PFDB: A standardized protein folding database with temperature correction

Balachandran Manavalan1, Kunihiro Kuwajima1,2,3 & Jooyoung Lee1

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 mechanisms1,2, at least 16 protein folding kinetics datasets have been reported3–19. 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, ACPro19 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 equation20, and the logarithmic rate constants of folding and unfolding, ln(kf) and ln(ku), 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...
of a particular protein is determined by two parameters of the folding or unfolding reaction, the activation heat temperature in PFDB, we developed a method for temperature correction. The predicted shape of the Eyring plot at different temperatures (\(T_x\)). The temperature correction.

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 (\(\alpha/\beta\), \(\alpha/\beta/\alpha\), and \(\alpha+\beta\)), (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) \(\ln(k_u)\) reported, (x) \(\ln(k_I)\) after temperature correction, (xi) \(\ln(k_f)\) (only for N2S proteins), (xii) \(\ln(k_f)\) reported, (xiii) \(\ln(k_u)\) after temperature correction, and (xiv) Tanford \(\beta(\tilde{\beta})\). The AG dataset is also included in our database for comparison. A comment section is provided in the final column.

Figure 2A shows a distribution of the temperature at which the \(\ln(k_u)\) 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 (\(T_f\)). The \(T_f\) 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 temperature...
The predicted logarithmic rate constant at $T_0$ (298.15 K) is thus given by the following equation:

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

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 ($\Delta C_p$) of the equilibrium protein unfolding. The $\Delta C_p$ is approximately proportional to the protein chain length in the PDB structure ($L_{PDB}$) and empirically given by:

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

Now, it follows that:

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

where $\beta$ is a proportionality constant. Therefore, once we have reasonable estimates of $T_H$ and $\beta$, 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, $\Delta C_p^\ddagger$ diminishes to zero when $L_{PDB}$ tends to zero. A regression equation between $\Delta C_p$ and $L_{PDB}$ with the zero intercept has thus also been reported in the original literature as given by

$$\Delta C_p = 0.058 \cdot L_{PDB}$$

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 $k_i$ 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. Both the $T_H$ and $\beta$ values for folding kinetics, $T_Hf$ and $\beta_f$, 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 and 1PIN), for which $-\Delta C_p$ for folding was larger than $\Delta C_p$. 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 $\Delta C_p^\ddagger$. Examples of the Eyring plot for three proteins (1APS, 1D6O, and 1AVZ) are shown in Figure S1. For folding kinetics, the Eyring...
Table 2. List of proteins used to estimate $T_H$ and $\beta_f$ for two-state proteins.

| PDB    | $I_{ref}$ | Temp. (K) | $\Delta H^p$ (kJ/mol) | $\Delta C_p^p$ (kJ/mol/K) | $T_H$ (K) | $\Delta C_p$ (kJ/mol/K) | $\beta_f$ |
|--------|-----------|-----------|------------------------|---------------------------|-----------|-------------------------|-----------|
| 1APS4  | 98        | 301.15    | 40.70                  | $-5.27$                   | 316.99    | 5.55                    | $-0.46$   |
| 1D6O4  | 107       | 298.15    | 48.53                  | $-2.80$                   | 315.46    | 6.10                    | $-0.46$   |
| 1EIG6  | 48        | 298.15    | 28.45                  | $-1.76$                   | 314.34    | 2.45                    | $-0.72$   |
| 1HDN2  | 85        | 293.15    | 86.10                  | $-3.22$                   | 319.89    | 4.74                    | $-0.68$   |
| 2VH72  | 94        | 301.15    | 23.60                  | $-2.48$                   | 310.67    | 5.30                    | $-0.47$   |
| 3C123  | 64        | 298.00    | 33.55                  | $-2.05$                   | 324.12    | 3.44                    | $-0.60$   |
| 1EHB2  | 82        | 298.15    | 42.40                  | $-3.60$                   | 309.93    | 4.55                    | $-0.79$   |
| 1CSP4  | 67        | 298.15    | 31.60                  | $-2.70$                   | 309.85    | 3.62                    | $-0.74$   |
| 1AYZ7  | 57        | 293.00    | 43.09                  | $-1.86$                   | 316.20    | 3.00                    | $-0.62$   |
| 1SHG2  | 57        | 298.00    | 37.00                  | $-2.30$                   | 314.09    | 3.00                    | $-0.77$   |
| 1HC1D  | 118       | 293.15    | 57.74                  | $-4.39$                   | 306.29    | 6.79                    | $-0.65$   |
| 2JMC3  | 77        | 298.15    | 45.00                  | $-2.20$                   | 318.60    | 4.24                    | $-0.52$   |
| Mean ± SE |           |           |                        |                           | 314.70 ± 1.44 | 3.52 ± 0.03 |           |

Table 3. List of proteins used to estimate $T_H$ and $\beta_f$ for non-two-state proteins.

| PDB    | $I_{ref}$ | Temp. (K) | $\Delta H^p$ (kJ/mol) | $\Delta C_p^p$ (kJ/mol/K) | $T_H$ (K) | $\Delta C_p$ (kJ/mol/K) | $\beta_f$ |
|--------|-----------|-----------|------------------------|---------------------------|-----------|-------------------------|-----------|
| 2CRO26 | 65        | 293.15    | 40.70                  | $-3.05$                   | 310.50    | 3.50                    | $-0.87$   |
| 1PGB4  | 56        | 298.15    | 16.80                  | $-1.90$                   | 306.99    | 2.94                    | $-0.64$   |
| 1L633  | 162       | 285.15    | 92.05                  | $-6.84$                   | 298.61    | 9.51                    | $-0.72$   |
| Mean ± SE |           |           |                        |                           | 305.369 ± 3.526 | $-0.746 ± 0.067$ |           |

plot is convexed, and hence, $T_H$ corresponds to the temperature of the maximum point in the Eyring plot. The $\Delta C_p^\beta_f$ is given by the curvature of the Eyring plot, and the $\beta_f$ was thus evaluated by $\beta_f = \Delta C_p^\beta_f / \Delta C_p$, where $\Delta C_p$ was obtained by Eq. (2); $\Delta C_p^\beta_f$ and $\beta_f$ are negative because the Eyring plot is convexed. The $T_H$ and $\beta_f$ values thus obtained were averaged for the 12 2S proteins and for the 3 N2S proteins (Tables 2 and 3). The $T_H$ and $\beta_f$ 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 $T_H$ and $\Delta C_p^\beta_f$ were not available directly, we employed Eqs (1) and (3) to predict $\ln[k(T_H)]$ by assigning the $T_H$ and $\beta_f$ values to $T_H$ and $\beta_f$ in the equations. However, for the proteins whose $T_H$ and $\Delta C_p^\beta_f$ were available (1E0G28, 1HDN30, 2VH729, 1EHB27, 1HCD31, and 2CRO26), we directly calculated the $\ln[k(T_H)]$ values by Eq. (1). To distinguish $\ln[k(T_H)]$ predicted by using the averaged $T_H$ and $\beta_f$ and that directly calculated by Eq. (1) with the known $T_H$ and $\Delta C_p^\beta_f$, the latter values are indicated in boldface type in our dataset. It should be also noted that the above $T_H$ and $\beta_f$ 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 $\ln[k(T_H)]$ after the temperature correction with the experimentally observed $\ln[k(T_H)]$. For 9 2S and 5 N2S proteins (Table 4), which were not included in those used for estimating $T_H$ and $\beta_f$, the experimental $\ln[k(T_H)]$ was available at both $T_H$ and $T_H$. We thus applied the temperature correction to the $\ln[k(T_H)]$ values using the above $T_H$ and $\beta_f$, and compared predicted $\ln[k(T_H)]$ with the experimentally observed $\ln[k(T_H)]$. From Fig. 2B, the predicted $\ln[k(T_H)]$ 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 $\Delta C_p$ of unfolding. Therefore, we can reasonably assume that $\beta_f$ is equivalent to $-\beta_f$ for 2S proteins. Therefore, for the 2S proteins for which the $\beta_f$ is available, we also calculated the $\ln[k(T_H)]$ values by assigning the $T_H$ and $-\beta_f$ values to $T_H$ and $\beta_f$ in Eqs (1) and (3). The $\ln[k(T_H)]$ 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 $T_H$ and $\beta_f$. As seen from the PFDB dataset, these two types of predicted $\ln[k(T_H)]$ are reasonably coincident with each other.

**Temperature correction for unfolding.** We introduced the temperature corrections into the proteins whose $k_i$ values were measured at a temperature other than the standard temperature (298.15 K), and the $T_H$ and $\beta_f$ values for unfolding kinetics. $T_H$ and $\beta_f$, respectively, were required for temperature correction. For unfolding kinetics, the Eyring plot is usually concaved with a positive $\beta_f$. For 2S proteins, there is only a single transition state between U and N with a $\beta_f$ of $-0.62 ± 0.03$, and we can reasonably assume that $\beta_f = 1 + \beta_f$. Therefore, we find that $\beta_f = 0.38 ± 0.03$. For N2S proteins, this simple relationship may not hold, because of a contribution from an intermediate (I) state. For the N2S proteins, however, $(1 - \beta_f)$ is expected to be equivalent to $\beta_f$, because $\beta_f$ represents the relative position of the transition state between U and N in terms of the denaturant $m$ values. The $\beta_f$ was
reported for 38 N2S proteins in PFDB, and their average was estimated at 0.79 ± 0.02, and hence βc = 0.21 ± 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 $T_{hu}$ and $\beta_c$. Therefore, we had to use a different way to estimate $T_{hu}$. We thus chose 6 2S proteins (1IMQ$^{13,43}$, 1K9Q$^{31,44}$, 1RFA$^{45}$, 1SS1$^{46}$, 1U4Q$^{47,48}$, and 2WXC$^{49,50}$) and 3 N2S proteins (1BNJ$^{51}$, 1EKG$^{52}$, and 1ENH$^{53}$), for which the experimental $\ln(k_u)$ is available at both $T_h$ and $T_s$ (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 $\beta_c$ values to $T_{hu}$ and $\beta$ in Eqs (1) and (3) to calculate 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 $\beta_c$ values, which indicates a reasonable coincidence between the experimental and predicted values.

For the proteins whose $T_{hu}$ and $\Delta C_m^i$ 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 $\beta_c$ values to $T_{hu}$ and $\beta$. However, for the proteins whose $T_{hu}$ and $\Delta C_m^i$ were available (1EBH$^{27}$ 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 $\beta_c$ and that directly calculated by Eq. (1) with the known $T_{hu}$ and $\Delta C_m^i$, the latter values are indicated in boldface type in our dataset.

For the 2S proteins for which $\beta_c$ is available, we also calculated the $\ln(k_u(T_0))$ values by assigning the $T_{hu}$ and $(1 - \beta_c)$ values to $T_{hu}$ and $\beta$ 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 $\beta_c$. As seen from the PFDB dataset, these two types of predicted $\ln(k_u(T_0))$ are reasonably coincident with each other.

| PDB   | $\ln[k_u(T_h)]$ | $T_h$ (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$^{31,44}$ | 10.92 | 351.15 | 6.66 | 6.30 |
| 1RFA$^{45}$ | -3.10 | 281.15 | -1.17 | -0.45 |
| 1SS1$^{46}$ | 7.40 | 323.15 | 3.40 | 4.20 |
| 1U4Q$^{47,48}$ | 0.92 | 283.15 | 3.40 | 2.61 |
| 2WXC$^{49,50}$ | -3.37 | 298.15 | 0.26 | 0.06 |
| 1BNJ$^{51}$ | -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 4. List of Proteins used for predicting $\ln(k_u)$ at 25°C. *$T_h$ for 1DWR was 299.15 K (26°C). Normal font and bold, respectively, represent the 2S and N2S proteins.

| PDB   | $\ln[k_u(T_h)]$ | $T_h$ (K) | $\ln[k_u(T_0)]$ observed | $\ln[k_u(T_0)]$ predicted |
|-------|----------------|----------|---------------------------|---------------------------|
| 1IMQ$^{13,43}$ | -2.66 | 278.15 | -0.92 | -0.12 |
| 1K9Q$^{31,44}$ | 7.09 | 283.15 | 7.33 | 8.69 |
| 1RFA$^{45}$ | 8.92 | 311.15 | 8.37 | 8.67 |
| 1K9Q$^{31,44}$ | 7.41 | 351.15 | 8.37 | 7.87 |
| 1EKG$^{52}$ | 4.40 | 281.15 | 7.00 | 6.11 |
| 1SS1$^{46}$ | 12.41 | 323.15 | 12.08 | 12.07 |
| 1K9Q$^{31,43}$ | 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 |
| 1BNJ$^{51}$ | 2.07 | 318.15 | 2.50 | 2.31 |
| 1DWR$^{15,43}$ | 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 5. List of proteins used for predicting $\ln(k_u)$ at 25°C. Normal font and bold, respectively, represent the 2S and N2S proteins.
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 ACPro19 and the Garbuzynskiy dataset17. 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, 2GAS, 2J5A, 2JMC, 2LLH, 2L6R, 2WQG,
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 \( \ln(k_i) \) values thus obtained are listed in our dataset.

The \( \ln(k_i) \), \( f_i \) and \( k_i \) 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 (1PHE (1–175)\(^{55} \), 1PHE (186–394)\(^{56} \), 1L63\(^{57} \), 1HNG\(^{58} \) and 1TTG\(^{59} \)), 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 \( \ln(k_i) \) 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 \( \ln(k_i) \).

### 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 \( (T_c) \) is known. The following section will describe the derivation of Eq. (1).

According to the Eyring–Kramers equation\(^{20} \), we find that:

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

where \( \Delta H(T) \) and \( \Delta S(T) \) are the activation enthalpy and the activation entropy, respectively, at a reference temperature \( T_i \), and \( \Delta C_p^\ddagger \) is the activation heat capacity; we assume that the \( \Delta C_p^\ddagger \) is a constant independent of temperature \( T \). When we set \( T_i \) to the temperature where \( \Delta H(T) = 0 \), i.e., the maximum or minimum point of the Eyring plot, Eq. (5) is rewritten as:

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

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

\[
\ln \left( \frac{k(T_0)}{T_0} \right) - \ln \left( \frac{k(T_i)}{T_i} \right) = \frac{\Delta C_p^\ddagger}{R} \cdot \left[ \frac{T_i}{T_0} - \frac{T_i}{T_0} + \ln \left( \frac{T_i}{T_0} \right) \right]
\]  

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., Ossypov, 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. Isomoto, 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/10.11002/ps.071324507 (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. Ivanov, D. N. et al. Contact order revisited: influence of protein size on the folding rate. Protein Sci 12, 2057–2062, https://doi.org/10.11002/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/10.1100/ps.041205045 (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/10.1100/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)57410-6 (2002).

16. Zou, T. & Oztan, 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., Ivanov, 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., Thangkani, A. M. & Selvaraj, S. FOLD-RATE: prediction of protein folding rates from amino acid composition. *Nucleic Acids Res* **34**, W70–74, https://doi.org/10.1093/nar/gkt245 (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. Bilel, 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/nir/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. Candell, 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/bi8922630 (1999).

30. Van Nuland, N. A. et al. Slow cooperative folding of a small globular protein HIP. *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 hscAphosphin. *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. Maen, 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. 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).

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.1999.0647 (1999).

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/10.1016/S0022-2836(01)004873 (2001).

42. 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).

43. Backer, P. 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).

44. Neuweller, 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).

45. Backer, P. et al. 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/jmb.1997.1547 (1998).

46. Faraj, S. E., Gonzalez-Lebrero, R. M., Roman, E. A. & Santos, J. Human Prataxis 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., Makki, 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 phosophglycerate 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, 13996–13405, 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. & Makki, 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, J. 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.