Uncertainty about Phi Values

Ingo Ruczinski [1] Miguel de los Rios [2] Kevin Plaxco [2]

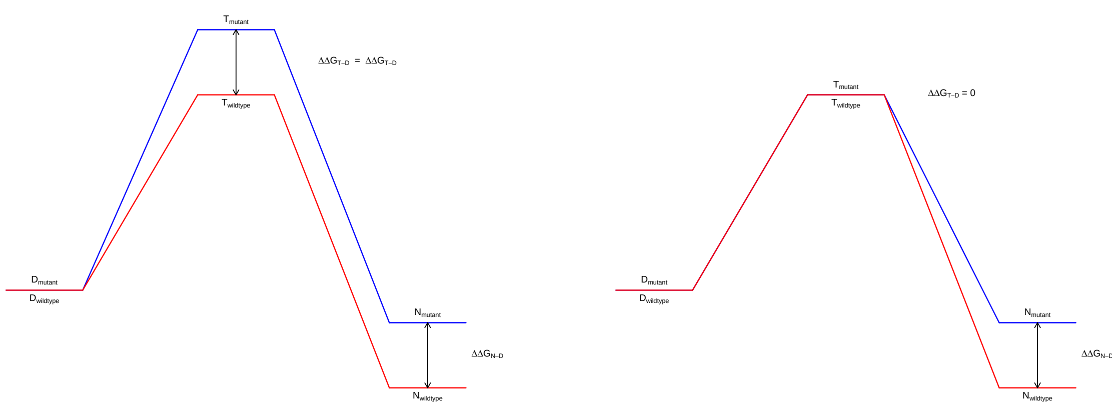
The computation of Φ-values from mutation experiments is a tool commonly used to examine the structures of transition states. However, there is substantial controversy over the precision with which phi-values can be determined experimentally. We replicated experiments independently in three laboratories to investigate variance components that influence the precision of the phi-value estimates. In particular, we addressed the question how the precision depends on the change in free energy between the mutant and the native protein in their folded states.

The poster is organized as follows:

- A very brief introduction of Φ-values (p. 2-3).
- How to calculate a confidence interval for a Φ-value (p. 4-6).
- On the precision of a Φ-value in a single experiment (p. 7).
- On the precision of Φ-values in general (p. 8-12).

[1] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University.
[2] Department of Chemistry and Biochemistry, University of California at Santa Barbara.

Energy Profiles and Phi Values

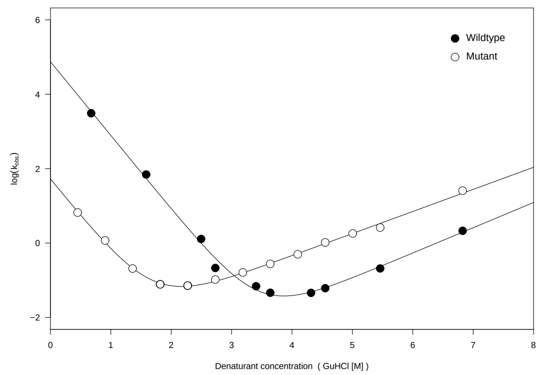


- The Φ-value is defined as the ratio $\Delta\Delta G_{\text{T-D}} / \Delta\Delta G_{\text{N-D}}$.
- If the part of the protein that contains the mutant amino acid is fully structured in the transition state, we have $\Delta\Delta G_{\text{T-D}} \approx \Delta\Delta G_{\text{N-D}}$, and hence $\Phi \approx 1$.
- If the part of the protein that contains the mutant amino acid is equal in denatured and the transition state, we have $\Delta\Delta G_{\text{T-D}} \approx 0$, and hence $\Phi \approx 0$.

At least this is the idea . . .

Phi-Value Estimation

We can estimate $\Delta\Delta G_{T\text{-}D}$, $\Delta\Delta G_{N\text{-}D}$, and hence Φ, using chevron curves:



$$\log(k_{obs}) = \log\left(\exp\left[\log(k_f) + m_f \times \frac{C_{GuHCl}}{RT}\right] + \exp\left[\log(k_u) + m_u \times \frac{C_{GuHCl}}{RT}\right] \right)$$

$$\Delta\Delta G_{T\text{-}D} = RT \times \left[\log(k_f^{wildtype}) - \log(k_f^{mutant}) \right]$$

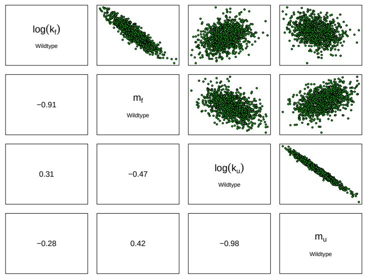
$$\Delta\Delta G_{N\text{-}D} = RT \times \left[\log(k_f^{wildtype}) - \log(k_u^{wildtype}) - \log(k_f^{mutant}) + \log(k_u^{mutant}) \right]$$

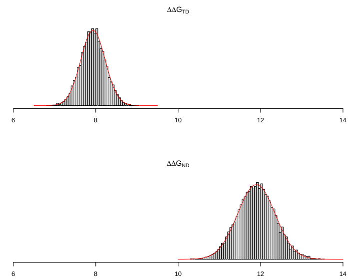
Confidence Intervals

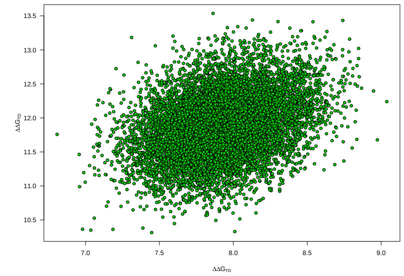
How do we get a confidence interval for our estimate of the Φ-value?

- We assume that the experimental errors for the wild-type and the mutant data are independent, since the data come from separate experiments (this holds unless for example our measurement device would always get too low readings for high concentrations of the denaturant). Therefore the parameter estimates between the two chevron curve fits (such as the folding rates for wild-type and mutant) are independent.
- However, the parameter estimates obtained from one chevron curve (such as the wild-type folding and unfolding rates) are not independent. Assuming the experimental error is gaussian, the parameter estimates from one chevron curve roughly follow a multivariate normal distribution though.
- The estimates for $\Delta\Delta G_{T\text{-}D}$ and $\Delta\Delta G_{N\text{-}D}$ are also not independent, since they have some common ingredients (namely the folding rates for the wild-type and the mutant). A bit of math shows that the estimates of $\Delta\Delta G_{T\text{-}D}$ and $\Delta\Delta G_{N\text{-}D}$ follow a bivariate normal distribution, with a variance-covariance structure that is a function of the variances and covariances of the parameter estimates in the chevron curves, which are also estimated in the fitting procedure.

Confidence Intervals







$$\begin{bmatrix}\widehat{\Delta\Delta G}_{\text{TD}}\\ \widehat{\Delta\Delta G}_{\text{ND}}\end{bmatrix} \sim N\left(\begin{bmatrix}\Delta\Delta G_{\text{TD}}\\ \Delta\Delta G_{\text{ND}}\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \sigma_3^2\\ \sigma_3^2 & \sigma_2^2\end{bmatrix}\right)$$

$$\begin{aligned}\sigma_1^2 &= \sigma^2_{\text{F}_W} + \sigma^2_{\text{F}_M}\\ \sigma_2^2 &= \sigma^2_{\text{F}_W} + \sigma^2_{\text{F}_M} + \sigma^2_{\text{U}_W} + \sigma^2_{\text{U}_M} - 2\rho_{\text{FU}_W}\sigma_{\text{F}_W}\sigma_{\text{U}_W} - 2\rho_{\text{FU}_M}\sigma_{\text{F}_M}\sigma_{\text{U}_M}\\ \sigma_3^2 &= \sigma^2_{\text{F}_W} + \sigma^2_{\text{F}_M} - \rho_{\text{FU}_W}\sigma_{\text{F}_W}\sigma_{\text{U}_W} - \rho_{\text{FU}_M}\sigma_{\text{F}_M}\sigma_{\text{U}_M}\end{aligned}$$

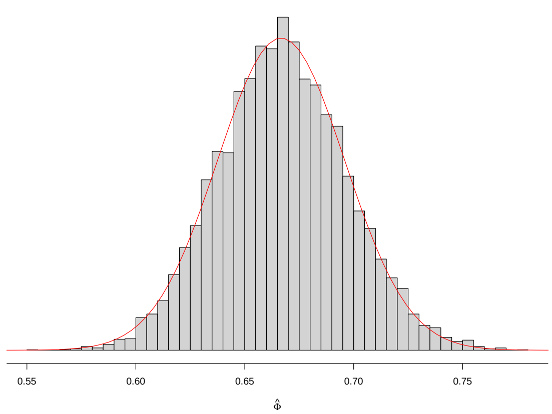
For sufficiently large $\Delta\Delta G_{\text{N-D}}$, even more math shows that the estimate for Φ is approximately normal (there is some slight abuse of a theorem called the “delta method” involved).

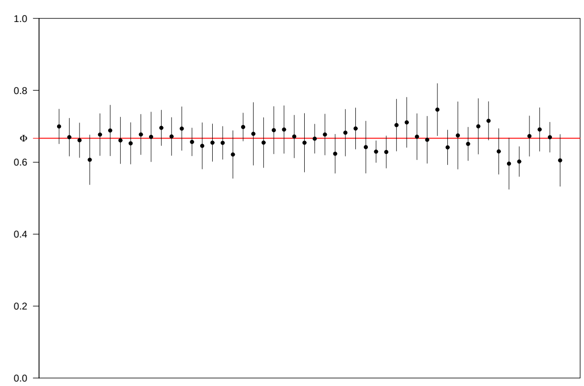
$$\widehat{\Phi} = \frac{\widehat{\Delta\Delta G}_{\text{TD}}}{\widehat{\Delta\Delta G}_{\text{ND}}} \approx N(\Phi, B) \qquad B = \frac{1}{(\widehat{\Delta\Delta G}_{\text{ND}})^4} \times (\sigma_1^2(\widehat{\Delta\Delta G}_{\text{ND}})^2 - 2\sigma_3^2\widehat{\Delta\Delta G}_{\text{TD}}\widehat{\Delta\Delta G}_{\text{ND}} + \sigma_2^2(\widehat{\Delta\Delta G}_{\text{TD}})^2).$$

Confidence Intervals

This variance can be used to derive confidence intervals for the true Φ-value.

$$I = \left[\hat{\Phi} - t^{0.975}_{n_1+n_2-10} \times \sqrt{B}\ ;\ \hat{\Phi} + t^{0.975}_{n_1+n_2-10} \times \sqrt{B}\right]$$

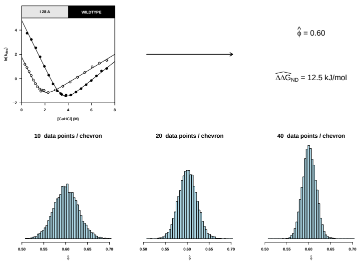


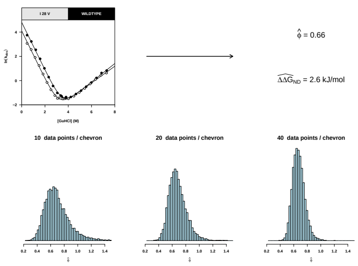


$\longrightarrow$ Software freely available soon!

It is not a priori clear what the degrees of freedom in the t-quantile should be. Adding the number of data points used to fit the two chevron curves (n_1 and n_2) and subtracting the number of parameters estimated in the fitting procedure (a total of 10) however gave 95% covergage for the confidence intervals in simulation studies.

What does the Precision Depend On?

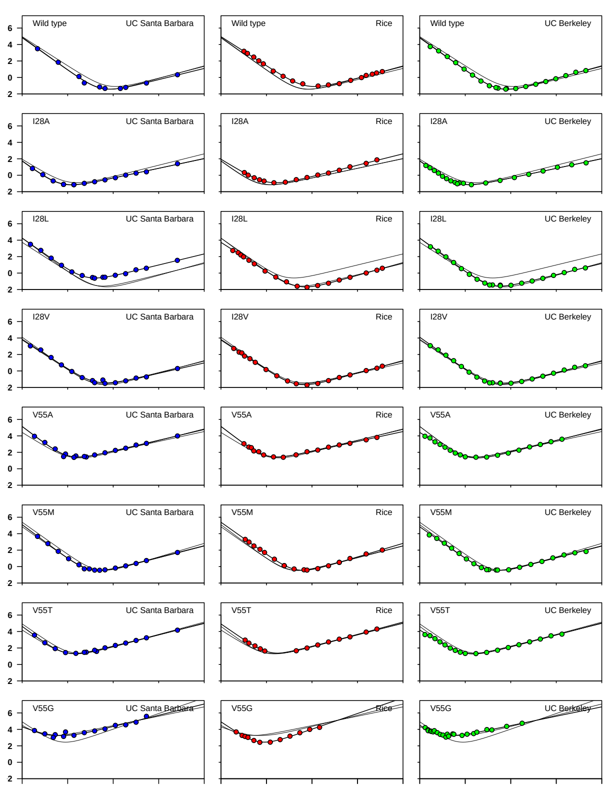




- For a particular experiment, the precision of the estimates for the folding and unfolding rates and therefore the Φ-value depend on the experimental error and the number of observations we have.
- The precision of the estimate for the Φ-value also heavily depends on $\Delta\Delta G_{ND}$.
- A guideline on which cut-off of $\Delta\Delta G_{ND}$ assures "enough precision to estimate Φ" (in some sense) needs to say something about the experimental error and the amount of data gathered!

We took data from an actual experiment, fi t the chevron curves using least squares, and estimated the experimental error, the changes in free energy, and the Φ-value. We then simulated some experiments: pick a certain number of data points (10, 20, or 40) on each chevron curve, for simplicity equally spaced along the x-axis, and add noise (gaussian with standard deviation equal to the experimental error, pretty much the same in all chevron curves). Calculate the Φ-value. Repeat 10,000 times. The above are histograms of those calculated the Φ-values. Pretty obvious are: a) the precision of the Φ-value estimate can go completely out the window when the change in free energy (between wild-type and mutant) in the folded state is small, and b) everything else being equal, the more data we have, the more precision we have for the Φ-value estimate (the deviation roughly scales with the square root of the number of data points).

Replication

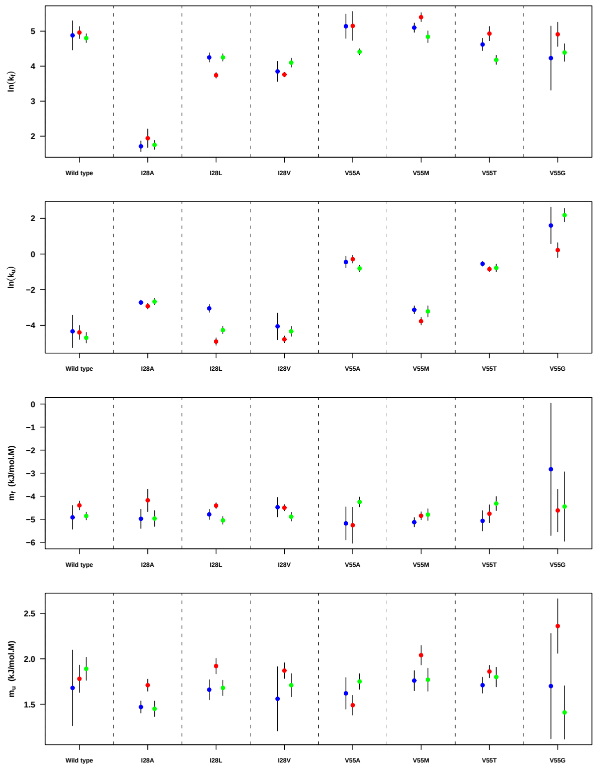


The previous slides addressed the precision in a Φ-value estimate from a single experiment. But is the variability due to the experimental error the same as the variability we see if we repeated the experiment several times, possibly in different labs?

To investigate this, three labs independently performed the same folding measurements on the wild-type FynSH3 domain and seven point mutants (mutations at sites 28 and 55).

Miguel de los Rios, Kevin Plaxco @ UC Santa Barbara
B K Muralidhara, Pernilla Wittung-Stafshede @ Rice
David Wildes, Susan Marqusee @ UC Berkeley

Replication

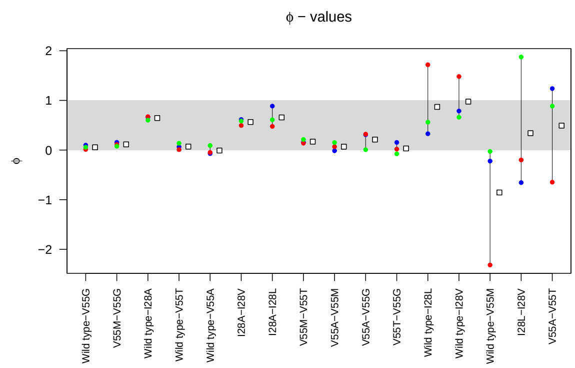


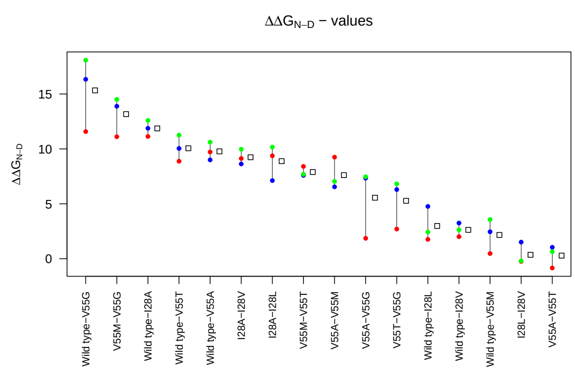
Just by looking at the confidence intervals for the parameters in the chevron curves, it is obvious that there is an extra variance component besides the experimental error!

So what is the variability in the actual Φ-values given our experimental settings?

Hardly surprising, it depends on the change in free energy in the folded states.

Replication

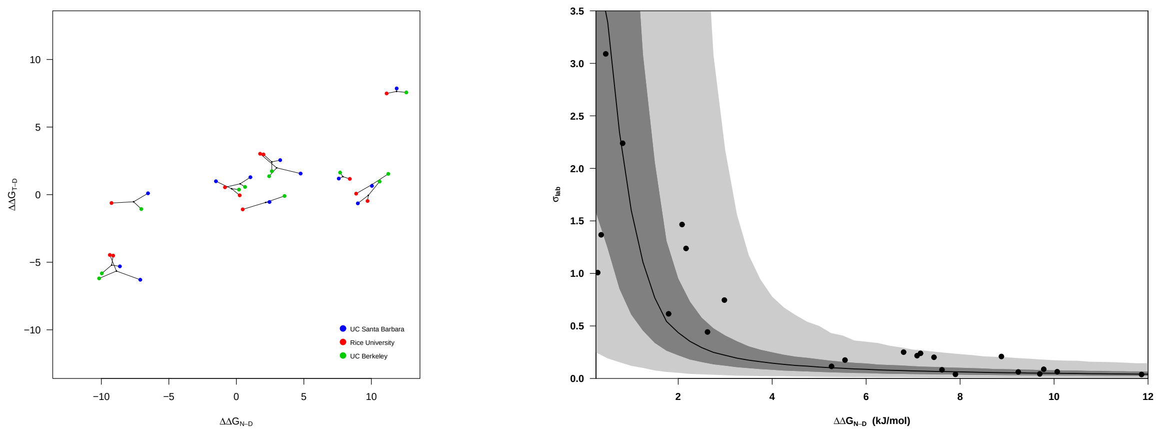




Below are the Φ-value estimates, sorted by average $\Delta\Delta G_{ND}$ (kJ/mol):

	Φ			$\Delta\Delta G_{ND}$		
	UCSB	RICE	CAL	UCSB	RICE	CAL
Wild type-V55G	0.10	0.01	0.06	16.34	11.58	18.07
V55M-V55G	0.16	0.11	0.08	13.88	11.11	14.50
Wild type-I28A	0.66	0.67	0.60	11.87	11.13	12.59
Wild type-V55T	0.06	0.01	0.14	10.04	8.87	11.25
Wild type-V55A	-0.07	-0.05	0.09	9.00	9.72	10.61
I28A-I28V	0.61	0.49	0.58	8.63	9.12	9.97
I28A-I28L	0.89	0.48	0.61	7.11	9.37	10.16
V55M-V55T	0.16	0.14	0.21	7.59	8.40	7.68
V55A-V55M	-0.02	0.07	0.15	6.54	9.25	7.04
V55A-V55G	0.31	0.32	0.01	7.34	1.86	7.46
V55T-V55G	0.15	0.02	-0.08	6.30	2.70	6.82
Wild type-I28L	0.33	1.72	0.56	4.76	1.76	2.43
Wild type-I28V	0.79	1.48	0.66	3.25	2.01	2.63
Wild type-V55M	-0.22	-2.32	-0.03	2.45	0.47	3.57
I28L-I28V	-0.66	-0.2	1.88	1.51	-0.25	-0.20
V55A-V55T	1.24	-0.65	0.88	1.04	-0.84	0.64

Some Simulations

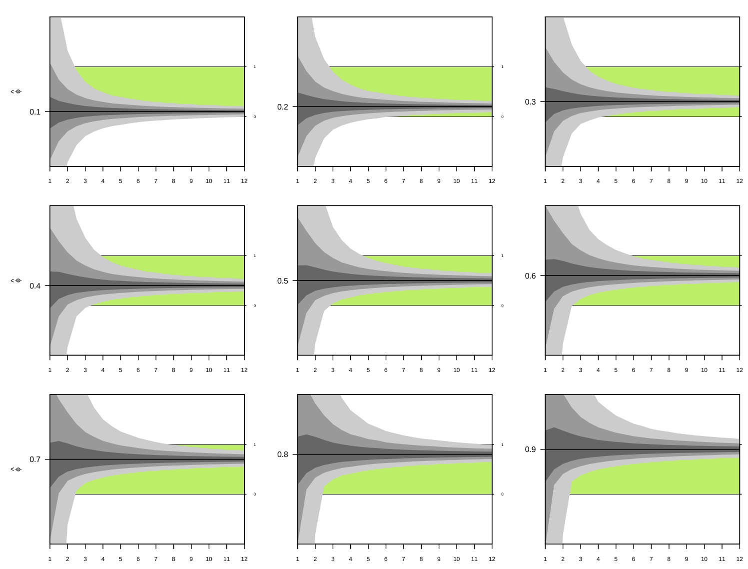


To assess what variability in the Φ-values we can expect under the given laboratory conditions in general, we carried out a simulation study. Using the variability we see in the estimates of $\Delta\Delta G_{TD}$ and $\Delta\Delta G_{ND}$ (left panel), we were able to predict what variability we expect to see between the replicate Φ-values of the three labs (right panel). This is to confirm that we captured the variance components adequately.

In the right panel, the points indicate the standard deviation between the three lab measurements for the single mutants. While double mutants do not make much scientifi c sense, they were added to the plot as additional data. The dark grey area indicates a 50% coverage region, the lighter grey area indicates a 95% coverage region (10,000 independent simulations for each value of $\Delta\Delta G_{ND}$).

Some Simulations

Using our variance components from the previous simulation reflecting true laboratory conditions, we generated plots for the variability of the Φ-value estimates as a function of $\Delta\Delta G_{ND}$.



The grey areas indicate 50%, 75%, and 95% coverage regions respectively (10,000 independent simulations for each value of $\Delta\Delta G_{ND}$). The true Φ-value in the simulation is indicated by a black horizontal line. The green area highlights the region between zero and one.