Supplementary Material

1. Derivation of the Distribution of Φ Estimates

Using least squares or maximum likelihood methods, we infer that the estimates of the kinetic parameters follow (asymptotically) a multivariate normal distribution. As the chevron curves for the two amino acid sequences are fit separately, kinetic parameter estimates from the same chevron fit will be correlated, while kinetic parameter estimates from different chevron fits will be independent. Mathematically, this means that

$$\boldsymbol{Y} = \begin{bmatrix} \widehat{\ln(k_f)} \\ \widehat{\ln(k_u)} \\ \widehat{\ln(k'_f)} \\ \widehat{\ln(k'_u)} \end{bmatrix} \sim N \left(\begin{bmatrix} \ln(k_f) \\ \ln(k_u) \\ \ln(k'_f) \\ \ln(k'_u) \end{bmatrix}, \begin{bmatrix} (\sigma_f)^2 & \rho\sigma_f\sigma_u & 0 & 0 \\ \rho\sigma_f\sigma_u & (\sigma_u)^2 & 0 & 0 \\ 0 & 0 & (\sigma'_f)^2 & \rho'\sigma'_f\sigma'_u \\ 0 & 0 & \rho'\sigma'_f\sigma'_u & (\sigma'_u)^2 \end{bmatrix} \right) \tag{1}$$

Here, $\ln(k_f)$ denotes the logarithm of the folding rate of the wild type, $\ln(k_u)$ denotes the logarithm of the unfolding rate of the wild type, and $\ln(k'_f)$ and $\ln(k'_u)$ denote the respective parameters for the mutant. The symbol σ is used for the respective standard errors, and ρ is used for the respective correlations. Figure 1 shows scatterplots for the kinetic parameter estimates of a chevron fit obtained in a statistical simulation study.

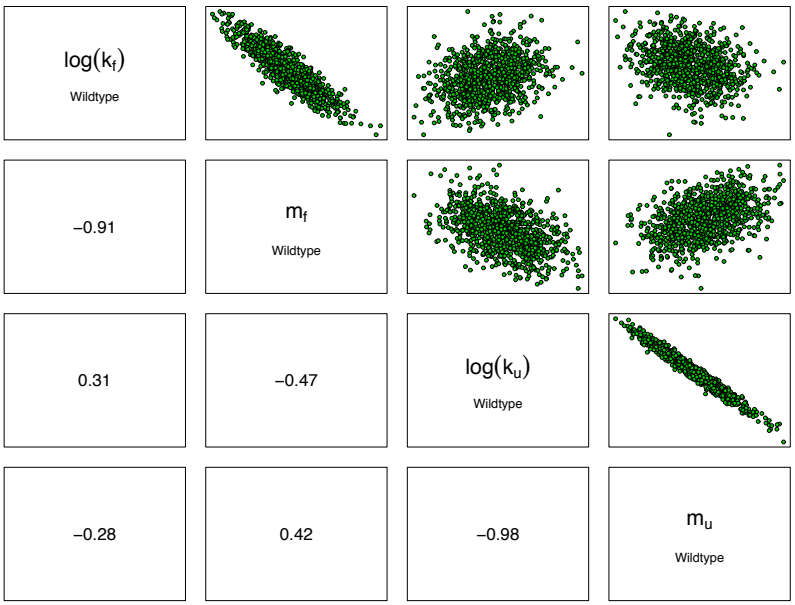

Figure 1: For 10,000 iterations, Gaussian noise with standard deviation typically seen for the experimental error in such kinetic studies (see for example `www.foldeomics.org`) was added to a pair of "synthetic" chevron curves, and the kinetic parameters were estimated by fitting the chevrons. For brevity, the distribution of the kinetic parameter estimates (including the estimates for the folding and unfolding arms, m_f and m_u respectively) are shown for only one chevron. This figure is only intended to highlight the distributions between the kinetic parameter estimates, and thus the axis labelings are omitted.

An elementary result of linear model theory (e. g. Seber [1]) states that if $\boldsymbol{Y}$ is a random vector of length n following a multivariate normal distribution, i. e. $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{C}$ is a matrix with n columns and p

rows, then $\boldsymbol{CY} \sim N_p(\boldsymbol{C\mu}, \boldsymbol{C\Sigma C}^T)$. As the estimates in the changes in free energy are linear functions of the kinetic parameters ($\widehat{\Delta\Delta G_\ddagger} = RT \times [\widehat{\ln(k_f)} - \widehat{\ln(k'_f)}]$, and $\widehat{\Delta\Delta G_\cup} = RT \times [\widehat{\ln(k_f)} - \widehat{\ln(k'_f)} - \widehat{\ln(k_u)} + \widehat{\ln(k'_u)}]$), we choose

$$\boldsymbol{C} = RT \times \begin{bmatrix} +1 & 0 & -1 & 0 \\ +1 & -1 & -1 & +1 \end{bmatrix}, \tag{2}$$

where R is the gas constant, and T is the absolute temperature. Therefore

$$\boldsymbol{CY} = RT \times \begin{bmatrix} +1 & 0 & -1 & 0 \\ +1 & -1 & -1 & +1 \end{bmatrix} \times \begin{bmatrix} \widehat{\ln(k_f)} \\ \widehat{\ln(k_u)} \\ \widehat{\ln(k'_f)} \\ \widehat{\ln(k'_u)} \end{bmatrix} = RT \times \begin{bmatrix} \widehat{\ln(k_f)} - \widehat{\ln(k'_f)} \\ \widehat{\ln(k_f)} - \widehat{\ln(k'_f)} - \widehat{\ln(k_u)} + \widehat{\ln(k'_u)} \end{bmatrix} = \begin{bmatrix} \widehat{\Delta\Delta G_\ddagger} \\ \widehat{\Delta\Delta G_\cup} \end{bmatrix}, \tag{3}$$

and its distribution is given by

$$\begin{bmatrix} \widehat{\Delta\Delta G_\ddagger} \\ \widehat{\Delta\Delta G_\cup} \end{bmatrix} \sim N\left(\begin{bmatrix} \Delta\Delta G_\ddagger \\ \Delta\Delta G_\cup \end{bmatrix}, \begin{bmatrix} \sigma_\ddagger^2 & \rho_\Delta \sigma_\ddagger \sigma_\cup \\ \rho_\Delta \sigma_\ddagger \sigma_\cup & \sigma_\cup^2 \end{bmatrix} \right), \tag{4}$$

where

$$\begin{aligned} \sigma_\ddagger^2 &= \mathsf{var}(\widehat{\Delta\Delta G_\ddagger}) \\ &= (RT)^2 \times \left[(\sigma_f)^2 + (\sigma'_f)^2 \right], \\ \sigma_\cup^2 &= \mathsf{var}(\widehat{\Delta\Delta G_\cup}) \\ &= (RT)^2 \times \left[(\sigma_f)^2 + (\sigma'_f)^2 + (\sigma_u)^2 + (\sigma'_u)^2 - 2\rho\sigma_f\sigma_u - 2\rho\sigma'_f\sigma'_u \right], \\ \rho_\Delta \sigma_\ddagger \sigma_\cup &= \mathsf{cov}(\widehat{\Delta\Delta G_\ddagger}, \widehat{\Delta\Delta G_\cup}) \\ &= \mathsf{corr}(\widehat{\Delta\Delta G_\ddagger}, \widehat{\Delta\Delta G_\cup}) \times \mathsf{se}(\widehat{\Delta\Delta G_\ddagger}) \times \mathsf{se}(\widehat{\Delta\Delta G_\cup}) \\ &= (RT)^2 \times \left[(\sigma_f)^2 + (\sigma'_f)^2 - \rho\sigma_f\sigma_u - \rho\sigma'_f\sigma'_u \right]. \end{aligned} \tag{5}$$

Here, var() denotes the variance, cov() the covariance, corr() the correlation, and se() the standard error of the respective arguments. Figure 2 shows the marginal and joint distributions of $\widehat{\Delta\Delta G_\ddagger}$ and $\widehat{\Delta\Delta G_\cup}$ derived from our simulation study.

As the Φ-value estimate is not a linear function of the estimates of the changes in free energy, we use a Taylor-series expansion to derive an approximate distribution. This method is also often referred to as *error propagation* in the life sciences, and *Delta method* in the statistical literature. In brief, it states that if we have an estimator $\hat{\boldsymbol{\theta}}$ for a parameter $\boldsymbol{\theta}$, following a multivariate normal distribution with mean $\boldsymbol{\theta}$ and variance covariance matrix Σ, then for a non-constant and differentiable function f,

$$f(\hat{\boldsymbol{\theta}}) \approx N\left(f(\boldsymbol{\theta}), \boldsymbol{V}\right). \tag{6}$$

where

$$\boldsymbol{V} = \left(\frac{\delta f}{\delta \boldsymbol{\theta}}\right)^T \boldsymbol{\Sigma} \left(\frac{\delta f}{\delta \boldsymbol{\theta}}\right) |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{7}$$

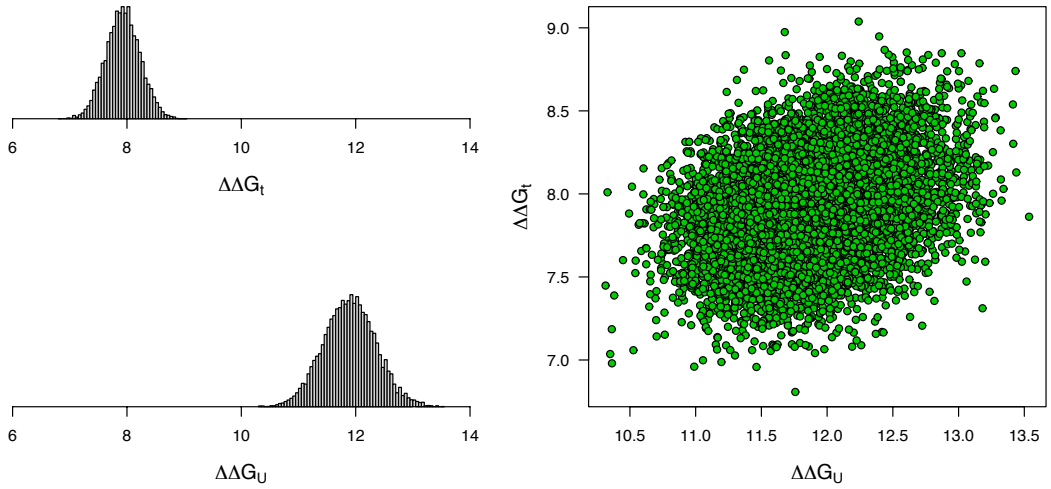

Figure 2: The marginal and joint distributions of $\widehat{\Delta\Delta G}_{\ddagger}$ and $\widehat{\Delta\Delta G}_{\mathsf{U}}$ derived from our simulation study, confirming the theoretical result that $\widehat{\Delta\Delta G}_{\ddagger}$ and $\widehat{\Delta\Delta G}_{\mathsf{U}}$ follow a bivariate normal distribution.

The theorem is often stated as a limit theorem, i. e. in terms of the sample size of observations, that give rise to the kinetic parameters, getting large. In this context, the above approximation holds if the absolute values of the estimates of the changes in free energy are large compared to their standard errors (see for example the discussions in Hinkley [2] and Marsaglia [3]).

To use this result to derive a distribution for $\widehat{\Phi}$, consider $f(\boldsymbol{\theta}) = f(\Delta\Delta G_{\ddagger}, \Delta\Delta G_{\mathsf{U}}) = \Delta\Delta G_{\ddagger}/\Delta\Delta G_{\mathsf{U}}$. Since

$$\frac{\delta f}{\delta \Delta\Delta G_{\ddagger}} = \frac{1}{\Delta\Delta G_{\mathsf{U}}} \qquad \text{and} \qquad \frac{\delta f}{\delta \Delta\Delta G_{\mathsf{U}}} = -\frac{\Delta\Delta G_{\ddagger}}{\Delta\Delta G_{\mathsf{U}}^2}, \tag{8}$$

it follows that

$$\widehat{\Phi} = \frac{\widehat{\Delta\Delta G}_{\ddagger}}{\widehat{\Delta\Delta G}_{\mathsf{U}}} \approx N\left(\Phi, \mathsf{var}(\widehat{\Phi})\right), \tag{9}$$

and the variance of the estimate of Φ is given by

$$\begin{aligned}
\mathsf{var}(\widehat{\Phi}) &= \frac{1}{(\Delta\Delta G_{\mathsf{U}})^2}[\Delta\Delta G_{\mathsf{U}}, -\Delta\Delta G_{\ddagger}] \begin{bmatrix} \sigma_{\ddagger}^2 & \rho_{\Delta}\sigma_{\ddagger}\sigma_{\mathsf{U}} \\ \rho_{\Delta}\sigma_{\ddagger}\sigma_{\mathsf{U}} & \sigma_{\mathsf{U}}^2 \end{bmatrix} \begin{bmatrix} \Delta\Delta G_{\mathsf{U}} \\ -\Delta\Delta G_{\ddagger} \end{bmatrix} \frac{1}{(\Delta\Delta G_{\mathsf{U}})^2} \\
&= \frac{1}{(\Delta\Delta G_{\mathsf{U}})^4}(\sigma_{\ddagger}^2 \times (\Delta\Delta G_{\mathsf{U}})^2 - 2\rho_{\Delta}\sigma_{\ddagger}\sigma_{\mathsf{U}} \times \Delta\Delta G_{\ddagger}\Delta\Delta G_{\mathsf{U}} + \sigma_{\mathsf{U}}^2 \times (\Delta\Delta G_{\ddagger})^2). \\
&= \left(\frac{\Delta\Delta G_{\ddagger}}{\Delta\Delta G_{\mathsf{U}}}\right)^2 \times \left[\frac{\sigma_{\ddagger}^2}{(\Delta\Delta G_{\ddagger})^2} - 2\rho_{\Delta}\frac{\sigma_{\ddagger}\sigma_{\mathsf{U}}}{\Delta\Delta G_{\ddagger}\Delta\Delta G_{\mathsf{U}}} + \frac{\sigma_{\mathsf{U}}^2}{(\Delta\Delta G_{\mathsf{U}})^2}\right] \\
&= \Phi^2 \times \left[\left(\frac{\sigma_{\ddagger}}{\Delta\Delta G_{\ddagger}}\right)^2 - 2\rho_{\Delta}\left(\frac{\sigma_{\ddagger}}{\Delta\Delta G_{\ddagger}}\right)\left(\frac{\sigma_{\mathsf{U}}}{\Delta\Delta G_{\mathsf{U}}}\right) + \left(\frac{\sigma_{\mathsf{U}}}{\Delta\Delta G_{\mathsf{U}}}\right)^2\right]
\end{aligned} \tag{10}$$

2. A Web based Implementation

The above algorithm has been implemented as part of a web server for use by the folding community, accessible through `biostat.jhsph.edu/~iruczins/software/phi/`. A snapshot of the website is shown in Figure 3. Chevron data can be uploaded as a spreadsheet or as tab delimited text file (there is no limit on the number of mutants that can be analyzed simultaneously). Several user options are available, including the possibility of fitting parallel chevron arms, and the option to measure the folding and unfolding rates at non-zero denaturant concentrations. Once executed, the script creates a webpage with tabulations of the relevant kinetic parameters and their standard errors, as well as estimates of the changes in free energy between mutants, and the estimate for Φ (including the standard error and the confidence interval). These tables can also be downloaded in spreadsheet format.

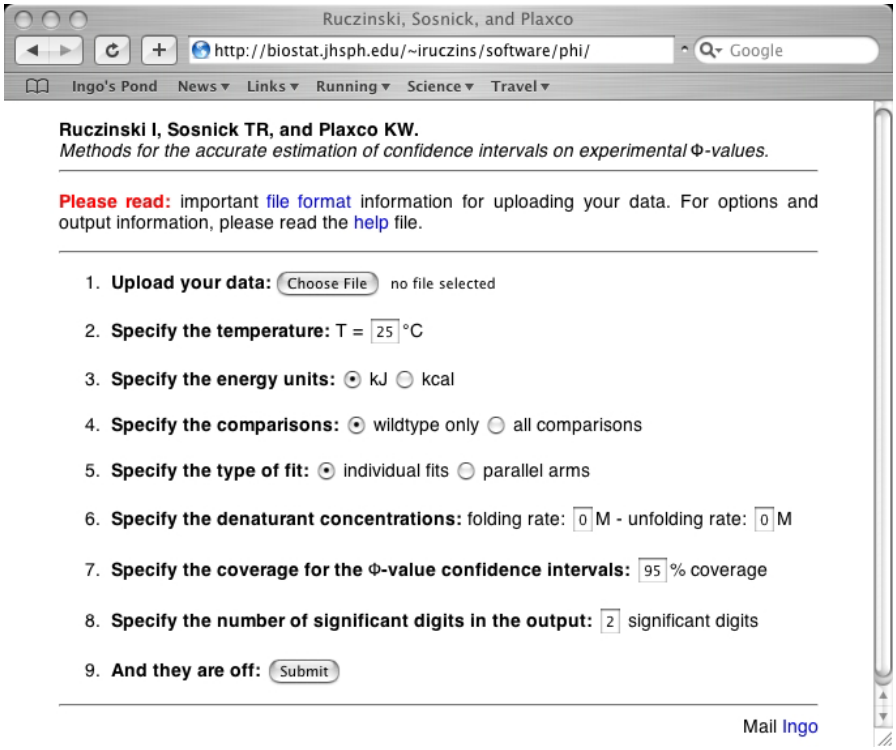

Figure 3: A snapshot of the web server. A perl/cgi script uploads and parses the kinetic data, and calls an R script for the statistical analysis. The estimates for the kinetic parameters and their standard errors and correlations are obtained using the R function `nls()`.

3. Effects of Ignoring the Covariance

When ρ_Δ is ignored (i. e. assumed to be zero), the estimate for the variability in $\widehat{\Phi}$ is simply

$$\mathrm{var}_{\rho_\Delta=0}(\widehat{\Phi}) \quad = \quad \Phi^2 \times \left[\left(\frac{\sigma_\ddagger}{\Delta\Delta G_\ddagger} \right)^2 + \left(\frac{\sigma_U}{\Delta\Delta G_U} \right)^2 \right]. \tag{11}$$

The absolute difference in those variabilities is

$$\mathrm{var}_{\rho_\Delta=0}(\widehat{\Phi}) - \mathrm{var}(\widehat{\Phi}) = \Phi^2 \times 2\rho_\Delta \left(\frac{\sigma_\ddagger}{\Delta\Delta G_\ddagger} \right) \left(\frac{\sigma_U}{\Delta\Delta G_U} \right) = 2\Phi \times \frac{\rho_\Delta \sigma_\ddagger \sigma_U}{(\Delta\Delta G_U)^2}. \tag{12}$$

This formula states that the smaller $\rho_{\Delta}, \sigma_{\ddagger}, \sigma_{\cup}$, and Φ (for a given $\Delta\Delta G_{\cup}$), the smaller the absolute difference in the estimated variabilities. Also, the larger $\Delta\Delta G_{\cup}$ (for a given Φ), the smaller that difference.

To investigate the relative increase in the estimate of the variability of $\widehat{\Phi}$ when ρ_{Δ} is ignored, we simplify the notation by writing

$$\begin{aligned} V_1 &= \Phi^2 \left(a^2 - 2\rho ab + b^2\right), \\ V_2 &= \Phi^2 \left(a^2 + b^2\right), \end{aligned} \tag{13}$$

where $a = \sigma_{\ddagger}/\Delta\Delta G_{\ddagger}$ and $b = \sigma_{\cup}/\Delta\Delta G_{\cup}$. Using $b = c \times a$, we define

$$\frac{V_1}{V_2} = 1 - \frac{2\rho_{\Delta} ab}{a^2 + b^2} = 1 - \frac{2\rho_{\Delta} ca^2}{a^2 + c^2 a^2} = 1 - \frac{2\rho_{\Delta} c}{1 + c^2} =: f(c) \tag{14}$$

The first derivative of this function f is

$$f'(c) = -\frac{2\rho_{\Delta}(1 - c)}{(1 + c^2)^2}, \tag{15}$$

which is equal to zero iff $c = 1$. The function f attains a global minimum at $c = 1$, i.e. when $\sigma_{\ddagger}/\Delta\Delta G_{\ddagger} = \sigma_{\cup}/\Delta\Delta G_{\cup}$. Hence V_1/V_2 is bounded below by $f(1) = 1 - \rho_{\Delta}$, and converges to 1 (i. e. towards equal variance estimates) as c moves away from this point in either direction. Equivalently, V_2/V_1 is bounded above by $1 - \rho_{\Delta}/(1 - \rho_{\Delta})$. The lengths of the confidence intervals scales with the standard errors of the free energy estimates, and therefore the ratio of confidence intervals is bounded by $\sqrt{V2}/\sqrt{V1} = \sqrt{V2/V1} = \sqrt{1 - \rho_{\Delta}/(1 - \rho_{\Delta})}$.

References

[1] Seber, G.A. (1977), *Linear Regression Analysis*, Wiley.

[2] Hinkley, D.V. (1969), *On the Ratio of Two Correlated Normal Random Variables*, Biometrika, 56, 635–39.

[3] Marsaglia, G. (1965), *Ratios of normal variables and ratios of sums of uniform variables*, Journal of the American Statistical Association 60, 193–204.