\ddagger}}{R} \cdot \left[\frac{T_H}{T_0} - \frac{T_H}{T_x} + \ln\left(\frac{T_0}{T_x}\right)\right] \tag{7}$$

Therefore, we obtain Eq. (1).

References

1. Dill, K. A. & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **338**, 1042–1046, https://doi.org/10.1126/science.1219021 (2012).
2. Englander, S. W. & Mayne, L. The nature of protein folding pathways. *Proc Natl Acad Sci USA* **111**, 15873–15880, https://doi.org/10.1073/pnas.1411798111 (2014).
3. Bogatyreva, N. S., Osypov, A. A. & Ivankov, D. N. KineticDB: a database of protein folding kinetics. *Nucleic Acids Res* **37**, D342–346, https://doi.org/10.1093/nar/gkn696 (2009).
4. De Sancho, D. & Munoz, V. Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys Chem Chem Phys* **13**, 17030–17043, https://doi.org/10.1039/c1cp20402e (2011).
5. Guo, J. & Rao, N. Predicting protein folding rate from amino acid sequence. *J Bioinform Comput Biol* **9**, 1–13 (2011).
6. Huang, J. T., Cheng, J. P. & Chen, H. Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics. *Proteins* **67**, 12–17, https://doi.org/10.1002/prot.21282 (2007).
7. Huang, J. T., Xing, D. J. & Huang, W. Relationship between protein folding kinetics and amino acid properties. *Amino acids* **43**, 567–572 (2012).
8. Istomin, A. Y., Jacobs, D. J. & Livesay, D. R. On the role of structural class of a protein with two-state folding kinetics in determining correlations between its size, topology, and folding rate. *Protein Sci* **16**, 2564–2569, https://doi.org/10.1110/ps.073124507 (2007).
9. Ivankov, D. N. & Finkelstein, A. V. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc Natl Acad Sci USA* **101**, 8942–8944, https://doi.org/10.1073/pnas.0402659101 (2004).
10. Ivankov, D. N. *et al.* Contact order revisited: influence of protein size on the folding rate. *Protein Sci* **12**, 2057–2062, https://doi.org/10.1110/ps.0302503 (2003).
11. Jung, J., Buglass, A. J. & Lee, E.-K. Topological quantities determining the folding/unfolding rate of two-state folding proteins. *Journal of solution chemistry* **39**, 943–958 (2010).
12. Jung, J., Lee, J. & Moon, H. T. Topological determinants of protein unfolding rates. *Proteins* **58**, 389–395, https://doi.org/10.1002/prot.20324 (2005).

13. Maxwell, K. L. *et al*. Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci* **14**, 602–616, https://doi.org/10.1110/ps.041205405 (2005).
14. Ouyang, Z. & Liang, J. Predicting protein folding rates from geometric contact and amino acid sequence. *Protein Sci* **17**, 1256–1263, https://doi.org/10.1110/ps.034660.108 (2008).
15. Zhou, H. & Zhou, Y. Folding rate prediction using total contact distance. *Biophys J* **82**, 458–463, https://doi.org/10.1016/S0006-3495(02)75410-6 (2002).
16. Zou, T. & Ozkan, S. B. Local and non-local native topologies reveal the underlying folding landscape of proteins. *Phys Biol* **8**, 066011, https://doi.org/10.1088/1478-3975/8/6/066011 (2011).
17. Garbuzynskiy, S. O., Ivankov, D. N., Bogatyreva, N. S. & Finkelstein, A. V. Golden triangle for folding rates of globular proteins. *Proc Natl Acad Sci USA* **110**, 147–150, https://doi.org/10.1073/pnas.1210180110 (2013).
18. Gromiha, M. M., Thangakani, A. M. & Selvaraj, S. FOLD-RATE: prediction of protein folding rates from amino acid sequence. *Nucleic Acids Res* **34**, W70–74, https://doi.org/10.1093/nar/gkl043 (2006).
19. Wagaman, A. S., Coburn, A., Brand-Thomas, I., Dash, B. & Jaswal, S. S. A comprehensive database of verified experimental data on protein folding kinetics. *Protein Sci* **23**, 1808–1812, https://doi.org/10.1002/pro.2551 (2014).
20. Bilsel, O. & Matthews, C. R. Barriers in protein folding reactions. *Adv Protein Chem* **53**, 153–207 (2000).
21. Andreeva, A. *et al*. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* **36**, D419–425, https://doi.org/10.1093/nar/gkm993 (2008).
22. Jackson, S. E. How do small single-domain proteins fold? *Folding and Design* **3**, R81–R91 (1998).
23. Sanchez, I. E. & Kiefhaber, T. Evidence for sequential barriers and obligatory intermediates in apparent two-state protein folding. *J Mol Biol* **325**, 367–376 (2003).
24. Robertson, A. D. & Murphy, K. P. Protein Structure and the Energetics of Protein Stability. *Chem Rev* **97**, 1251–1268 (1997).
25. Candel, A. M., Cobos, E. S., Conejero-Lara, F. & Martinez, J. C. Evaluation of folding co-operativity of a chimeric protein based on the molecular recognition between polyproline ligands and SH3 domains. *Protein Eng Des Sel* **22**, 597–606, https://doi.org/10.1093/protein/gzp041 (2009).
26. Laurents, D. V. *et al*. Folding kinetics of phage 434 Cro protein. *Biochemistry* **39**, 13963–13973 (2000).
27. Manyusa, S. & Whitford, D. Defining folding and unfolding reactions of apocytochrome b 5 using equilibrium and kinetic fluorescence measurements. *Biochemistry* **38**, 9533–9540 (1999).
28. Nickson, A. A., Stoll, K. E. & Clarke, J. Folding of a LysM domain: entropy-enthalpy compensation in the transition state of an ideal two-state folder. *J Mol Biol* **380**, 557–569, https://doi.org/10.1016/j.jmb.2008.05.020 (2008).
29. Taddei, N. *et al*. Thermodynamics and kinetics of folding of common-type acylphosphatase: comparison to the highly homologous muscle isoenzyme. *Biochemistry* **38**, 2135–2142, https://doi.org/10.1021/bi9822630 (1999).
30. Van Nuland, N. A. *et al*. Slow cooperative folding of a small globular protein HPr. *Biochemistry* **37**, 622–637, https://doi.org/10.1021/bi9717946 (1998).
31. Wong, H. J., Stathopulos, P. B., Bonner, J. M., Sawyer, M. & Meiering, E. M. Non-linear effects of temperature and urea on the thermodynamics and kinetics of folding and unfolding of hisactophilin. *J Mol Biol* **344**, 1089–1107, https://doi.org/10.1016/j.jmb.2004.09.091 (2004).
32. Alexander, P., Orban, J. & Bryan, P. Kinetic analysis of folding and unfolding the 56 amino acid IgG-binding domain of streptococcal protein G. *Biochemistry* **31**, 7243–7248 (1992).
33. Chen, B. L., Baase, W. A. & Schellman, J. A. Low-temperature unfolding of a mutant of phage T4 lysozyme. 2. Kinetic investigations. *Biochemistry* **28**, 691–699 (1989).
34. Chiti, F. *et al*. Structural characterization of the transition state for folding of muscle acylphosphatase. *J Mol Biol* **283**, 893–903, https://doi.org/10.1006/jmbi.1998.2010 (1998).
35. Main, E. R., Fulton, K. F. & Jackson, S. E. Folding pathway of FKBP12 and characterisation of the transition state. *J Mol Biol* **291**, 429–444, https://doi.org/10.1006/jmbi.1999.2941 (1999).
36. Martinez, J. C., Pisabarro, M. T. & Serrano, L. Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat Struct Biol* **5**, 721–729, https://doi.org/10.1038/1418 (1998).
37. Plaxco, K. W. *et al*. The folding kinetics and thermodynamics of the Fyn-SH3 domain. *Biochemistry* **37**, 2529–2537, https://doi.org/10.1021/bi972075u (1998).
38. Schindler, T. & Schmid, F. X. Thermodynamic properties of an extremely rapid protein folding reaction. *Biochemistry* **35**, 16833–16842, https://doi.org/10.1021/bi962090j (1996).
39. Tan, Y. J., Oliveberg, M. & Fersht, A. R. Titration properties and thermodynamics of the transition state for folding: comparison of two-state and multi-state folding pathways. *J Mol Biol* **264**, 377–389, https://doi.org/10.1006/jmbi.1996.0647 (1996).
40. Crane, J. C., Koepf, E. K., Kelly, J. W. & Gruebele, M. Mapping the transition state of the WW domain beta-sheet. *J Mol Biol* **298**, 283–292, https://doi.org/10.1006/jmbi.2000.3665 (2000).
41. Jager, M., Nguyen, H., Crane, J. C., Kelly, J. W. & Gruebele, M. The folding mechanism of a beta-sheet: the WW domain. *J Mol Biol* **311**, 373–393, https://doi.org/10.1006/jmbi.2001.4873 (2001).
42. Myers, J. K., Pace, C. N. & Scholtz, J. M. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Science* **4**, 2138–2148 (1995).
43. Friel, C. T., Capaldi, A. P. & Radford, S. E. Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: similarities and differences in the folding of homologous proteins. *J Mol Biol* **326**, 293–305 (2003).
44. Ferguson, N., Johnson, C. M., Macias, M., Oschkinat, H. & Fersht, A. Ultrafast folding of WW domains without structured aromatic clusters in the denatured state. *Proc Natl Acad Sci USA* **98**, 13002–13007 (2001).
45. Vallee-Belisle, A., Turcotte, J. F. & Michnick, S. W. raf RBD and ubiquitin proteins share similar folds, folding rates and mechanisms despite having unrelated amino acid sequences. *Biochemistry* **43**, 8447–8458, https://doi.org/10.1021/bi0359426 (2004).
46. Dimitriadis, G. *et al*. Microsecond folding dynamics of the F13W G29A mutant of the B domain of staphylococcal protein A by laser-induced temperature jump. *Proc Natl Acad Sci USA* **101**, 3809–3814, https://doi.org/10.1073/pnas.0306433101 (2004).
47. Scott, K. A., Batey, S., Hooton, K. A. & Clarke, J. The folding of spectrin domains I: wild-type domains have the same stability but very different kinetic properties. *J Mol Biol* **344**, 195–205, https://doi.org/10.1016/j.jmb.2004.09.037 (2004).
48. Wensley, B. G., Gartner, M., Choo, W. X., Batey, S. & Clarke, J. Different members of a simple three-helix bundle protein family have very different folding rate constants and fold by different mechanisms. *J Mol Biol* **390**, 1074–1085, https://doi.org/10.1016/j.jmb.2009.05.010 (2009).
49. Neuweiler, H. *et al*. Downhill versus barrier-limited folding of BBL 2: mechanistic insights from kinetics of folding monitored by independent tryptophan probes. *J Mol Biol* **387**, 975–985, https://doi.org/10.1016/j.jmb.2008.12.056 (2009).
50. Neuweiler, H. *et al*. The folding mechanism of BBL: Plasticity of transition-state structure observed within an ultrafast folding protein family. *J Mol Biol* **390**, 1060–1073, https://doi.org/10.1016/j.jmb.2009.05.011 (2009).
51. Dalby, P. A., Clarke, J., Johnson, C. M. & Fersht, A. R. Folding intermediates of wild-type and mutants of barnase. II. Correlation of changes in equilibrium amide exchange kinetics with the population of the folding intermediate. *J Mol Biol* **276**, 647–656, https://doi.org/10.1006/jmbi.1997.1547 (1998).
52. Faraj, S. E., Gonzalez-Lebrero, R. M., Roman, E. A. & Santos, J. Human Frataxin Folds Via an Intermediate State. Role of the C-Terminal Region. *Sci Rep* **6**, 20782, https://doi.org/10.1038/srep20782 (2016).

53. Mayor, U., Johnson, C. M., Daggett, V. & Fersht, A. R. Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc Natl Acad Sci USA* **97**, 13518–13522, https://doi.org/10.1073/pnas.250473497 (2000).
54. Enoki, S., Saeki, K., Maki, K. & Kuwajima, K. Acid denaturation and refolding of green fluorescent protein. *Biochemistry* **43**, 14238–14248, https://doi.org/10.1021/bi048733+ (2004).
55. Kamagata, K., Sawano, Y., Tanokura, M. & Kuwajima, K. Multiple parallel-pathway folding of proline-free Staphylococcal nuclease. *J Mol Biol* **332**, 1143–1153 (2003).
56. Patra, A. K. & Udgaonkar, J. B. Characterization of the folding and unfolding reactions of single-chain monellin: evidence for multiple intermediates and competing pathways. *Biochemistry* **46**, 11727–11743, https://doi.org/10.1021/bi701142a (2007).
57. Parker, M. J., Spencer, J. & Clarke, A. R. An integrated kinetic analysis of intermediates and transition states in protein folding reactions. *J Mol Biol* **253**, 771–786, https://doi.org/10.1006/jmbi.1995.0590 (1995).
58. Parker, M. J. & Marqusee, S. The cooperativity of burst phase reactions explored. *J Mol Biol* **293**, 1195–1210, https://doi.org/10.1006/jmbi.1999.3204 (1999).
59. Parker, M. J. *et al*. Domain behavior during the folding of a thermostable phosphoglycerate kinase. *Biochemistry* **35**, 15740–15752, https://doi.org/10.1021/bi961330s (1996).
60. Parker, M. J., Dempsey, C. E., Lorch, M. & Clarke, A. R. Acquisition of native beta-strand topology during the rapid collapse phase of protein folding. *Biochemistry* **36**, 13396–13405, https://doi.org/10.1021/bi971294c (1997).
61. Cota, E. & Clarke, J. Folding of beta-sandwich proteins: three-state transition of a fibronectin type III module. *Protein Sci* **9**, 112–120, https://doi.org/10.1110/ps.9.1.112 (2000).
62. Manyusa, S. & Whitford, D. Defining folding and unfolding reactions of apocytochrome b5 using equilibrium and kinetic fluorescence measurements. *Biochemistry* **38**, 9533–9540, https://doi.org/10.1021/bi990550d (1999).
63. Plaxco, K. W., Spitzfaden, C., Campbell, I. D. & Dobson, C. M. A comparison of the folding kinetics and thermodynamics of two homologous fibronectin type III modules. *J Mol Biol* **270**, 763–770, https://doi.org/10.1006/jmbi.1997.1148 (1997).
64. Mizukami, T., Abe, Y. & Maki, K. Evidence for a Shared Mechanism in the Formation of Urea-Induced Kinetic and Equilibrium Intermediates of Horse Apomyoglobin from Ultrarapid Mixing Experiments. *PLoS One* **10**, e0134238, https://doi.org/10.1371/journal.pone.0134238 (2015).
65. Uzawa, T. *et al*. Collapse and search dynamics of apomyoglobin folding revealed by submillisecond observations of alpha-helical content and compactness. *Proc Natl Acad Sci USA* **101**, 1171–1176, https://doi.org/10.1073/pnas.0305376101 (2004).
66. DeVries, I., Ferreiro, D. U., Sanchez, I. E. & Komives, E. A. Folding kinetics of the cooperatively folded subdomain of the IkappaBalpha ankyrin repeat domain. *J Mol Biol* **408**, 163–176, https://doi.org/10.1016/j.jmb.2011.02.021 (2011).

Acknowledgements

The work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017R1E1A1A01077717), and by JSPS (Japan Society for the Promotion of Science) KAKENHI Grant Numbers JP25440075 and JP16K07314. The authors thank KIAS Center for Advanced Computation for providing computing resources for this work.

Author Contributions

B.M., K.K. and J.L. designed and performed research, analyzed the data and wrote the paper.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-36992-y.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019